Multiple-encodings frameworks for explainable multimedia representation and retrieval

Varsha Devi

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ GRENOBLE ALPES

École doctorale : MSTII - Mathématiques, Sciences et technologies de l'information, Informatique
Spécialité : Informatique
Unité de recherche : Laboratoire d'Informatique de Grenoble

Encodages multiples pour une représentation et une recherche par le contenu explicables de documents multimédia

Multiple-encodings frameworks for explainable multimedia representation and retrieval.

Présentée par :

Varsha DEVI

Direction de thèse :

Georges QUENOT DIRECTEUR DE RECHERCHE, CNRS DELEGATION ALPES	Directeur de thèse
Philippe MULHEM CHARGE DE RECHERCHE HDR, CNRS DELEGATION ALPES	Co-directeur de thèse

Rapporteurs :

JEAN MARTINET
PROFESSEUR DES UNIVERSITES, UNIVERSITE COTE D'AZUR
STEPHANE AYACHE
PROFESSEUR DES UNIVERSITES, POLYTECH MARSEILLE

Thèse soutenue publiquement le **12 décembre 2023**, devant le jury composé de :

GEORGES QUENOT DIRECTEUR DE RECHERCHE, CNRS DELEGATION ALPES	Directeur de thèse
PHILIPPE MULHEM CHARGE DE RECHERCHE HDR, CNRS DELEGATION ALPES	Co-directeur de thèse
JEAN MARTINET PROFESSEUR DES UNIVERSITES, UNIVERSITE COTE D'AZUR	Rapporteur
STEPHANE AYACHE PROFESSEUR DES UNIVERSITES, POLYTECH MARSEILLE	Rapporteur
ALEXANDRE BENOIT PROFESSEUR DES UNIVERSITES, POLYTECH ANNECY-CHAMBERY	Examinateur
DANIELLE ZIEBELIN PROFESSEURE DES UNIVERSITES, UNIVERSITE GRENOBLE ALPES	Présidente

Abstract

This dissertation focuses on the field of multimedia information retrieval, and more specifically on video-text retrieval. In the era of vast and diverse multimedia content on the internet, finding relevant videos or texts has become a challenging problem. Before 2012, traditional keyword-based approaches for video-text retrieval were inefficient and relied heavily on human annotations. With the advent of deep learning models, the performance of video-text retrieval systems has greatly increased. In this thesis, we delve into three approaches for video-text retrieval: the concept based approach, using pre-defined visual concepts and concept bank; the concept free approach, which directly extracts patterns from multimedia data; and the hybrid approach, which combines elements from both concept-based and concept-free strategies.

The core objective of this thesis is to train and develop a hybrid model based on latent space and concept space that performs simultaneously retrieval and classification tasks, while providing the causal explanations for retrieved results. This causality-based retrieval model aims to enhance user understanding of the decision-making process without impacting accuracy.

We explore the fundamental elements of video-text retrieval, highlighting the challenges of aligning and retrieving the information across multiple modalities. In video-text retrieval, ambiguity in the queries and pre-defined concept banks can make it difficult to accurately understand the user's intention. In order to better understand the intention of user query and to overcome the issue of ambiguity, in the second part of this thesis, we extend a hybrid state of the art approach by integrating the Part-of-Speech (PoS) tags into their dual encoding model for video-text retrieval. We explore the impact of PoS-tags on the performance and explainability of video-text retrieval results and show the advantages of using PoS-tags to enhance retrieval accuracy, precision, and overall system performance.

In the third part of PhD, we introduced a general framework for analyzing the relationships and the complementarity between different representation spaces in hybrid approaches, namely the (non-explainable) latent space and the (explainable) concept space, in a way to assess to which extend explainable spaces differ from non-explainable ones. Additionally, the thesis embarks on a comprehensive exploration of the complementarity between these spaces. Hybrid models like dual encoding model [1] or interpretable embedding model [2] train two common spaces (latent and concept) in order to find the similarity between video and text for retrieval purpose, but these models lack the analysis of inter and

intra-relationship between those spaces.

In the pursuit of explainability, the concept based part of the hybrid model plays a critical role. It identifies and extracts semantic concepts from video and textual data, visualized through tag clouds, making the retrieval process more interpretable and comprehensible. Current state of the art models provide explanations using tag-clouds, but the provided explanations are not causal. We addressed the problem of providing causal and interpretable visual explanations for video-text retrieval. By providing a visual representation of the causal relationships between the query and the retrieved results, tag clouds enhance user trust and support applications where accountability and insight are paramount.

Résumé

Cette thèse se penche sur le domaine de la recherche d'informations multimédias, avec un accent particulier sur la recherche conjointe de vidéos et de textes, appelée vidéo-texte. À l'ère du contenu multimédia vaste et diversifié sur Internet, la recherche de vidéos ou de textes pertinents est devenue un problème difficile. Avant 2012, les approches traditionnelles basées sur les mots-clés pour la recherche de vidéos et de textes étaient inefficaces et dépendaient fortement des annotations humaines. Avec l'avènement des modèles d'apprentissage profond, la performance des systèmes de recherche de vidéos et de textes a largement augmenté. Dans cette thèse, nous examinons trois approches pour la recherche de vidéo-textes : l'approche basée sur les concepts, qui utilise des concepts visuels prédéfinis et une banque de concepts ; l'approche sans concept, qui extrait directement des modèles à partir de données multimédias ; et l'approche hybride, qui combine des éléments des stratégies basées sur les concepts et des stratégies sans concept.

L'objectif principal de cette thèse de doctorat est de former et de développer un modèle hybride basé sur l'espace latent et l'espace conceptuel et d'effectuer les tâches de recherche et de classification simultanément, tout en fournissant des explications causales pour les résultats obtenus. Ce modèle de recherche basé sur la causalité vise à améliorer la compréhension du processus de prise de décision par l'utilisateur sans impacter négativement les performances.

Cette thèse explore les éléments fondamentaux de la recherche vidéo-texte, en soulignant les défis que posent l'alignement et la recherche d'informations à travers des modalités multiples. Dans la recherche vidéo-texte, l'ambiguïté des requêtes et les banques de concepts prédéfinies peuvent rendre difficile la compréhension précise de l'intention de l'utilisateur. Afin de mieux comprendre l'intention de la requête de l'utilisateur et de surmonter le problème de l'ambiguïté, dans la deuxième partie de cette thèse, nous étendons une approche hybride état de l'art en intégrant les balises Part-of-Speech (PoS) dans le modèle d'encodage double pour la recherche de vidéotexte. Nous étudions l'impact des balises PoS sur les performances et l'explicabilité des résultats de la recherche de vidéo-texte et montrons les avantages de l'utilisation des balises PoS pour améliorer la précision de la recherche, la précision et les performances globales du système.

Les modèles hybrides tels que le modèle d'encodage double [1] ou le modèle d'intégration interprétable [2] forment deux espaces communs (latent et conceptuel) afin de trouver la similarité entre la vidéo et le texte à des fins d'extraction, mais ces modèles manquent d'analyse

des relations inter et intra entre ces espaces. Dans la troisième partie de ce travail, nous proposons un cadre général pour l'analyse des relations entre les différents espaces de représentations d'approches hybrides, l'espace (non-explicable) latent et l'espace (explicable) conceptuel, afin de comprendre dans quelle mesure les espaces explicables se différencient des non-explicables. En outre, la thèse se lance dans une exploration complète de la complémentarité entre ces espaces.

Dans l'exploration sur l'explicabilité, la partie des modèles hybrides basée sur les concepts joue un rôle primordial. Elle identifie et extrait des concepts sémantiques à partir de données vidéo et textuelles, visualisées par des nuages de tags, ce qui rend le processus de recherche plus facile à interpréter et à comprendre. Les modèles actuels de l'état de l'art fournissent des explications à l'aide de nuages de mots-clés, mais les explications fournies ne sont pas causales. La dernière partie de cette thèse aborde le problème des l'explications visuelles causales et interprétables pour la recherche vidéo-texte. En fournissant une représentation visuelle des relations causales entre la requête et les résultats récupérés, les nuages de tags renforcent la confiance des utilisateurs et soutiennent les applications où la responsabilité et la perspicacité sont primordiales.

Acknowledgement

After the genuine period of 3 years (Oct, 2020 – Dec, 2023), this is the day composing this note of obliged, is the wrapping up touch on my Ph.D. dissertation. It has been a period of unequivocally learning for me, not only on technical level, but also at personal level. Composing this proposal has had a gigantic influence on me. I would like to reflect on the people who have supported and helped me so much all through this period.

First and foremost I am extremely grateful to my advisors, Prof. Georges Quénot (Director de Recherche CNRS), and Dr. Philippe Mulhem (Chargé de Recherche CNRS), for their invaluable advices, continuous support, and patience during my Ph.D. studies. Their immense knowledge and plentiful experience have encouraged me in all the time of my academic research and daily life.

Secondly, I am really very grateful to Jean Martinet, Stéphane Ayache, Alexandre Benoit, and Danielle Ziébelin for accepting to be the part of my thesis defence jury, and freeing themselves from their busy schedule to review my thesis manuscript. Their positive comments on my thesis encourage pursuing the further research in the field of explainable cross-modal video-text retrieval.

I am also very thankful to the Government of Pakistan, and the Higher Education Commission of Pakistan (HEC) for such a wonderful opportunity and providing the financial support to continue my doctoral studies under AHBP Project. Moreover, I would also like to express my gratitude to the French Government, and Campus France for facilitating the scholars from Pakistan, and providing the assistance whenever in need.

I would like to say heartfelt thanks to my beloved parents, brother, sister, and my relatives for always believing in me and encouraging me to follow my dreams. I would also like to thank my friends overseas in Pakistan and here in France for their presence, support, encouragement, and help in whatever way they could in this challenging period.

Finally, I would like to say thank you to my better half (husband) Dr. Pardeep KUMAR, from the core of my heart. A few words or sentences can't express his efforts, encouragement, and support through thick and thin. In short, without her, it would not have been an easy journey.

List of Publications

Article in International Conference

Devi, Varsha, Philippe Mulhem, and Georges Quénot. “Analysis of the Complementarity of Latent and Concept Spaces for Cross-Modal Video Search”. *Proceedings of the 19th International Conference on Content-based Multimedia Indexing. 2022*, Sept. 14-16, 2022 Graz, Austria.

DOI:10.1145/3549555.3549600.

Devi, Varsha, Georges Quénot, and Philippe Mulhem. "Improving Causality in Interpretable Video Retrieval." *Proceeding of 20th International Conference on Content-based Multimedia Indexing. 2023*, Sept. 20-23, 2023 Orléans, France.

Article in International Journal

Devi, Varsha, Georges Quénot, and Philippe Mulhem. "Evaluating and Enhancing Causality in Tag-Based Video Retrieval Systems" submitted in Multimedia Tools and Applications (MTAP) Journal.

Contents

List of Figures

List of Tables

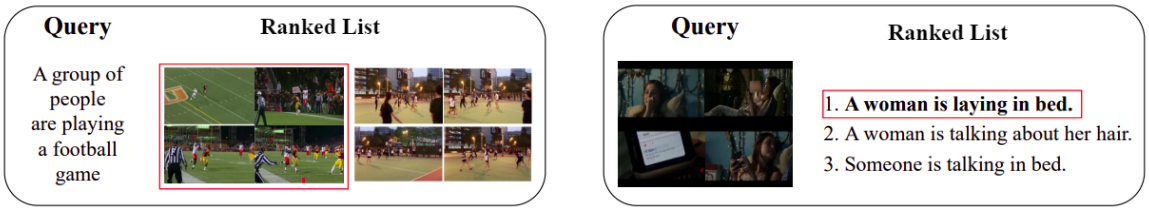
Chapter 1

Introduction

In the current era of big data, we are surrounded by information systems and applications that generate and collect millions of data from diverse sources at an unprecedented rate. On the one hand, this big data can help us better understand the world and foster innovations in various domains, whereas on the other hand, is making it more challenging and time-consuming to locate the desired content. Indeed, such data are of no value if it cannot be searched efficiently; for that, it must be appropriately indexed to be retrievable among billions of other videos in a matter of seconds.

To leverage big data, machine learning techniques, especially deep learning, emerged as powerful tools for analyzing and utilizing data on various tasks, such as image classification, information retrieval, and natural language processing. This thesis focuses on one such task: multimedia information retrieval for multi-modal data. Cross-modal retrieval allows users to find relevant data of different modalities (e.g., texts, videos) given a query of another modality (e.g., texts or videos). This thesis is in particular related to video-text retrieval. In the past, video representations primarily relied on keywords provided by content creators. However, relying on such information is no longer suitable as malicious uses have become increasingly prevalent on the internet: for instance, a user uploading a promotional video could purposely use incorrect keywords to attract more viewers. Moreover, given the scale and time constraints involved, manual operations like keyword-based annotation, representation, and searching of videos are not feasible. Deep learning models gain a lot of attention due to its efficiency and effectiveness in cross-modal retrieval field. Typically, a deep retrieval model would map data items from their original modal spaces to a common space, where their similarity can be easily computed. This way, users can retrieve desired data from different modalities with a query of one modality. Due to its high practical value, deep learning-based multi-modal retrieval has attracted much attention and become a fundamental task in both industry and academia.

The deep learning field has proposed several methods for computing compact data rep-



(a) Text-to-Video Retrieval **(b)** Video-to-Text Retrieval

Figure 1.1: Video-Text Retrieval

resentations, including designing visual descriptors such as Convolutional Neural Networks (CNNs) [1, 14, 2], Recurrent Neural Networks (RNNs) [1, 15], and Transformers [16, 17] for extracting spatial or temporal information from video frames. Moreover, for textual representations Word2Vec [18], Glove [19], LSTMs [20], and sequence models [1, 21] have been proposed. We will provide an overview of these approaches in this report.

1.1 Video-Text Retrieval

To address the exponential growth of multimedia data (i.e., videos and texts) for efficient retrieval, Video-Text Retrieval (VTR) approaches are proposed. VTR [1, 8, 2, 16, 22, 23] involves analyzing a given sentence to identify the most suitable video from a collection (namely text-to-video, or TTV), and vice versa (namely video-to-text, or VTT), as illustrated in Figure 1.1. Figure 1.1a presents one example of text-to-video retrieval: a given text query is used to retrieve a list of videos and rank them according to their relevancy score. The Figure 1.1b shows a video-to-text retrieval example, where the query is a video, and the goal is to retrieve pre-existing relevant captions and rank them in the same manner in any retrieval task. A any retrieval task, the overall objective is to retrieve relevant elements, and Figure 1.1 presents the relevant videos for text-to-video retrieval and the relevant captions for the video-to-text within red boxes.

Any VTR requires analyzing a vast number of video-text pairs, extracting multi-modal information, and determining whether the two modalities can be aligned. While visual tasks such as visual classification, object detection, and semantic segmentation have been extensively studied, and have achieved remarkable results, VTR is still a relatively new area that requires further exploration. Although the performance of VTR has improved gradually, there are still challenges and issues that require further investigation.

In recent years, deep neural networks have received significant attention from the research community due to their outstanding performance in various fields, particularly in computer vision. For cross-modal retrieval in particular VTR, deep neural networks have resulted in considerable performance improvements, making them the primary choice for solving retrieval problems. TTV and VTT retrieval involve retrieving information between

text and video modalities. As described earlier, a TTV retrieval is the task of retrieving a relevant video based on a textual query (shown in Figure 1.1a). For example, given a textual description of a scene, the system should retrieve a video clip that best matches the description. Such task is commonly used in applications such as video search, video summarization, and video recommendation. For a VTT retrieval, on the other hand, the goal is to find relevant textual information based on a video query (shown in Figure 1.1b). For example, given a video clip, the system should retrieve relevant textual description for the video. This task is commonly used in applications such as video captioning and video indexing.

These two tasks involve matching the features extracted from the text and video modalities and ranking the results based on their relevance to the query. TTV and VTT are challenging tasks due to the heterogeneity and high dimensionality of the modalities involved. State of the art VTR systems require sophisticated techniques such as deep learning-based approaches to achieve high performance. More generally, cross-modal retrieval relies mainly on three types of approaches: *concept-based approach* – a method that uses **concept Bank or vocabulary** (which is a set of pre-defined concepts and thier classifiers) to build the concept representations of video and text using a classification task. The similarity between the videos and texts is measured using a concept space, where video & text is represented by a prediction vector, *concept-free approach* – a method that learns and extracts embedding directly from the video and text and matches the features in common **latent space**, and *hybrid approach* – a method that combines both concept-based (concept space) and concept-free (latent space) methods. We will review these approaches into detail in the next upcoming chapter 2.

1.2 Explainable Video-Text Retrieval

A compelling need of video-text cross-modal retrieval is the pursuit of explainability – the ability to justify why certain retrieval results are obtained. As the complexity of models and the amount of data increase, it becomes crucial to provide non-technical or technical users with insights into the decision-making process of cross-modal retrieval systems. Explainable video-text retrieval strives to offer transparency and interpretability, enabling users to understand which aspects of the video or text influenced the retrieval outcome. This not only enhances user trust but also supports applications where accountability and insight into system behavior are essential, for instance, medical diagnosis and patient education, and recommendation systems.

In traditional video-text retrieval systems, the matching between the query and the item is based on the similarity between the low-level features of both. However, these methods lack *transparency* and do not provide insights into how the results are generated. Explainable video-text retrieval aims to overcome this limitation by using advanced techniques, such as

deep learning, natural language processing, and attention mechanisms, to not only retrieve relevant items but also provide explanations for why the retrieved items are relevant to the query. The explanations generated by explainable video-text retrieval systems can take different forms, such as tag clouds (shown in Figure 1.2), texts or heatmaps, and can provide insights into the specific aspects of the video/text content that match the query. An explanation can also highlight the objects, scenes, or actions in the video frames that are relevant to the query and show how they relate to the query using heatmaps.

1.3 Objective

In this thesis, our focus is to train an explainable hybrid model comprised of a *dual* space, composed of one latent space and one concept space, and able to perform dual tasks i.e. Retrieval and Classification. To be more precise, we aim at designing and training multiple encoding networks based on deep learning, capable of efficient video-text retrieval and justification of retrieved results at the same time. For instance, when provided with the text query, we would like to retrieve the relevant videos from the dataset, and vice versa. Additionally, the system should provide an explanation for the retrieved results in the form of tag clouds. The explanation should be more reliable and satisfiable than current state-of-the-art systems. Given a query, our model should therefore be able to explain why the list of videos is similar to the given query and vice versa, and the explanation provided should resemble the major portion of the decision-making process of the retrieval system.

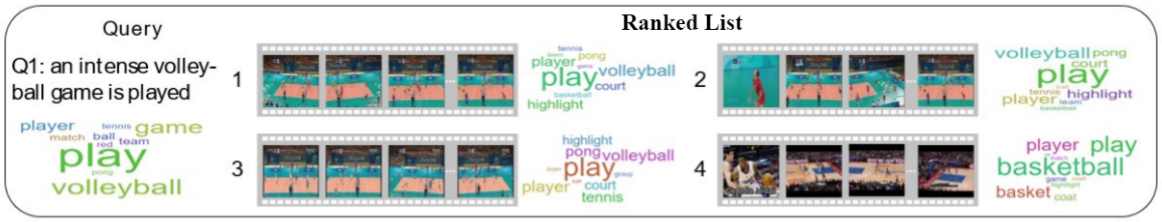


Figure 1.2: Text-to-video Retrieval. Tag clouds in front of each (query, video item) for justifying the retrieved results for one query (from [1]).

For explaining the retrieved results of a hybrid model, the concept space part of the hybrid approach model is used. The explainable video-text retrieval model involves identifying and extracting semantic concepts from both video and textual data to enable more interpretable and explainable retrieval of information. The approach aims to improve transparency and interpretability of the retrieval process by associating the retrieved results with specific concepts, making it easier for users to understand how the system arrived at the results. For example, a concept-based retrieval system may analyze video and text data to identify concepts such as “play”, “volleyball,” “basketball”, and “game,” and use these concepts to guide the retrieval process. By highlighting predicted concepts in query and videos

using tag clouds, as shown in Figure 1.2, users can better understand how the system works, why certain results were returned, and why a certain video is matched to the query. This approach can be particularly useful in applications where transparency and interpretability are important, such as in legal, medical, or security contexts.

We will evaluate the retrieval and explanation capabilities of our models on two tasks: text-to-video retrieval and video-to-text retrieval. In the remainder of this section, we will introduce these tasks and their associated challenges.

1.4 Challenges & Motivation

In the field of video-text retrieval, we face challenges that involve processing the two heterogeneous modalities: video and text. Videos capture dynamic, temporal sequences, while text provides a static, sequential representation. Bridging this gap requires addressing issues related to the fundamentally distinct nature of these modalities, as well as the need to effectively align and interpret them. Additionally, the intrinsic complexity of multimedia data further complicates the task of extracting meaningful features from visual and textual modalities in large-scale video-text datasets.

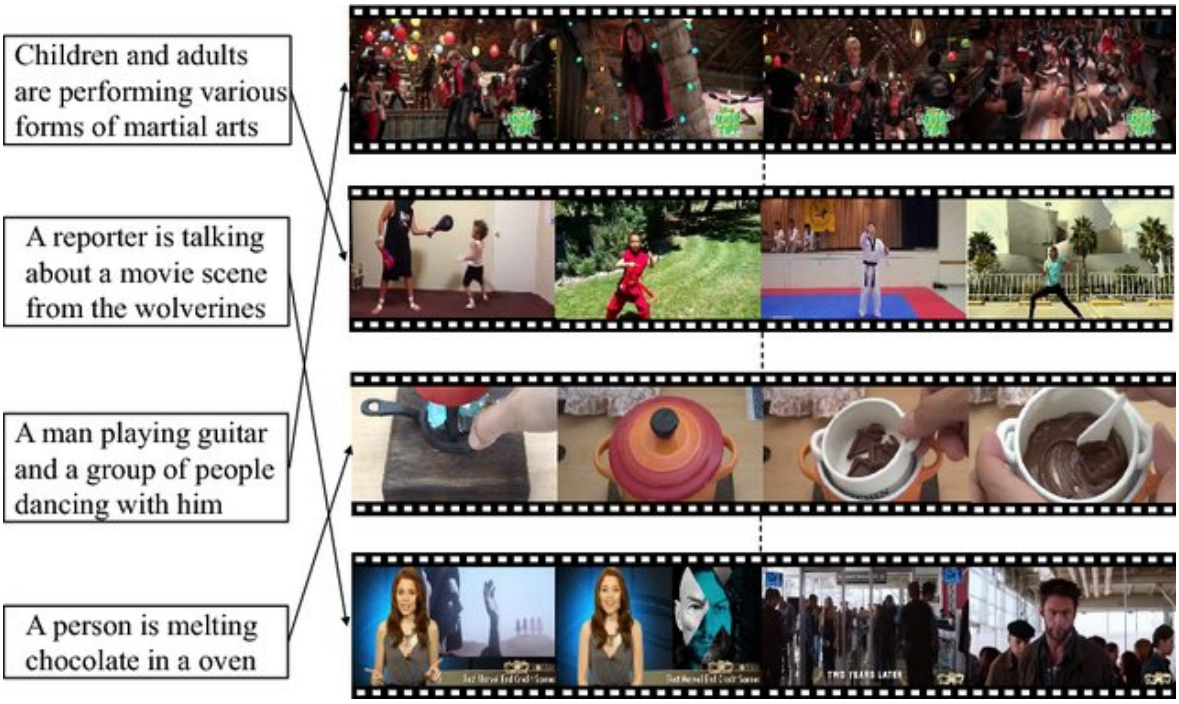


Figure 1.3: Video-text Annotated Dataset (Image courtesy of Lin et al. [3])

One of the primary challenges in video-text retrieval is ***scale***, as the video and text collection to retrieve from can contain millions or billions of videos and text, and the best video candidates should be provided to the user in less than a second. To overcome this constraint, offline computation of video and text representations is used that could be re-used for each

query (text or video). Mapping text and video to a joint embedding space allows computation of a similarity score via a single dot product, hence achieve retrieval speed [1, 24].

Another challenging aspect of this field is the ***temporality*** of video. It is important to pay attention to the relative order of events in a video to understand a scene. Videos can also vary greatly in duration, and it is necessary to encode videos of variable durations into a fixed-size representation without discarding the temporal information. This can make it difficult to encode and retrieve captions for videos with variable durations and complex temporal structures.

Last but not least, ***ambiguity*** is another challenge for video-text retrieval task especially in the case of concept-based and hybrid-based approaches where the concepts may or may not contain multiple meanings. This can result in incorrect classification, and because of which irrelevant videos or captions will be retrieved for the query. The performance of the model trained for either of the task is evaluated on a video-caption dataset as shown in Figure 1.3. Hence, efficient VTR system requires encoding models trained on a large dataset of video-captions pairs, which is both ***difficult and expensive to annotate***.

This thesis aims to address the critical challenges of scale, temporality, and ambiguity in video-text retrieval, advancing the efficiency, accuracy and explainability of multimedia search systems. By developing innovative techniques that leverage shared embedding spaces, offline computations, and cross-modal understanding, this research seeks to enhance both quality of retrieved results and explanation provided by the retrieval system, ultimately improving user accessibility, search experiences, and the interpretation of complex multimedia content.

1.5 Contributions

The following contributions will be presented in this document:

1.5.1 Development of PoS-tag dual encoding model

Chapter 3 of this dissertation significantly extends the research presented in *Dong et al.*'s work [1] by integrating Part-of-Speech (PoS) tags into their dual encoding model for video-text retrieval. Unlike many state-of-the-art retrieval models [1, 2, 25, 26] that overlook syntactic information during the encoding of high-level semantics, this chapter recognizes the value of syntactic cues in pinpointing specific user needs and intentions. The contributions encompass: i) the integration of PoS tags in textual encoding pipeline and alongside class labels as shown in Figure 1.4, ii) an exploration of the impact of PoS-tag integration on performance and explainability, and iii) a comparative evaluation that underscores the advantages of using PoS tags to enhance retrieval accuracy, precision, and overall system performance.

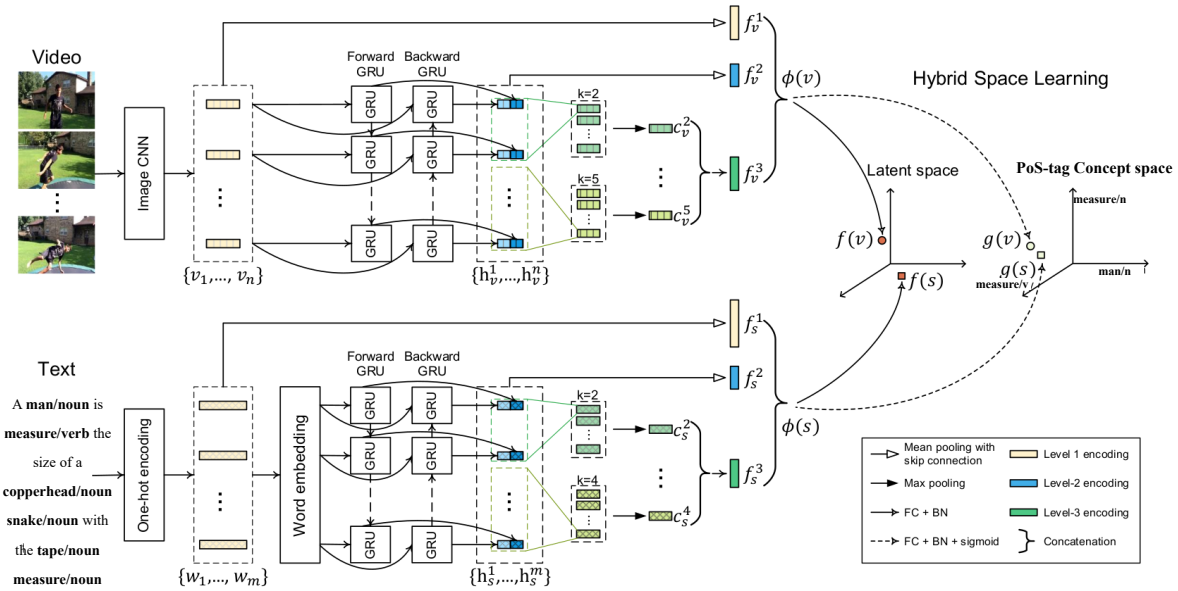


Figure 1.4: Proposed PoS-tag based Dual Encoding Architecture with PoS-tagged "text" & "concept space" (inspired from [1])

These contributions are pivotal in bridging the gap between linguistic structures and visual content, leading to a more robust and interpretable video-text retrieval system capable of capturing the nuanced semantics of language and visuals. This, in turn, is expected to elevate user satisfaction and the quality of retrieval outcomes.

1.5.2 General framework for analysis of complementarity in hybrid approach

The analysis of relationships between different spaces, such as the latent space and concept space (as shown in Figure 1.5), bears significance in investigating complex data patterns and designing efficient retrieval systems. Chapter 4 brings forth noteworthy contributions in this realm. Firstly, it introduces a general analysis framework that acts as a navigational tool for investigating the relationship between latent space and concept space, and how these spaces work together, particularly within the domain of cross-modal video search employing a dual encoding model as its baseline. This framework offers strategic guidance to explore the collaboration between spaces in hybrid approaches, ensuring their effectiveness and interpretability. Secondly, the chapter embarks on a comprehensive exploration of the complementarity between these spaces. This investigation pursues two discrete pathways: analyzing the representation power of both spaces, and complementarity analysis via Canonical Correlation Analysis (CCA) and ensemble learning. This general framework showcases the complementarity between the latent and concept spaces in cross-modal video search tasks.

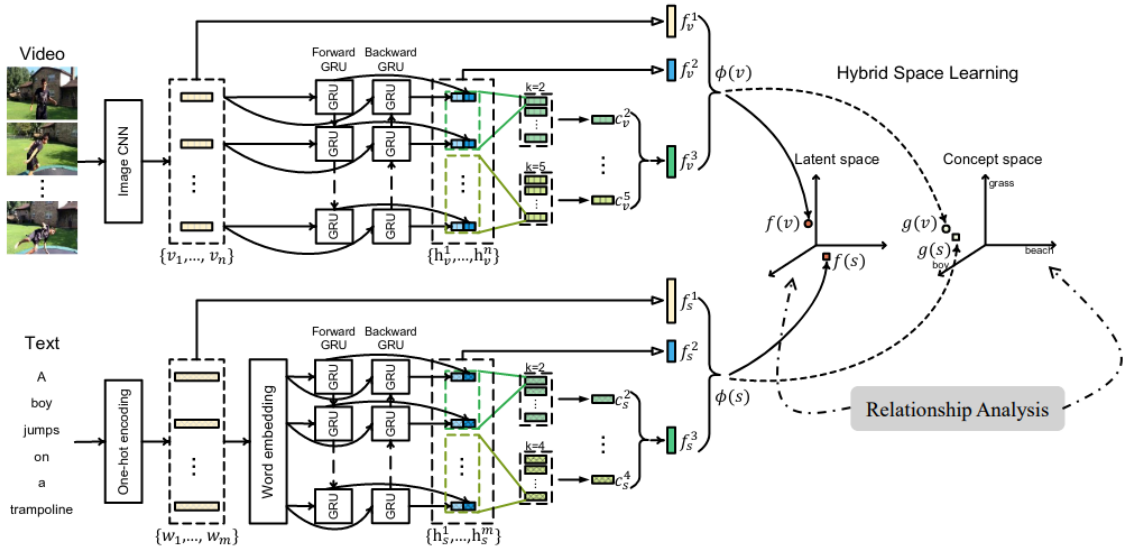


Figure 1.5: Intra and Inter relationship analysis in Dual Encoding Architecture [1]

1.5.3 Causal inference in Video-Text Retrieval

Chapter 5 addresses the problem of providing causal and interpretable visual explanations for multimedia retrieval systems that use human-readable tags. We first introduce a new evaluation measure that quantifies the degree of causality (how many tag(s) in concept based representation of video and text contributed to the retrieval decision) in visual tag-cloud explanation for concept based dual encoding video retrieval. Then, we apply this measure to a state-of-the-art video retrieval system that uses a deep neural network to generate tags from videos. We propose to enhance the causality of the state-of-the-art system by modifying the way the tag scores are computed, using a generalized sigmoid function that increases the relative effect of the top relevant tags. We conduct experiments on benchmark datasets and show that our method improves the causality measure by up to an order of magnitude, while maintaining comparable accuracy to the original system. We also analyze the trade-off between causality and accuracy, and discuss the challenges and limitations of achieving 100% causality in visual explanations. This research on causal based explanation is a preliminary work that opens new perspectives for improving the interpretability and trustworthiness of multimedia retrieval systems.

1.6 Thesis Organization

Chapter 2

This chapter surveys the existing methods and techniques for representing and retrieving multimedia content, as well as for providing explanations for the retrieval results. It also

identifies the limitations and challenges of the current state of the art, and highlights the research gaps that this thesis aims to fill.

Chapter 3

In Chapter 3, we introduce a new model for video-text retrieval that uses part-of-speech (PoS) tags to encode the syntactic information of text queries. It shows that by using PoS tags, the model can better capture the semantic relevance between videos and text queries, and improve the retrieval performance over existing models.

Chapter 4

A general framework for analyzing the complementarity of different approaches for video-text retrieval, namely concept-based, concept-free and hybrid approaches, is presented in Chapter 4. It also evaluates the effect of different fusion strategies on the retrieval results, and provides insights into the strengths and weaknesses of each approach.

Chapter 5

This chapter proposes a new evaluation measure, which is a way of providing causality in visual explanations for video-text retrieval systems. It shows that by using counterfactual reasoning, the measure can quantify the degree of causality between the tags prediction and the retrieval decision. It also applies the measure to a state-of-the-art video retrieval system, and proposes a method to enhance the causality of the system by modifying the tag scores computation. It discusses the trade-off between causality and accuracy, and the challenges and limitations of achieving 100% causality in visual explanations.

Chapter 6

We summarize in this chapter the main findings and contributions of the thesis, and suggest some possible directions for future research in video-text retrieval and explainability.

Chapter 2

State of the Art

As described in the Section 1.1, cross-modal retrieval aims at retrieving a ranked list of relevant items in one modality from a dataset for a given query in another modality. These modalities can be text, images, or videos. Most of the proposed models in cross modal retrieval field fall under one of the following approaches: (i) *latent based approach*, (ii) *concept-based approach*, or (iii) *interpretable embedding based hybrid approach*.

The end goal of this dissertation is to be able to enhance the richness in justification of retrieval models, specifically *video-to-text (VTT) retrieval* and *text-to-video (TTV) retrieval*, without any loss in accuracy. This requires understanding the content and semantics of both video and text modalities that have different representations and challenges. Video is a rich and complex medium that contains visual, audio, and temporal information, whereas text is a symbolic and structured medium that conveys semantic and syntactic information. To bridge the gap between these two modalities, various techniques have been proposed (i) for representing the content of both modalities individually, and (ii) strategies to build a model for finding the match between two heterogeneous data.

Objective

In this chapter, we will begin by examining the state-of-the-art techniques employed in image, video, and text representation. Subsequently, we will explore the approaches introduced in the field of cross-modal retrieval. Additionally, we will delve into the methods utilized for explaining the decisions made by deep learning and retrieval models. The entirety of this discussion directly relates to the contributions made in this thesis.

2.1 Content Representation

2.1.1 Image Representation

An image is a 2D matrix of pixels. The fundamental information in an image is contained in the gradient of pixel intensity, which is highest at the corners, edges, and areas with high contrast in the 2D pixel grid. Before 2012, conventional methods for image representation relied on manually crafted visual descriptors. Examples include the Histogram of Oriented Gradients (HOG) [27] and the Scale Invariant Feature Transform (SIFT) [28], which computed gradient histograms for different image regions. These descriptors were subsequently fed into classifiers such as the k-nearest neighbors (k-NN) [29] and Support Vector Machine (SVM) [30] algorithms for image classification.

Around a decade ago, Artificial Neural Networks (ANNs) regained interest due to recent advancements in large training datasets and computational power. ANNs were initially introduced with the Perceptron model [31] in 1958 by *Rosenblatt* inspired by the biological neural network, and have evolved into Multilayer Perceptrons [32] in late $'80s$, which are considered the first deep learning architectures capable of approximating any function. Back-propagation [33], a technique for training ANNs, was popularized in 1986 and involves optimizing model parameters using gradient descent. These advancements have revolutionized image representation and paved the way for further developments in deep learning-based approaches. With diversified research directions, Convolutional Neural Networks (CNNs) have emerged as a powerful variant of ANNs for image processing, utilizing convolution operations to reduce parameters while maintaining input transformation equivariance, an example of which is the Neocognitron model [34], an early CNN model. *Le Cun et al.* introduced LeNet [35], a CNN model that was specifically designed to identify handwritten alphanumeric characters. CNN models are trained in an end-to-end fashion, where the model weights are optimized using back-propagation [33], while the design of the architecture is a manual task. Unlike handcrafted descriptors, CNNs can automatically determine the optimal filters. In a multi-layer CNN, the first layers usually identify low-level features like edges, enabling the subsequent layers to learn higher semantics such as recognizing a car (Figure 2.1). The advent of large-scale datasets, such as ImageNet [36], has paved the way for deep learning approaches in computer vision. AlexNet [37], introduced in the 2012 ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [38], was one of the first deep learning models to surpass traditional approaches in image classification. Since then, deep learning approaches, including deeper CNNs, with skip connections in residual blocks [39], have dominated the field of computer vision.

In addition to CNNs, the Transformer architecture [40], originally designed for language processing, has recently been adapted to computer vision. The Transformer model uses self-

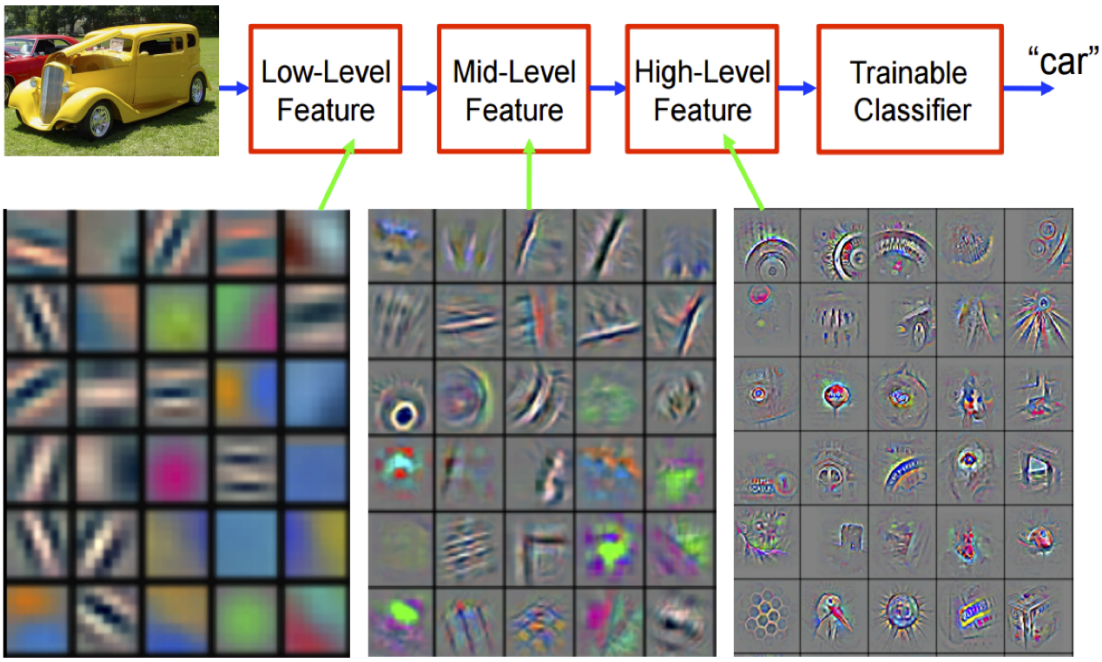


Figure 2.1: The reconstructed images are the result of projecting the feature maps of a Convolutional Neural Network (CNN) to pixel space using the deconvolutional network method [4].

attention instead of convolutions, allowing for less inductive biases. Various strategies, such as the Vision Transformer (ViT) [41] and the Perceiver [42], have been proposed to process pixels directly with Transformers at a reasonable computational cost.

To summarize, classical handcrafted descriptors like HOG [27] and SIFT [43] have been widely used but struggle to capture complex patterns and variations in data. Whereas, deep learning-based approaches, particularly CNNs, automatically learn hierarchical features and excel at capturing complex patterns, although they require large labeled datasets and high computational power. Additionally, transformers offer an alternative approach, but they are still a relatively new area of research in computer vision and often require more data and computational resources [41]. To date, deep learning methods, especially CNNs, have emerged as the leading choice due to their ability to achieve state-of-the-art results [1, 2], while the full potential of Transformers is still being explored.

2.1.2 Video Representation

Video features extraction refers to a method of representing the content of video. As video is the sequence of frames/images, the video processing methods have been heavily influenced by image representation techniques that were introduced in the previous section. However, processing videos is much more computationally expensive than processing images, as it involves analyzing many frames/images. Additionally, neighboring frames in videos are often highly similar, resulting in a highly redundant representation. The evolution of video

processing methods has followed a similar path than still images, starting with manually designed descriptors and moving towards end-to-end trained convolutional neural networks (CNNs). Compared to images which only contain spatial data, videos contain additional information due to the inclusion of a temporal dimension. This provides short-term motion information as well as long-term scene transitions, making it possible to capture dynamic changes over a period of time. Current research focuses on learning spatial and temporal representations, as well as multi-modal feature extraction due to the multiple modalities contained in videos [8, 14, 44, 45]. With the increase in computing resources and the availability of large-scale data, deep learning methods have become a popular approach for capturing video features. We discuss below deep learning approaches for video features extraction as they are the bases of this dissertation, including spatial features (section 2.1.2.1), temporal features (section 2.1.2.2).

2.1.2.1 Spatial Features Extraction

To represent a video, the first step typically involves identifying its *keyframes* (Figure 2.2). Several methods can be used for this, including random sampling (e.g., taking multiple frames per second) [46, 5, 47], uniform sampling of a fixed number of frames [48], or sparse sampling [49]. Once the keyframes are selected, the next step is to extract frame level features from them. The spatial features from such frames can be extracted similarly as described in Section 2.1.1, which are then concatenated in order to obtain video level features (as depicted in Figure 2.2). Our literature review focuses on CNNs for spatial feature extraction [50, 51, 5, 8], while also studying "transformer-based" spatial feature extractors.

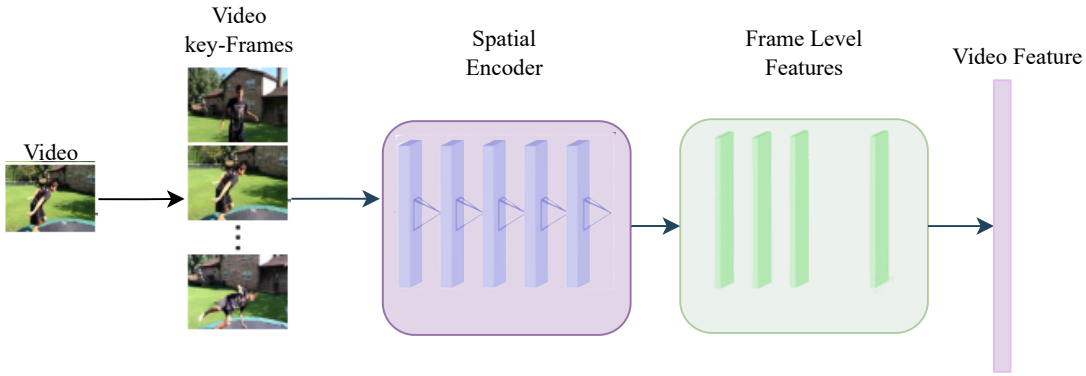


Figure 2.2: Video Feature Extraction

CNN-based methods have emerged as the state-of-the-art for various computer vision tasks, such as object recognition, image classification, and scene understanding. These methods leverage pre-trained CNNs to extract features from images, proving highly effective in capturing spatial patterns and features. Several CNN architectures, including AlexNet [37], VGGNet [52], GoogLeNet [53], ResNet [39], ResNeXt [54, 55], and DenseNet [56], are

commonly used for feature extraction from video frames. VGGNet [52], proposed by *Simonyan and Zisserman* in 2014, is widely used and consists of 16 and 19 layers of convolutional and pooling layers. Another popular CNN architecture is ResNet [39], which introduced residual connections to facilitate gradient propagation in deep networks. Inception [53] is another widely adopted architecture that combines convolutional layers with different filter sizes to capture features at multiple scales. Other CNN architectures for video spatial feature extraction, such as MobileNet [57], and EfficientNet [58] have been designed to be computationally efficient, making them suitable for resource-constrained devices like smartphones. Additionally, some methods learn feature representations from scratch using autoencoders or generative adversarial networks (GANs), which generate images from latent vectors that can serve as feature representations for downstream tasks [59, 60].

Indeed, CNN-based spatial feature extraction methods have revolutionized computer vision by providing powerful feature representation tools in the field of computer vision. However, transformer-based methods [40], particularly in the realm of visual transformer networks, have also made significant progress in recent years. Transformers differ from traditional neural networks by employing stacked encoder-decoder blocks that utilize multi-head self-attention, multi-layer perceptron, and layer normalization. Various transformer-based models, such as VSRNet [61], BiC-Net [62] and ViT [41], have been developed for video or image feature extraction, and classification. ViT [41], in particular, has gained popularity due to its exceptional performance on different datasets. However, transformer-based architectures are computationally expensive, making them less suitable for large-scale video datasets [1]. So, we focus in this dissertation on CNN-based methods for the task of video spatial representation, because of their lower complexity and computation time while building cross-modal video retrieval system.

2.1.2.2 Temporal Feature Extraction

Temporal feature extraction from videos involves capturing the temporal patterns and dynamics present in a sequence of frames. These features provide information about the motion, action, and temporal context within the video data. Once the spatial features have been extracted from videos, mean pooling or max pooling can be used to model their temporal interaction. However, more advanced architectures like, recurrent neural networks (RNNs), and convolutional neural networks (CNNs) including 2D and 3D CNNs are being used to capture temporal features. Moreover, transformers have also been utilized in recent studies to generate more complex temporal features.

In the case of RNNs, researchers have drawn inspiration from their success in natural language processing (NLP) and applied them to capture long-range temporal features in videos [8, 63, 1]. For example, *Dong et al.* [8, 1] employed bidirectional Gated Recurrent Unit (biGRU) networks to extract temporal information from both forward and backward

directions, followed by mean pooling to generate a comprehensive representation of video along the temporal dimension. Similarly, *Yang et al.* [64] used ResNet152 [39] to extract spatial features, which were then transformed into GRU to capture temporal features from video frames, and an attention module was applied to aggregate these features. However, RNNs are time-consuming to train, particularly for videos with long duration.

Other than RNNs, CNNs have also demonstrated remarkable performance in encoding temporal information. The Temporal Shift Module (TSM) [65] was introduced to shift channels between frames in the temporal dimension, effectively fusing temporal information from multiple frames. The Separable Self-Attention Network (SSAN) [66] learns spatial correlations before extracting temporal correlations, leading to improved video feature representation. For spatio-temporal feature extraction, 3D CNNs have also been widely employed. Res3D [67], SlowFast [68], and I3D [69] are examples of methods that extend 2D CNNs to 3D CNNs. They incorporate a temporal dimension while traversing channels, allowing them to capture spatio-temporal information effectively. Since 3D CNNs are computationally very intensive, pseudo-3D CNNs which contain spatial 2D CNNs and temporal 1D CNNs [70, 71, 72] replace 3D CNNs as an alternative to reduce the computational load.

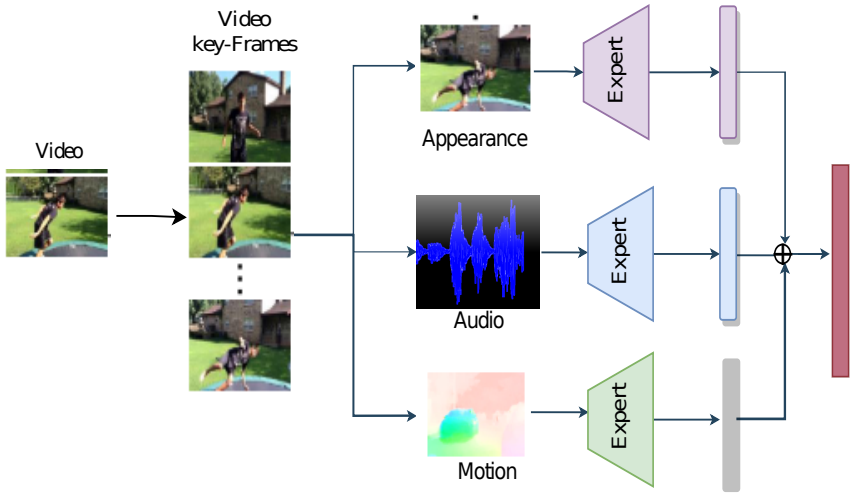


Figure 2.3: Multi-modalities representation of a video (inspired from [5])

Beside all kinds of CNNs, transformer-based models also have a strong ability to capture temporal relationships across long distances. CLIP2Video [73] models motion features between adjacent frames using a temporal transformer. COOT [46] introduces a temporal transformer to successively capture frame and clip feature interactions. X-CLIP [74] uses a three-layer transformer to encode frame features, averaging them to obtain video features. All these transformer-based models are used in various research works, for instance, *Han et al.* proposed BiC-Net [62], utilizes a multi-layer transformer to capture spatial features between adjacent frames and uses an attention-aware feature aggregation layer to fuse fea-

tures into a comprehensive representation. Additionally, CLIP [75] provides four layers of temporal transformer blocks (with frame position embedding and residual connection) that are widely used in various works [76, 77, 78]. For the past few years, researchers have explored novel transformer networks that learn both spatial and temporal features. Frozen [45] proposes stack of space-time transformer blocks that learn both temporal and spatial positions, while TimeSformer [79] investigates different spatio-temporal combinations and finds the divided space-time scheme to achieve superior performance. *Ge et al.* [47] apply divided space-time self-attention blocks to obtain fine-grained video information.

Videos contain not only spatio-temporal characteristics, but also various types of information, such as audio, optical characters, and motion. Hence, various models namely "experts" (as shown in Figure 2.3) are used to extract relevant features from each modality, which are then concatenated to form the final representation of a video, known as "Multi-Modal video feature extraction method" [80, 22, 14, 16]. For instance in [16], authors included DenseNet161 [56] for scene embeddings, SSD [81] for face features, S3D [82], SlowFast [68], I3D [69], a 34-layer R(2+1)D [72] for motion features, VGGish [83] for audio features, pixel-link text detection [84] for OCR features, and Google Cloud Speech to Text API for speech transcripts, and aggregate all of them to generate video representations composed of multiple experts features.

The discussion of encoding spatial, temporal, and multi-modal information in videos is crucial for effective semantic representation in cross-modal video-text retrieval systems. Different techniques, such as CNNs, RNNs (including biGRU networks), and transformers, have been explored. CNNs excel at capturing spatial and temporal patterns but struggle with short-term dependencies across frames. RNNs, particularly biGRU networks, are effective in modeling long-range dependencies. Transformers offer complex temporal relationship capture but come with increased computational complexity. Additionally, the multiple experts approach, which extracts features from videos and merges multiple sources, has shown promise but requires high computational resources. Understanding these strengths and limitations helps in selecting appropriate methods for video-text representation. In summary, for the goal of this thesis, transformers and multiple experts approach is not preferred because of the need of high computational resources for large-scale video-text datasets.

2.1.3 Text Representation

Textual representation aims to extract features from language sentences. The primary challenge is to model the sequential relationships accurately to capture comprehensive semantic information. For text representation, earlier bag-of-words (BoW) remains common [85, 86, 87]. BoW consists in counting the occurrence frequency of each word (token) in the text to

obtain a histogram. This histogram can then be used as a text-level representation to calculate similarities between documents, or provided as input to a classifier. The simplicity of BOW is counter-balanced by several drawbacks, mainly sparsity and insensitivity to word order. With the advent of deep neural networks in NLP, the capability of processing text has

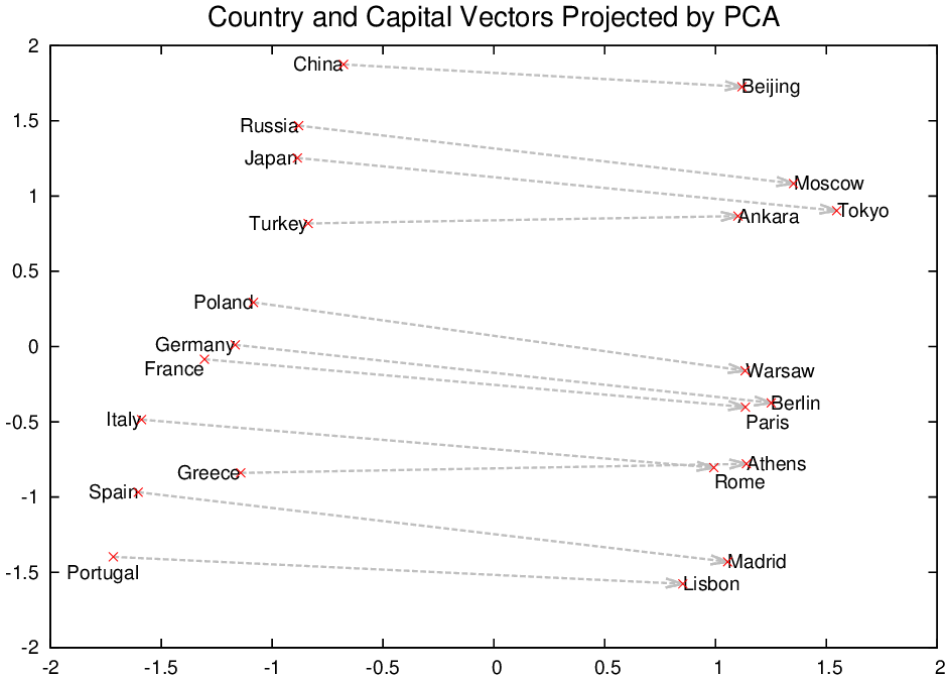


Figure 2.4: Two-dimensional PCA projection of Word2Vec embeddings of countries and their capital cities [6].

improved significantly. Word embeddings have been developed to obtain text representations at the word level instead of at the text level. Word embeddings associate each word with a vector that encodes its semantics, such that relative similarity between embeddings in vector space correlate with words' semantic similarity. An unsupervised approach for learning word embeddings from the unlabeled text was introduced by *Mikolov et al.* with Word2Vec [88]. Leveraging the skip-gram architecture, Word2Vec embeddings are randomly initialized and iteratively optimized using the representation of a given word to predict its context words. Using word embeddings, a text can then be transformed into a sequence of vectors to be processed by a language model to obtain a text-level representation. Word embeddings have been used in two ways; (i) as input features to a language model (frozen features), and (ii) as initialization of its first layer lookup table (fine-tuned features). Figure 2.4 demonstrates how the Word2Vec model effectively positions words in a vector space, allowing for concepts like 'city = capital(country)' to be represented through vector addition. This illustration highlights the capability of Word2Vec to capture semantic relationships between words.

Despite their performance, the main drawback of these methods is ignoring the sequential order in text. Another way to text representation is by using Recurrent neural networks

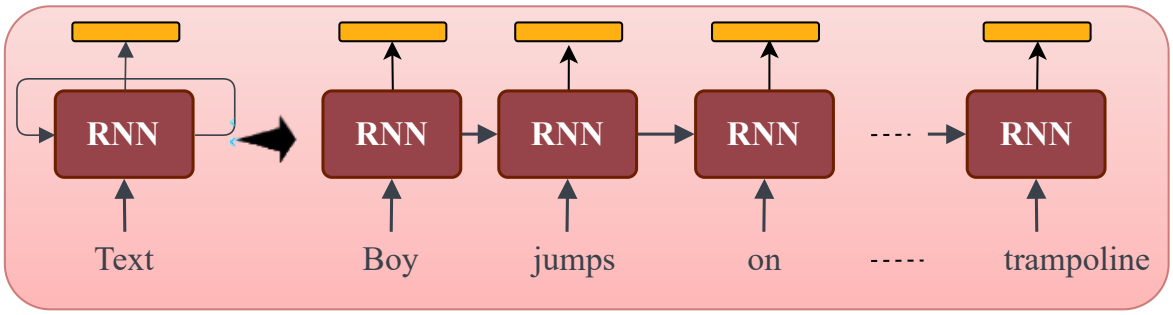


Figure 2.5: RNN-based textual feature extraction

(RNN) based techniques, which have been instrumental in capturing *long-term dependencies* in sequential data (as shown in Figure 2.5) and are well-suited for tasks where the context and order of the input sequence matter. RNNs, including variants like Long Short Term Memory (LSTM) and GRU, have proven effective in modeling the sequential relationships of text and capturing comprehensive semantic information. LSTM, introduced by *Hochreiter and Schmidhuber* [20], addresses the vanishing gradient problem associated with traditional RNNs. It allows the network to retain important information and discard irrelevant information, making it suitable for processing long sequences. GRU, proposed as a simplified version of LSTM by *Chung et al.* [89], combines input and forget gates into an update gate, offering computational efficiency while still capturing long-term dependencies. Researchers have explored various approaches to leverage RNNs for text representation, such as using bidirectional LSTMs for context-aware word embeddings and employing hierarchical attention networks (HANet) [25] to select features corresponding to verbs and nouns at individual-level representations. The combination of modified relational graph convolutional networks (GCNs) with bidirectional LSTMs has been effective in obtaining local and global-level representations [90]. Additionally, the Tree-structured LSTM [91] which captures semantic features based on the relationship between word nodes in a tree structure, has also been successful in extracting meaningful textual representations. These RNN-based techniques are especially useful for tasks such as language modeling and machine translation, where capturing long-term dependencies and context is crucial.

In recent years, transformers emerged as an alternative to RNNs in the NLP field. Transformers address the limitations of RNNs and capture global semantic information by employing self-attention mechanisms. With the introduction of Bidirectional Encoder Representations from Transformer (BERT) [92], the development of NLP has taken a huge leap forward. BERT utilizes a combination of Multi-Head Attention, Add & Norm, Feed Forward, and Residual Connection, and has become a popular choice for extracting textual features in many applications, including video-text retrieval tasks [22, 14], with BERT-Base being the most commonly applied encoder [14, 46, 93]. Due to the excellent performance achieved by BERT and its variants, several BERT-like architectures have been proposed,

such as RoBERTa [94], ALBERT [95], and DistilBERT [96].

As temporal modeling in text refers to capturing and modeling the sequential or temporal relationships between elements in a sequence. Among so many different techniques for text representation in cross-modal video-text retrieval system, RNN-based textual representation techniques could be preferable over CNN text representation, due to their ability to capture long sequential dependencies and contextual information in videos [97, 98]. Whereas, transformers excel in capturing the overall context or relationships between different elements in a sequence and are designed to process sequences as a whole, allowing them to capture long-range dependencies and contextual information effectively. However, transformers may not possess the same level of inherent temporal modeling capabilities as RNNs [92, 40]. This limitation becomes significant in cross-modal video-text retrieval, where the temporal alignment between video frames and text is crucial. Studies such as [99, 100] have demonstrated the effectiveness of RNN-based models in capturing temporal dynamics and generating accurate video descriptions.

With the noticeable introduction of image, video, and text representation, we have observed in recent years a convergence of machine learning models to process these different modalities. In the next section, we study the training of cross-modal video-text retrieval systems in the state-of-the-art based on some of these representation techniques.

2.2 Cross-Modal Retrieval

In the previous sections, we extensively investigated visual and textual representation or encoding techniques utilizing CNNs, RNNs, and transformers for both modalities. Building upon these foundational encoding methods, our focus now transitions to an in-depth study of the state-of-the-art cross-modal (video-text) retrieval models proposed in recent years. As already mentioned, such models can be categorized into three main approaches: (i) ***concept-based approach***, (ii) ***concept-free approach***, and (iii) ***hybrid approach***. The examination of these models serves several purposes: (a) to identify the cutting-edge techniques employed, (b) to establish performance benchmarks, and (c) to gain insights into the current landscape of cross-modal retrieval systems. This knowledge is critical for the rigorous evaluation of our proposed system and for meaningful comparisons against existing methods in subsequent chapters.

2.2.1 Concept-based Approaches

Concept-based video-text retrieval aims to bridge the gap between textual queries and video content by utilizing a predefined set of visual concepts (textual keywords to describe the image or video). The main idea behind concept-based approach is to build ***"concept bank"***, consisting of tens of thousands of pre-trained or newly trained Convolution Neural Networks

detectors trained on visual concepts i.e. $\mathbb{C} = \{c_1, c_2, c_3, c_4, ..., c_d\}$. These detectors are responsible for detecting visual concepts in videos or text e.g. "book, cat, dog, dance and so on", in order to generate "concept-based video representation" and "concept-based text representation" as shown in Figure 2.6.

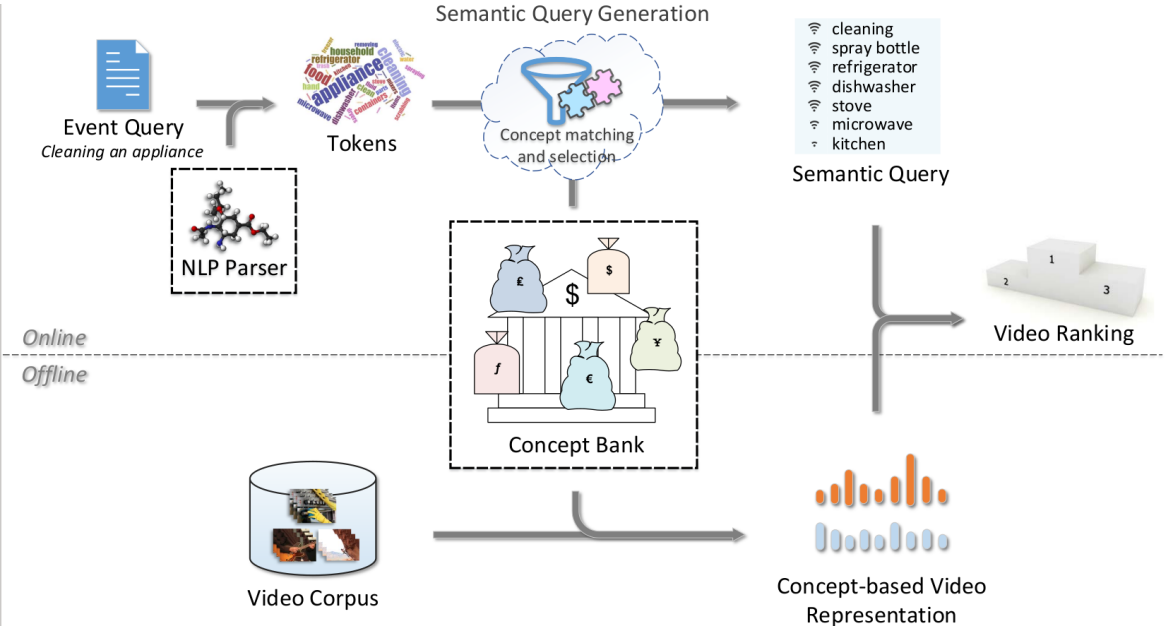


Figure 2.6: concept-based video retrieval framework proposed by *Lu et al.* [7].

Let $\mathbb{V} = \{v_1, v_2, ..., v_m\}$ be the set of m videos in the database and each video v_i is indexed by visual concept present in $\mathbb{C}$ and represented by a vector as $v_i \in \mathbb{R}^d$, where d is the total number of visual concept detectors in Concept Bank $\mathbb{C}$. The core of the framework is a *large concept bank* containing detectors of objects, scenes, actions, and activities.

There are two phases "offline" and "online": In the offline phase, all videos in $\mathbb{V}$ are represented by visual concepts in concept bank using trained concept classifiers. Each dimension in $\mathbb{R}^d$ indicates the likelihood of the presence of a specific concept in the video v_i. In the online phase, when a text query is provided, such as "cleaning an appliance" with a detailed description, the noun and verb phrases within the query are extracted as individual tokens. These tokens are then matched to a predefined set of concepts to select as many relevant concept classifiers as possible, resulting in a concept-based query representation also known as a "semantic query" (see Figure 2.6). Semantic query generation is the process, that selects the concepts from concept bank $\mathbb{C}$ relevant to the query and generate internal query representation, namely "semantic query". Finally, for each video, a score is calculated for the query sentence by integrating the scores from multiple concept classifiers which are detected for the given query [7, 101].

In many real-world video retrieval systems, such as YouTube[1], text matching plays a

[1] https://www.youtube.com/

significant role, in which titles and tags provided by video uploaders facilitate matching and retrieval. The multimedia research community has been actively promoting the advancement of video content understanding, aiming to enable machines to search based on visual concepts and high-level semantics without relying solely on human annotations. Notable efforts in this direction include TRECVid AVS challenge that conducts benchmark evaluation for this task [102]. In this challenge, among top performing solutions often employed concept-based approach by using visual concept classifiers to describe video content and linguistic rules to identify concepts in textual queries [101, 103, 104, 105, 106].

Concept-based approaches rely heavily on the trained detectors, where each detector has to be learned on large training sets. Examples of concept-based approaches are [101, 103, 106], in which authors utilize multiple pre-trained Convolution Neural Networks (CNN) models to detect multiple objects and activities in videos. As for concept-based query representation, the approaches design relatively complex linguistic rules to extract relevant concepts for a given query.

Other than user-defined linguistic rules for concept-based query representation, the visual concept classifiers for a given textual query, are also either selected manually (humanly annotated) or automatically (text embedding like Word2Vec [88], BERT [92] or using part-of-speech (POS) tagging). *Waseda et al.* [103] used both methods; manual selection of the related visual concept classifiers for the query keywords, and automatic selection using "Word2Vec" algorithm [88], to represent query in concept-based representation. They expanded their concept bank to include over fifty thousand ($50K$) concepts, and they trained SVM classifiers to automatically annotate video content along with the pre-trained CNN models. *Snoek et al.* [107] utilized a model called VideoStory [108] to represent videos using ImageNet concepts [36], and then embed concept-based video representation on a concept space by a linear transformation, while they still represented the textual query by selecting concepts based on part-of-speech tagging heuristically. Then, the video-text similarity is implemented as the cosine similarity between their concept vectors. In [26], the authors use query interpretation to perform the concept-based search. The idea is to use the BERT-based concept selection module to extract semantic concepts from a text query, and then use the query encoder to embed these concepts into the common latent space where the similarity between video and query can be calculated.

Part-of-speech tagging attempts to determine which tag a word has in the sentence e.g. "*I play/verb the main character/noun in a play/noun*". This can be used to build a concept bank as well as for concept-to-query mapping. In [109], *Shen et al.* assign a part-of-speech label to each word in a sentence, and filter out words that are unlikely to be visual objects to let their proposed text-guided object detector (TGOD) focuses more on distinguishing between possible object words. Other than just identifying the POS tags for words in order to build a concept bank, it also helps in disambiguation while mapping visual concepts to queries. The

problem is very challenging and complicated, as the part of speech can often be dependent on the meaning of the word in a sentence. For example, the sentence "*I play/verb the main character/noun in a play/noun*" highlights how the word "*play*" is used twice each having a different meaning, and parts of speech tag. The first occurrence of play is a "*verb*", and the second is a "*noun*". The concept to query mapping can be more accurate when knowing the POS tag of each concept in the concept bank, and each word in a sentence in case of mapping.

No doubt, concept-based approaches have good performance in TTV tasks when the concepts required for the textual query can be accurately identified. However, human intervention is often required in practice, in order to filter concepts after automatic mapping [7, 110], and it is also challenging to capture the rich sequential information in both video and query using only a few selected concepts. Moreover, [111] proposed to train classifiers for a combination of concepts (one joint-classifier) composed by Boolean logic operations such as "AND", "OR", etc. They call these logical combinations "composite concepts" and defined them as the logical composition of primitive concepts. The combination of concepts with Boolean logic operations makes the results of the retrieval even more explainable. However, the problem of ambiguity occurs when considering a large vocabulary of concepts to represent the complex information in the query. Despite these disadvantages, the paradigm has the merit of making retrieved results justifiable, as both the video and text modality is represented by concepts that are human-understandable.

2.2.2 Concept-free Approaches

To address the challenges posed by the ambiguity of large concept vocabularies in concept-based approaches, subsequent models have emerged based on the concept-free approach. The core idea of concept-free approach is to map the encoded (vectorized) representation of sentence/query (s), and video $(v \in \mathbb{V})$ onto the learned common embedding space (present in the right part of Figure 2.7).

Given a query s expressed by a natural-language sentence of l words ($w \in \mathbb{W}$, where $\mathbb{W} = \{w_1, w_2, ..., w_l\}$), the aim is to build a video search system that retrieves videos relevant to the query from a collection of m unlabeled videos $\{v_1, v_2, ..., v_m\}$. The key problem is to construct a cross-modal similarity function $f(s,v) \in \mathbb{R}$, such that the similarity score of a relevant sentence-video pair (s, v^+) will be larger than the similarity score of an irrelevant sentence-video pair (s, v^-), and then v^+ will be ranked ahead of the irrelevant v^- in the search results.

For these kinds of approaches, the model designed for encoding the video and text plays a vital role and is as important as learning latent space for similarity calculation. A good video-text encoder should be chosen for more accurate retrieval results. The detailed liter-

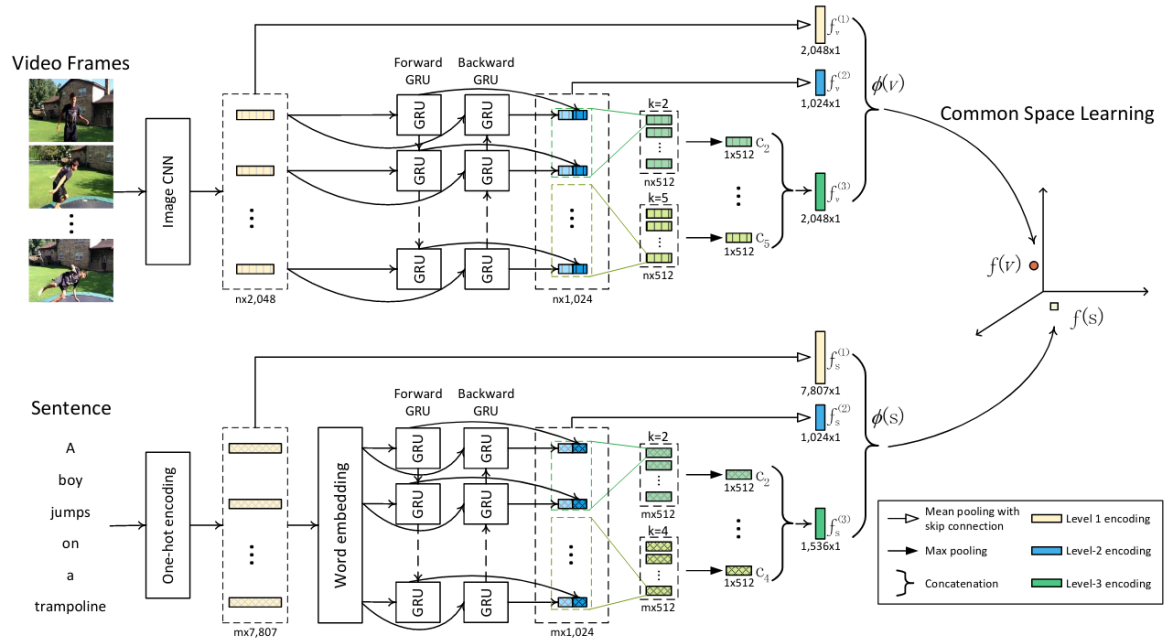


Figure 2.7: Proposed methodology using concept-free approach [8].

ature review on visual and textual representation is also given in Section 2.1.2 and 2.1.3. concept-free approaches have been proven effective in improving retrieval performance in terms of accuracy and efficiency and a significant amount of research has been carried out in the cross-modal retrieval based on concept-free approach [1, 8, 14, 22, 112, 113]. All the approaches used various video-text encoders (see Section 2.1.2 and 2.1.3) and their combinations to encode spatial and temporal information of video and text. The dual encoding model proposed by *Dong et al.* [8] aims to address zero-example video retrieval. The goal is to retrieve videos that are semantically relevant to a given text query without having any positive example of the target video. The authors propose a dual encoding framework that leverages multi-level encoding to capture global, local, and temporal patterns in both videos and sentences (Figure 2.7). At the first level, a pre-trained 2D-CNN such as ResNet-152 is used to extract frame-level features. These features are then processed through a multi-scale temporal pooling module to aggregate information from multiple frames and obtain a fixed-length video-level representation. The second level of encoding further refines the video-level representation by incorporating temporal information using a Bi-directional GRU network. Whereas, the final level of encoding involves 1-D convolutional networks on top of the Bi-directional GRU to enhance local patterns. By concatenating the outputs from all three levels, the model achieves multi-level encoding of the input video.

Furthermore, *Torabi et al.* [113] propose a joint embedding model for video understanding, and retrieval. Their approach focuses on two main tasks; video annotation and retrieval, and text generation for videos. To achieve this, the authors introduce a joint neural network architecture that consists of two sub-networks: a visual network, and a language network.

The visual network extracts spatial features from video frames, while the language network generates embeddings from natural language descriptions using a recurrent neural network (RNN). A joint attention mechanism allows the visual and language networks to attend to each other's outputs during training. *Mithun et al.* [112] presents an approach for learning a joint embedding space that captures the relationship between video and text modalities. Their objective is to retrieve relevant videos given a text query and vice versa. Their multimodal embedding model consists of a visual feature extractor and a text feature extractor, with a multimodal fusion approach to learn a joint embedding space.

Another concept-free approach "collaborative experts" [80] (also known as *multi-modal video feature extraction* as discussed in Section 2.1.2), utilizing multiple pre-trained classifiers, is used to extract specific video features such as scenes, objects, faces, and speech. Collaboration among the experts enhances the importance of relevant features, resulting in improved accuracy. The video-query similarity is calculated by a weighted sum of each expert's video-query similarity. Multi-Modal Transformers (MMT), proposed by *Gabeur et al.* [14], proposed to use BERT for query representation and a Multi-Modal Transformer with stacked transformer layers. This approach jointly encodes diverse video features for video representation. They used the collaborative experts approach [80] and generated the representation of video and text by using different experts

Concept-free approaches leverage large video captioning datasets such as MSCOCO, MSVD, MSR-VTT, and TGIF. Various models based on the concept-free approach, including Video Story [87], VSE++ [114], Word2VisualVec (W2VV) [115], and dual encoding model [8], encoded videos and text into a common latent space. These models use triplet ranking loss functions or their variants for space training. Additional loss functions, such as contrastive loss and reconstruction loss, have also been utilized to further constrain the latent space [15]. Moreover, recent research work has focused on using multiple CNN feature extractors to encode video and text, learning multiple latent spaces for each CNN encoder [16, 21, 80, 112, 116].

All these approaches show promising results but lack the interpretation and justification of retrieved results, as the latent space does not directly correspond to specific concepts or semantic meanings. Moreover, as the models are completely black box models, hence detailed analysis of the learned embeddings or joint embedding space could be valuable in understanding how the model is encoding and using the multimodal information.

2.2.3 Hybrid Approach

The combination of Concept-Based and Concept-Free approaches in video retrieval has shown potential for improving retrieval results [1, 2, 104]. *Ueki et al.*[104] conducted a study comparing the effects of these approaches, finding that their combination led to im-

proved video retrieval performance. *Dong et al.* proposed a dual encoding model [1] based on *dual space* (latent and concept space) and *dual task* (classification and retrieval) in 2021, which is the extension of their prior work proposed in 2019 [8].

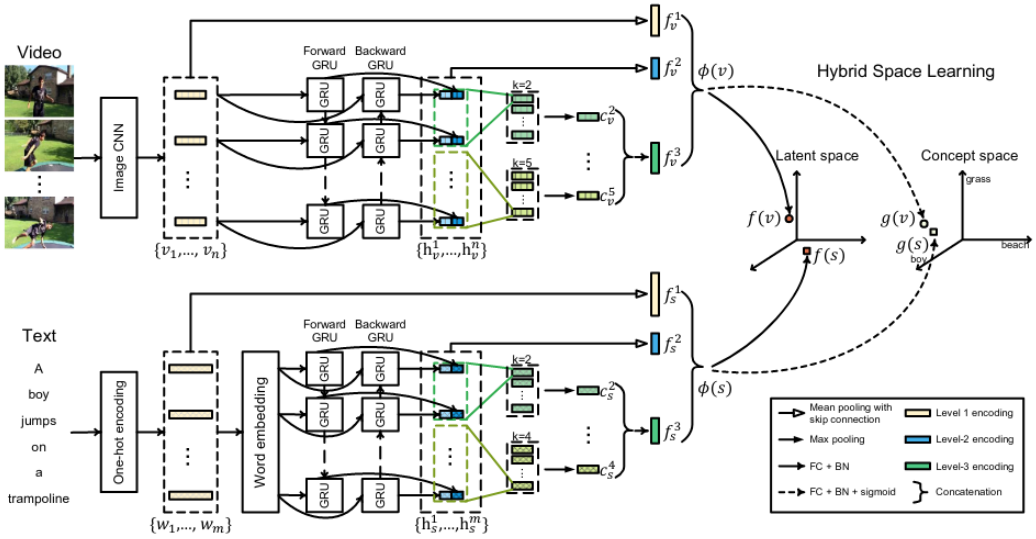


Figure 2.8: Dual encoding model based on hybrid approach [1]

This extended dual encoding model [1] is based on a hybrid approach, which extracts the video-text embedding using multi-level encoding model and simultaneously performs *video-text retrieval* and *video and text classification* tasks, as shown in Figure 2.8. This approach have shown to enhance retrieval accuracy of retrieved results [1]. The dual encoding model handles two distinct modalities: videos and sentences and has two different pipelines, i.e. i) video encoding pipeline, and ii) text encoding pipeline. It employs a dual encoding network to encode videos and sentences in parallel, facilitating latent space learning and concept space learning. Multi-level encodings are performed for each modality in their respective pipelines, including global, temporal, and local encodings. The resulting encodings, $\phi(v)$ for videos and $\phi(s)$ for sentences, describe the modalities in a coarse-to-fine fashion (Figure 2.8). These multi-level encodings ($\phi(v)$ and $\phi(s)$) are mapped to latent space using affine transformation, where the similarity between these two heterogeneous data can be computed. As the concept space is based on multi-label classification learning whose dimensions are aligned with a set of "concepts" or "tags" defined in vocabulary, the encoding ($\phi(v)$ and $\phi(s)$) are mapped to concept space using a sigmoid function. The extended dual encoding network is trained by minimizing the combination of the latent-space loss and the concept space loss. The final similarity between a video and a query is computed as the weighted sum of their latent-space similarity and concept-space similarity in order to perform the video-text retrieval.

Another similar hybrid space model proposed by *Wu et al.* [2], who uses dual encoding network proposed in 2019 [8] for the multi-level representation of video and text and latent

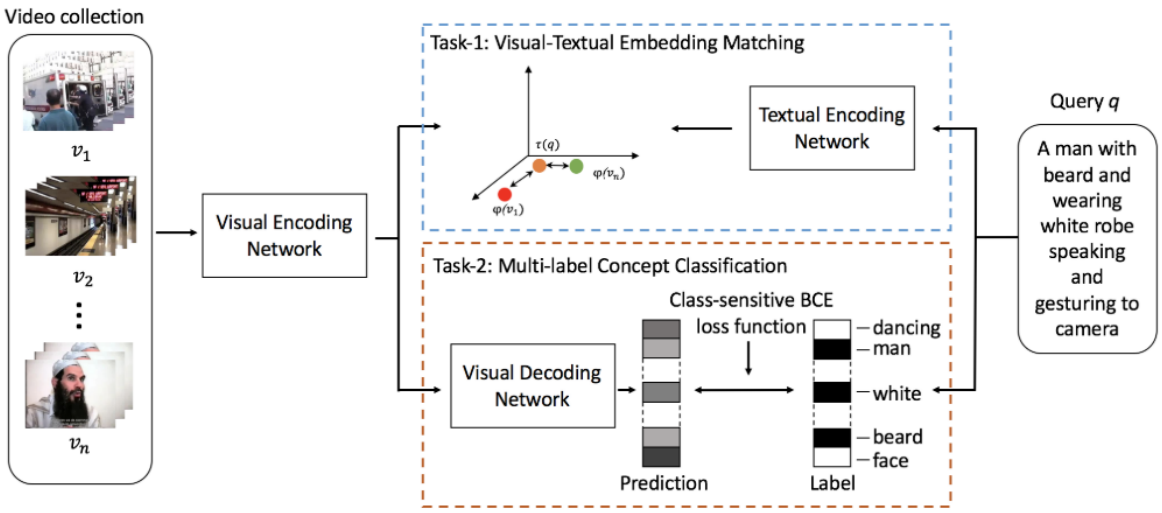


Figure 2.9: Interpretable embedding model based on hybrid approach [2]

space learning as shown in Figure 2.9 (Task-1). *Wu et al.* learned concept space and extended latent space based dual encoding model (Task-2). The difference between *Wu et al.*'s [2] and *Dong et al.*'s work in 2021 [1] lies in the training strategy of concept space. The final similarity between a video and a query is the weighted sum of both spaces (latent and concept space).

As already mentioned, TRECVid AVS conducts benchmark evaluation for video-text retrieval task, in TRECVid-2019, several proposed video-text retrieval models were based on hybrid approach [2, 15, 104]. *Ueki et al.* [104] compared the usage of concept-based approach and concept-free approach individually, and their combined effect for video retrieval from large-scale video databases using textual query sentences. For the former approach, the authors also built the concept bank comprising several concept types such as persons, objects, scenes, and actions to deal with various forms of query sentences. Using this concept bank, all concept scores for all videos were calculated. They also experimented with the latter approach and even also with the combination of these two approaches, and showed that the video retrieval results can even improve with a combination of these two techniques.

To sum up, the hybrid approach combines the strengths of concept-based and concept-free approaches, making it more comprehensive and effective. It can enhance retrieval accuracy and provide some level of interpretability and explainability. However, the hybrid approach may introduce additional complexity and computational overhead. It also relies on the availability and quality of concept annotations and concept vocabulary, which can vary in different datasets or domains. Thus, careful consideration and evaluation are necessary when adopting the hybrid approach in video-text retrieval tasks. Next, we will now explore various approaches proposed for explaining the decision-making process of deep learning models across different tasks.

2.3 Explainability

In recent years, the sophistication and complexity of machine learning models have increased significantly, leading to challenges in understanding their behavior and decision-making processes. These models are often viewed as black boxes, where only the input and output are known, and the internal process remains obscure. Therefore, the field of explainability in machine learning and/or deep learning models has gained considerable attention. The concept of explainability in machine learning is explained in different ways, as evidenced by the examples in [12, 117, 118, 119]. In [118], the authors stated that "... explainability is a broader concept referring to all actions to explain...". *Arrieta et al.* state in [117] that "... explainability is associated with the notion of explanation as an interface between humans and a system that is, at the same time, both an accurate proxy of the system and comprehensible to humans...". Another distinction is drawn in [119], stating that "...explanation provides information that gives insights to users as to how a model came to a decision...". These points of views highlight conflicts in the specific definitions and understanding of the concept of explainability: While *Markus et al.* [118] emphasize the broad nature of explainability, *Arrieta et al.* [117] emphasize on accuracy and human comprehensibility, and *Akata et al.* [119] focus on providing insights into the decision-making process. In [12], *Saeed et al.* aims to clarify the notion of explainability by offering the following definition which is similar to *Akata et al.* [119]: ***"...Explainability provides insights to a user to fulfill a need..."***. In this dissertation, we will stick to the definition of explainability provided by *Saeed et al.* [12].

This approach of explainability in machine learning is formally referred to as Explainable AI (XAI), which applies to all areas of artificial intelligence. The ultimate goal of XAI is to enable researchers, developers, domain experts, and users to better understand the complex and non-linear behavior of deep learning models while preserving their high accuracy and performance. However, achieving explainability in deep learning models remains a significant challenge due to the complexity and non-linearity of these systems. As a result, researchers and developers are actively exploring novel methods and techniques to enhance the explainability of these models. The growing demand for XAI emphasizes the importance of understanding the decision-making process of deep learning models, particularly as they continue to play an increasingly important role in our lives.

There is usually a tradeoff between model accuracy and model explainability [117]. However, various XAI techniques can be found in the literature, and most of these are dedicated to deep learning models. Generally, machine learning models can be classified into two major categories (as shown in Figure 2.10): (i) interpretable/transparent models, and (ii) opaque models. Approaches such as *Decision Trees, K-Nearest Neighbors, Bayesian Models* etc. are part of transparent models and can easily achieve explainability, while opaque mod-

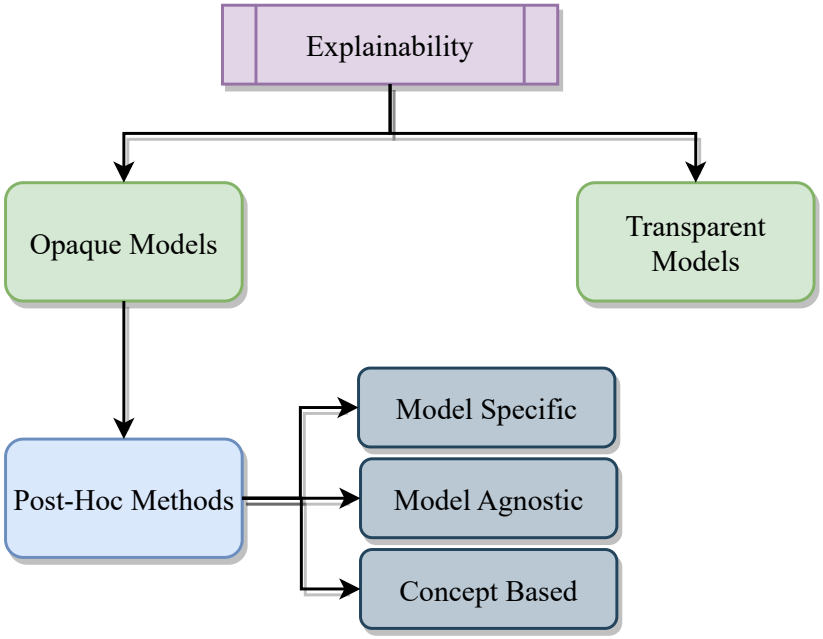


Figure 2.10: The modified high-level ontology of explainable artificial intelligence approaches inspired from [9]

els consisting of approaches like *Ensemble Method, Support Vector Machine, Deep Neural Network* etc. require post-hoc approaches to make them explainable. Post-hoc explainability refers to the process of explaining the behavior and decision-making process of a machine learning model after it has been trained and deployed. Figure 2.10 depicts the ontology of the XAI taxonomy, along with the categories of post-hoc approaches. Moving forward, we will review some of the main XAI techniques and methods that can be used to explain different types of opaque ML models, with respect to their characteristics provided in Table 2.1. These methods are:

- **Model specific methods**: These methods are tailored to specific types of ML models. These methods are not applicable to all ML models, but only to a specific type or group of models, such as *DeepLIFT [120], GRAD-CAM [121]*, by exploiting the inherent structure or properties of the models to generate explanations (see Section 2.3.1).

- **Model agnostic methods**: These are methods that can be applied to any type of ML model, regardless of their internal structure or complexity. They treat the models as black boxes and generate explanations based on their inputs and outputs. *LIME [122], Anchor [123], Counterfactual [124]* etc. are some examples of model agnostic methods (see Section 2.3.2).

- **concept-based methods**: These are methods that generate explanations based on high-level concepts or features that are meaningful to humans. They aim to bridge the gap

between the low-level representations used by the models and the high-level semantics understood by humans (see Section 2.3.3).

2.3.1 Model Specific Methods

Model-specific methods are tools for explanation of a machine learning model, but their *scope* is limited to specific types of models for which the explanation methods are developed. These methods exploit the structure or properties of the model to generate explanations, hence they are not inherently *coherent* because their coherence or consistency depends on the specific characteristics and decision-making processes of each individual model being explained. Imagine you have a convolutional neural network that can classify images into different categories, such as animals, plants, or vehicles. How do you know which regions in the image are most relevant for the network to make its predictions?. One way to find out is to use model-specific methods with heatmaps-based visualization techniques. For instance, DeepLIFT [120] assigns importance scores to pixels in an image by comparing the activation of each neuron in the network with the input image to a reference image. Backpropagation is then used to calculate the contribution of each pixel to the difference in activation, revealing the relevant pixels for the network's decision. Similarly, Class Activation Mapping (CAM) [125] identifies discriminative regions in convolutional neural networks (CNNs) by applying global average pooling to create a class activation map. This map highlights the regions used for classification.

Figure 2.11: Activation Map for Class Siberian Husky (ImageNet Class #250)[2]

As shown in Figure 2.11, we can interpret the class activation map as a heatmap in which the regions in red are the most salient for a particular prediction, and the regions in blue are the least salient. While CAM has limitations for tasks like visual question answering (VQA) because of global average pooling layer, Gradient-weighted Class Activation Mapping (Grad-CAM)[121] overcomes them by extending CAM using gradient information to generate a localization map.

[2] https://www.pinecone.io/learn/class-activation-maps/

Other model-specific methods offer alternative perspectives; such as Guided Backpropagation [126] that visualizes regions that activate specific neurons within a CNN, but may overlook negative input features and produce noisy heatmaps. Integrated Gradients [127] attribute importance to pixels or features by integrating gradients along a path from a baseline to the target image, but it can be computationally expensive, and is limited to retrospective explanations. SmoothGrad [128] aims to reduce noise in attribution maps but has downsides like over-smoothing and inefficiency, particularly for larger datasets or complex models.

Overall, all these methods, while *trustworthy* and valuable for explaining image-based tasks, have limitations in terms of high time complexity and no *simplicity* in understanding the explanations, user-friendliness, and applicability to videos. Moreover, these methods also require a certain level of *domain expertise* to interpret the provided explanation. Hence, the research lays the foundation for exploring alternative approaches to achieve interpretability in video-text retrieval systems.

2.3.2 Model Agnostic Methods

Model-agnostic methods for explainable AI (XAI) aim to provide insights into the inner workings of machine learning models by producing quantitative visualizations of how the model predictions are calculated. The *scope* is not limited and can be applied to any machine learning model, regardless of its internal structure or algorithm. These methods aim to produce quantitative visualizations of how model predictions are calculated, making it easier and *simple* for humans to understand and somewhat *trust* the decision-making process of these models. LIME [122] and Kernel SHAP [129] are two of the most commonly cited post-hoc model-agnostic techniques in the literature [130]. Local Interpretable Model-agnostic Explanations (LIME) [122] is a technique that aims to provide explanations for the predictions made by any classifier, irrespective of its complexity. It achieves this by creating an interpretable model that approximates the original classifier, specifically for a given input, and generates local explanations by perturbing the input sample within a neighborhood of a local decision boundary. This allows for the identification of the most influential features that contributed to the classifier's decision for that particular input.

As illustrated in Figure 2.12, the prediction for the original image is first calculated using a deep learning model i.e. P(tree frog = 0.54), then it is divided into interpretable components (contiguous superpixels), and a dataset of perturbed instances is generated by turning some of the interpretable components "off" (in this case, making them gray). For each perturbed instance, the probability is calculated according to the deep learning model. After that, a new locally weighted linear model is learned on this perturbed dataset. In the end, the superpixels with the highest positive weights are presented as an explanation graying out everything else. Moreover, LIME enables the replacement of the underlying "black box"

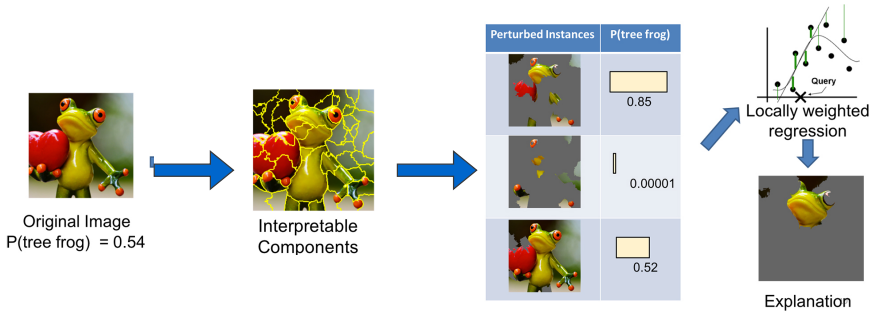


Figure 2.12: Local Interpretable Model-Agnostic Explanations (LIME)[3]

model while keeping the same local interpretable model for the explanation. LIME can work for various types of data, including tabular, text, and images. However, as LIME is an approximation model, and when learned it can provide a good approximation of local behavior, but it may not have a good global approximation, a characteristic known as local fidelity. Additionally, there is no consensus on the boundary of the neighborhood for the local model, and sometimes, it provides very different explanations for two nearby data points. Besides, an extension of LIME called Anchor [123] uses a rule-based approach to overcome some of the limitations of LIME. Anchor maximizes the likelihood of how a certain feature contributes to a prediction, and it introduces "IF-THEN" rules as explanations, along with the concept of coverage, which helps the decision-maker understand the range within which the generated explanations are valid. *SHAP (SHapley Additive exPlanations)* proposed by *Lundberg et al.* [129] is another local explanation based model-agnostic method. It is based on game theory and shapley values for model interpretability. The SHAP method explains the contribution of individual features to a model prediction by treating the data features as players in a coalition game, and using shapley values to distribute the payout fairly. This method can be applied to individual categories in tabular data or groups in images. One limitation is that it requires calculating all possible permutations of the input features, which can become computationally expensive for high-dimensional datasets. Researchers have also improved the SHAP method by addressing its limitations, such as generating counter-intuitive explanations and handling dependent features [131, 132, 133].

Counterfactual Explanations utilize the General Data Protection Regulation (GDPR) to provide counterfactual explanations for automated decisions made by machine learning models [124]. The authors, *Wachter et al.*, propose an intervention-based framework that generates explanations without accessing the black box model. By identifying interventions that could have led to different outcomes and comparing them with the actual outcome, transparency and accountability in automated decision-making systems can be improved while addressing GDPR-related concerns [124]. However, challenges such as the need for accu-

[3] https://tinyurl.com/3e8w6j6b

rate data, reliability, and the identification and testing of interventions exist. Nonetheless, Counterfactual Explanations contribute significantly to the field of explainable AI, particularly in the context of data privacy and protection [124].

In short, model-agnostic approaches have drawbacks that can affect their effectiveness and reliability. Firstly, these techniques often rely on approximations, which means they may not accurately represent the behavior of the original model. Secondly, defining the boundary of the local model can be challenging, leading to inconsistent and moderately *coherent* explanations for nearby data points. They may also sacrifice performance for interpretability, resulting in decreased predictive accuracy. Additionally, model-agnostic methods can be computationally expensive *with low-to-moderate time complexity*, especially for high-dimensional video datasets. Visualization of such explanations is also not feasible in the case of video-text retrieval based on classification tasks.

2.3.3 Concept-based Methods

concept-based methods, based on classification task, are also used in the literature to provide justification for retrieved results.

Section 2.2.3 already mentioned that the combination of Concept-Based and Concept-Free approaches (section 2.2.1 and 2.2.2) holds promise for improving video retrieval results while achieving *explainability* of retrieved results. In concept-based approaches (section 2.2.1), the video-text features are in-fact probability scores corresponding to visual concepts or tags, which are understandable to human. Using probability scores for similarity calculation, provide some level of explanation of retrieved results with *low-to-moderare* complexity and high *simplicity*. Hence, the late fusion of concept-based and concept-free spaces, namely hybrid approach (section 2.2.3), has become the norm in TRECVid benchmarking [1, 2, 103, 104, 134, 135, 136] that support partial explainability (as results coming from latent space are opaque). The *scope* of these kinds of explanation methods are not limited to one model only. These methods are generalized and can be applied to model based on classification task.

Liao et al. [10] addressed this issue and presents an explainable hybrid based method for retrieving fashion products based on a combination of visual and textual information. The main idea is to use a hierarchical structure of fashion concepts, called EI (Exclusive and Independent) tree, to guide the end-to-end learning of the model. The EI tree organizes the fashion concepts into multiple semantic levels and imposes exclusive and independent constraints on the sibling concepts. For example, at the category level, a product can only belong to one category (exclusive), and the categories are independent of each other (independent). The EI tree helps to learn an explicit hierarchical similarity function that can measure the semantic similarities among fashion products at different levels. The paper also integrated

the concept-level feedback from users in their interpretable fashion retrieval model. For example, if a user searches for a dress with a floral print, the system can show the most relevant results and also allow the user to refine the search by selecting or deselecting some concepts, such as color, style, or material.

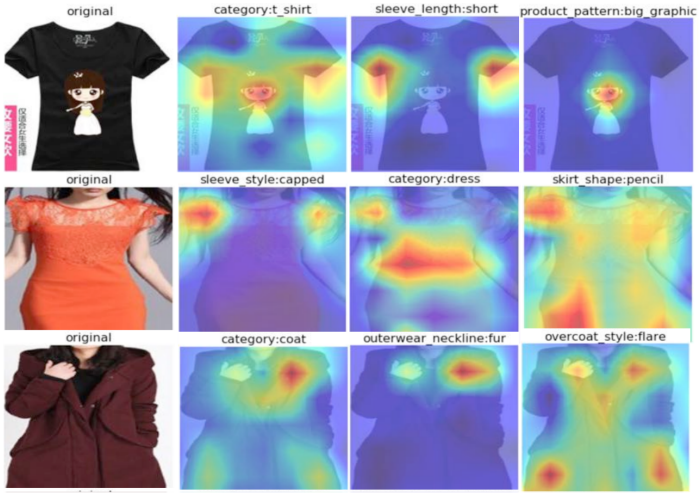


Figure 2.13: Concept localization examples [10]

The proposed approach [10] contains two main components: a "joint embedding module" and a "weight matrix module". The joint embedding module is responsible for generating embeddings that capture both visual and textual features of fashion products. The weight matrix module is used to assign weights to different features learned from EI concept tree and used to weigh the contributions of different features to the embeddings. This enables the generation of explainable embeddings. By analyzing the weight matrix, one can gain insights into which features are most important for the embeddings and how they contribute to the retrieval process. The authors validated the learning of multi-level concepts using heatmaps, as shown in Figure 2.13. The authors argue that this provides a more explainable approach to concept-based multimodal retrieval. The explainability visualization technique of [10] only explains concepts on one image at a time: in case of concept explanation on video-level, such approach is not feasible, as highlighting regions in all frames of videos can be computationally expensive.

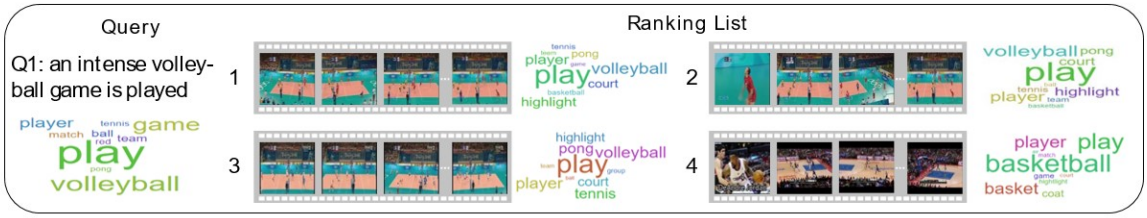


Figure 2.14: Tag clouds for justifying the retrieved results for one query [1]

Wu et al.[2] and *Dong et al.* [1] propose an interpretable hybrid space model for video-text retrieval which support explainability. In the concept space of hybrid model, video-text matching is implemented using *only* the concept classification scores, the system operation is then explainable as retrieval decisions (based on similarities) use only classification scores corresponding to tags meaningful to humans. These kinds of concept-based justification can be visualized with the help of tag clouds (as shown in Figure 2.14). By looking at the tag clouds for query and each retrieved video in Figure 2.14, the user can evaluate the common visual concepts between video and query, and on the basis of similar concepts, the user can judge why such video is retrieved. As the visual concepts in the tag cloud are human-understandable, a user can interpret the results and is not required to have *domain expertise*.

2.4 From Explainability to Interpretability and Causality

As seen above with the concept of explainability, there is also a lack of consensus in the literature regarding the meaning of the term "interpretability". While the term "explainability" and "interpretability" are frequently used interchangeably, some papers [117, 118, 119, 137, 138, 139] make a distinction between them. Four such definitions from [117, 118, 119, 12] are already presented in Section 2.3. The complete definition for explainability and interpretability are as follows: In [118], *Markus et al.* stated that "we consider interpretability a property related to an explanation and explainability is a broader concept referring to all actions to explain". *Arrieta et al.* [117] stated that, "interpretability is the ability to explain or to provide the meaning in understandable terms to a human, while explainability is associated with the notion of explanation as an interface between humans and a system that is, at the same time, both an accurate proxy of the system and comprehensible to humans". Another distinction is drawn in [119] by *Akata et al.*, in which the authors stated that "In the case of interpretation, abstract concepts are translated into insights useful for domain knowledge (for example, identifying correlations between layers in a neural network for language analysis and linguistic knowledge). An explanation provides information that gives insights to users as to how a model came to a decision or interpretation".

The definitions provided in these papers are not entirely clear, and general for all users and applications. There is still a significant amount of uncertainty. In [12], *Saeed et al.* aims to clarify the difference between explainability and interpretability by offering the following distinction: ***"Explainability provides insights to a user to fulfill a need, whereas interpretability is the degree to which the provided insights can make sense for the user's domain knowledge"***. In this dissertation, we will use this definition of explainability and interpretability provided by *Saeed et al.* [12].

As the definition provided by *Saeed et al.* [12], clearly states that the aim of explainability is to provide "insights" that help the "users" to fulfill a specific "need", here **"Insights"** refer

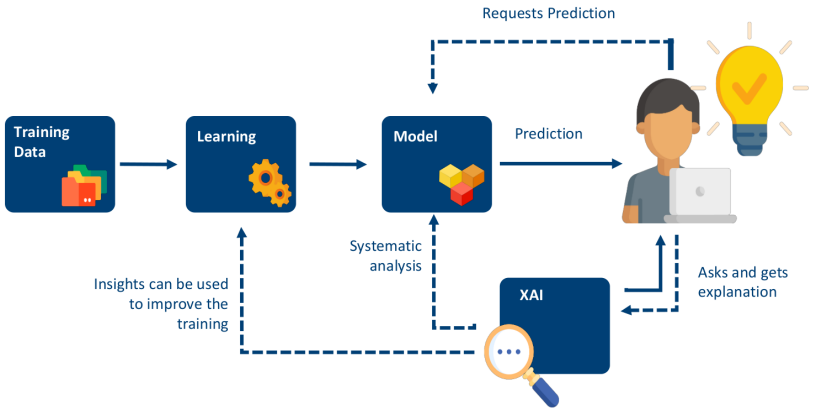


Figure 2.15: Process of machine learning based prediction with an additional XAI component to explain the results to the users (taken from [11])

to the output generated from explainability techniques like tag-clouds, text explanations, feature relevance, and local explanations. These insights are provided to a **"user"**, which could include domain experts or non-technical/common users of the application. The **"need"** for these insights could be to address various concerns, such as justifying decisions, discovering new knowledge, improving black-box AI models, or ensuring fairness. In Figure 2.15, XAI component, which tries to use models to explain the results to the *end user*. Here *insights* are being used to fulfill the *need* of improving the training of model. On the other hand, interpretability is concerned with whether the provided explanations are consistent with the user's knowledge?, if they make sense to the user?, if the user can reason and infer based on the explanations to support decision-making?, and whether the provided explanations are reasonable for the model's decision?. To ensure interpretability, a model must provide explanations that are logical to decision-makers and accurately reflect the true reasons for the model's decisions, and those should make sense to the end user.

Currently, XAI models that attempt to interpret a pre-trained black-box model (known as model-agnostic models) build interpretable models around local interpretations by approximating the predictive black-box instead of reflecting the true underlying mechanisms of the black box [122, 129]. This approximation is based on computing correlations between individual features, and it may lead to suboptimal or even erroneous explanations for decision-makers because it cannot disentangle correlation from causation [140]. However, finding causal relationships between features and predictions in observational data is a challenging task that is essential for explaining predictions [141]. Humans rely heavily on causality in their understanding of the environment [142].

Causability is defined as ***"the extent to which an explanation of a statement achieves a specified level of causal understanding with effectiveness, efficiency, and satisfaction in a specified context of use"*** [141]. Thus, *causality*, the inherent relationship between cause and effect, can be viewed as a property of human intelligence, whereas explainability is a

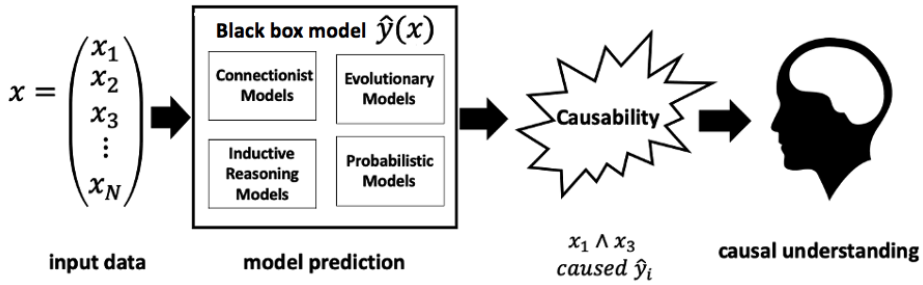


Figure 2.16: The notion of causality: given a predictive black box model, the goal is to create interpretable and explainable methods that will provide the user a causal understanding of why certain features contributed to a specific prediction [12]

property of artificial intelligence [143]. Figure 2.16 illustrates the concept of causability and causality in the context of XAI. In contrast, AI systems have been primarily designed to solve pattern recognition problems, rather than building causal models of the world that support explanation and understanding [144]. Therefore, there is a growing emphasis on developing AI systems that can build causal explainable models to support explanation and understanding [145]. However, the challenge lies in making computer-generated explanations causally understandable to humans [141].

Causality is explored in different research domains, such as activity classification from video pairs and providing explanations for machine learning model decisions [146, 147, 148]. [148] introduced a causality-inspired Video-Moment Retrieval (VMR) framework that builds a structural causal model to capture the true effect of query and video content on the prediction. Moreover, along with causality, interpretability is also necessary for explainability, since a clear understanding of the model's decision-making process and the factors that influence it are essential for providing a meaningful explanation. The concepts of interpretability, causality, and explainability are not mutually exclusive, and they can coexist in the same model or system. However, there may be cases where a model is interpretable and causal, but still difficult to explain due to the complexity or size of the model, or the difficulty of translating technical information into a more accessible format.

In the case of cross-modal video-text retrieval, the end user can be a computer scientist or non-technical user, the way of explaining the retrieved results, for example using heatmaps in model-specific methods (Figure 2.13), or just showing tag-clouds in concept-based methods (Figure 2.14) with misleading text sizes or colors, the explanation may not represent the *causality* between tag-clouds and decision-making process of system in the explanation. Moreover, it is up to the user to compare the tag-clouds of queries and videos and build an understanding of the justification provided by the system. The authors of [1] provided justification with the help of tag-clouds, but there is no information present in the tag-cloud related to the contribution of each tag in making the whole retrieval decision. No causality

analysis has been conducted between (i) the classification module of the system, and (ii) the retrieval results based on concept space.

2.5 Discussion

In this chapter, we discuss methods for generating content representations of images, videos, and text for cross-modal retrieval (Section 2.1). Various approaches are employed to obtain video representations, considering spatial and temporal aspects. Architectures based on CNNs, Transformers, or RNNs are used to process features extracted from video frames, with the aim of capturing fine-grained interactions. RNNs extract long-range features but suffer from gradient issues, while CNNs enable parallel computation but struggle with long-distance features. Transformers are powerful but lack interpretability. For text representation, RNNs suffer from context dependence, while CNNs are efficient but struggle with long-distance features. Transformers are preferred for their ability to extract long-term features, but they are not explanation-friendly.

The cross-modal retrieval system focuses on tasks like TTV and VTT, employing concept-based or concept-free approaches. While concept-based approaches provide explanations, they may not capture rich sequential information and pre-defined concept bank introduce ***ambiguity***. Nonetheless, recent studies [7, 101] have expanded the concept bank to include more concepts and utilized text embedding algorithms like Word2Vec and BERT to automatically select related visual concept classifiers for the query keywords to overcome these issues. Furthermore, hybrid approaches that combine different spaces and tasks have emerged as promising solutions in the field of video retrieval. These approaches have demonstrated that the combination of the latent space and the concept space leads to improved accuracy compared to using each space individually. It is claimed that the concept space not only provides interpretability to the retrieved results but also complements the latent space, enhancing the overall performance of the retrieval system [1]. Hence, the potential weakness to consider is the ***lack of detailed analysis of the learned embedding spaces or feature representations*** to support such a claim. This detailed analysis could also be valuable in understanding how the hybrid model is encoding and using multimodal information, leading to high accuracy. However, despite these advancements, there is still an issue of ***ambiguity*** when defining concepts within the concept vocabulary. This ambiguity not only affects the classification of video and text but also hampers the interpretation of the retrieved results.

For “explainability”, “interpretability”, and “causality”, various methods were discussed based on three main methods: model-specific, model-agnostic, and concept-based methods. Explainability helps developers understand how a model works, while interpretability provides users with transparent and trustworthy results by explaining the causal mechanisms behind the model’s output. Each explainability method has its own advantages and limitations

Characteristic	Model-Specific	Model-Agnostic	concept-based
Explainability	Heatmaps, decision trees	LIME, SHAP, anchor explanations	Tag clouds, topic models
Scope	Limited to specific models	Applicable to any model	Applicable to any domain
Coherence	N/A	High	High
Simplicity	N/A	High	High
Granularity	High	Low	Low
Causality	N/A	Moderate	**N/A**
Trust	High	Moderate	Low
Domain Expertise	High	Low to Moderate	High
Scalability	Low to Moderate	High	High
Time Complexity	High	Low to Moderate	Low to Moderate
Quantitative Analysis	Limited to model-specific metrics	Can be applied to both model-specific and model-agnostic metrics	**N/A**

Table 2.1: Comparison of model-specific, model-agnostic, and concept-based explainability techniques in machine learning and AI.

(see Table 2.1) and can be applied to different scenarios and applications. Model-specific methods identify important features or inputs for a model's predictions, providing insights into the decision-making process and highlighting potential biases or areas for improvement. However, they may not capture complex feature interactions or the non-linearity of the model, thus limiting their ability to provide ***causal explanations***. On the other hand, Model-agnostic methods can be applied to any machine learning model, offering local explanations for individual predictions or global explanations for overall model behavior. However, these methods may not provide faithful explanations or a complete understanding of how the model works. Explanations from model-agnostic methods are approximations, making them less reliable, which also means the original model cannot be fully trusted. Additionally, model-agnostic methods may not generate intuitive or human-interpretable explanations.

To generate more human-friendly explanations, concept-based methods have been proposed. These models use high-level concepts or abstract representations to provide insights into the decision-making process. They focus on semantically meaningful explanations rather than low-level features or parameters. Some concept-based methods, such as those based on hybrid space [1, 2], utilize visual concepts to guide learning and retrieval mechanisms of model, offering interpretable and semantic explanations for retrieved results. Current concept space methods for generating tag-cloud explanations [1] lack thorough causality-based explanations, including the extent of tag(s) contribution to retrieval decisions and how the concepts are derived or learned.

To sum up, there is no single method that can provide perfect explanations for all scenar-

ios and applications. Therefore, it is important to consider the trade-offs between different methods and techniques, and choose the most suitable one for the task of TTV and VTT retrieval for instance tag-cloud based explanations [1]. Furthermore, it is also important to evaluate the quality and effectiveness of the explanations generated by different methods, and compare them with human expectations and preferences.

As a result, several questions arise with this discussion:

1. Can we enhance the concept vocabulary to minimize ambiguity in visual concepts?

2. Is the idea of complementarity between the concept space and the latent space truly valid?

3. What will be the effect of causality integration in tag-cloud based explanation for video-text retrieval?

In the upcoming chapters, we will delve into these questions and explore the evidence that supports or challenges the notion of complementarity between the concept space and the latent space. By doing so, we aim to gain a deeper understanding of the strengths and limitations of these hybrid approaches in video retrieval.

Chapter 3

The Role of PoS-tagging in Multimedia Retrieval and Explainability

3.1 Introduction

This chapter builds upon the dual encoding model based on hybrid approach proposed by Dong et al. [1], which provides a framework based on latent space and concept space for cross-modal retrieval (see Section 2.2.3). In this chapter, we extend the dual encoding model [1] by exploring the impact of incorporating Part-of-Speech (PoS) tags in the text encoding pipeline for training the dual encoding model in order to overcome the issue of ambiguity in vocabulary of concept space. PoS-tagging is a crucial process that assigns specific tags to words, representing their syntactic categories. By incorporating PoS-tags, we aim to leverage the syntactic and grammatical information they provide to overcome the ambiguity and enhance the performance, relevance, and explainability of video-text retrieval.

For instance by utilizing PoS-tags in the query *"A man is measuring the size of a copper-head snake with the tape measure"*, the visual concept "measure" is present as both a verb and a noun. The presence of "measure" as a verb and noun in the query, forces the retrieval system to focus on the videos where the measuring activity and measuring objects both are present and are likely to show relevant videos. The PoS-tags provide clarity regarding the intended action and highlight the relevance of videos that depict this specific activity.

Similarly, in other cases for instance "*a person is watering his flowers while people walk under the water*", and *"a man and a woman cooking on a cooking show"*, the presence of PoS-tags helps disambiguate the verb or noun words like water in former and cooking in latter, and also helps in identifying the singular or plural nouns e.g. people or person. This is expected to lead to more precise and relevant video retrieval results aligned with the intended meaning of the query.

From the above examples, it is clearly shown that the inclusion of PoS-tags offers several

benefits in the context of video-text retrieval. Firstly, it helps address ***ambiguity*** in concepts within the vocabulary of concept space. For instance, if the "measure" concept will not be distinguished according to its PoS-tag (verb and noun), it is likely that the retrieval system will focus more on the measuring objects rather than the measuring activities while retrieving the videos. By considering the syntactic categories of words, we can disambiguate their meanings and improve the accuracy of classification. This is particularly useful in scenarios where multiple interpretations or senses are possible. Secondly, PoS-tagging contributes to the ***interpretability*** of the retrieved results, currently bounded to technical users only. By incorporating PoS-tags, we can analyze and explain the influence of words with PoS-tags in retrieval process. Furthermore, the integration of PoS-tags enables a deeper analysis of the textual content. By considering the syntactic structure of sentences, we gain insights into the ***relationships between words*** and their roles within the sentence. This additional information can aid in capturing the nuances and context of the text, thereby improving the performance of video-text retrieval.

Through a comparative evaluation, we will assess the impact of incorporating PoS-tags in the dual encoding model for video-text retrieval. We will compare the performance of the dual encoding retrieval system with and without the utilization of PoS-tags, focusing on metrics such as accuracy, precision, and explainability. This investigation will contribute to a better understanding of the potential benefits and implications of incorporating PoS-tags in the dual encoding model, advancing the existing knowledge presented by Dong et al. [1].

In summary, this chapter aims to extend the dual encoding model by incorporating PoS-tags in the text encoding pipeline. By leveraging the syntactic and grammatical information provided by PoS-tags, we aim to improve the accuracy, precision, and explainability of the video-text retrieval system. The investigation will specifically address ambiguity within the concept vocabulary and enhance the system's ability to provide meaningful explanations.

Objective

The goal of Chapter 3 is to investigate the potential benefits and implications of integrating Part-of-Speech (POS) tags in text encoding pipeline of the dual encoding model for video-text retrieval with the idea of reducing the ambiguity in visual concepts bank. The chapter aims to analyze how the inclusion of PoS-tags impacts the retrieval accuracy, precision, explainability, and overall performance by leveraging syntactic and grammatical information.

3.2 Methodology

Formally, we are given a set of videos $\mathbb{V} = \{v_1, v_2, ..., v_n\}$ and a set of captions $\mathbb{S} = \{s_1, s_2, ..., s_m\}$ where $n \leq m$, and represents the total number of videos $\mathbb{V}$ and captions $\mathbb{S}$ in the dataset re-

spectively. Each video v_i is described by set of captions $\mathbb{C}_p^i = \{c_1^i, c_2^i, ..., c_p^i\}$ where $\mathbb{C}_p^i \subset \mathbb{S}$ belongs to one video v_i. The primary objectives of dual encoding model [1] are

1. To learn two mapping functions $f()$ and $g()$ for visual and textual encodings in two spaces i.e. latent space encodings $(f(v_i), f(s_i))$ and concept space encodings $(g(s_i), g(v_i))$.

2. To learn two similarity functions in order to compute the similarity between video $v_i \in \mathbb{V}$ and caption $s_j \in \mathbb{S}$ such that it yields a high value if $i = j$ (i.e., the caption $s_j \in C_p^i$ correspond to the same video v_i) and a low value if $i \neq j$ (i.e. caption s_j does not correspond to video v_i) using $sim_{lat}(v_i, s_j)$ for similarity in latent space and $sim_{con}(v_i, s_j)$ for similarity in concept space.

Accurate representations of both the video and caption are vital for accurately estimating this similarity, as detailed in Sections 3.2.2.1 and 3.2.2.2. The main aim of this chapter is to integrate the PoS-tags in text encoding pipeline of dual encoding model [1] and learn a similarity function $sim(v, s)$ to determine the similarity between text s and video v in latent space $sim_{lat}(v, s)$ and concept space $sim_{con}(v, s)$.

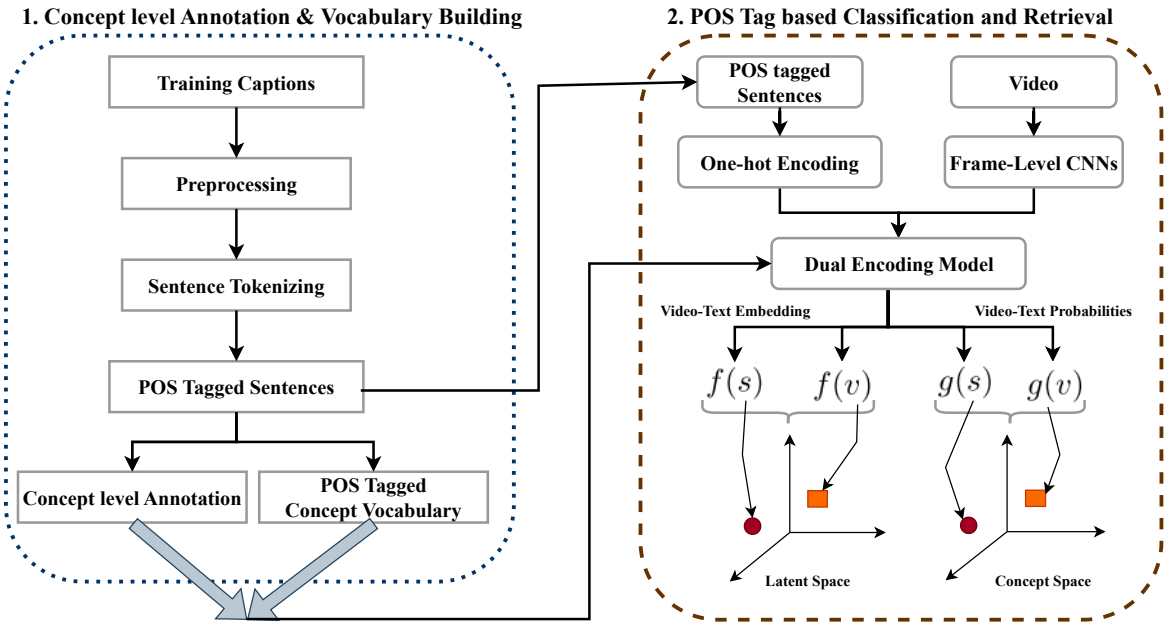


Figure 3.1: Proposed Architecture for Dual Encoding with Part-of-Speech (POS) Tagging in Concept Classification and Video-Text Retrieval

Our overall approach is centered around two steps as shown in Figure 3.1, 1) PoS-tag-based concept level annotation & vocabulary building, and 2) PoS-tag-based classification and retrieval. The first step is responsible for annotating the videos with concepts or tags along with their corresponding PoS-tags. In the second step, the concept level annotation and vocabulary are then used to train the concept space of the hybrid model using PoS-tagged captions/sentences. As depicted in step 2 of Figure 3.1, the video and text modality

are first represented on frame and word level respectively, which are then passed to dual encoding model to generate video and sentence level encodings respectively. Subsequently, during text-to-video (or video-to-text) retrieval, we rank all the videos (or captions) in the dataset based on their combined weighted similarity in latent space and concept space. In the following discussion, our proposed model will be referred to as a "PoS-tag based dual encoding model".

Following, we provide a detailed description of all the steps and processes essential for dual encoding model [1] and integrating PoS-tags within the process. The dual encoding model is based on two spaces (latent and concept). To train the concept space, we need concept-level annotation and vocabulary. So, in Section 3.2.1, we will first discuss the procedure we used to build the: a) "PoS-tags based concept level annotations", and b) "PoS-tags based tag/concept-vocabulary". Then, we provide a comprehensive overview of the feature extraction process from both video and text data (Sections 3.2.2.1 and 3.2.2.2), as described by *Dong et al.* [1]. Finally, we will see in detail the hybrid space learning and evaluation in Sections 3.2.3 and 3.3, respectively.

3.2.1 Concept-level Annotation & Concept Vocabulary building

Concept-level annotations and vocabulary building is done automatically for each training video. As already mentioned that for a specific training video v_i, we have access to p sentences/captions $\{c_1^i, ..., c_p^i\}$, that describe the video v_i content.

For concept-level annotations or ground truth building for each video, the method involves analyzing the p sentence descriptions and determining the frequency of each concept/word except stopwords to that video based on its occurence in those p sentences. For instance, *Dong et al.* [1] annotated the videos v_i with visual concepts by computing the relevance of a specific concept by its occurrence in p sentences corresponds to video v_i. *Li et al.* [149] suggest that a concept appearing in multiple sentences is usually more important than those presented once. Hence, rather than using binary labels as in [2], *Dong et al.* used soft labels based on concept frequency to obtain a more nuanced understanding of the importance of different concepts. So, the K-dimensional ground-truth vector for video v is represented by y_v where $y_v = [y_{v1}, y_{v2}, ..., y_{vi}, ..., y_{vk}]$. The value of its i^{th} dimension, i.e., $y_{vi} \in \mathbb{R}^+$), is defined as the frequency of the i^{th} concept divided by the maximum frequency of all concepts within the p sentences of a video v_i, and k represents the numbers of concept classes in vocabulary.

To obtain the concept vocabulary, *Dong et al.* [1] conducted part-of-speech tagging by NLTK toolkit on all training sentences, and only keep the nouns, verbs and adjectives. All the English stopwords also removed. Besides, words are also lemmatized, making *dog* and *dogs* to be a same concept. Finally, the top $k = 512$ frequent words are selected as the final

concept vocabulary.

Instead of using soft labels, another way of building concept vocabulary and ground-truth is to use binary labels. *Wu et al.* [2] constructed a k-dimensional concept vocabulary from the training set, which included words that appeared in at least five descriptions and excluded words from the NLTK stopword list. For a video v, the words that appear in its captions are identified as ground-truth labels $y_v = [y_{v1}, y_{v2}, ..., y_{vi}, ..., y_{vk}]$, where $y_{vi} \in \{0, 1\}$ indicates whether a word is present in the captions of v. The k-dimensional concept vocabulary is compiled from the training set, by including the words that appear in at least five descriptions and removing the words in the NLTK stopword list, where and $k = 11,147$.

Both of these approaches then extend relevant video-sentence to a triplet training instance by adding the K-dimensional ground-truth vector, creating a supervised learning framework for concept space learning of the hybrid space model which is described into detail in Section 3.2.3.

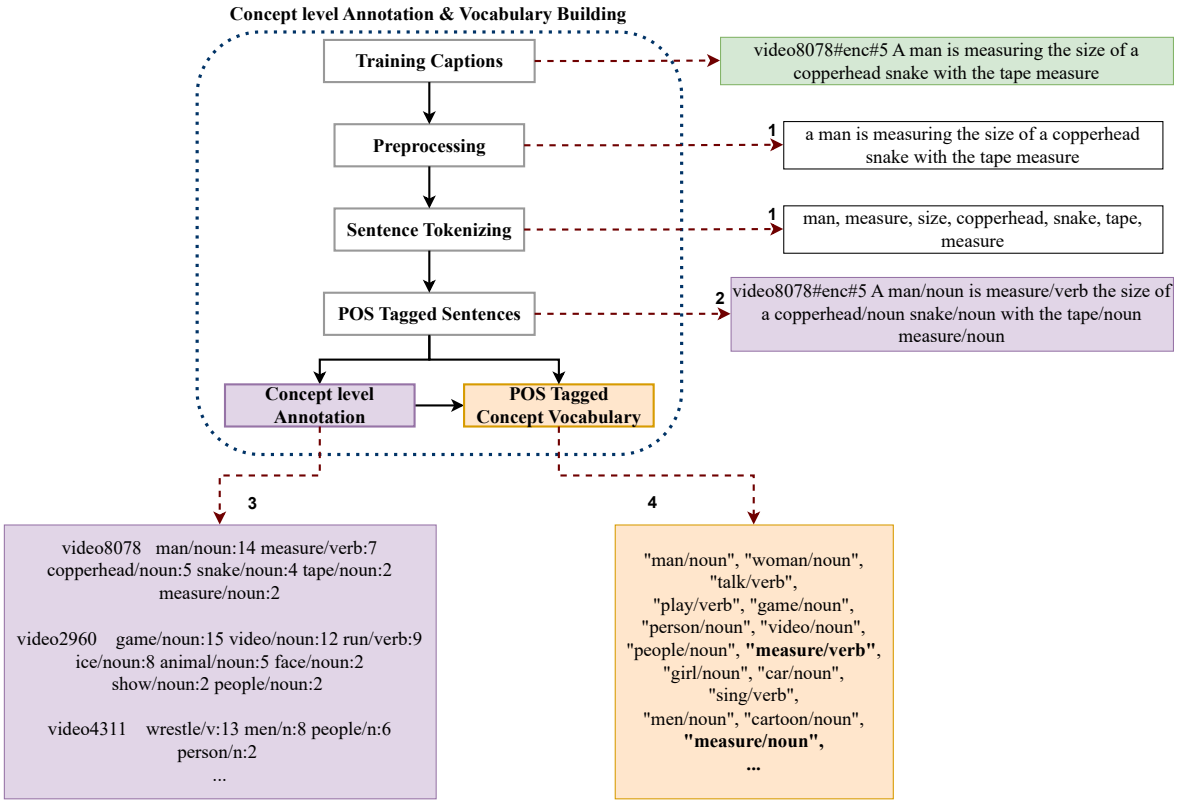


Figure 3.2: Workflow for Building a PoS-tag-based Vocabulary and Concept-Level Annotation

PoS-tag based Annotation & Vocabulary: In our case, the difference lies in annotating the videos with visual concepts along with their PoS-tags. An example of this is given in Figure 3.2. Given an input sentence s, we apply preprocessing on sentence s and tag each word w of the sentence with its PoS-tag and only keep the nouns (n), verbs (v), adverb (a), adjectives (j), and preposition (p). (see step1-2 in Figure 3.2). English stopwords are also

excluded in this pre-processing. Besides, we also lemmatize the words, making "scene/n" and "scenes/n" to be the same concept. For concept-level annotations (step-3 in Figure 3.2) of each video v, the process is the same as in dual encoding model [1] i.e. let y be a K-dimensional ground-truth vector for video v. The value of its i^{th} dimension, i.e., y_i, is defined as the frequency of the i^{th} PoS-tagged concept divided by the maximum frequency of all PoS-tagged concepts within the p sentences. As the words in sentences are tagged with their PoS-tags, the annotation is based on PoS-tagged words. Finally, the top K frequent words selected from step-3 are chosen as the final PoS-tag based concept vocabulary in step 4.

3.2.2 PoS-tag based classification and retrieval

In this section, we discuss the representation procedure of video and text in PoS-tag dual encoding model.

3.2.2.1 Video Representation

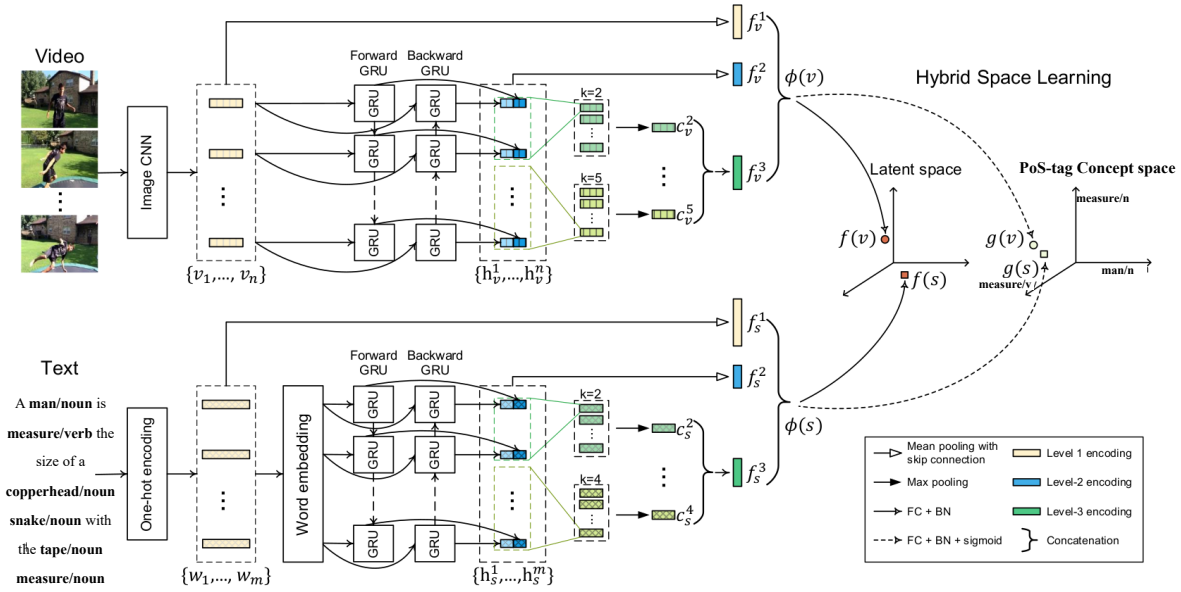


Figure 3.3: Proposed PoS-tag based Dual Encoding Architecture with PoS-tagged "text" & "concept space" (inspired from [1])

The second part of our approach, as shown in Figure 3.1, achieves training dual encoding network for classification and retrieval. We are based on the same architecture as the dual encoding network [1] to extract the multi-level video features. The video encoding pipeline consists of three levels, and each serving a specific purpose, as shown in Figure 3.3. Initially, for a given video v, a sequence of n frames is selected, with each frame spaced at a predetermined interval. For each frame, the spatial features are extracted using deep CNN [85, 106, 150]. As a result, the video from the collection can be represented as a sequence of feature vectors $\{v_1, v_2, ..., v_n\}$, where v_t represents the deep feature vector of the t^{th}

frame. Then these frame-level video features are passed to a multi-level encoding network. It is important to note that this multi-level representation captures the visual characteristics of the video at different time points.

Level 1. Global Encoding by Mean Pooling: At this stage, the objective is to capture the overall information encompassing all frames in a video using mean pooling. Mean pooling has been widely favored for video-text retrieval, as highlighted in the literature review by Dong et al. [1]. This pooling technique represents a video by computing the average of the features extracted from its frames. Consequently, it effectively captures visual patterns that consistently appear throughout the video content. Such patterns typically have a global scope. The resulting encoding at this level is denoted as f_v^1.

$$f_v^1 = \frac{1}{n}\sum_{t=1}^{n} v_t \tag{3.1}$$

Level 2. Temporal-Aware Encoding by biGRU: In the dual encoding model, the second level of visual encoding pipeline focuses on extracting the temporal information of the video using the bidirectional GRU (biGRU) [151]. Let $\overrightarrow{h_t}$ and $\overleftarrow{h_t}$ be their corresponding hidden states at a specific time step $\{t = 1,...,n.\}$. The hidden states are generated as

$$\begin{aligned} \overrightarrow{h_t} &= \overrightarrow{GRU}(v_t, \overrightarrow{h_{t-1}}) \\ \overleftarrow{h_t} &= \overleftarrow{GRU}(v_{n+1-t}, \overleftarrow{h_{t-1}}) \end{aligned} \tag{3.2}$$

The forward-GRU and backward-GRU, denoted as $\overrightarrow{GRU}$ and $\overleftarrow{GRU}$ respectively, capture the past and future contextual information in the video sequence. The hidden states of the forward and backward GRUs at a specific time step t, represented by $\overrightarrow{h_t}$ and $\overleftarrow{h_t}$ respectively, contain the encoded information from their respective directions. To combine the information from both directions, $\overrightarrow{h_t}$ and $\overleftarrow{h_t}$ are concatenated, resulting in the biGRU output h_t. This concatenated vector is denoted as $v = [\overrightarrow{h_t}, \overleftarrow{h_t}]$. By considering all the time steps, we obtain a d-dimensional feature map $H_v = \{h_v^1, h_v^2, ..., h_v^n\}$ where each $h_v^j \in \mathbb{R}^d$. The feature map H_v then has a shape of $d \times n$, where n represents the number of time steps in the video. To obtain the single vector temporal representation of the video, mean pooling to the feature map H_v along the row dimension is applied. This mean-pooled representation, denoted as f_v^2, serves as the biGRU-based encoding of the video.

$$f_v^2 = \frac{1}{n}\sum_{t=1}^{n} h_t \tag{3.3}$$

Level 3. Local-Enhanced Encoding by biGRU-CNN: The final layer further extracts the local information in the temporal sequence of the video by performing 1D CNN on top of the biGRU features from the second level. The input of the 1D CNN is the feature map H_v generated by the previous biGRU module. Let $\text{Conv1d}_{(k,r)}$ be a 1D convolutional block that contains $r = 512$ filters of size k, with $k \geq 2$. Feeding H_v, after zero padding, into $\text{Conv1d}_{(k,r)}$ produces an $n \times r$ feature map. Non-linearity is introduced by applying the ReLU activation function on the feature map. The number of frames n varies for videos, max pooling to compress the feature map to a vector c_k of fixed length r is further applied. More formally, the above process can be expressed as:

$$c_k^v = \text{max-pooling}(\text{ReLU}(\text{Conv1d}(k,r)(H_v))). \tag{3.4}$$

A filter with $k = 2$ allows two adjacent rows in H_v to interact with each other, while a filter of larger k means more adjacent rows are exploited simultaneously. In order to generate a multi-scale representation, multiple 1D convolutional blocks with $k = 2,3,4,5$ are employed. Their outputs are concatenated to form the biGRU-CNN based encoding.

The multi-level encoding of the input video is achieved by combining the outputs from different levels of encoding pipeline i.e. global (f_v^1), temporal (f_v^2), and local (f_v^3) extracted from these sub-networks. The final visual embedding is denoted as

$$\phi(v) = [f_v^1, f_v^2, f_v^3]. \tag{3.5}$$

3.2.2.2 Text Representation

The multi-level encoding network described above for video representation is adapted for the PoS-tag based text modality. In our case, each word w of the sentence s is first tagged with its respective PoS-tag as $\{w_1/PoS, w_2/PoS, \ldots, w_m/PoS\}$ as shown in step 2 of Figure 3.2. A sequence of one-hot vectors is generated for s, where w_t indicates the vector representation of the t-th word.

For "Level 1. Global Encoding by Mean Pooling (f_1^s)", the text representation is obtained by averaging all the individual vectors in the sequence of one-hot vectors, which corresponds to the classical bag-of-words representation. To compute the "Level 2. Temporal-Aware encoding by biGRU", each word is first converted to a dense vector by multiplying its one-hot vector with a word embedding matrix. The matrix is initialized using a pre-trained word2vec model [88] provided by [115], which was trained on English tags from 30 million Flickr images. The subsequent steps follow a similar approach as in the video counterpart. Denoting the level 2 encoding of the sentence as f_2^s, and the Local-Enhanced Encoding by biGRU-CNN as level 3 encoding f_3^s, where three 1D convolutional blocks with $k = 2,3,4$ is utilized for the latter. The multi-level encoding of the sentence is obtained by concatenating the

encoding results from all three levels in the dual network. Specifically, it is expressed as:

$$\phi(s) = [f_1^s, f_2^s, f_3^s]. \tag{3.6}$$

3.2.3 Hybrid Space Learning

As $\varphi(v)$ and $\varphi(s)$ have not been correlated, they are not directly comparable. For video-text similarity computation, the vectors need to be projected into a common space, so the next step is to train the model in order to map the visual and textual encoding to the hybrid space i.e. d-dimensional latent space and K-concept space. It is worth noting that except for the image CNNs used as an input for the dual encoding network, the dual encoding network is trained in an end-to-end manner.

3.2.3.1 Learning a Latent Space

Given the encoded video vector $\varphi(v)$ and PoS-tag based encoded sentence vector $\varphi(s)$, the encodings are mapped to the latent space of d-dimension by using affine transformations (i.e. fully connected layer and a batch normalization layer) as:

$$\begin{aligned} f(v) &= BN(W_1\varphi(v) + b_1) \\ f(s) &= BN(W_2\varphi(s) + b_2) \end{aligned} \tag{3.7}$$

where W_1 and W_2 are the weights and b_1 and b_2 are the bias of the fully connected layer in the visual and encoding network respectively. Both $f(v)$ and $f(s)$ are d-dimensional vectors and are directly comparable. Finally the similarity of the video-text in latent space $sim_{lat}(v,s)$ is computed using cosine similarity as:

$$sim_{lat}(v,s) = \frac{(f(v) \cdot f(s))}{(||f(v)||||f(s)||)} \tag{3.8}$$

The visual-textual encoding matching task is done in latent space, and it is trained in an end-to-end manner by using the improved triplet ranking loss function. Given a relevant video-sentence pair (v,s) in a mini-batch, its loss $\mathcal{L}_{lat}(v,s)$ is:

$$\begin{aligned} \mathcal{L}_{\text{lat}}(v,s) = &\, max(0, m + sim_{lat}(v,s^-) - sim_{lat}(v,s)) \\ &+ max(0, m + sim_{lat}(v^-,s) - sim_{lat}(v,s)) \end{aligned} \tag{3.9}$$

where m is the margin, s^- and v^- is the negative sample for video v and sentence s respectively which is not chosen randomly from the current mini-batch, but instead, the most similar yet negative sentence and video are chosen.

3.2.3.2 Learning a Concept Space

As video and text can be described using multiple concepts, hence the task of learning concept space is a multi-label classification problem. Given the encoded video vector $\varphi(v)$ and PoS-tag based encoded sentence vector $\varphi(s)$, the encodings are mapped to the concept space of K-dimension by using a similar network to the network used for latent space learning. That is,

$$\begin{aligned} g(v) &= \sigma(BN(W_3\varphi(v)+b_3)) \\ g(s) &= \sigma(BN(W_4\varphi(s)+b_4)) \end{aligned} \tag{3.10}$$

In Eq. 3.10, the sigmoid function is used to generate the probabilistic output of K-dimensions, where each dimension of $g(v)_i$ denotes the probability of the concept being relevant to the video v. Similarly, $g(s)$ for sentence s is obtained. The similarity between video and text concept representation vector in concept space $sim_{con}(v,s)$ is computed using generalized jaccard similarity as:

$$sim_{con}(v,s) = \frac{\sum_{i=1}^{K} min(g(v)_i, g(s)_i)}{\sum_{i=1}^{K} max(g(v)_i, g(s)_i)} \tag{3.11}$$

Learning concept space is done using binary cross-entropy (BCE) loss function. Given the video-sentence pair (v, s), their probability scores $(g(v), g(s))$, and its ground truth y (see Section 3.2.1), the loss is computed as:

$$\begin{aligned} \mathcal{L}_{\text{bce}}(v,s,y) = -(&\sum_{i=1}^{K} [y_i \log(g(v)_i) + (1-y_i)\log(1-g(v)i)] \\ &+ \sum_{i=1}^{K} [y_i \log(g(s)_i) + (1-y_i)\log(1-g(s)_i)]) \end{aligned} \tag{3.12}$$

As the purpose of concept space is not limited to multi-label classification problem, but it is also used for improving video-text matching in case of video-text retrieval. Hence, improved triplet ranking loss function is also minimized in concept space.

$$\begin{aligned} \mathcal{L}_{\text{con,rank}}(v,s) = &\, max(0, m + sim_{con}(v,s^-) - sim_{con}(v,s)) \\ &+ max(0, m + sim_{con}(v^-,s) - sim_{con}(v,s)) \end{aligned} \tag{3.13}$$

The concept is learned in parallel to latent space, and is learned by minimizing the combination of $\mathcal{L}_{\text{bce}}$ and $\mathcal{L}_{\text{con,rank}}$

$$\mathcal{L}_{\text{con}} = \mathcal{L}_{\text{bce}} + \mathcal{L}_{\text{con,rank}} \tag{3.14}$$

3.2.3.3 Joint Learning & Inference

The dual encoding network is trained by minimizing the combination of the latent space loss $\mathcal{L}_{\text{lat}}$ and the concept based loss $\mathcal{L}_{\text{con}}$. In particular, given a training set $D = (v,s,y)$, the aim

is to find the optimal parameters θ that minimize the objective function:

$$\arg\min_{\theta} \left(\mathcal{L}_{\text{lat}}(v,s) + \mathcal{L}_{\text{con}}(v,s,y) \right), \tag{3.15}$$

In the inference stage, for each video $v \in V$ in the collection, the dual encoding model is used to extract its latent space embedding and concept space embedding using Eq. 3.7 and Eq. 3.10. As a result, each video v in the dataset is associated with an embedding $f(v) \in \mathbb{R}^d$ and a predicted concept vector $g(v) \in \mathbb{R}^K$. Given a test query s in textual form, the text pipeline of the dual encoding model is utilized to encode the query as a textual embedding in latent space as $f(s) \in \mathbb{R}^d$ and in concept space as $g(s) \in \mathbb{R}^K$. The similarity scores will be calculated in each space sim_{lat} and sim_{con} between each video v in collection and query s. Hence, the overall similarity is a weighted sum of similarity in two spaces i.e. latent space and concept space, given as:

$$sim(v,q) = \alpha.sim_{lat}(v,s) + (1-\alpha).sim_{con}(v,s) \tag{3.16}$$

The rank-list of videos for the given query is obtained based on the combined similarity score obtained. where α is a hyper-parameter to balance the importance of two spaces, ranging within [0, 1]. Note that raw values of $sim_{lat}(v,s)$ and $sim_{con}(v,s)$ reside in distinct scales. Hence, they are rescaled separately by min-max normalization before being combined. Also note that in the inference stage, the multi-level encoding at the video side i.e. $f(v)$ can be performed independently. Hence, for a large-scale video collection, their hybrid-space features $f(v), g(v)$ can be pre-computed, allowing to answer ad-hoc queries on the fly.

3.3 Experiments & Evaluation

In our evaluation, we compared our PoS-tagged dual encoding model with the dual encoding model [1] on five video-text datasets, covering text-to-video retrieval and video-to-text retrieval tasks. To ensure a comprehensive analysis, we carefully selected diverse datasets and utilized rigorous evaluation metrics. Additionally, we focused on assessing the impact of the PoS-tagged model on both retrieval accuracy and explainability. We also conducted comparable experiments using different PoS-taggers, namely TreeTagger (TT)[1], WordNet (WN) [152], and Spacy[2]. By investigating these factors, we aimed to gain valuable insights into the performance and potential advantages of our PoS-tagged dual encoding model for video-text retrieval. The datasets used in the experiments are listed in Table 3.1. Our proposed PoS-tag model is trained and validated on captioning datasets, which are also utilized by other approaches but not limited to W2VV++ [18], dual encoding model [1, 8] and inter-

[1] https://pypi.org/project/treetaggerwrapper/
[2] https://github.com/explosion/spaCy

Dataset	#Video	#Caption	AVS Test Query
Training set			
MSR-VTT	10,000	200,000	-
TGIF	100,855	124,534	-
Validation set			
TV2016TRAIN	5	200	400
AVS test set			
IACC.3	335,944		90
V3C1	1,082,659		90

Table 3.1: Datasets information.

pretable embedding model [2]. The number of captions per video varies across the datasets, ranging from 2 in TV2016TRAIN [102] to as many as 20 in MSR-VTT [13]. The MSR-VTT dataset [13] was initially created for video captioning and comprises 10,000 Web video clips along with 200,000 natural sentences that describe the visual content of these clips. Each video clip has around 20 sentences associated with it. Different versions of data partition have been proposed in the literature for this dataset [5, 51, 13]. The official partition [13] divides the dataset into 6,513 clips for training, 497 clips for validation, and the remaining 2,990 clips for testing. In our experiments, we used official partition of MSRVTT [13] for training and evaluation of POS-tagged dual encoding model.

To assess the performance of POS-tagged dual encoding model on Ad-Hoc Video Search (AVS) model, we evaluate it on three large video collections. These collections are part of the TRECVid benchmarked datasets, with i) **IACC.3** [102] a dataset used from 2016 to 2018, consists of approximately 4,600 Internet Archive videos. It has a size of around 144 GB, totaling 600 hours of video content. These videos are in MPEG-4/H.264 format and come with Creative Commons licenses. The duration of the videos ranges from 6.5 minutes to 9.5 minutes, with an average duration of almost 7.8 minutes. Additionally, most videos in this dataset have metadata provided by the donor, such as title, keywords, and description. ii) **V3C1** [153] is a subset derived from a larger V3C video dataset [154], used from 2019 to 2021. V3C1 comprises 7,475 Vimeo videos with Creative Commons licenses. The dataset size is approximately 1.3 TB, spanning a total duration of 1,000 hours. The videos in V3C1 have a mean duration of 8 minutes. Metadata, including title, keywords, and description, is available for all videos in JSON files. The dataset is further segmented into 1,082,659 short video segments based on provided master shot boundary files. Keyframes and thumbnails for each video segment have also been extracted and made available.

3.3.1 Implementation Details

Prior to conducting our experiments, we provide a comprehensive overview of the common implementations employed in our study, including text preprocessing, video feature extrac-

tion, concept vocabulary extraction, and model training. Most of these implementations were adopted from the dual encoding model [1].

For sentence preprocessing, we performed several steps. Firstly, we converted all words to lowercase and then assigned each sentence its corresponding Part-of-Speech (PoS) tags. Additionally, we replaced words occurring less than five times in the training set with a special token to handle infrequent words effectively. Regarding video features, we utilized frame-level ResNeXt-101 [54, 55], and ResNet-152 [39] features provided by *Dong et al* [1]. These features were concatenated to create a combined 4,096-dimensional CNN feature known as ResNeXt-ResNet. To construct the concept vocabulary as already explained in Section 3.2.1, we employed part-of-speech tagging using different taggers such as WordNet (WN), TreeTagger (TT), and the Spacy toolkit on all training sentences. From the tagged sentences, we retained nouns, verbs, adverbs, adjectives, prepositions, and conjunctions while removing all English stopwords. Additionally, we performed lemmatization to ensure words with similar meanings, such as "dog/noun" and "dogs/noun", were treated as the same concept. Finally, we selected the top $K = 512$ most frequent words as our concept vocabulary, with the option of including or excluding PoS-tags. These detailed implementations lay the foundation for our experiments and enable us to evaluate the effectiveness of our PoS-tagged dual encoding model accurately. For frame level visual encoding, n frames are extracted from each video, spaced at a predetermined interval of 0.5 seconds, similar to dual encoding model [1].

3.3.2 Impact on Video-Text Retrieval

3.3.2.1 MSR-VTT Experiments

Metrics. In order to evaluate the accuracy of TTV and VTT retrieval system, we used the proportion of the queries for which at-least one correct document is retrieved among top-K results (R@K, with $K = 1, 5, 10$), the median rank of first relevant item (Med R), the mean Average Precision (mAP) and sum of R@K for TTV and VTT (SumR). Higher R@K and lower median rank represents better performance of system.

Results. Table 3.2 presents the results of the MSR-VTT experiments, focusing on text-to-video (TTV) and video-to-text (VTT) retrieval tasks. Each method is evaluated based on metrics defined above, which measure the accuracy and ranking performance of the retrieval system. The first section in the *"Hybrid training"* of Table 3.2 presents the reported results by *Dong et al.* [1] in various platforms and papers. The 2^{nd} and 3^{rd} section showcases various configurations of the Dual Encoding model, including different dimensions and taggers. The $(1536 + 512)$-d Dual Enc. configurations represent the Dual Encoding model evaluated using a weighted similarity approach. This approach combines the 1536-dimensional latent space and the 512-dimensional concept space to leverage their respective strengths.

Method	Text-to-Video Retrieval					Video-to-Text Retrieval					SumR
	R@1	R@5	R@10	Med r	mAP	R@1	R@5	R@10	Med r	mAP	
Hybrid training											
Dual Encoding TPAMI [1]	11.6	30.3	41.3	17	21.2	22.5	47.1	58.9	7	10.5	211.7
Dual Encoding GitHub [1]	11.8	30.6	41.8	17	21.4	21.6	45.9	58.5	7	10.3	210.2
Dual Enc. (Conf. Ver. 2048-d)[1])	11.0	29.2	39.8	19	20.2	18.8	42.7	56.2	8	9.3	197.7
Dual Enc. (Conf. Ver. 1536-d)[1])	11.0	29.3	39.9	19	20.3	19.7	43.6	55.6	8	9.3	199.0
(1536+512)-d Dual Enc. Re-Run	11.78	31.00	42.08	16.10	21.52	20.70	45.10	57.74	7.00	10.13	**208.40**
(1536+512)-d Dual Enc. (TT)	12.09	31.39	42.50	16.10	21.87	20.68	45.32	58.12	6.90	10.32	210.10
(1536+512)-d Dual Enc. (WN)	12.02	31.44	42.59	16.00	21.84	20.56	45.60	58.28	6.90	10.36	**210.50**
(1536+512)-d Dual Enc. (Spacy)	11.84	31.07	42.08	16.40	21.59	20.29	45.00	57.78	7.00	10.22	208.05
512-d Dual Enc. (Concept Space)	9.9	26.8	37.4	23	18.7	17.9	41.5	53.9	8	9.0	**187.40**
512-d Dual Enc. (Concept Space - TT)	10.10	27.09	37.31	22.65	18.86	18.64	41.83	54.79	8.50	9.14	189.74
512-d Dual Enc. (Concept Space - WN)	10.14	27.23	37.53	22.40	18.94	18.54	42.25	55.19	8.00	9.11	**190.87**
512-d Dual Enc. (Concept Space - Spacy)	9.85	26.71	36.76	23.55	18.54	18.15	41.45	54.22	8.30	9.02	187.15
Concept only training											
512-d Dual Enc. (Concept Space)	9.84	26.79	37.02	23.30	18.57	18.57	41.85	54.22	8.40	8.88	**188.28**
512-d Dual Enc. (Concept Space - TT)	10.15	27.33	37.65	22.10	18.98	18.82	42.32	55.22	8.15	9.17	**191.49**
512-d Dual Enc. (Concept Space - WN)	10.14	27.30	37.58	22.40	18.96	18.69	42.13	54.99	8.00	9.15	190.82
512-d Dual Enc. (Concept Space - Spacy)	9.96	26.92	37.18	23.00	18.71	18.39	41.59	54.52	8.50	9.04	188.55

Table 3.2: MSR-VTT experiments – Averages. Official full-size test set [13].

On the other hand, the 512-d Dual Enc. configurations in the hybrid training section indicate that the model is trained in hybrid mode, but only evaluated in the 512-dimensional concept space. These evaluations provide valuable insights into the performance of the Dual Encoding model in different spaces, helping to identify the most effective configuration for text-to-video and video-to-text retrieval tasks. It provides a comparative analysis of their performance in both TTV and VTT retrieval tasks. The "SumR" column represents the SumR score, which is the sum of R@1, R@5, R@10, reflecting the overall retrieval accuracy.

The *"Concept only training"* section explores the performance of the Dual Encoding model using Concept Space training. Similar to the previous section, it includes different taggers, but trained and tested on only 512-d concept space, providing insights into their effectiveness.

Overall, the Table 3.2 allows for a comprehensive evaluation of different configurations of the Dual Encoding model, highlighting the impact of dimensionality and the choice of PoS-taggers on the retrieval performance. The metrics provide a clear indication of which configurations yield better results, assisting in the selection of the most suitable setup for text-to-video and video-to-text retrieval tasks. The results demonstrate the effectiveness of our proposed PoS-tagged dual encoding model on the official MSR-VTT split, achieving a SumR of 191.49 from 188.28. Among the three PoS-taggers, TreeTagger (TT) outperforms the others. Spacy did not outperform other taggers in terms of accuracy because it exhibited a relatively higher number of mistakes compared to the other taggers. This lower accuracy can be attributed to the errors made by Spacy in assigning the correct PoS-tags to words in the text.

However, the improvement, though positive, is not particularly significant, with a SumR

of 191.49 for TreeTagger from 188.28 for original dual encoding model. Upon investigating the reasons for this modest increase in accuracy, we noticed that only approximately $\approx 1\%$ of the queries in the dataset contain ambiguous words (i.e., words with multiple PoS-tags). Nevertheless, due to the low proportion of queries affected by PoS-tags, the results in Table 3.2 do not exhibit the significant improvements we anticipated. Our intention is to integrate PoS-tags in a way to overcome the ambiguity in queries, so we chose to focus on the set of queries that contain ambiguous words and evaluate the overall accuracy and explainability of retrieval system on this set of queries.

For further insights, we extracted a subset $\mathbb{C} = \{s_1, s_2, ..., s_t\}$ of queries from the MSR-VTT official split, which contains ambiguous words along with their corresponding videos. Here, t represents the number of captions affected. We then evaluated the original and POS-tagged dual encoding systems on this query set $\mathbb{C}$ for the TTV and VTT tasks.

Method	Text-to-Video Retrieval					Video-to-Text Retrieval					SumR
	R@1	R@5	R@10	Med r	mAP	R@1	R@5	R@10	Med r	mAP	
Concept only Training											
512-d Dual Enc. (Concept Space)	29.25	56.65	67.41	4.00	41.94	26.59	49.35	58.88	5.80	34.62	288.13
512-d Dual Enc. (Concept Space - WN)	30.39	57.39	68.23	3.80	43.14	28.11	49.45	58.97	5.70	35.41	292.53
512-d Dual Enc. (Concept Space)	31.18	59.56	70.62	3.40	44.33	30.00	55.14	66.09	4.20	39.89	312.59
512-d Dual Enc. (Concept Space - TT)	32.26	60.69	71.66	3.10	45.49	31.35	56.00	67.20	4.10	40.87	319.17

Table 3.3: PoS-tagged Impacted Captions. Official full-size test set [13].

For this evaluation, we exclusively utilized the two best-performing POS-tagger (i.e. TreeTagger (TT) and WordNet (WN)) chosen from Table 3.2. Notably, as shown in Table 3.3, the POS-tagged dual encoding system demonstrated a significant improvement in accuracy compared to the original system. Specifically, the SumR score increased from 312.59 to 319.17 on the $\mathbb{C}$ subset. This improvement underlines the effectiveness of incorporating POS tagging in enhancing retrieval performance.

Moreover, the reason for the larger difference in SumR values between Table 3.2 and Table 3.3 is due to the varying number of candidate videos/queries to be retrieved in different partitions. The official MSR-VTT partition has a larger number of candidates, making it more challenging for the models. As a result, the performance scores of models evaluated on the official partition are lower compared to their evaluated performance on other subset $\mathbb{C}$. It is to be noted that the PoS-tag encoding plays a vital role in concept space and is useless for latent space, that is why in Table 3.3 and Table 3.4 we compared the results evaluated solely on concept only training.

In Figure 3.4, we present examples of text-to-video retrieval on the MSR-VTT testing split to visually demonstrate the impact of the POS-tagged dual encoding system. For a query, the top 3 ranked videos and the ground-truth video (boundary with green) are shown. In case the ground-truth video is among the top three, the fourth video will be included as

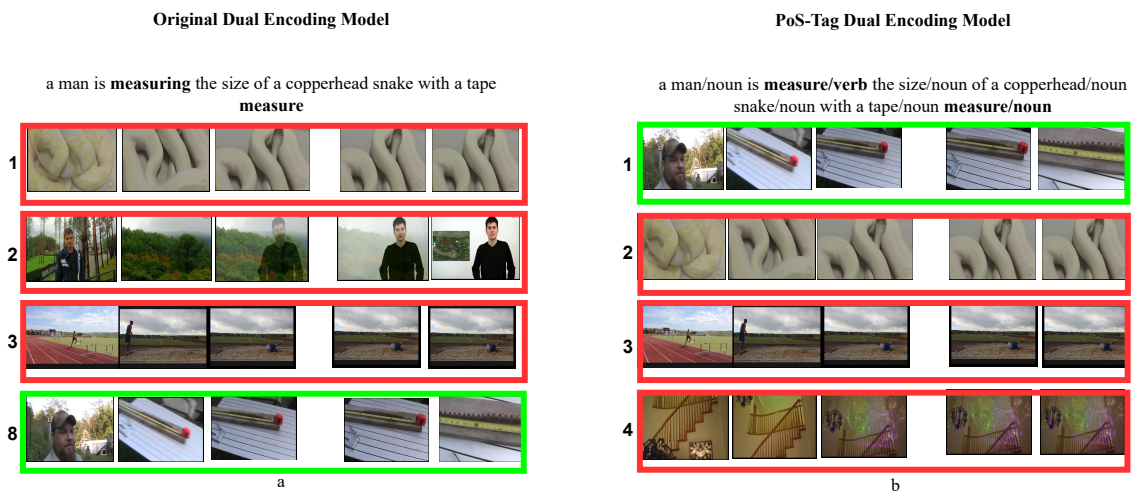


Figure 3.4: Text-to-video retrieval examples on MSR-VTT testing set (subset $\mathbb{C}$).

well. By definition, each query has only one ground-truth video. Number on the left hand side of each video indicates the video's rank in the retrieval result.

Figure 3.4 clearly shows that our model (POS-tagged dual encoding on the right side) successfully ranks the correct video at the top position. This achievement can be attributed to its accurate capture of all the described actions and entities in the query sentence. In contrast, the original dual encoding model [1] on the left side ranks the correct video at the 8th index, indicating a lower retrieval accuracy. These examples serve as compelling evidence of how our POS-tagged dual encoding model improves the ranking of relevant videos, leading to enhanced accuracy in retrieval tasks. The performance enhancements observed on the MSR-VTT dataset [13], along with the visual examples, highlight the effectiveness of our POS-tagged dual encoding model in augmenting accuracy and overall performance in text-to-video retrieval scenarios.

3.3.2.2 Experiments on Ad-hoc Video Search (AVS)

The earlier experimental results on MSR-VTT dataset in Table 3.2 demonstrate that Tree-Tagger (TT) outperformed other POS-taggers. Therefore, in our evaluation of the AVS task, we only compared the TreeTagger-Dual encoding model with the original dual encoding model [1]. Another model known as "Interpretable Embedding" proposed by *Wu et al.* [2] based on same principle as of original dual encoding model, is also compared against our proposed PoS-tag dual encoding model.

Metrics. The performance of the retrieval system is evaluated using the inferred Average Precision (infAP), which is the official performance metric used in the TRECVID AVS task. The overall performance is measured by averaging the infAP scores across all the queries, with scores reported as percentages (%).

Methods	IACC.3			V3C1			Overall
	2016	2017	2018	2019	2020	2021	
Concept only Training							
512-d Dual Encoding (Re-Run)	11.30	17.18	8.60	8.19	10.23	10.17	10.98
Interpretable Embedding ($ST_{concept}$) [2]	13.40	13.70	6.80	10.40	-	-	11.10
512-d Dual Enc. (Concept Space - TT)	11.07	**18.12**	**8.66**	7.46	**11.38**	**11.16**	**11.31**

Table 3.4: Experiments on TRECVID AVS 2016 - 2022. Results (infAP) are presented in %

Results. Table 3.4 shows the performance of the proposed POS-tagged dual encoding model on TRECVID Ad-Hoc video search (AVS) 2016-2021 tasks, and the overall performance is the mean score of the these years. The table compares different methods and their performance on specific datasets from each year in concept only training and evaluation setting. The Dual Encoding model with a 512-dimensional concept space is re-trained and re-evaluated on our systems, as the authors had not reported the results with this specific configuration in their paper. Table 3.4 also includes the results reported by *Wu et al.* [2], who employed interpretable embedding with *single-task* ($ST_{concept}$) setting (system is trained and evaluated on concept space only). Additionally, the table highlights the performance of the PoS-tag Dual Encoding model with a 512-dimensional concept space using TreeTagger (TT). These findings provide insights into the relative performance of these methods in AVS, emphasizing the importance of the proposed PoS-tag dual encoding model.

The proposed method performs the best, with overall infAP of 11.31 for TTV task, showing the slight improvement of proposed POS-tagged dual encoding model. In some years (2017-2018, 2020-2021), the PoS-tagged dual encoding model exhibited slight improvements compared to the original dual encoding model [1] and interpretable model [2], indicating its effectiveness in enhancing performance. However, in certain years (2016, 2019), there was no improvement observed. One possible reason for this lack of improvement is the absence of ambiguous queries in those particular years. As the POS-tags are primarily utilized to disambiguate ambiguous words, the lack of such queries may limit the benefits of the POS-tagged dual encoding model. Additionally, it is worth noting that the introduction of POS-tags introduce some errors in the tagging process as well, which impact the performance in certain years (2016, 2019), leading to slight drop in performance. Overall, while the POS-tagged dual encoding model showed slight and not significant improvement in AVS task performance, played a crucial role in determining the extent of improvement.

3.3.3 Impact on Explainability

The integration of PoS-tagging in our dual encoding model has a significant impact on the explainability of the retrieved results. By incorporating PoS-tags into the encoding process, we enhance the interpretability of the model by providing insights into the syntactic struc-

ture and grammatical relationships within the text. The PoS-tags enable us to identify the different parts of speech in the input text, such as nouns, verbs, adjectives, adverbs, prepositions, and conjunctions. This linguistic information offers a deeper understanding of the underlying meaning and semantic composition of the text.

The PoS-tagged dual encoding model enhances explainability by enabling technical users to analyze the influence of different parts of speech on the retrieval process. We examine how the presence or absence of certain PoS-tags affects the relevance of retrieved videos, thereby gaining insights into the model's decision-making process. For instance, in the query *"a man/noun is **measure/verb** the size/noun of a copperhead/noun snake/noun with a tape/noun **measure/noun**"* in Figure 3.4(b), the presence of PoS-tags in the query, such as "measure/verb" and "measure/noun", provides significant benefits in the explainability of the retrieval process. When we have distinct PoS-tags for different senses of a word, such as "measure" being tagged as both a verb and a noun, it allows us to capture the intended meaning more accurately.

If we only have the word "measure" without any specific PoS-tags in vocabulary, the interpretability of the retrieval results is compromised. Without the PoS-tags, the model treat "measure" as a generic term without considering its verb or noun sense. This lead to mixed or less precise retrieval results. As shown in Figure 3.4(a), the model does not fully understand the intended action associated with "measure", and listed the videos related to "snake", "man" on the top of list. However, with the PoS-tagged dual encoding model (Figure 3.4(b)), the presence of PoS-tags like "measure/verb" and "measure/noun" provides clarity and precision. By considering these specific PoS-tags, the model distinguish between the verb and noun senses of "measure" and retrieve videos that align more closely with the intended meaning of the query. This targeted retrieval based on the PoS-tags improves the overall accuracy and relevance of the results, leading to a more effective and interpretable retrieval system.

By incorporating PoS-tags, technical users or developers better understand whether the relevance of the retrieved videos is based on the verb or noun. This information allows us to assess the accuracy and alignment of the results with the query's intended meaning. When retrieving results using our PoS-tagged dual encoding model, the presence of PoS-tags in the output help explain why certain videos are considered relevant or irrelevant. Overall, the incorporation of PoS tagging in our dual encoding model enhances the explainability of the retrieved results by providing linguistic context, highlighting relevant linguistic features, and enabling the analysis of the impact of different parts of speech on retrieval decisions. This improved explainability can be valuable in various applications, such as content recommendation systems, information retrieval, and understanding user preferences and intents.

While the PoS-tagged dual encoding model provides valuable insights for technical users or developers, visualizing PoS-tags in tag-cloud based explanations for normal end users can

pose challenges. The main issue is the potential confusion that arises from displaying words with multiple PoS-tags like "measure/verb" and "measure/noun" in the tag-cloud. For non-technical users, the inclusion of PoS-tags in the tag-cloud can be difficult to interpret without proper guidance. The presence of multiple PoS-tags for a single word may lead to confusion about the different senses and interpretations associated with that word.

Imagine a tag-cloud that includes both "measure/verb" and "measure/noun" as prominent tags. Non-technical users may struggle to comprehend the distinction between these senses and how they relate to the retrieved videos. This complexity can hinder their understanding of the underlying meaning and make the explanation less accessible and intuitive.

To address this challenge, it becomes crucial to develop user-friendly visualization techniques that simplify the presentation of PoS-tags in tag-cloud explanations in a more user-friendly manner, so that the non-technical users can more easily grasp the meaning behind the retrieval results and make informed interpretations without being overwhelmed by the complexity of PoS-tags. Currently, the challenge of effectively visualizing PoS-tags in tag-cloud based explanations for non-technical users remains unaddressed. While we recognize the potential confusion that can arise from displaying PoS-tags like "measure/verb" and "measure/noun" in the tag-cloud, we have not yet developed specific solutions to simplify the presentation of these tags for better user understanding.

3.4 Discussion

In this chapter, we study the impact of incorporating PoS-tags within the dual encoding model proposed by *Dong et al.* [1] for TTV and VTT tasks. This research aims to provide valuable insights into the significance of considering syntactic information for enhancing the accuracy and relevance of video-text retrieval systems. The PoS-tagged dual encoding model represents an innovative approach to enhancing the accuracy and explainability of video-text retrieval systems. The experiments conducted on various datasets, including the MSR-VTT dataset for TTV and VTT retrieval tasks, as well as the AVS datasets IACC.3 and V3C1, provide valuable insights into the performance and potential advantages of the PoS-tagged model.

One of the key findings from the experiments is the improvement in retrieval accuracy achieved by the PoS-tagged dual encoding model. In the **MSR-VTT experiments**, the model exhibited a modest increase in the SumR score, indicating a higher proportion of queries for which at least one correct document is retrieved among the top-K results. Although the improvement was not particularly significant, it is worth noting that only a small percentage of queries in the dataset contained ambiguous words that could benefit from PoS tagging. This suggests that the impact of PoS tagging on retrieval accuracy may be more pronounced when dealing with queries that have a higher proportion of ambiguous words or complex

linguistic structures.

However, when evaluating a subset of queries specifically designed to contain ambiguous words, the PoS-tagged model demonstrated a substantial improvement in retrieval accuracy. The SumR score increased significantly, highlighting the effectiveness of incorporating PoS tagging in enhancing the ranking of relevant videos. This finding is important as it demonstrates that the PoS-tagged model is particularly beneficial when dealing with queries that require disambiguation or where the presence of specific parts of speech is crucial for accurate retrieval. Visual examples of text-to-video retrieval in Figure 3.4(b) further illustrated the enhanced ranking accuracy achieved by the PoS-tagged model.

Furthermore, the experiments conducted on the **AVS datasets** provide additional evidence of the PoS-tagged dual encoding model's effectiveness. The model outperformed the original dual encoding model in terms of inferred Average Precision (infAP) for the TRECVID AVS 2018 tasks. Although the performance varied across different years, with the PoS-tagged model achieving the best performance in 2018, the overall results indicate that incorporating PoS tagging can lead to improved retrieval accuracy in the AVS task as well.

In addition to enhancing retrieval accuracy, the integration of PoS tagging in the dual encoding model has a significant impact on the **explainability** of the retrieved results. By providing linguistic context and highlighting relevant linguistic features, the PoS-tags enable a deeper understanding of the underlying meaning and semantic composition of the text. This improved explainability is valuable in various applications where understanding the reasons behind the retrieval decisions is important. For instance, in content recommendation systems, being able to explain why certain videos were recommended based on specific linguistic features can enhance user trust and satisfaction. Moreover, the PoS-tagged dual encoding model enables analysis of the impact of different parts of speech on retrieval decisions. By examining how the presence or absence of certain PoS-tags affects the relevance of retrieved videos, insights can be gained into the model's decision-making process. This level of analysis and interpretability is especially valuable in domains where precise control over the retrieval process is required, such as in specialized video search applications or when catering to specific user preferences.

Despite the promising results and advantages of the PoS-tagged dual encoding model, there are certain **limitations** and potential areas for improvement. One limitation is the reliance on external PoS taggers, such as TreeTagger and WordNet, which may introduce errors or inconsistencies in the tagging process. Integrating more advanced and accurate PoS tagging techniques or exploring domain-specific PoS taggers could potentially improve the overall performance of the model. Additionally, the experiments showed that the improvement in retrieval accuracy achieved by the PoS-tagged model was more pronounced when dealing with queries containing ambiguous words. Further research could explore techniques

to automatically identify and prioritize queries that are likely to benefit from PoS tagging, allowing for more targeted and efficient use of the model.

Moreover, in future research, it is also crucial to consider the expansion of the concept vocabulary by incorporating meaningful phrases rather than solely relying on single-word concepts. For instance, phrases like “video game” or “measuring tape” carry specific semantic associations that are only captured when these words are analyzed together. By incorporating noun-verb phrases in the concept vocabulary, the retrieval system can achieve more accurate classification and provide clearer explanations to non-technical end- user. However, expanding the concept vocabulary to include phrases on a large scale requires efficient parsing techniques to avoid introducing unnecessary noise. By addressing this challenge, we can significantly enhance the accuracy and explainability of video-text retrieval systems, ultimately improving the user experience and enabling applications in diverse domains.

In conclusion, the PoS-tagged dual encoding model proposed by Dong et al. [1] demonstrates its effectiveness in enhancing the accuracy and explainability of video-text retrieval systems. The experiments conducted on multiple datasets provide evidence of improved retrieval performance, particularly for queries containing ambiguous words, and highlight the model’s ability to provide linguistic context and insights into the decision-making process. The integration of PoS tagging in the dual encoding model opens up new directions for enhancing the interpretability of video-text retrieval systems.

However, despite this improvement, *the concept space does not surpass the latent space in terms of performance*. To address this limitation, further analysis and exploration are required to understand the relationship between the latent space and concept space more deeply. This will be explored in the next chapter.

Chapter 4

A General Framework for Complementarity Analysis of Dual Space Models

4.1 Introduction

Hybrid models that combine the similarity results between video-text from latent space and concept space representations in order to get final weighted similarity between video-text pair have been introduced in multimedia processing and retrieval [1, 2] (see Section 2.2.3). The results of hybrid approaches have demonstrated improved performance compared to using either concept-based (Section 2.2.1) or concept-free (Section 2.2.2) models alone. This suggests that concept-based and concept-free approaches are taking benefit from each other [1]. However, no specific study has thoroughly investigated the complementarity or relationship of such hybrid approaches. Understanding their complementarity, by exploring how they can mutually enhance each other's effectiveness and insights, is of paramount importance (see Section 2.5).

We have seen in Chapter 3.3 that, despite the complementarity between these two spaces, the concept space alone does not surpass the accuracy of latent space. Hence, the major drawback of the proposed models (i.e. dual encoding model [1, 2]) is the lack of analysis and human interpretability in the visual or textual encoding process, making it challenging to analyze the representations in latent or concept space. This lack of interpretability and analysis poses issues in debugging the embedding model, understanding relationships between spaces, detecting biases, analyzing their representation power, and explaining system decisions.

To address the challenge of interpretability, this chapter aims to provide a general framework for a detailed analysis of hybrid approaches, employing various visual and statistical

methods. The analysis explores the inter and intra-relationships of latent spaces and concept spaces. Inter-relationship analysis investigates connections and similarities between different spaces i.e latent space and concept space.

In [155], *Dosovitskiy and Brox* propose a method to generate images using perceptual similarity metrics based on deep networks, highlighting the inter-relationship between visual representations and perceptual qualities. However, intra-relationship analysis focuses on the internal structure of individual space, revealing patterns, clusters, and relationships between data points within the same space. The methods to interpret the internal structure and representation includes t-Distributed Stochastic Neighbor Embedding (t-SNE) [156], Principal Component Analysis (PCA), or Uniform Manifold Approximation and Projection (UMAP) to project high-dimensional data into a 2D space [17]. This allows observation of clusters and comparison between the original and latent spaces to verify if the properties of the original data are preserved in the latent vectors. Additionally, domain-specific methods are used to visualize semantic meanings in latent space, such as creating attribute vectors for opposing concepts [157] or clustering latent variables based on hierarchical structures in documents [158].

Our hypothesis aims to verify the complementarity between latent space and concept space in the proposed hybrid-based retrieval models [1, 2]. Additionally, we seek to investigate whether these hybrid models are indeed based on ensemble learning algorithms. Ensemble learning [159, 160] is a general and reliable technique, mostly used for classification, which co-ordinates the outputs of multiple supervised learning algorithms with the same architecture trained with different initializations using diversified data. Different initializations allow the machine learning models to have different learning paths, reducing the overall error by averaging out the individual error due to diversity of results and errors. There are two ways to design ensemble learning algorithms. The first approach is to train the machine learning models independently several times in such a way that the resulting set of models are accurate and diverse. The second approach [161] designs the ensemble algorithm in a coupled fashion, where the models are trained jointly and weighted scores for each model gives a good fit to the data [162, 163]. As our analysis is based on a dual encoding pipeline, which consists of CNN, hence, the ensemble learning technique in our case is applied to CNN-based models. In the case of hybrid approach models, we want to see if the latent space and concept space can be the same space with different initialization.

Despite its importance, the analysis of latent spaces and concept spaces poses several challenges. One fundamental challenge is the selection of an appropriate embedding technique. Numerous methods exist for constructing latent spaces, such as principal component analysis (PCA), autoencoders, and generative adversarial networks (GANs). Each technique has its strengths and limitations, and the choice depends on the specific data characteristics and analysis objectives. As we are based on dual encoding model proposed by *Dong et*

al. [1], we will be using the same encoding models and training strategy as the authors used. However, another model i.e. "Interpretable Embedding model" proposed by *Wu et al.* [2] is also based on a similar technique as of dual encoding model, hence the observations made could be valid on their interpretable model as well.

Another challenge is evaluating and validating the quality of the embeddings. It is crucial to assess how well the latent space or concept space captures the relevant information and preserves the essential structure of the data. This accuracy is critical because any shortcomings in this representation could potentially result in inaccurate or ineffective outcomes in subsequent tasks, including classification and retrieval. Additionally, the curse of dimensionality poses a challenge in analyzing latent spaces. As the dimensionality of the data increases, the available sample density decreases exponentially, making it difficult to accurately capture the underlying distribution. Understanding this challenge underscores the importance of maintaining data quality and interpretability. It helps to identify potential issues that arise when dealing with high-dimensional data. Techniques like dimensionality reduction aim to address this challenge by mapping high-dimensional data into lower-dimensional spaces while preserving relevant information. However, striking a balance between dimensionality reduction and retaining meaningful features remains a constant challenge.

Objective

The goal is to explore the intra and inter-relationships between latent space and concept space. By examining how these spaces interact and complement each other, we can gain deeper insights into the strengths and limitations of the dual encoding model [1]. This analysis will provide valuable guidance for developing more effective cross-modal retrieval systems.

4.2 General framework for Latent and Concept Space Analysis

In this section, we present the methodology employed to analyze the complementarity between the latent space and concept space in the dual encoding model [1]. The aim is to understand the extent to which these models complement each other in terms of representation capabilities and information content. The methodology is designed to address the research questions, which are designed to provide insights into the intra and inter-relationship between latent space and concept space, as well as their synergy.

4.2.1 Research Questions

4.2.1.1 Optimal Dimensionality Analysis

The first research question (R1) aims to determine if the concept and latent spaces have similar optimal dimensions when learned independently. By exploring the optimal dimensions of each space separately, the chapter investigates whether these spaces possess similar representation capabilities. This question is important for understanding if the complementarity between these spaces arises from inherent differences in their representation, or if they share common characteristics. This leads to our first research question:

> *R1: Is the number of optimum dimensions the same in both the concept and latent spaces for two subtasks of cross-modal video-text retrieval, Text-to-Video (TTV) and Video-to-Text (VTT)?*

To answer R1, we employed two techniques: one that does not rely on PCA (Principal Component Analysis) and one that does. Without PCA, we trained the latent space and concept space with varying number of dimensions in order to find the optimal dimensions in both space. In the case of with PCA, we utilized the initially learned high dimensional latent and concept space and employs Principal Component Analysis to explore the salient linear dimensions and their variance in lower dimensional space.

4.2.1.2 Complementarity Analysis

Moving on to the complementarity analysis between the latent and concept spaces, we study the **complementarity** from two points of view, i) Correlation Analysis, and ii) Ensemble learning. The second research question (R2) focuses on assessing whether these spaces represent complementary information considering the correlation between them. Hence, our 2^{nd} research question is as follows:

> *R2: Do the latent and concept spaces represent complementary information?*

This question aims to understand if the latent and concept spaces capture different aspects of the data or if they represent similar kinds of information. To address R2, the chapter employs Canonical Correlation Analysis (CCA) to study the correlation between the two high-dimensional feature spaces. By examining the correlation, the chapter can determine the extent to which these spaces complement each other in terms of the information they capture.

The third research question (R3) explores the complementarity of the latent and concept spaces through **ensemble learning**. This question investigates whether using ensemble

learning techniques in these spaces results in improved performance and demonstrates their complementarity. Ensemble learning allows for the combination of multiple models or representations with different initialization, and if the performance of the ensemble model is comparable to using either space independently, it suggests that the spaces are indeed similar and do not exhibit strong complementarity. The 3^{rd} and last research question to be answered is:

> *R3: Does ensemble learning exhibit complementarity on the latent and concept spaces?*

By addressing R3, the chapter can assess the effectiveness of ensemble learning in exploiting the complementarity between the latent and concept spaces

4.3 Experiments

In this section, we provide an overview of the experimental context and the conducted experiments in our study. The experiments consist of three main components: the dimension study, the correlation study using Canonical Correlation Analysis (CCA), and the ensemble learning analysis.

4.3.1 Experimental Context

Dataset & Evaluation Measures. Similar to Section 3.3.2.1, we performed all of our experiments on the official split of the MSR-VTT dataset [13] and used all important performance evaluation metrics provided for the evaluation of MSR-VTT dataset.

Implementation Details. We chose the dual encoding model proposed by *Dong et al.* [1] for the mapping of visual and textual representation in hybrid space, i.e. shared latent and concept space, as it achieves state-of-the-art performance. We use the PyTorch code provided by the dual coding model[1] to set up the basic architecture of a visual encoding network and a textual encoding network. We employ the frame-level video features of 4,096 dimensions for each video frame provided by the authors, extracted using the pre-trained ResNet-152 and ResNext-101. For the video-text concept features, the concept list is compiled from the training set captions of MSR-VTT. We use a learning rate of 0.0001 and Adam optimizer to train the model, with a batch size of 128.

4.3.2 Optimal Dimension Study without PCA

In the initial phase of our analytic experiments we aim to answer the first research question **(R1)**, we focus on exploring the optimal number of dimensions for both the latent and

[1] https://github.com/danieljf24/hybrid_space

concept spaces. These optimal regions indicate the dimensions at which the performance of latent and concept space reaches its peak. The goal is to identify these optimal regions by evaluating the dual encoding model across different numbers of dimensions in the latent space and concept space. Subsequently, we compare the two spaces to determine if they exhibit similarity in terms of their respective optimal dimensions.

To carry out this analysis, we employ the dual encoding model proposed by Dong et al. [1]. This model enables us to extract video and text features and map them into both the latent and concept spaces. For the latent space investigation, we systematically vary the number of dimensions and evaluate the performance of the dual encoding model at each dimension. This process allows us to pinpoint the optimal regions where the model achieves its highest performance.

Similarly, for the concept space examination, we make modifications to the provided code to focus solely on training and testing the dual encoding model using concept space features. We vary the number of dimensions in the concept space and assess the model's performance to identify the regions of optimal performance.

By conducting these experiments and comparing the performance of both spaces, we aim to gain deeper insights into how the dimensionality affects the effectiveness of the latent and concept representations. Additionally, we seek to determine whether both spaces share similar optimal dimensions or if they diverge in their ideal configurations. This comprehensive investigation will also provide valuable knowledge on how to fine-tune and leverage the latent and concept spaces effectively in multimedia processing and retrieval tasks, ultimately leading to more accurate and interpretable models.

4.3.3 Optimal Dimensionality Study using PCA

In the second part of our optimum dimensionality study, we utilize PCA for re-verification of the optimum dimensions in the latent and concept spaces found in Section 4.3.2. PCA allows us to identify the most significant linear dimensions that capture the variability of the data in these high-dimensional feature spaces.

We apply PCA to the high-dimensional latent and concept features, projecting them onto a lower-dimensional space. From this lower-dimensional representation, we extract the top $\mathbb{K}$ principal components for both the latent and concept spaces. These components capture the most significant variability in the data. By estimating the variance of these top $\mathbb{K}$ principal components, we can evaluate whether the previously identified optimum dimensions in section 4.3.2 remain consistent when considering the dominant dimensions found through PCA.

The combination of the initial analytic experiments (Section 4.3.2) and the PCA-based re-verification provides us with a comprehensive understanding of the optimal dimensionality

in the latent and concept spaces. Overall, this thorough investigation contributes to building more accurate and interpretable deep learning models for practical applications.

4.3.4 Complementarity Analysis using CCA and Ensemble Learning

We study here the complementarity between the latent and concept spaces. This is done using two approaches. The first one is to analyze the correlation between two different high dimensional feature spaces using *Canonical Correlation Analysis (CCA)*. This will evaluate the similarity in their representations and will answer the question **R2**. The second one is to compare the performances of these two spaces used independently and jointly, using ensemble learning: if the results using *ensemble learning* are the same than assuming complementarity, this will show that the two spaces hold same information.

Correlation between Vector Spaces using CCA. Using the notation of [1], consider the set of video features $f(v)$ and text features $f(s)$ in latent space as $\mathbb{X}_l \in \mathbb{R}^{N\times p}$ and the set of video features $g(v)$ and text features $g(s)$ in concept space as $\mathbb{X}_c \in \mathbb{R}^{N\times q}$, with dimensions p and q respectively, where N number of videos and text/captions in the dataset. The feature vector for i^{th} video or text in latent and concept spaces can be denoted as $\mathbb{X}_l^i$ and $\mathbb{X}_c^i$ respectively. In this section, we investigate the relationships between the feature vectors of all video-text in latent space ($\mathbb{X}_l$) and in concept space ($\mathbb{X}_c$) using Canonical Correlation Analysis (CCA).

Consider $\mathbb{M}$-Dimensional CCA transformed latent space features $\mathbb{X}_l^i \in \mathbb{X}_l$ and concept space features $\mathbb{X}_c^i \in \mathbb{X}_c$ for i^{th} video-text as $\tilde{\mathbb{X}}_l^i \in \mathbb{R}^{N\times\mathbb{M}}$ and $\tilde{\mathbb{X}}_c^i \in \mathbb{R}^{N\times\mathbb{M}}$ respectively, where $\mathbb{M}$ being chosen as minimum of the latent space dimension p and concept space dimension q, mathematically $\mathbb{M} = min(p,q)$. We define the highly correlated feature vectors $(\tilde{\mathbb{X}}_l, \tilde{\mathbb{X}}_c)$ as the projection of $\mathbb{X}_l$ and $\mathbb{X}_c$ onto CCA basis vectors, along with which the correlation was above the threshold $\mathbb{T}_h$. Let us denote two linear transformation matrices corresponding to these i^{th} correlated basis vectors ($\tilde{\mathbb{X}}_l^i$ and $\tilde{\mathbb{X}}_c^i$) by A_l^i and A_c^i respectively for latent space and concept space. The correlated projections of $\tilde{\mathbb{X}}_l^i$, and $\tilde{\mathbb{X}}_c^i$ are given by:

$$\begin{aligned} \tilde{\mathbb{X}}_l^i &= A_l^i \mathbb{X}_l^i \\ \tilde{\mathbb{X}}_c^i &= A_c^i \mathbb{X}_c^i \end{aligned} \tag{4.1}$$

Here $\tilde{\mathbb{X}}_l^i$, and $\tilde{\mathbb{X}}_c^i$ can be considered as correlated components embedded in $\mathbb{X}_l^i$, and $\mathbb{X}_c^i$. Hence, using these correlated components, we want to observe the correlation measure between variables of latent space vectors and concept space vectors.

If there are groups with high correlation amongst variables which cover a good amount of variance, then there might be overlapped feature representation amongst the spaces, which answers **R2**.

Ensemble Learning. Ensemble learning of neural networks is a reliable technique to increase the performance of models for a task. Due to the presence of several local minima, multiple trainings of the exact same neural network architecture with and without same hyper-parameters can reach a different distribution of errors in model training. Hence, combining their outputs lead to improved performance on the overall task. In order to answer the **R3**, i.e. to study the reason of performance gain in hybrid approaches, we deeply analyze the model [1] with ensemble learning approach.

The complementarity analysis consists of training the dual encoding model in three different ways: (i) *Homogeneous non-coupled model*: where the latent space model and concept space model are trained independently with same configuration and hyper-parameters, then the retrieval accuracy is combined using weighted average of two models; (ii) *Homogeneous coupled model*: where the latent space and concept space are trained jointly with same hyper-parameters; and (iii) *Heterogeneous coupled model*: in which the latent space and concept space are trained jointly with a) similar hyper-parameters for both spaces, b) different hyper-parameter for each space (similar to the hyper-parameter setting in baseline model [1]). Hence, with all these model training scenarios, we want to observe the correlation in the behaviour of latent space and concept space when learned in similar or different settings. We also want to see if the performance gain in TTV and VTT task is either due to ensemble learning or something else.

4.4 Results & Discussion

Within this section, we provide a comprehensive analysis of the results obtained, and the observations made during our investigation into latent and concept space exploration. We delve deeply into the discourse surrounding the complementarity that exists between two separate high-dimensional feature vectors. This rigorous examination enables us to effectively address and answer the three specific research questions that were posited as the focus of this study.

To effectively address our research question (**R1**) on optimum dimensionality (presented in Section 4.2.1.1), we systematically manipulate the dimensions of both the latent space and the concept space. The aim here is to identify the regions that yield the most optimal results, as elaborated in Section 4.3.2. The dynamics of the mean Average Precision (mAP) for both the TTV and VTT retrieval tasks, with respect to the varying number of dimensions in the latent space, are graphically depicted in Figure 4.1(a) and (b). Furthermore, Figure 4.1(c) provides an insightful overview of the average mAP across both the TTV and VTT retrieval tasks.

In order to comprehensively explore the impact of dimensions, we sample latent dimen-

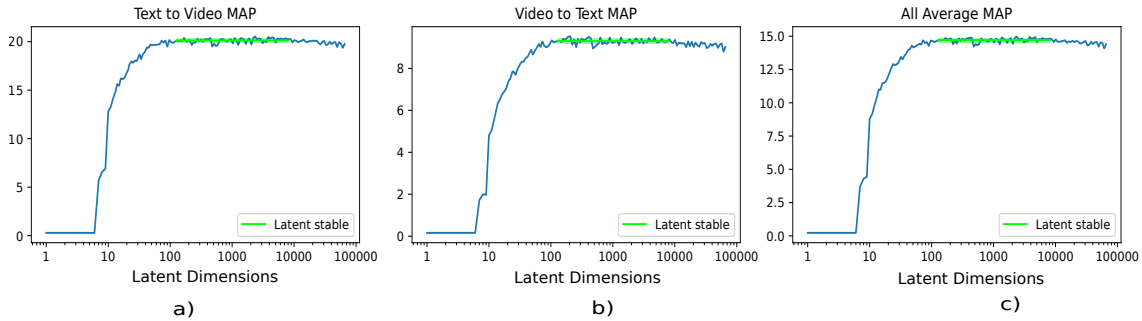


Figure 4.1: Results for latent-only training and latent-only decoding.

sions across a logarithmic scale range, spanning from 16-D to 65536-D. This sampling involves 10 measurements per octave, resulting in a sequence such as 16, 17, 18, 19, 21, 22, and so forth, up to 65536. Mean and standard deviation in the stable region are shown with the green line and box. It's intriguing to note that the performance stabilizes and reaches a plateau around the 200-dimension mark, maintaining this level of stability until approximately, 8000 dimensions (Figure 4.1(c)). Beyond this threshold, there's a gradual decline in performance, at a relatively slow rate. This detailed analysis allows us to gain valuable insights into the behavior of the model across varying dimensions and sheds light on the optimal region for our analysis.

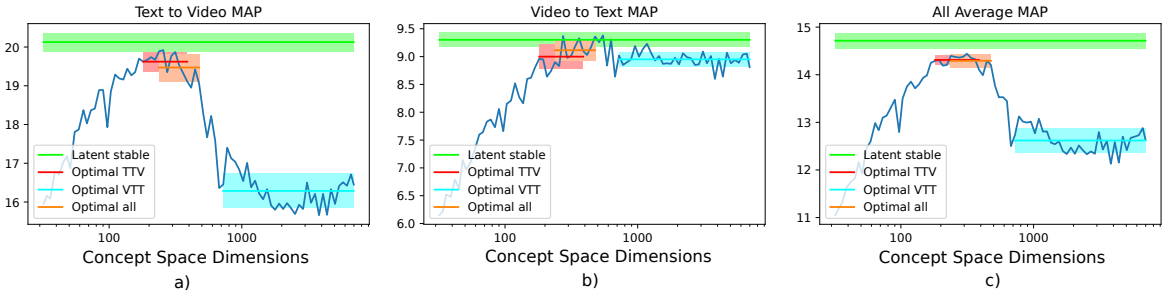


Figure 4.2: Results for concept-only training and concept-only decoding. Mean Average Precision (mAP) metric as a function of the number of concept dimensions.

Figure 4.2(a) and (b) shows the evolution of the mAP for the TTV and VTT retrieval tasks as a function of the number of ***concept space dimensions***. Figure 4.2(c) shows the evolution of the avegare mAP for the combined TTV and VTT retrieval tasks. This number is sampled between from 16-D to 6983-D (i.e., the maximum number of selected concepts), also on a log scale with 10 samples per octave. Mean and standard deviation in the stable region of the *latent space* are shown for comparison with the green line and box. Means and standard deviations in the optimal regions for the Text-to-Video and Video-to-Text retrieval tasks in the *concept space* are shown with lines and boxes in red and cyan, respectively.

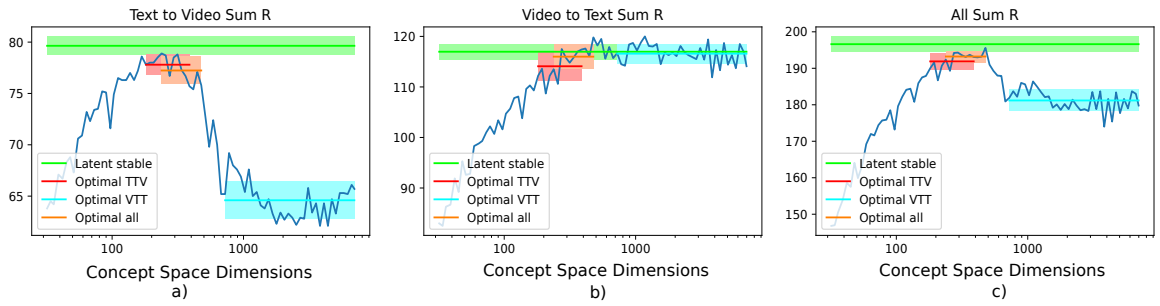


Figure 4.3: Results for concept-only training and concept-only decoding. SumR metric as a function of the number of concept dimensions.

This time, the plots for the TTV and VTT tasks are very different and are shown in Figure 4.2(a) and (b). There are different optimal regions for the TTV and VTT tasks and these optimal regions are much narrower, around 200 for the TTV task and overall and beyond 500 for the VTT task. The difference is even clearer on the sumR metric shown in Figure 4.3. The asymmetry between the TTV and VTT task's behaviors can be explained by the asymmetry in the ratio of captions to videos in the MSR-VTT dataset [13].

Overall, the optimum number of dimensions is around 200 for both the latent and concept spaces, and this is much lower than the numbers used by *Dong et al.* in dual encoding model [1] (1536-D and 512-D). The performance using concepts is a bit lower, but not much for the optimal values. The fact that the values are lower and even more outside the optimal region can be explained by the fact that the classification task associated with the concept space places strong constraints on it. On all curves, fluctuations are observed due to the effect of the random initialization in the training, and they are at the same level as what is observed when running the same experiments multiple times.

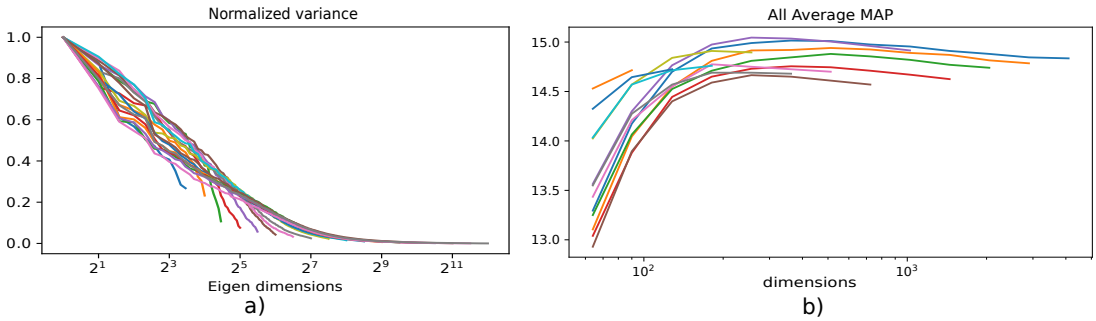


Figure 4.4: PCA performance analysis for latent space. The X-axis represents the number of principal components, whereas Y-axis represents the normalized variance in (a) and the all-average mAP in % in (b).

The research question R1 is also studied using PCA to re-verify the number of optimum

dimensions in latent space, as mentioned in Section 4.3.3. For the sake of simplicity, we only analyse here the latent space. Figure 4.4(a) shows the decrease of the values of the normalized variances as a function of the eigen dimensions for multiple latent space trainings with a variable number of latent dimensions, ranging from 11 to 4096 on a log scale and with two samples per octave (11, 16, 22, 32, 45, 64 ... 1024, 1448, 2048, 2896, 4096). Curves are shown with different colors and the number of latent dimensions used for training can be inferred from the point at which the curve stops on the right side. As can be seen, whatever the initial number of latent dimensions, there is no significant variances beyond a few hundred eigenvalues (around $2^7 - 2^9$ eigenvalues). The cumulative variance continues to increase beyond (not shown) but this likely corresponds to noise.

Figure 4.4(b) shows the evolution of the performance (all averag mAP) of the same multiple trainings when reducing the size of the latent representations by keeping only the components with the highest variances. The difference in the initial performance (at the point which is most on the right for each curve) likely comes again from the random initialization and is also in the standard deviation obtained with multiple identical experiments. However, the variation of performance on each single curve corresponds to a same initialization and is expected to be significant. We see that for those starting with a high number of dimensions there is a slight increase in performance, confirming that the components with lowest variances contain mostly noise. The best performance seems to be reached by training with a number of dimensions larger than the optimum value found either directly (without PCA) or indirectly (with PCA) and then reducing the space size using PCA, *e.g.*, 4096 $\rightarrow$ 256.

As the number of optimum dimensions with and without PCA are very close for the latent space, these experiments show that even with dimensionality reduction of latent representation, the optimum regions of both spaces are still the same with the slight increase in performance due to noise reduction, which shows that (i) the properties of original high dimensional latent space are preserved in reduced dimensional space, and (ii) the observation holds that these two spaces may have a lot in common, leading to answer **yes to R1**.

To answer this question **R2** as described in Section 4.2.1.2, we rely on a CCA study, in order to find out the correlation and complementarity between two spaces.

One large difference between the latent space and the concept space is that the concept space is associated with a classification task, while the latent space is not. If the classification task is removed, the concept space becomes just a second latent space with different characteristics (e.g., using a Jaccard similarity instead of a cosine). So, in our analysis, we consider four possible combinations of latent and/or concept spaces (as also mentioned in Section 4.3.4): (i) two identical latent spaces independently trained and lately fused (i.e., **homogeneous non-coupled**), (ii) two identical latent spaces jointly trained (i.e., **homogeneous coupled**), (iii) two different latent spaces (respectively with cosine and Jaccard similarities)

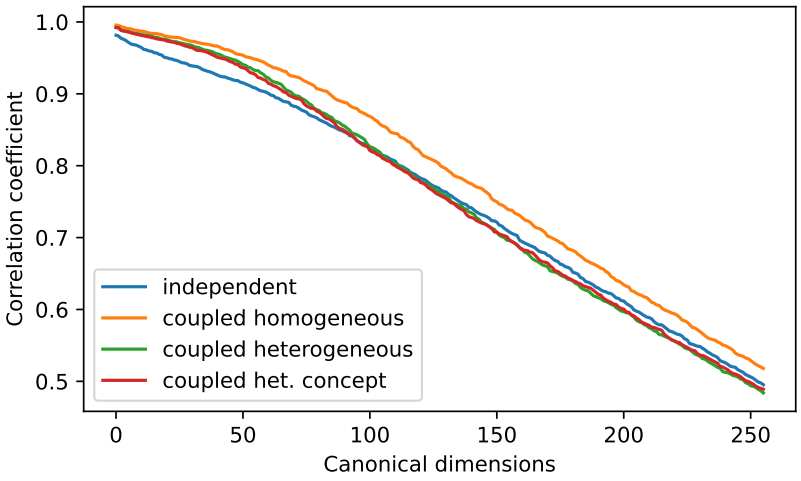


Figure 4.5: CCA analysis: Top 256 canonical correlation of independent (non-coupled), coupled homogeneous, coupled heterogeneous and coupled heterogeneous concept training's.

jointly trained (i.e., **heterogeneous coupled**), and (iv) the same with additionally a classification task on the second latent space, which turns it into a concept space, and the overall system as the regular hybrid one (**coupled heterogeneous concept**). For better comparisons, we used 512 as the dimension for all spaces. Figure 4.5 plots the top-256 canonical correlations of the latent-latent or latent-concept mappings in the four configuration just mentioned. Most correlated is coupled homogeneous training of two spaces; whereas least correlated is independent training of two spaces. The correlations are all quite high with similar profiles but still small differences. The independent training is the least correlated, the coupled homogeneous one is the most correlated, the other two, coupled heterogeneous and coupled heterogeneous concept (hybrid) are in between and almost identical, indicating that the classification task makes no difference in the correlation.

The experiments with CCA showed that there is high correlation between the latent space and concept space when considering the same hyperparameters, for instance same distance metrics for calculating similarity in two spaces. There is the least correlation when considering independent training of two spaces, as the spaces are not optimized with the constraints present in the other space. Overall, we can answer **No to R2**.

To answer question **R3** (Section 4.2.1.2), we report a quantitative evaluation of the four combinations described above used in ensemble learning experiments.

The first part of the table 4.1 represents the evaluation of dual encoding model only on one space (i.e. latent space), whereas the second part of table 4.1 shows the evaluation of the four combinations on the both spaces fused using the standard MSR-VTT metrics for both first and second part, the last row of the second part corresponding to the regular dual encoding hybrid system [1]. The values correspond to the average of 10 identical runs with

Table 4.1: Ensemble learning Experiments on MSR-VTT. Larger R@{1,5,10}, mAP, and smaller Med r indicate better performance.

Method	Text-to-Video Retrieval					Video-to-Text Retrieval					SumR
	R@1	R@5	R@10	Med r	mAP	R@1	R@5	R@10	Med r	mAP	
Evaluation on a single space:											
Latent independent 1536 (10)	10.94	29.12	39.73	19.30	20.21	19.00	42.60	54.81	8.10	9.26	196.20
Latent independent 512 (20)	10.88	29.06	39.73	19.25	20.15	19.42	42.91	55.20	7.95	9.30	197.21
Latent-latent coupled homogeneous (10)	11.17	29.83	40.58	18.10	20.63	19.95	43.77	56.31	7.60	9.57	201.61
Latent-latent coupled heterogeneous (10)	11.37	30.25	41.11	17.80	20.93	19.85	43.70	56.26	7.60	9.63	202.54
Latent-concept coupled heterogeneous (10)	11.42	30.29	41.16	17.70	21.00	19.65	43.24	55.84	7.90	9.57	201.60
Evaluation on two spaces:											
Latent-latent indep. homogeneous (10)	11.50	30.30	41.22	17.55	21.04	20.92	44.78	56.99	7.25	9.89	205.71
Latent-latent coupled homogeneous (10)	11.41	30.31	41.16	17.70	20.98	20.30	44.57	56.85	7.10	9.80	204.60
Latent-latent coupled heterogeneous (10)	11.78	31.12	42.28	16.20	21.60	21.23	45.65	58.08	7.00	10.36	210.14
Latent-concept coupled heterogeneous (10)	11.76	30.98	42.10	16.40	21.52	20.25	44.74	57.48	7.10	10.09	207.31

different random initialization so that we can get an estimate of the statistical significance of the differences between the experiments using a Z-test. The first part of the table shows the performance when performing the task using the first latent space only and the first line is inserted for showing that there is no statistically significant difference between a 1536-D latent space and a 512-D one.

When considering evaluation on fused spaces, there is no statistically significant difference between the independent and coupled trainings for the homogeneous latent spaces. The main best performance is achieved by the latent-latent coupled heterogeneous method that uses latent spaces of different types (with cosine and Jaccard similarities). The experiments on latent-latent coupled homogeneous underperform latent-latent coupled heterogeneous: there is a decrease in performance if cosine similarity is used in concept space. There is a slight decrease in performance when adding the classification task but the statistical significance is marginal.

The ensemble learning experiments show that there is no significance difference in performance of two independent latent space with different dimensions. The significant improvement in retrieval comes from training two latent spaces with different similarity computation techniques. This analysis leads us also to answer **No to R3**.

4.5 Discussion

The conducted study in this chapter aimed to delve into the relationship and complementarity between latent and concept spaces within the context of cross-modal video-text retrieval. The comprehensive analysis and experimental results shed light on the optimal dimensions of these spaces, their potential complementarity, and the impact of ensemble learning. The discussion below summarizes the key findings and their implications.

4.5.1 Optimal Dimensions and PCA Analysis (R1)

The investigation into the optimal dimensions of both the latent and concept spaces revealed intriguing insights. Through a systematic exploration of the mAP and SumR metric with varying dimensions, it was established that the optimal regions of both spaces align around 200 dimensions. This optimal dimensionality is significantly lower than the dimensions used in the dual encoding model, highlighting the potential for dimensionality reduction without substantial performance loss.

The PCA-based analysis provided further validation of these findings. By reducing the dimensionality of the latent space and observing the preservation of optimal regions and performance and even a slight gain in performance because of noise reduction using PCA. It became evident that the key properties of the original latent space are retained even in lower dimensions. This conclusion reaffirmed that the latent and concept spaces share a significant commonality in terms of their optimal dimensions.

4.5.2 Correlation and Complementarity (R2)

To assess the complementarity between the latent and concept spaces, Canonical Correlation Analysis (CCA) was employed. The results of the CCA analysis indicated a high correlation between the latent and concept spaces when considering the same hyperparameters. However, the correlation decreased when spaces were independently trained, emphasizing that the spaces are not optimized to complement each other when trained separately.

The finding that the correlation between the spaces increases when they are trained jointly highlights the potential for these spaces to leverage shared information for cross-modal retrieval tasks. However, the lack of significant differences between coupled homogeneous and coupled heterogeneous settings suggested that the classification task in the concept space does not strongly contribute to the complementarity between the spaces.

4.5.3 Ensemble Learning and Complementarity (R3)

The study further explored ensemble learning to determine whether it exhibits complementarity between latent and concept spaces. The results showed that there is no significant difference in performance between independent latent spaces with different dimensions. The most substantial improvement was observed when training two latent spaces using different similarity computation techniques (i.e Latent space with cosine and Concept space with Jaccard Coefficient). This indicates that the unique characteristics of the latent and concept space contribute more significantly to complementarity than the classification task of concept space.

By conducting rigorous research in the analysis of latent spaces and concept spaces, the

aim is to enhance the understanding of data distribution. The analysis of the latent space and concept space provides deeper insights into the underlying structure and distribution of the data. By exploring these spaces, researchers can gain a better understanding of how data points are represented, organized, and related to each other. This understanding is crucial for various tasks such as data visualization, clustering, classification, and retrieval. The proposed framework in this chapter allows for exploring the complementarity between the latent space and concept space in order to leverage complementarity for improved performance. By understanding how these spaces can mutually enhance each other's effectiveness, researchers can develop hybrid models or fusion techniques that leverage the strengths of both spaces. This complementary approach has the potential to significantly improve the performance and accuracy of various tasks, such as cross-modal retrieval, recommendation systems, and data fusion. The analysis of latent space and concept space using the proposed framework contributes to the advancement of research and development in various domains. By exploring the inter-relationships, complementarity, and redundancy between these spaces, researchers can uncover new insights, develop novel techniques, and propose improvements to existing models. This contributes to the overall progress of the field, leading to more efficient and effective solutions in areas such as machine learning, computer vision, natural language processing, and data analytics.

In summary, analyzing the latent space and concept space using the proposed framework is important for gaining a deeper understanding of data distribution, improving model selection and configuration, leveraging complementarity for enhanced performance, achieving explainability and interpretability of results, and advancing research and development in relevant domains. This analysis provides valuable insights and guidance for developing more accurate, robust, complementary and interpretable models and systems.

Chapter 5

Causal Inference in Video-Text Retrieval

5.1 Introduction

So far we have learned, especially from the chapter 4, that in hybrid approaches like [1] the concept and latent spaces share similar representation capabilities. Given that the concept space is solely accountable for the explainability of retrieved results in these hybrid approaches, it is plausible to assume that relying solely on the concept space may be able to support both high quality retrieval and explainability. However, the actual causal effect of explanation on retrieval decision has not been discussed or analyzed in state-of-the-art. This chapter introduces two elements to address this: firstly, a method to assess the causal contributions of concept detection scores in the retrieval decision of state-of-the-art system [1]; and secondly, a strategy to enhance the tag(s) contributions in retrieval decision using modified concept probability scores, thus generating more causal explanation in this way.

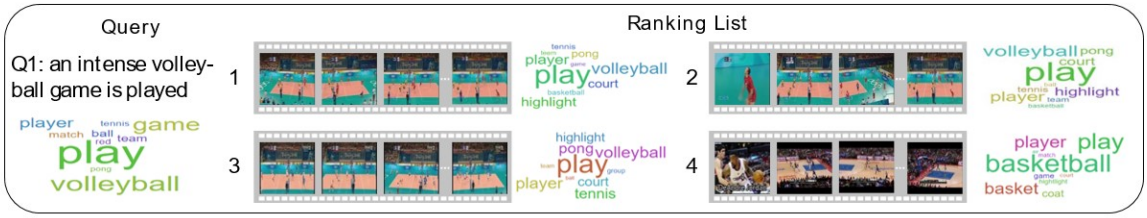


Figure 5.1: Tag clouds for justifying the retrieved results for one query (from [1]).

Regarding current state-of-the-art methods, Figure 5.1 (excerpts from [1]) illustrates how explanation/justification can be provided to a user using a hybrid approach: tag clouds show the concepts found to be the most relevant (with sizes related to their estimated importance) for the query and for the 4 top-ranked retrieved documents. A user is then supposed to evaluate to what extent these tag clouds are actually relevant to the query and to the documents, and to what extent they match. However, it's important to note that these tag-clouds do not provide information about their relative contribution to the overall retrieval decision, and are

then far from reflecting the retrieval causality. Hence, the work in this chapter takes place before such displays: we study how to measure the concept's detections scores' causal contribution in retrieval decisions, and we propose ways to evaluate and improve such causality on such state-of-the-art system, when focusing primarily on the concept representation space.

Therefore, this chapter focuses on evaluation and enhancement of the causality in tag-cloud-based explanations of dual encoding model [1], by quantifying and augmenting the causal contribution of tags, specifically those employed in tag-cloud explanations. By assigning greater weight to fewer relevant concepts, we seek to amplify their causality in the retrieval decision-making process without impacting retrieval accuracy.

Objective

This chapter aims to assess the causal impact of concept classification on retrieval decisions within dual encoding model [1], with the aim of enhancing the quality of explanations provided to users within the system. By introducing a new method for assessing causal contributions and proposing strategies to optimize tag contributions, the objective is to offer more interpretable and effective retrieval models while considering the broader principles of parsimony and model interpretability in the machine learning context.

The concept of assigning varying degrees of importance to elements within a model's decision-making process aligns with the broader goals of *parsimony*. Similar to our concept weighing approach proposed in this chapter, a few papers [164, 165] propose methods to enhance the plausibility of attention maps in RNN or transformer-based models, which are commonly used to explain classification model decisions. The approaches aim to provide more reliable explanations using fewer important words/tokens by giving them greater weight, thus addressing the concept of "parsimony" in attention maps. Additionally, *Liao et al.* [166] develop frameworks for automatically learning compact and parsimonious representations by focusing on a small subset of informative features while disregarding irrelevant or redundant ones. This leads to more compact and interpretable representations. Although *Liao et al.*'s work differs in the specific context of classification or retrieval, it aligns with the idea of achieving parsimony and interpretability in models.

Apart from the parsimonious models, researchers have explored causality in machine learning to gain a deeper understanding of the classification models [148, 167, 168, 169]. Our study also differs from *Yang et al.* [148], who propose a causality-inspired framework for Video-Moment Retrieval, employing a structural causal model to analyze the impact of queries and video content on prediction outcomes. However, we aim to quantify the true causal effect of a set of predicted concepts on retrieval decisions, rather than focusing on the effect of queries and video content on prediction. To the best of our knowledge, we are the

first to provide a quantitative measure of the causal contribution of visual concept classes in retrieval explanations, contributing to a better understanding and interpretation of retrieval models.

For the study of causality in visual explanation in the form of tag-clouds for video-text retrieval, we rely again on the dual stream implementation of [1], which uses a *dual space*, and a *dual task* learning approach, where the system simultaneously performs video-text retrieval and video and text classification tasks (see more details in Section 2.2.3). Additionally, we detail some problems in the dual task approach [1], *e.g.* issues in learning of several detectors and that the target task ground truth features lower the evaluation measures.

5.2 Analysis of causality

5.2.1 Quantifying causality

If we consider the interpretable embedding model of [2] or only the concept space part of a dual space dual encoding model of [1], video and text samples are eventually represented *only* by the detection scores of the selected concepts / tags. Then, for either a VTT or a VTT retrieval task, the ranking of the test samples is performed *only* on the basis of the similarity of these concept-based representations to that of the query in the other modality. [1] uses by default the Jaccard similarity function for the concepts between a video sample v and a text sample s:

$$sim_{con}(v,s) = \frac{\sum_{i=1}^{i=K} \min(g(v)_i, g(s)_i)}{\sum_{i=1}^{i=K} \max(g(v)_i, g(s)_i)} \tag{5.1}$$

with g being the function projecting v and s into the concept space with $g(x)_i$ being the detection score for tag i for the sample x, and K being the number of dimensions of the concept space, which is also the number of selected tags. The g function contains a final sigmoid function that normalizes the concept detection scores between 0 and 1 (used also in the binary cross-entropy during the concept classification training).

The cosine similarity function may also be considered, as it is already used by default on the latent space of [1]. Such cosine similarity is defined as:

$$sim_{con}(v,s) = \frac{h(v).h(s)}{\|h(v)\|.\|h(s)\|} = \frac{\sum_{i=1}^{i=K} h(v)_i.h(s)_i}{\|h(v)\|.\|h(s)\|} \tag{5.2}$$

with h being the function projecting v and s into the concept space without the final sigmoid function. In principle, It is also possible to use the cosine similarity on the post-sigmoid detection scores but this would not be consistent with how the cosine similarity is used in the latent space and the pre-sigmoid (logit) values equally represent the likeliness of presence of a tag in a sample in a way is easily understandable by humans.

We now propose to quantify the *causality* of a group of tags (which may be those presented using clouds as in figure 5.1) in a similarity value used for ranking of retrieved results by the sum of their *relative* overall contribution to this similarity value. The idea of measuring the contribution of each tag aligns with the notion of "feature importance" [170, 171]. Feature importance methods, as defined in [170], quantify the contribution of a feature to the model performance. However, in our particular case the feature corresponds to tags, and we seek to quantify the relative contribution of tags through which we will be able to assess the contribution of tags to the overall retrieval decision. Consequently, we consistently refer "relative contribution" as "feature importance" in the following. We observe that in the *sim* functions presented above (equations (5.1), and (5.2)), the numerators are based on a sum of per-tag terms. As we are interested in the relative importance of individual tags or of a group of tags, we may get rid of the *sim* denominators and normalize the terms so that the sum of their absolute values is equal to one (all values are positive in the Jaccard case but not necessarily in the cosine one). This gives, respectively for the Jaccard and cosine, individual tag contributions or feature importance:

$$w_i(v,s) = \frac{\min(g(v)_i, g(s)_i)}{\sum_{j=1}^{j=K} \min(g(v)_j, g(s)_j)} \quad \text{or} \quad \frac{|h(v)_i.h(s)_i|}{\sum_{j=1}^{j=K} |h(v)_j.h(s)_j|} \tag{5.3}$$

Based on equation (5.3), the causal effect, defined in $[0,1]$, of a set of tags G in the similarity computed between a video sample v and a text sample s is defined as:

$$c(G,v,s) = \sum_{i \in G} w_i(v,s) \tag{5.4}$$

We define the "causality at k" as the causality defined as in equation (5.4) with G corresponding to the k tags contributing the most to the computation of the similarity score:

$$c_k(v,s) = \max_{G \subset [\![1,K]\!], |G|=k} \sum_{i \in G} w_i(v,s) \tag{5.5}$$

From this measure defined for one pair (v,s), we derive global statistical measures on a whole cross-modal collection by computing statistics such as the mean (equation (5.6)) and the standard deviation of this value on a set of pairs P.

$$C_k(P) = \frac{1}{|P|} \sum_{(v,s) \in P} c_k(v,s) \tag{5.6}$$

P may be the set of all possible pairs in the collection or only the set of matching pairs. We can also consider the set of pairs obtained using all the text queries and, for each of them, the top-n retrieved videos, or the opposite using video queries and retrieved texts.

In our case, causality in explanations/justifications relies only on the detection scores for the displayed tags. This is the case by design for the dimensions in a concept space, but not for the dimensions in a purely latent space as these have no meaning for humans. The causal weight of any element coming from the latent space in the concept-based visual explanation/justification should then be strictly zero. In the latent-space-only approach, no concept detection scores are available anyway for displaying tag clouds. However, such scores are available in hybrid approaches, as the decision is made partly on similarities $sim_{lat}(v,s)$ coming from the latent space and partly on similarities $sim_{con}(v,s)$ coming from the concept space. The overall similarity is a weighted sum (after a global scale normalization) $sim(v,s) = \alpha.sim_{lat}(v,s) + (1-\alpha).sim_{con}(v,s)$. The overall causality should logically be a weighted sum based only on the concept scores multiplied by the $(1-\alpha)$ factor, as the causality on the latent part, should be zero.

5.2.2 Evaluating causality of the target system

We have evaluated the tag-detection-score-to-similarity causality using the pre-trained hybrid model provided by the authors of [1] on the MSR-VTT dataset [13]. In this hybrid model, the concept-based similarity accounts for 40% of the global score, As described above, the causal weight of the concepts is reduced accordingly.

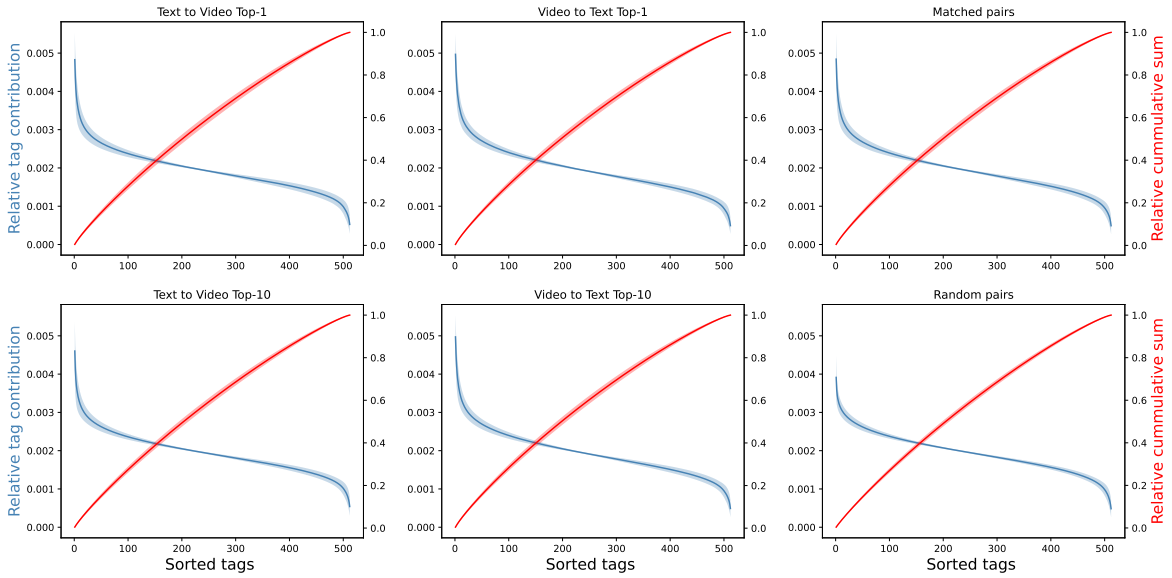


Figure 5.2: Individual and cumulative contribution (mean $\pm$ standard deviation), of the tags ranked by decreasing contributions.

Figure 5.2 shows the mean and standard deviations of the both individual contribution $w_k(v,s)$ and cumulative $c_k(v,s)$ contributions of the tags. The tags are ranked by decreasing contributions for the matched pairs (associated videos and captions). The same curves (individual and cumulative contribution of the tags) for "Text to Video Top-n" and "Video to Text

Top-n" looks very similar, while the curves for all possible combinations gives lower and more spread out values. It's important to note that six distinct sets, denoted as P, containing pairs of (v,s) are considered in this analysis. It's worth emphasizing that the mean value of $c_k(v,s)$ across a set P aligns with the concept of $C_k(P)$ as defined in equation (5.6).

The "Text to Video Top-n" (resp. "Video to Text Top-n") corresponds to the sets of the pairs formed from all the caption queries (resp. video queries) with the corresponding top-n retrieved videos (resp. retrieved captions), with $n = 1$ and $n = 10$. "Matched" corresponds to the set of all videos with their associated captions and "Random" to a set of $100,000$ randomly samples (v,s) pairs from all possible combinations. We observe from these curves that:

- The sets of pairs labeled as "Text to Video Top-n", "Video to Text Top-n", and "Matched" produced highly similar outcomes. This alignment in results was anticipated because, in each of these scenarios, the textual content and the associated videos in the pairs share similarities. On the other hand, when we "randomly" selected pairs, a different pattern emerged. In these cases, the texts and videos generally lacked substantial similarity, leading to a more evenly distributed pattern of individual contributions. This outcome was in line with our expectations. For the remainder of our analysis, we will focus exclusively on examining causality within the "Matched" pairs, as explanations primarily related to pairs with significant similarities, while the other scenarios yield somewhat lower causality values due to their inherent dissimilarity.

- Even when considering the Matched (v,s) pairs in Figure 5.2, it is noteworthy that the individual and cumulative causal contributions of the first few tags (i.e. *top-k*) are very small. Specifically, the first tag's contribution amounts to less than 0.5%, and the accumulated contribution of the first 10 tags remains below 4%. This observation highlights a very important point: that the actual causality in visual explanations such as illustrated in figure 5.1 is very limited. when considering solely concept space of dual encoding model, the overall causality is only 4%. In the case of the comprehensive hybrid approach, this causality figure further reduces to a mere 1.6%. The same behavior is observed for all the Pairs P.

- A large majority of the tags have a significant contribution to the similarity measure and therefore play a crucial role in the ranking decision. Notably, we also observed that, if we only include the initial tens of tags within Jaccard similarity computation, the retrieval performance is very degraded (as illustrated in Figure 5.3). For instance, considering top-10 tags showed in tag-cloud (Figure 5.1) for similarity computation and retrieval, we see that mean average precison (mAP) is downgraded significantly i.e. close 0.3 from 14.46 (considering all 512 concepts in concept vocabulary) (see

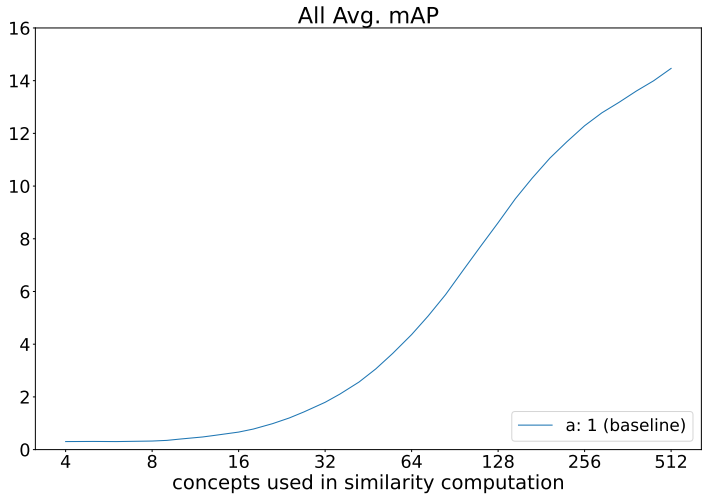


Figure 5.3: Effect of considering limited tags for Jaccard similarity computation on retrieval performance

Figure 5.3).

Even if the displayed tags seems relevant to both the caption query and the retrieved video, it's essential to recognize that the primary factors influencing the ranking decision extend beyond the initial set of tags. In other words, the ranking determination relies heavily on terms and elements beyond the first few tens of tags.

Later on, we computed the causality for the baseline dual encoding model [1] using Equation (5.6) for quantitative evaluation. Table 5.1 presents the retrieval performance and the causality at 10 and at 30 of the original hybrid approach from [1], as well as of a number of variants aiming at improving the performance and/or the causality values. "2048-d latent" corresponds to a latent-space only version; "1536d+512d hybrid" is the original hybrid (GitHub) version "512-d (hyb. train.)" is the same hybrid system in which only the concept-based part is used for the ranking.

Table 5.1: Comparison on the MSR-VTT task [13] for the original hybrid approach [1] and for some selected variant. mAP (3^{rd} last column) represents average of the TTV and VTT mAPs and last two "C@n" columns for the causality at n on the matched pairs. Metrics are same as in [1] except "C@n", and described in Section 5.4.

Method	**Text-to-Video Retrieval**					**Video-to-Text Retrieval**					**SumR**	**mAP**	**C@10**	**C@30**
	R@1	R@5	R@10	Med r	mAP	R@1	R@5	R@10	Med r	mAP				
2048-d latent [1]	11.0	29.2	39.8	19	20.2	18.8	42.7	56.2	8	9.3	197.7	14.0	n/a	n/a
1536d+512d hybrid	11.8	30.6	41.8	17	21.4	21.6	45.9	58.5	7	10.3	210.2	15.8	1.6	4.0
512-d (hyb. train.)	9.7	26.2	36.2	25	18.2	19.3	43.8	56.0	8	9.2	191.1	13.7	3.9	10.0
512-d Jaccard	10.4	28.2	39.1	20	19.5	19.8	42.4	55.0	8	9.4	194.8	14.5	4.0	10.0
256-d Jaccard	10.9	29.2	40.2	18	20.3	18.2	41.3	53.4	9	9.1	193.2	14.7	8.2	19.6
512-d cosine	10.6	28.8	39.2	20	19.9	20.3	44.0	56.6	7	9.9	199.6	14.9	10.4	23.5
256-d cosine	11.0	29.5	40.4	18	20.5	19.8	44.2	56.2	8	9.8	201.0	15.1	17.4	37.8

In the quantitative experiments, we explored the concept of causality in various scenarios

using five distinct variations. These variations involved two important aspects: the inclusion and exclusion of latent space from dual encoding model [1], and the use of different vocabulary size for the concept space, specifically 512-d Jaccard (original dual encoding configuration for concept space), and 256-d Jaccard. Additionally, we conducted experiments with concept-only training, replacing the Jaccard similarity metric with cosine similarity, while still using the same vocabulary sizes of "512-d cosine" and "256-d cosine," which are typically used in the context of latent space.

We considered five different variants in order to assess the behavior of causality in different scenarios. We looked at five different approaches, considering whether to include a latent space and using different vocabulary sizes. We also replaced one similarity metric with another while keeping the vocabulary sizes the same. This allowed us to analyze the impact of these variations on our results.

Regarding causality, we chose the causalities at 10 and 30 as they correspond to practically useful values in the sense that 10 is the number of tags that a user can grasp simultaneously, e.g.; [172] mentions that human are unable to process more than a few, typically 7 ± 2, stimulus at one time, and 30 is a reasonable bound on the number of tags that could be validly assigned to a given caption or video. In fact, explanations involving more than these numbers are unlikely to be causally correct and the components beyond them would likely be used just as latent dimension in a quite opaque way.

The “512-d (hyb. train.)” case corresponds to the curves displayed in figure 5.2; the causality for “2048-d latent” would be 0 (or rather n/a); and the causality for “1536d+512d hybrid” is in between. The causality for “512-d cosine” is significantly higher because the decreasing of the sorted component values happens to be much faster in this case. For both the Jaccard and cosine versions there is a significant increase in the causality when the vocabulary size is decreased, which is expected as the relative weight values automatically increase when their count decreases.

The causality analysis study reported here focuses only on the dual encoding model proposed by *Dong et al.* [1] in 2021. However, there are other interpretable models, *Wu et al.* [2], that use similar methods. So, it’s likely that these alternative models will show similar behavior.

Our approach, focused on causality, is generalized and can be applied to retrieval models based on classification tasks, offering a broader understanding of their behavior and performance.

5.3 Improving causality

We have seen above that the causality in the actual visual explanations is very low, because instead of having the causal weights mostly distributed on only a few tags as it would be

expected if only relevant tags were detected with significant scores, we have quite the opposite with most tags being detected with similar and non-negligible scores, as can be seen in Figure 5.2. While this effect is somewhat reduced when using cosine similarity function in concept space instead of Jaccard similarity, but the causality at 10 and 30 is still relatively low. However, when we reduced the number of tags from 512 to 256, we noticed a significant improvement in causal relationships, without a notable drop in performance and sometimes even a slight improvement. Nonetheless, it's important to highlight that even with this reduction, the causal relationships are still not particularly strong.

The reason behind the spread of causal weight across all the tags instead of having the causal weights mostly distributed on only a few tags is because nearly all tags are consistently detected with average probability score of 0.4. On an average, about 200 concepts are always detected with high probability out of 512, which should not be the case as it is much higher than what we would expect based on average tag frequency in training data. This phenomenon likely occurs because the detection scores depend on two different loss functions: one for the classification task and another for the retrieval task. The latter loss function appears to disrupt the balance of the former, leading to an ***over-detection of tags***. More generally, the detectors are not trained independently, and their scores are highly influenced by both an estimated tag probability and latent component (unrelated to the classification task). This dual influence add noise to the detection scores. Additionally, we also observed that several detectors are quite bad (see section 5.6), possibly due to insufficient and/or inconsistent training data, or to the fact that the retrieval loss function forced the detector to learn something useful for the retrieval task regardless of the detector classification performance, or due to both.

There are several ways through which the causality can be improved but, unsurprisingly, they have some impact on the retrieval accuracy. Generally, users are interested in explainability, but not at the expense of the accuracy or with only a negligible reduction in accuracy. Therefor, we will first propose methods for increasing the causality without sacrificing the accuracy. We will also propose methods for increasing further the causality, possibly up to 100%, with possibly also a significant drop in accuracy. This might be interesting for users insisting for having a fully causal explanation (for this part) and accepting the associated cost in accuracy, and for researchers interested in understanding the limits of the approach.

5.3.1 Improving causality by tag detection score transformation

In order to improve the causality from the first few tags, we propose to modify the detection scores by applying a transformation function to them so that the causal weight becomes more concentrated on the first few tags. There are several ways to do this. First, considering the tag probabilities used in the Jaccard similarity (equation (5.1)), simply applying a power

transformation with an exponent p greater than 1 automatically increases the relative weights of the first terms. Second, the tag probabilities $g(v)$ or $g(s)$ are obtained by applying a sigmoid function to "raw" detection scores $h(v)$ or $h(s)$; we can then apply a bias b (shift) and/or a gain a (scale) to these raw scores before applying the sigmoid function, performing a kind of Platt normalization [173], possibly correcting the influence of the retrieval loss in the classification calibration. Combining transformations, we replace probability score i.e. $g(x)_i = \sigma(h(x)_i)$ by:

$$\left(g_{(a,b,p)}(x)\right)_i = (\sigma(a(h(x)_i - b)))^p \tag{5.7}$$

with σ being the sigmoid (expit) function and x being either a video sample v or a text sample s. The original function in case of original dual encoding model [1] corresponds to $(a,b,p) = (1,0,1)$. In practice, we did not investigate all the possible combination and considered varying either only a and b or only p, as p and a have similar effects. Similarly, in order to improve the causality from the first few tags with the cosine similarity (equation 5.2), we replace $h(x)_i$ (which is non-sigmoid detection scores) by:

$$\left((h_{(a,b,p)}(x)\right)_i = (a(h(x)_i - b))^p \tag{5.8}$$

The main difference with formula (5.7) being that the sigmoid transform is not used with the cosine similarity. Again, the original function corresponds to $(a,b,p) = (1,0,1)$ but it can be noted that, as a scale factor, the a parameter has no effect in the cosine similarity, which is related to an angle between vectors. We will then keep $a = 1$ in this case.

For appropriate values of the a, b and p parameters, the transformations described in equations (5.7) and (5.8) increase the contrast between the values used for the similarity computation and therefore increase the causality over the first few most contributing tags. Indeed, these transformations do impact the retrieval performance of the system as well, sometimes positively and sometimes negatively, depending upon the choice of the a, b and p parameters. These parameters should then be chosen in order to obtain the best compromise between causality and accuracy. This is done by giving preference first to the accuracy −as we generally do not want to sacrifice it to causality− and second to the causality as long as this does not hurt accuracy. The corresponding optimal a, b and p parameters are obtained by direct search on the validation set, one at a time, and iteratively. Optimizations are done on the overall mAP (mean of TTV and VTT mAPs) as it is more stable than the SumR metrics, and it leads to very similar results.

5.3.2 Improving causality further by dropping tags

We will explore three main and simple ways through which the causality can be improved with a cost in accuracy:

1. The first one is applicable when a hybrid system combining a latent (fully opaque) space and a concept (partly explainable) is used. In this case, dropping the latent component and using the "concept only" part of the system will significantly increase the causality with a possible non-negligible cost in accuracy;

2. The second way consists in applying the approach described in section 5.3.1 while increasing the values of the p, a and/or b beyond the optimal values for maintaining the accuracy;

3. The third way consists in modifying the Jaccard or cosine score used for ranking the retrieved results so that it takes into account fewer tags.

5.4 Experiments

Dataset. Similar to Section 3.3.2.1, we performed all of our experiments on the official split of the MSR-VTT dataset [13] for the experimentation and evaluation of causality based dual encoding model.

Implementation details. We used PyTorch code[1] provided by the authors of [1]. In order to assess the performance and quality of retrieval results and explanation of baseline model [1] using causal parameters, we used the five variants of concept space which are described in Section 5.2.2. More specifically, we trained and evaluated the concept space for causality and retrieval performance in the following different settings: (i) *Concept-Hybrid*: where the concept space is trained in hybrid mode (latent and concept both), for testing of causality and retrieval, only concept space part is used, (ii) *Concept-Jaccard*: In this setting the dual encoding model is trained and tested on concept space part only with Jaccard coefficient as a similarity metric along with two different vocabulary sizes and dimensions of concepts space (512-d, 256-d), and (iii) *Concept-cosine*: As we have seen in Table 5.1 (row-8), there is slight improvement in accuracy when using cosine similarity in concept space, so we also reported results for "concept-cosine" setting again with two different vocabulary sizes (512-d, 256-d).

Performance Metrics. Similar to Section 3.3.2.1, we evaluated all of our experiments by using all important retrieval performance evaluation metrics provided for the evaluation of MSR-VTT dataset i.e. the mean Average Precision (mAP) and sum of R@K for TTV and VTT (SumR). To evaluate the *causality*, we proposed the formula (Equation 5.6) for calculating the averaged causal effect of a group of concepts G over the top-n results of all the queries of the dataset (refer to Section 5.2.2).

We now explore the impact of our score modifications on the causality of different variants of the system. Then, we check the impact of the proposed modifications on the accuracy.

[1] https://github.com/danieljf24/hybrid_space

We finally discuss the trade-off between these two criteria.

5.4.1 Improving causality by tag detection score transformation

Impact on causality. For each combination of the a, b and p parameters, it is possible to compute the modified tag "probabilities" or scores and to compute from them similarity values, causalities as displayed in figure 5.2 and performance metrics as displayed in table 5.1.

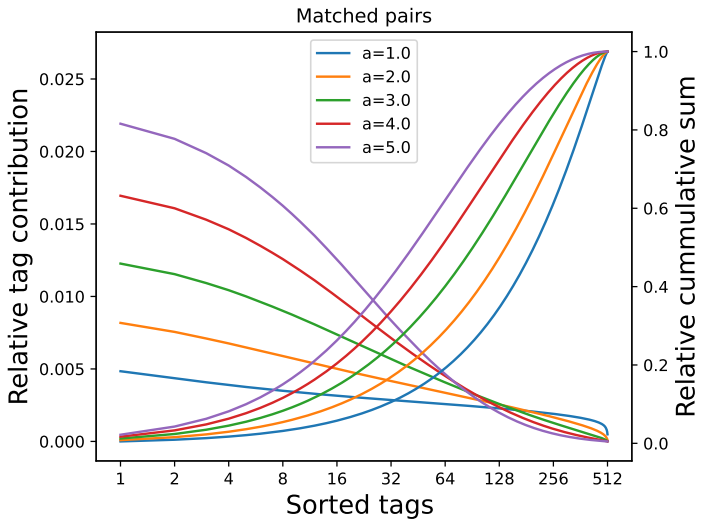


Figure 5.4: Per-tag (decreasing curves) and cumulative (increasing curves) causality for different values of scale a.

Figure 5.4 shows how per-tag and cumulative causality curves evolve according to the values of the scale parameters for the "512-d (hyb. train.)" system. The baseline for $a = 1$ correspond exactly to the "matched pairs" case of figure 5.2, except that (i) the standard deviation is not shown (ii) a log scale is used in order to better illustrate what happens for small numbers of selected tags and (iii) the vertical scale of the relative tag contribution is adjusted so that all curves fit in the window. As expected, the causality always increases with the value of the a (scale) parameter. We also observed that it similarly increases with the values of the p (power) and b (shift) parameters. This remains for all combinations of these parameters that we tried and is also the same for the other systems using a Jaccard similarity ("512-d Jaccard" and "256-d Jaccard"). Regarding the systems using a cosine similarity ("512-d cosine" and "256-d cosine"), the same behavior is observed for the p and b parameters and, as expected, the a parameter has no effect. As we are interested in values as high as possible for the causality at a few tens of tags, for all the systems, we should use values as high as possible for the p, a (if applicable) and b parameters.

Impact on accuracy. Choosing values as high as possible for the p, a and b parameters is likely to have a negative impact on retrieval accuracy. Figure 5.5 shows the evolution

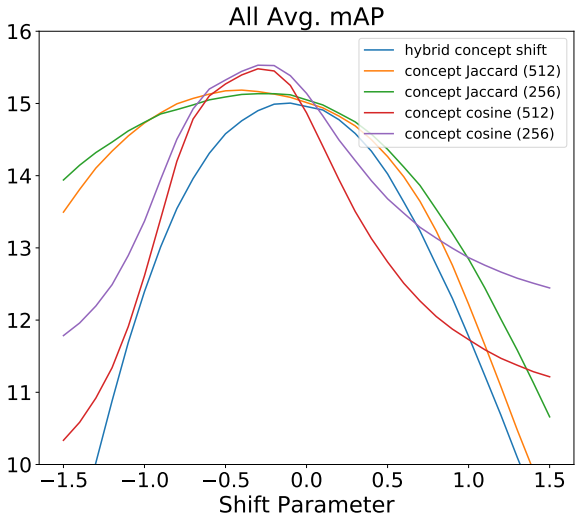


Figure 5.5: Global mAP evolution for the shift (for optimal scale) parameters for the five considered system variants.

of the global mAP according to the parameter b (shift) of equations (5.7) and (5.8). The baseline value is of 0.0 for those parameters We observe that, except (as expected) for the scale parameter with cosine similarity, there is an optimum value for each parameter for the global mAP. The optimum value generally gives a slight performance improvement over the baseline, sometimes significant. Regarding the p and a (when applicable) parameters, the optimum value is significantly higher than the baseline, indicating that it is possible to have a gain simultaneously on the causality and on the accuracy. On the opposite, the optimum value for the accuracy for the b parameter corresponds to a value lower than the baseline so that we loose on one criterion if we optimize on the other.

Joint optimization. As previously mentioned, we favor accuracy over causality as users generally do not want to sacrifice the former to the latter. Here, we even try to further improve the accuracy even if we improve less on the causality. This means that we choose the optimum values obtained from the functions displayed in the curves of Figure 5.5 except where the curve is rather flat and the optimum value is close the baseline one, in which case we keep the latter, which is better for the causality. Also, when relevant, we optimize jointly the b parameter and the p or the a parameter. We don't jointly optimize the p and the a parameters as they have a similar effect and keep the other to the baseline value. The optimization is done on the validation set and causalities and accuracies are measured on the test set. We have also checked that the optimal values are quite close on the validation set and on the test set. Table 5.2 shows the optimum value combinations found on the validation set for the five system variants considered, and Table 5.3 compares the original and improved accuracy values for these cases and the original hybrid version.

Discussion. We found out that there are many ways to improve the actual causality in

Table 5.2: Optimal values for the p (power), a (scale) and b (shift) parameters on the validation set for five system variants.

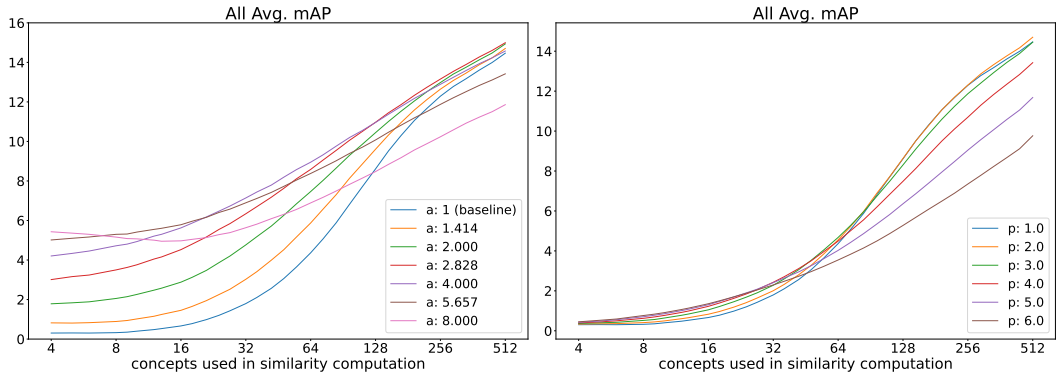
Training	p	a	b
512-d (hyb. tr.)	1.00	2.7	0.0
512-d Jaccard	1.00	2.9	0.0
256-d Jaccard	1.00	1.8	0.0
512-d cosine	1.07	n/a	-0.25
256-d cosine	0.98	n/a	-0.24

visual explanations: by using only a concept space for the retrieval, either with a hybrid training or with a concept-only training, by using a cosine similarity instead of a Jaccard one, by using a smaller tag vocabulary size, and finally by using a transformation on the tag probabilities or scores with optimized parameters. All of them may lead to a significant improvement in the causality on the first few tags or tens of tags without sacrificing on the retrieval accuracy or with even a slight increase in accuracy too with much *less training parameters*, except in the first considered step which is to drop the use of the purely latent space in the retrieval step (2^{nd} row of Table 5.3).

Table 5.3: Causality and performance with and without our improvements for five training conditions. C@10 and C@30 are the causality respectively for the top-10 and top-30 contributing tags. mAP is the mean of the TTV and VTT mAPs. SumR is as defined in [1]. All values are in percentages.

Model Variants	Tr. Params (in Millions)	inference	C@10	C@30	mAP	SumR
1536d+512d hyb.	69.17	original	1.6	4.0	15.8	210.2
512-d (hyb. tr.)	69.17	original	3.9	10.0	13.7	191.1
		improved	10.9	25.5	15.0	203.0
512-d Jaccard	42.23	original	4.0	10.0	14.5	194.8
		improved	16.0	29.7	15.0	198.7
256-d Jaccard	37.74	original	8.2	19.6	14.7	193.2
		improved	32.0	51.8	15.3	200.8
512-d cosine	42.23	original	10.4	23.5	14.9	199.5
		improved	15.6	31.3	15.5	207.0
256-d cosine	37.74	original	17.4	37.8	15.1	201.0
		improved	22.3	44.1	15.5	206.7

Regarding the transformations, we found that a scale-only transformation was the best for systems using the Jaccard similarity and that a transformation based on both shift and power was best for systems using the cosine similarity. The use of cosine similarity may lead to better accuracy for the improved version but with a slightly lower improvement in causality (Table 5.3 Row 5(b)). The accuracy of the improved cosine versions using a concept space



(a) Scale Parameter **(b)** Power Parameter

Figure 5.6: Impact of the *causal* parameter on accuracy while evaluating varying numbers of 'K' concepts using the Jaccard similarity function

only is closed to that of the original full hybrid version.

Regarding the size of the tag vocabulary, the accuracy (mAP and SumR) is comparable for 512-tag and 256-tag versions in all cases, while the causality is greatly improved for the latter. We tried to reduce further the tag vocabulary size in order to find the optimal values as discussed in Section 4.4 (R1) of Chapter 4, but the accuracy begins to drop significantly for sizes going below about 200 tags (see Figure 4.2).

One might question whether the modified tag probabilities or scores still represent well the detection scores from the tag classifiers. Both the Jaccard- and cosine-specific transformations are actually doing a *re-calibration* of these. In fact, the original "tag probabilities" are unlikely to be well calibrated because they correspond to an average detection of 40% of the tags (i.e. $\sim$ 200 concepts), which is much larger than the actual average tag annotation in the training data (i.e. ***over-detection of tags***), and as already mentioned that the issue of over-detection is because the calibration is biased due to the fact that the tag probabilities are subject to two different and competing loss functions (for classification and for retrieval). By reducing the average detection rate of the tags, it is likely that the proposed transformations actually leads to a *better* calibration of the detection scores and to more meaningful "tag probabilities".

5.4.2 Improving causality further by dropping tags

To enhance causality further, we undertook three main and simple ways as described in Section 5.3.2:

- Hybrid system combining a latent (fully opaque) space and a concept (partly explainable) for training is used. Then later on, while evaluating latent component is dropped and using the "concept only" part of the system.

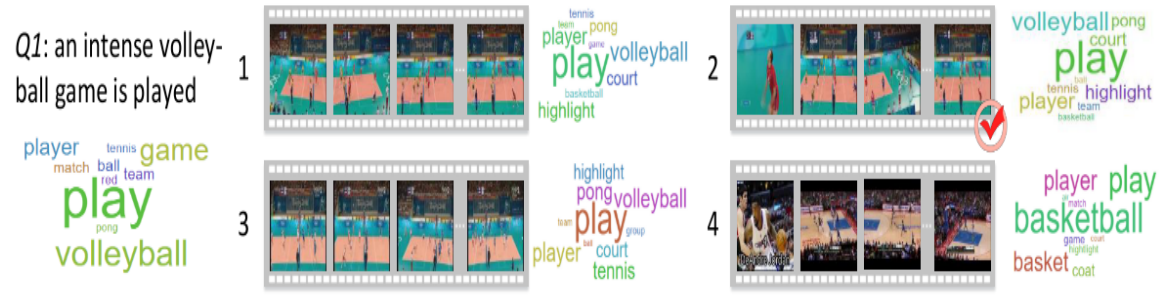
- Experimenting with higher values for the causal parameters, specifically the scale and power parameter, beyond the optimal range to maximize causality.

- Focusing on a selected subset of the most important concepts, denoted as the top-k concepts, when calculating the similarity between a video sample (v) and a text sample (s) for ranking.

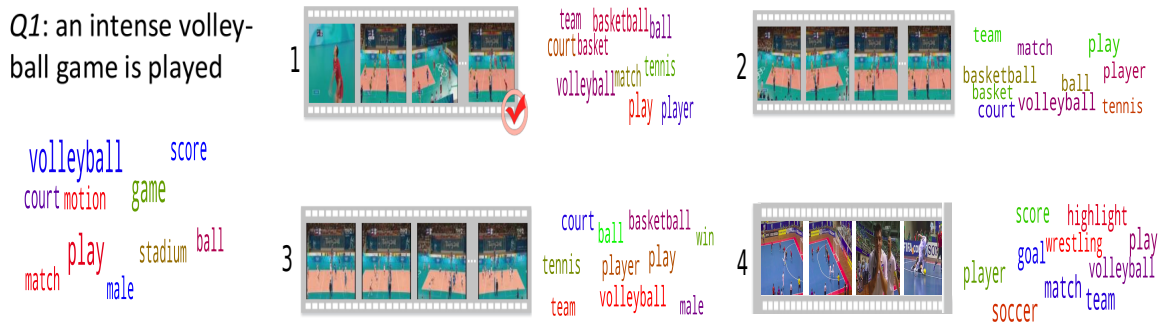
In Figure 5.6, we present two subfigures: Figure 5.6a illustrates our exploration of various values for the scale parameters and top-k concepts, where $k \in K$ signifies the number of dimensions within the concept space. In this figure, as we progressively increased the values of the scale parameter while considering top-k concepts for similarity measurement (e.g., top-10 concepts), we observed that the accuracy initially improved. However, beyond $k \geq 8$, the accuracy started to decline. Notably, maintaining values for the causal parameter a within the optimal range, which transforms the probability scores before considering the top-k concepts, proved to enhance accuracy even when using a few relevant concepts (k). Figure 5.6a demonstrates that without applying any transformation or causal parameter for normalizing probability scores (baseline $a = 1$), the accuracy for top-10 concepts was at its lowest, at 0.3mAP. Conversely, with scale values of $a = 2.8$ and $a = 4$, we achieved substantial accuracy improvement for top-10 concepts, elevating it from 0.3mAP to 5.8mAP.

However, the pattern observed in the case of the power parameter (p) (Figure 5.6b) is different from scale parameter a, as increasing the power $p \geq 2$ seems to not have significant difference in accuracy as compared to baseline system $p = 1$. The accuracy improved a little bit when considering power transformation values greater than 1 for top-k concepts ranges approximately from (1-60) but started decreasing significantly when going beyond $k \geq 60$.

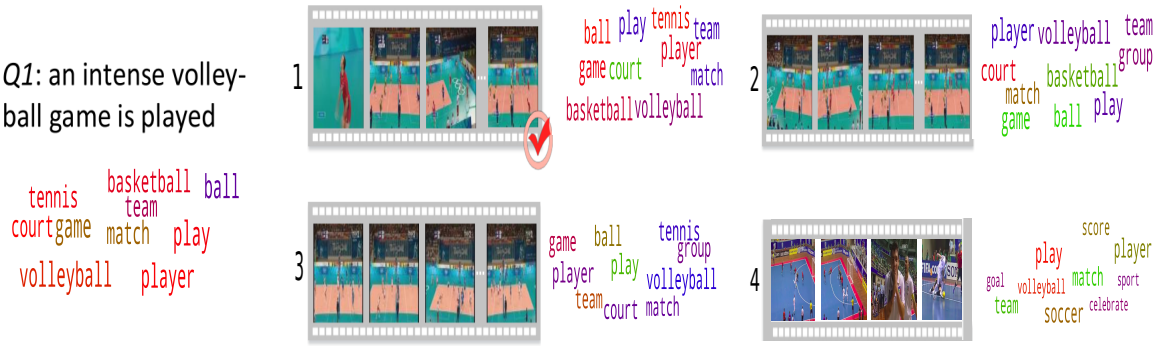
From the experiments and observations several conclusions can be drawn: i) Adjusting the scale parameter (a) can significantly impact accuracy when considering a select subset of the most relevant concepts (top-k concepts). Increasing the scale parameter value initially improved accuracy, but beyond a certain point (around $k \geq 8$), accuracy began to decline. Maintaining values for the causal parameter a within an optimal range proved to be effective in enhancing accuracy, even when using a limited number of relevant concepts (k). ii) Power Parameter: Unlike the scale parameter, increasing the power parameter (p) beyond 1 did not have a significant impact on accuracy, except for a slight improvement in the range of top-k concepts from 1 to 60. However, beyond $k \geq 60$, accuracy started to decrease significantly. This suggests that the power parameter may not be as influential in enhancing causality as the scale parameter.



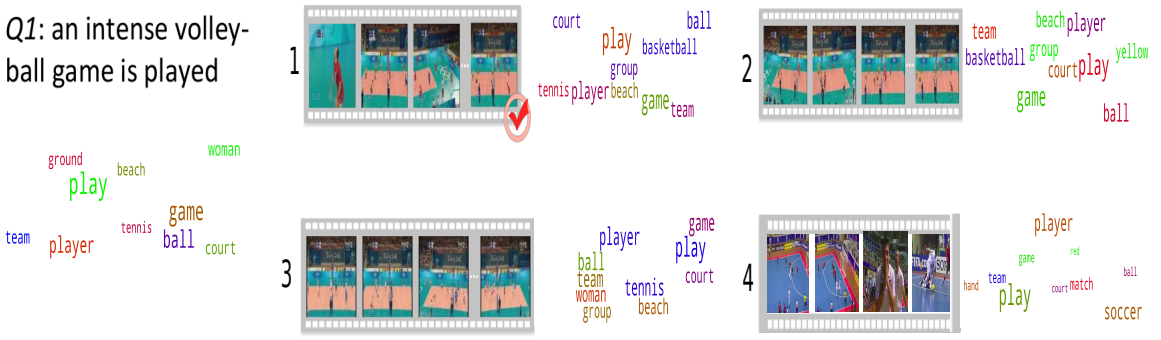
(a) 512-d Baseline [1] – Hybrid space Evaluation



(b) 512-d Baseline [1] – Concept space only Evaluation (re-run)



(c) 512-d Jaccard Improved – Concept space only Evaluation



(d) 256-d Jaccard Improved – Concept space only Evaluation

Figure 5.7: Tag Clouds with Text Sizes Proportional to Prediction Score (C@10)

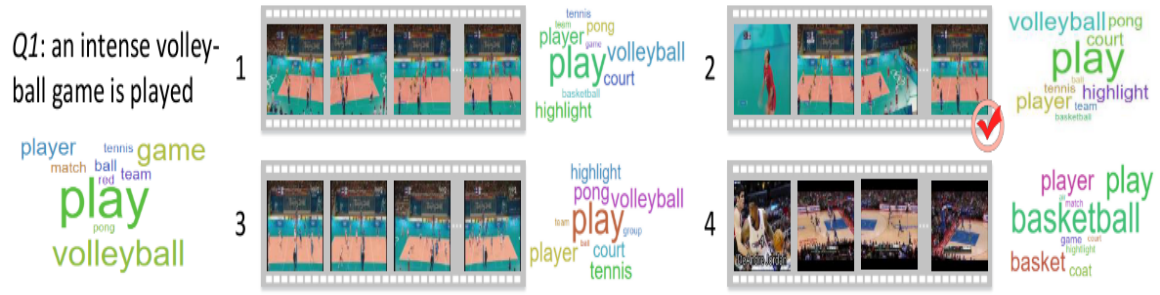
5.5 Improved Tag-Cloud-Based Result Interpretation

All the experiments above consider that displays if tags will support the explanation during retrieval. This section delves into the enhancements made using proposed causal parameters to the tag-cloud-based explanations provided by the dual encoding model [1]. These improvements are illustrated through a series of tag clouds, showcasing a visual comparative analysis of the causality in explanation between the original and improved systems with respect to causality, across various training conditions. As presented in Table 5.3, the causality based dual encoding model demonstrates a significant increase in causality (C@10, c@30) for the improved versions, along with the increase in mean Average Precision (mAP). In this section, we visually validate this enhanced causality (C@10 and C@30) using tag clouds as depicted in Figures 5.7 for the top 10 tags and 5.8 for the top 30 tags. We analyze three different trained systems: i) 512-d Hybrid Training: Representing the explanatory quality of the baseline system, ii) 512-d Jaccard (Improved), iii) 256-d Jaccard (Improved). Since these explanations are exclusively based on the concept space, we do not consider the "1536d+512d hyb" hybrid evaluation with 1536-d latent space and 512-d concept space for visualization.

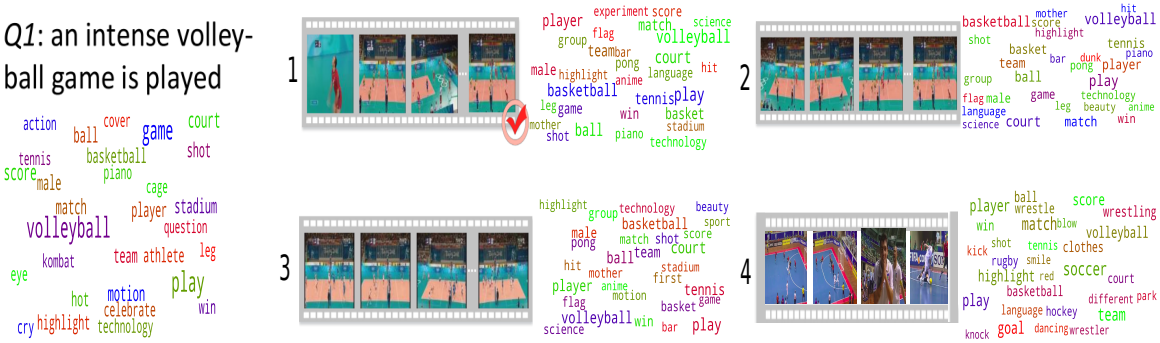
In Section 5.1, we already discussed the usage of tag clouds to explain video retrieval results, as presented by *Dong et al.*[1]. The tags in tag clouds (Figure 5.1) are meant to highlight the most relevant tags with high prediction scores (i.e. large font size for more relevant and important tags and small font size for less relevant tags) for a query and top four retrieved videos. However, it's essential to notice that the sizes of the tags in these tag clouds provided by *Dong et al.*, do not accurately depict their actual relative importance in the matching process. As shown in Figure 5.7a and 5.8a, the tag sizes are very misleading and do not truly reflect the prediction scores of tags. For instance there is very large difference in the sizes of tags "play" and "team", but in fact the prediction scores of these tags are not very different with prediction score of 0.992, and 0.919 respectively. Moreover, there are some issues of over or bad detection in original dual encoding model as well. As shown in Figure 5.7b the detected tag "pong" is not relevant to the textual query or video, but it is highlighted with high prediction scores in 3^{rd} ranked video.

In Figure 5.7b, we represent the tags in tag cloud using baseline dual encoding model [1], by actually displaying tag sizes proportionally to their prediction scores. We observe that almost all the tags are of the same sizes with negligible difference, as they are equally relevant and important for retrieval decision. The pattern is also evident in Figure 5.8b, where the top 30 concepts are visualized in tag clouds for explanation, which should not be the case. This illustrates the low causality in the dual encoding model from a visual perspective, as causal weights are distributed across a large number of tags and all tags are treated as equally important in the retrieval decision process.

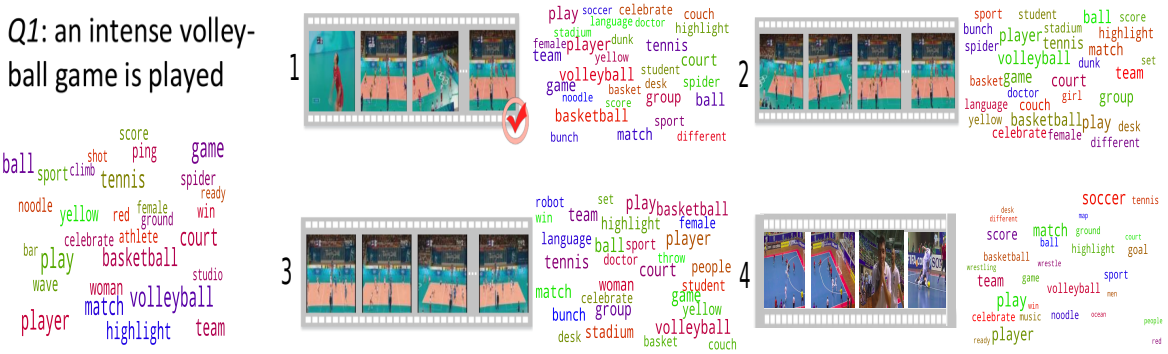
However, in Figure 5.7c, we present tag clouds based on optimal values of causal param-



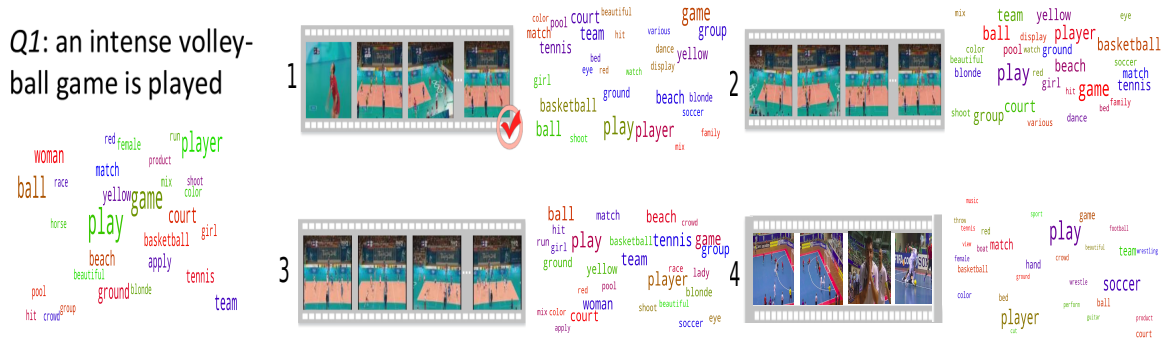
(a) 512-d Baseline [1] – Hybrid space Evaluation



(b) 512-d Baseline [1] – Concept space only Evaluation (re-run)



(c) 512-d Jaccard Improved – Concept space only Evaluation



(d) 256-d Jaccard Improved – Concept space only Evaluation

Figure 5.8: Tag Clouds with Text Sizes Proportional to Prediction Score (C@30)

eters ($p = 1.00$, $a = 2.9$, $b = 0.0$) for a 512-dimensional Jaccard Improved model (selected from Table 5.2). Here, we can see that not all tags have similar sizes, indicating that each tag contributes differently, and not all concepts have nearly identical importance. The most effective causality is observed when using the 256-dimensional Jaccard Improved system, as indicated in Table 5.3. This is achieved when employing causal parameters ($p = 1.00$, $a = 1.8$, $b = 0.0$), as visually depicted in Figure 5.7d. In this figure, we observe that causal weights are primarily distributed among the most relevant tags, indicating higher causality in the explanation compared to Figure 5.7b. This pattern is consistent for both the 512-dimensional and 256-dimensional Jaccard Improved versions, as seen in Figure 5.8c and Figure 5.8d.

Moreover, by accurate recalibration of the tags using transformation function given in Equation 5.7 and 5.8, the issue of over detection of tags are tackled, and detected tags for each query and videos are relevant with respect to the text and video content. For instance, in Figure 5.7d, for the 4^{th} ranked video the tag "soccer" is highlighted with high prediction score. Even though the tag "soccer" is not relevant to the query, but it is relevant to the video content. In the case of 256-d concept space, we do not even have the tag "volleyball" in the vocabulary, which leads to the retrieval of soccer game video because of high similarity with respect to other concepts in query i.e. play, player, ball, game, court etc.

Based on these findings, it is evident that generating causality based tag-cloud explanations more accurately mirrors the decision-making process of retrieval systems. Furthermore, the comparative analysis between the original dual encoding model [1] and the causality-based dual encoding model highlights a significant enhancement in causality without compromising the accuracy of the retrieval systems. The optimal level of causality is achieved with the 256-dimensional Jaccard Improved model. The choice of the best model should be made based on the preferences of the researcher or developer, considering the trade-off between causality and accuracy.

5.6 Additional Challenges and Issues

In this section, several major problems with original dual encoding model [1] are discussed. To do that, we use screen dumps (as shown in Figure 5.9) of an interface specifically developed for visual analysis and explanation. In this interface, for each textual query, the ground truth (G.T) video rank is shown. Below this, each row represents the information of retrieved video, its ID, concept space similarity score. The first column of the table represents the keyframe of the video retrieved, concept representation of video and textual query (2nd and 3rd column), along with the contribution (4th column) and prediction (5th column) of each similar concept between video and query.

In addition to these features, the font size of each tag in every column, except for the

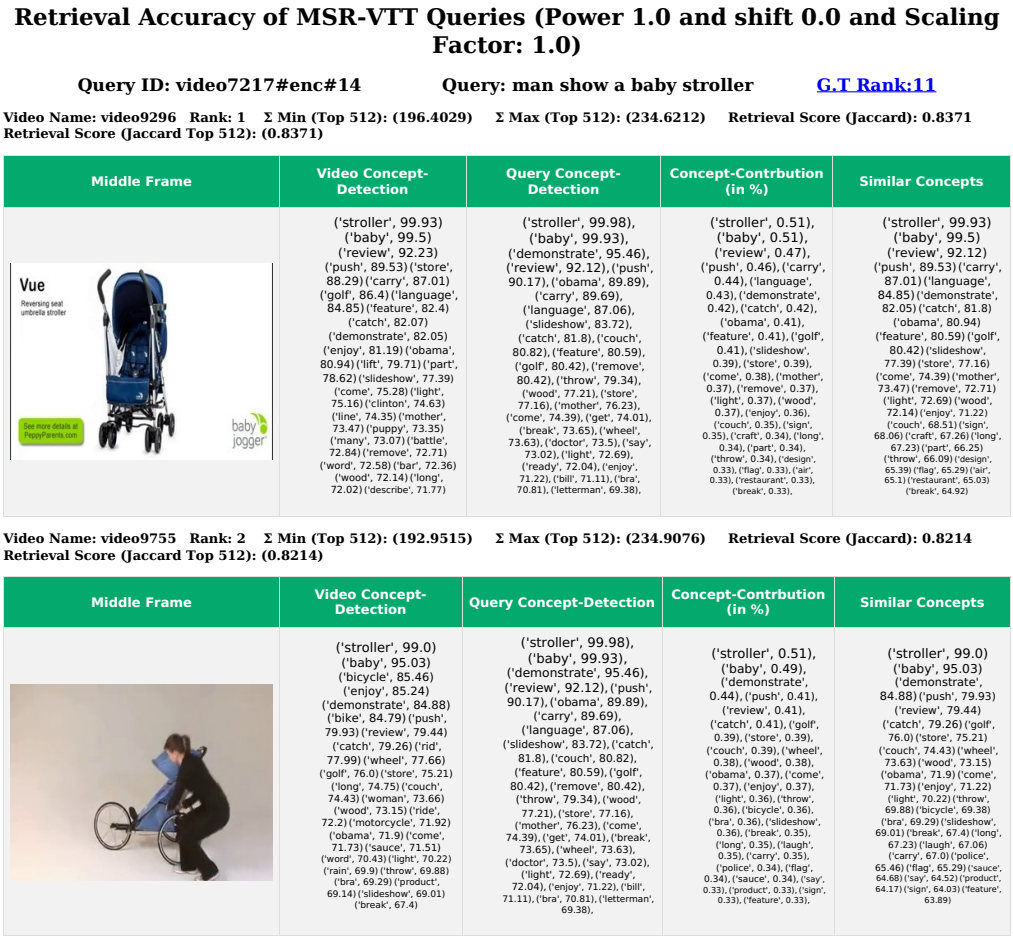


Figure 5.9: Visual Interface for Causality based Video-text retrieval

"concept-contribution" column, is directly proportional to the relative score of each tag provided within them, following the same approach proposed by *Dong et al.* as in [1]. For the "concept-contribution" column, however, the font size corresponds to the relative contribution score without percentage.

5.6.1 Problems with concept selection and annotation

The process of selection of concepts and the annotation of video and text are important steps in different domains, from image and video analysis to natural language processing. The process of concept selection and annotation involves assigning the labels to videos and texts to perform the retrieval and analysis. However, the inherent intricacies in these tasks can significantly impact the quality and completeness of concept representations of video and text. In the experiments conducted, i) one assumption is that the concepts used in the concept space representation are *consistent* with the textual annotations. Here, such consistency refers to the mapping between the text and the concepts selected. For instance, video annotated with text containing the word "stroller" is expected to have a concept representation that

assigns a large prediction score to the specific concept "stroller", leading to a comprehensive representation. ii) Second assumption is the selection of one-word concepts classifier for annotation and representation. For instance, consider the phrase "baby stroller", while based on the assumption, the phrase is divided into two visual concepts "baby" and "stroller" treated independently. Based on these assumptions, there are two main issues that can significantly impact the quality of these representations: specificity and exhaustivity.

- **Specificity:** Consider a query "a man shows the baby stroller", as shown in Figure 5.9, the system activated the visual concept "baby" as the second most detected concept for video-concept detection column even though there is no baby in the video. Activating the "baby" detector is inaccurate and impacts the retrieval accuracy and causal explanation quality. This approach fails to capture the specific meaning or intent of the combined term "baby stroller".

  Similar cases like "video game" or "beach balls" refers to digital entertainment and inflatable balls on the beach respectively, treating video and game alone would give high importance to other indoor and outdoor games instead of specific digital games or balls in general. All these cases highlights the issue of specificity in concept selection and annotation.

- **Exhaustivity:** Let's consider the example of "stroller" concept again. Experimentally it is found that the concept "stroller" is highly correlated with concept "baby" with the correlation of 55.4%, as both of these concepts are often seen together in the videos, images, and textual sentences. In such cases the concept "baby" will also play a large role even if the video does not contain the "baby". This issue is more related to the biases inherent to the collection itself and may be tackled by compensating for correlations between concepts.

So, the "specificity" is more due to limitations in extraction of precise concepts, where the "exhaustivity" is more related to the biases inherent to the collection itself that may be tackled by compensating correlations between concepts. The assumptions of consistency and one-word concept classifiers underscore the critical role of specificity and exhaustivity in these representations, with potential implications for retrieval accuracy and causal explanation quality. These challenges underscore the necessity for further research to refine methods for concept selection and annotation, ultimately advancing the accuracy and interpretability of multimedia retrieval systems.

5.6.2 Detectors' Limitations

The tag-based visual explanation relies on textual and visual detectors. The visual presentation of tags associated to a text queries and to associated retrieved video samples allow a user

to appreciate separately i) how well the detected tags match the text and the video samples and ii) how well the detected tags support the retrieval decision. As this chapter evaluates the causality in the explanation only for the second part, it does not take into account the causality in the tag detection process. Assessing the causality of this process is beyond the scope of this chapter, but in this section, we are going to point out some problems with it. The two main problems that we found are that (i) some detectors learn poorly what they are supposed to learn and (ii) what they are supposed to learn does not always correspond to what humans would intuitively understand by the tag label. Though this does not necessarily impact the retrieval performance (that might even be the opposite), it does undermine the principle of tag-based visual explanations. Whether or not the implicit annotations used for training the

Table 5.4: Performance (Mean Average Precision) of visual and textual tag detectors respectively on the train, test and val splits of MSR-VTT.

MAP	train	test	val
text	0.571	0.552	0.553
video	0.361	0.043	0.050

tag detectors correspond to an intuitive understanding of them by humans, it is possible to evaluate the tag detectors relatively to these implicit annotations. Table 5.4 shows the mean Average Precision (mAP) of visual and textual tag detectors respectively on the train, test and val splits of MSR-VTT. The performance on the text part is far from perfect but not too bad and quite consistent between the training set and the validation and test sets, indicating a quite good learning and a good generalization capability. However, the performance on the visual part is much lower, with a significantly lower mAP even on the training set and a quite catastrophic generalization capability. The most frequent tags can still be well learned and predicted, leading to good visual tag presentations in general.

Figure 5.10 shows three examples for which, respectively (top: *man*) the detector learns well and generalize well, (middle: *guitar*) the detector learns something but generalize very poorly, probably due to over-fitting on very small numbers of positive samples, and (bottom: *engine*) the detector does not learn anything, even on the training set, probably due to the fact that the classification signal is too weak compared to the retrieval one and that the corresponding dimension is used a purely latent one. We did not make a detailed analysis for all concepts, but our observations showed that (i) not all cases clearly correspond to one of these three categories but (ii) we were able to easily identify at least half a dozen tags in each of these categories (by looking for most separated or most overlapping histograms for the positive and negative classes in the training set and/or in the validation set).

Some detectors seem to learn and to be consistent between the training and the validation and test sets, but the concept they are supposed to detect is unclear (e.g. the concept “air”).

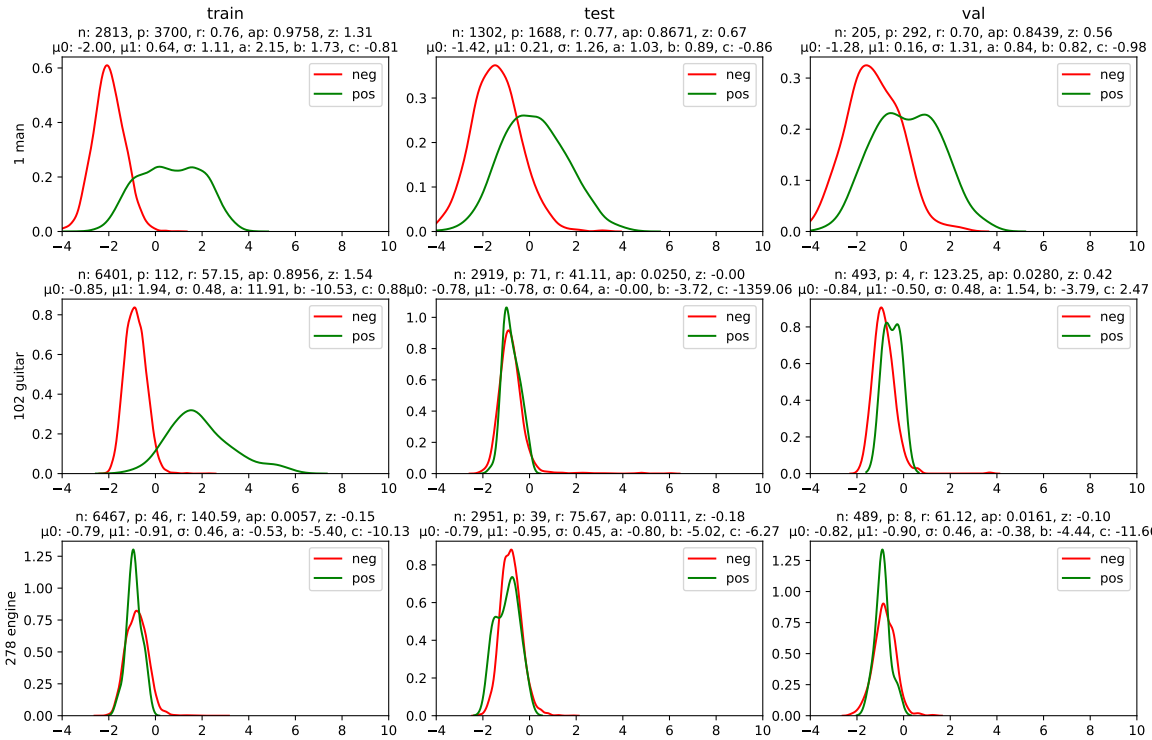
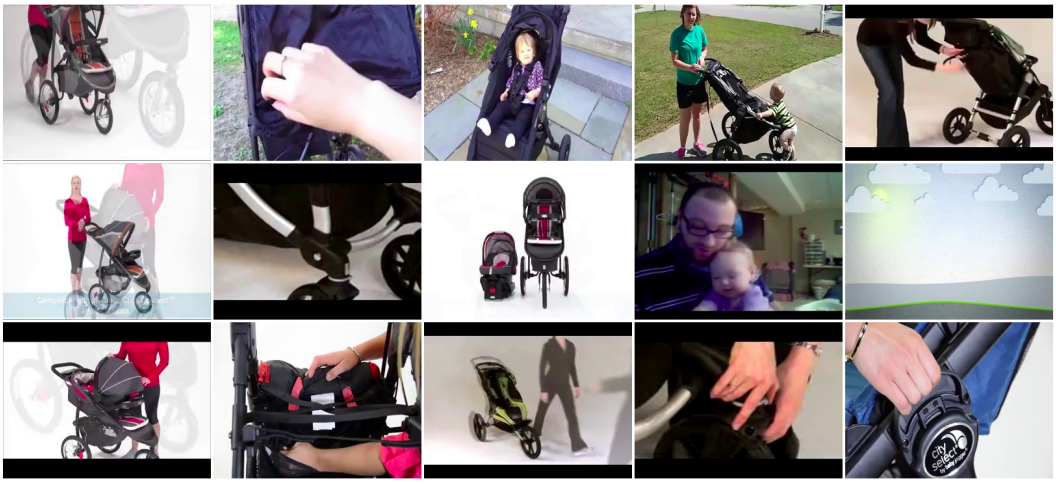


Figure 5.10: Learning behavior of some tag detectors.

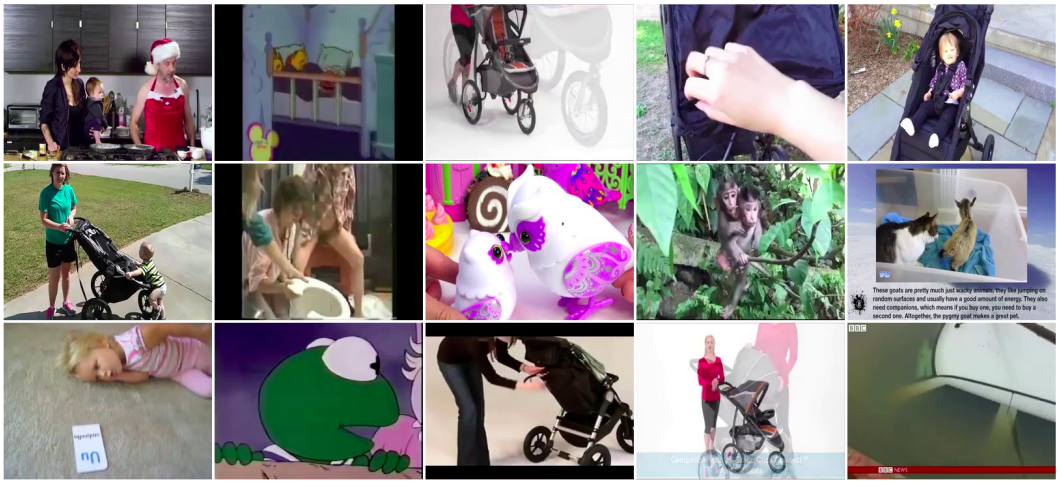
Finally, some detectors detect well the concept they are supposed to identify, but they additionally reliably detect obviously unrelated concepts, probably due to a few coincidences appearing in the training set.

For instance, if a *stroller* visual detector detects also *baby*, it is because in the training set of ***stroller*** class there are some co-occurrences of baby and stroller (as shown in Figure 5.11a 2^{nd} row 4^{th} video frame), then for a query on *stroller* we might have a high causality for the *baby* tag, even for videos that show a stroller and not a baby. Same issue can be observed with concept ***baby***. As in Figure 5.11b, it can be seen in 1^{st} row 3^{rd} video frame and last row 3^{rd} and 4^{th} video frames, the stroller images with no baby at all (checked all the video frames), affect the detector accuracy of prediction only baby with high probability (as shown in Figure 5.11b). These problems happen mostly with the visual detectors. Even if they are not directly related to the tag-to-similarity causality, they may clearly impact this causality, and therefore the global validity or soundness of the explanations/justifications.

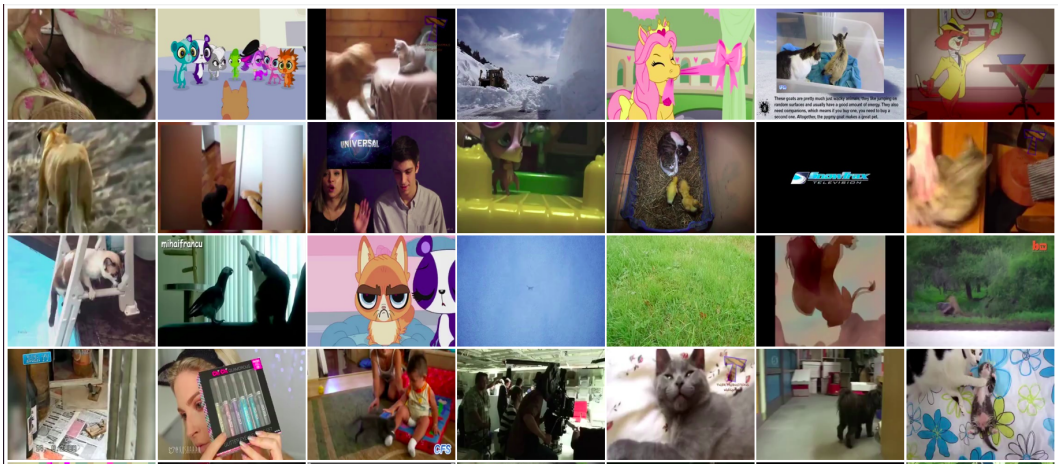
These issues are not limited but can be extended for the query related to "cat". For instance, if a *cat* visual detector detects also men or women because in the training set there are some training images including video frames of men or women (Figure 5.11c 10^{th} and 23^{rd} video). This happens because of ambiguity in annotation of video with cat-related terms like 'cat makeup, cat outfit, cat hat, and cat snowmobile etc. For example, in case of the query *"a woman is applying make up onto her face for a cat outfit"*, we might have a



(a) *Stroller* classifier



(b) *Baby* classifier



(c) *Cat* classifier

Figure 5.11: Training Instances of Classifiers. The grid in a) and b) represents the training images for the classifiers *stroller* and *baby*. The images are the first image of each video and top 15 videos are shown for a) and b), where as first image of each video and top 28 videos are shown for c).

high causality for the *cat* tag even for videos that does not even show a cat (Figure 5.11c last row 2^{nd} video]). All these problems happen mostly with the visual detectors. Even if they are not directly related to the tag-to-similarity causality, they may clearly impact this causality, and therefore the global validity or soundness of the explanations/justifications.

5.6.3 Problems with the task's ground truth

We identified many cases in which several videos correspond to the same given caption and vice versa, with more captions than the 20 associated ones corresponding to a given video. For instance, for the text query "an intense volley-ball game is played" in Figure 5.1, the "ground truth" video is ranked 2 but before it in the ranked list, the top video do match the query and in many other cases. This indeed comes from how the collection was built and annotated. This issue was also partly observed and discussed in [1]. The negative effect of this is that many retrieved items are wrongly counted as false positive, while they should instead be counted as true positive. This leads to a (possibly strong) underestimation of recall and mean average precision (mAP) evaluation measures. This is a problem both for the test (underestimation) and the training (noisy annotations). That might not affect the ranking of the different approaches, as all are likely to be impacted in comparable ways, but it is more problematic when analyzing the errors and their explanations. Correcting or evaluating this would require some adjudication and a modification of the evaluation procedures.

Building automatically ground truth from test collections also necessitates to avoid, when possible, biases that come from the data. In the MSR-VTT dataset, the ground truth is composed of couples (v,S), representing for each video v its 20 textual annotations $S = \{s_1,...,s_{20}\}$. The input used during training and the tests come strictly from this raw ground-truth: only the (v,s) from such ground-truth are positive and all the other pairs are negative. This means that, even if two videos v_1 and v_2 have one of their descriptions s that is the same (according to string equality), the ground-truth relating v_1 and s on one side and v_2 to s on the other side do not consider such overlap. We computed such overlap on MSR-VTT collection based on strict string quality, is equal to 8.1% in training set, 6.6% in test set and 3.8% in validation set. Several overlaps may be considered in a way to handle more or less strict overlaps between annotations, for instance using relaxed equality of strings or Part-of-Speech. If we compute such overlap considering Part-of-Speech (PoS) tags in the captions, by finding a semantic match between verb-verb and object-object in textual annotations, the percentage would be higher. It is worth noting that this issue is also addressed and tackled using automatic process by *Wray et al.* [174].

Such feature has a negative impact on both the train and the test datasets. A way to circumvent such problem should be to create datasets with overlap of texts without any uncertainty. Going back to our example above with v_1 and v_2, these two videos may be considered

both relevant to the text annotation *s*. Then, less negative samples will be considered during training, and for the test data the score will be higher.

5.7 Discussion

In this chapter, our research of video-text retrieval is centered around on fundamental concept of *Causality*. Causality plays a very vital role in explainability of video-text retrieval by determining the extent to which the specific visual concepts or tags in video and query has contributed in retrieval decision process. The contribution in the chapter lies in the proposal of a novel evaluation measure to quantify this causality in state of the system, and enhance it by using the proposed transformation function in order to generate causal explanations which should resemble the major portion of the retrieval decision process. This chapter described an extension of state-of-the-art system, specifically by generalizing the prediction scores for Jaccard and cosine similarity, and increasing the weight of top relevant tags for generating the more causal explanations. The results indicate a substantial increase in causality without a significant loss of accuracy, with potential applicability to other interpretable multimedia retrieval systems.

5.7.1 Causality vs. Accuracy: A Uncertain Equilibrium

Our research has revealed the trade-off between causality and retrieval accuracy. In the experiments, we have carefully examined the impact of causal function parameters i.e. power (p), scale (a), and shift (b), in our pursuit for increasing causality. Achieving the right balance between accuracy and causality requires critical analysis. Joint optimization of parameters offer a good understanding and allowed us to improve the causality in visual explanation of video-text retrieval models without sacrificing the retrieval accuracy.

5.7.1.1 Optimizing Causality and Accuracy

Joint Optimization: Moving towards the goal of increasing the causality in visual explanations (Figure 5.1), we first discussed the joint optimization of the parameters i.e. p, a, and b when transforming tag probabilities or scores of visual concepts or tags. This approach aims to strike an equilibrium between the improvement in accuracy of retrieved results and causality of visual explanation. By optimizing the causal parameters on a validation set and measuring causality and accuracy on a test set, researchers can fine-tune the system to achieve the desired level of causality without sacrificing retrieval accuracy.

Impact on Accuracy: While experimenting with the causal parameters, we discuss the importance of choosing the values for causal parameters i.e. power (p), scale (a), and shift (b) within the optimal range, when transforming the probability scores of tags. This transformation function (Equations (5.7) and (5.8)) aim at improving the causality of explanations

and perform better calibration of probability score, ensuring that the presence of specific relevant concepts aligns more closely with the textual query and contribute more in overall retrieval decision. However, we also acknowledge the delicate balance between optimizing for causality and accuracy. While higher values of p and a may enhance causality, they could potentially have a negative impact on retrieval accuracy. The Table 5.3 highlights the trade-off between causality and accuracy in multimedia retrieval systems.

Further Enhancing Causality: We then tried a few methods for further enhancing causality, even at the cost of a slight reduction in accuracy. This includes experimenting with higher values of the causal parameters, such as scale and power, and focusing on a select subset of the most relevant concepts (top-k concepts) during similarity measurement while also including and excluding the latent space (non-explainable) while evaluation of model. These approaches emphasize the pursuit of stronger causality, ensuring that retrieved videos are more directly related to the queried concepts. The experiments in Figure 5.3 revealed that only considering concept space while evaluation, along with careful scale parameter tuning, can enhance causality without sacrificing performance. The improved models achieve higher causality and comparable accuracy while employing fewer training parameters. For instance, the 256-dimensional Jaccard Improved model in Table 5.3, utilizes only 37.74 million parameters, 53% fewer than the 1536d+512d hybrid model with 69.17 million parameters. This reduction in complexity is a cost-effective approach without sacrificing performance. This reduction in parameter can also lead to saving energy and it is very important when training and using retrieval model on large scale data.

We also observed that causality while considering only top-k concepts for similarity between video and text is improved by using causal parameters (Figure 5.6). However, adjusting the scale parameter improved accuracy (Figure 5.6a), while the power parameter had a limited impact, emphasizing the importance of parameter selection in video-text retrieval systems.

Visualization of Causal based Tag-cloud Explanation: The visual comparison of causality in original dual encoding model explanation and causality based dual encoding model explanation is visualized through set of tag-clouds. The Visual comparisons of causality are presented for both the original and improved systems under various training conditions. The results, as shown in Table 5.3 and visually represented in Figure 5.7 and Figure 5.8, indicate a notable increase of causality (C@10 and C@30) in explanations for the improved versions. This enhancement is particularly evident in the 256-dimensional Jaccard Improved system, suggesting that refining the concept space by reducing the dimensions can lead to more meaningful and explanatory retrieval results.

5.7.2 Addressing Critical Challenges

As our aim is to improve explainability in video-text retrieval systems, which is completely based on concept space part of hybrid systems, we face two significant challenges. The first challenge involves the limitations in training of visual detectors which are essential for concept based video-text representations, retrieval and explanation. The second issue concerns the complexities of concept selection & annotation and ambiguities in ground-truth annotation in datasets, which can make evaluations difficult and uncertain.

Limitations of Visual Detectors: While improving the causality in video retrieval model, we found some serious limitations associated with visual detectors, which are pivotal in identifying visual concepts within videos. These limitations span various categories:

- **Failure to Learn:** Some visual detectors fail to learn, even on the training set, potentially because the classification signal is too weak compared to the retrieval one. The weak classifiers are then treated as a purely latent, which can highly impact the concept representation and subsequently, causality based explanations.
- **Overfitting:** Certain detectors were performing well on training datasets but were failed in generalization because of the limited positive instances, which caused overfitting and impact the reliability and validity of explanations.
- **Unclear Concept Identification:** The issue of learning the vague visual concepts also raised a lot of questions about the learning of the classifier because the concept they are suppose to identify is unclear. For example, detecting "air" may raise questions about the precise representation of this concept.
- **False Detections:** Detectors may detect unrelated concepts due to coincidental occurrences in the training set. These spurious detections can introduce inaccuracies into explanations (shown in Figure 5.11b and 5.11a).

While these limitations may not directly affect tag-to-similarity causality, they can substantially impact the overall validity and reliability of explanations. Future research endeavors should prioritize improving detector learning, enhancing generalization, and addressing issues related to concept identification and vocabulary building.

5.7.3 Issues with the Ground Truth

The following issues discussed below are related to the ground truth annotation and concept annotation.

- Concept Selection and Annotation: The consistency of concept selection and annotation with training caption and relying only on single concept classifiers highlights

the issues in the concept based video-text representations, with potential implications for retrieval accuracy and causal explanation quality. These challenges underscore the necessity for further research to refine methods for concept selection and annotation, ultimately advancing the accuracy and interpretability of multimedia retrieval systems.

- Ambiguity in Ground Truth: Instances where multiple videos correspond to the same textual caption and vice versa introduce ambiguity in the ground truth. Annotating only one video as relevant for one caption highly impacts retrieval accuracy of the system which is also discussed by *Wray et al.* [174]. This ambiguity can lead to inaccurate evaluations of system performance. Considering the overlap between captions of multiple videos may lead to better ground-truth annotation.

To mitigate these issues, future research could explore methods for creating ground truth datasets with reduced ambiguity and controlled overlap. Adjudication processes and modifications to evaluation procedures may be necessary to provide more accurate assessments of system performance.

Chapter 6

Conclusion and Future Work

In this dissertation, we tackle the problem of explainable cross-modal retrieval systems, in particular video-text retrieval. This research has led to a deeper understanding of cross-modal video-text retrieval systems, focusing on the working of dual space (latent and concept) models, complementarity between latent and concept spaces, and the explainability of video-text retrieval systems. More specifically, we studied the importance and complexity of causal inference in providing meaningful explanations. The following synthesis summarizes the key contributions and insights garnered from this research and outlines their implications in Section 6.1 and Section 6.2 respectively, and perspectives for future research in Section 6.3.

To build a more satisfiable explainable cross-modal retrieval system, we summarize our contributions as follows:

6.1 Summaries of Contributions

6.1.1 Extension of Dual Encoding Model with PoS-Tags

This contribution of the dissertation focuses on enhancing the concept space part of the hybrid model by enhancing the richness of the information represented in the concept space. In this contribution, we particularly extended the dual encoding model proposed by *Dong et al.* [1] by incorporating the Part-of-Speech (PoS) tags in the text encoding pipeline of the model and with the visual concept classes. The aim is to enhance the linguistic and syntactic information with the text and classes in order to improve the classification accuracy, which directly improves the retrieval accuracy. By incorporating the PoS-tags into the model, we have improved its ability i) to understand the true intent of the textual query and ii) to categorize more effectively both the text and video into relevant visual concepts. This contribution opens many paths to explore the integration of additional linguistic features in the text encoding process and deepening the connection between two heterogeneous modalities i.e. video and text.

6.1.2 Complementarity Analysis in Dual Space Models

The integration of PoS-tags in the aforementioned contribution revealed that despite the introduction of additional linguistic features within the text encoding pipeline, there was no substantial improvement in accuracy. This observation prompted us to conduct an in-depth analysis of latent and concept space in hybrid models. Thus, in this contribution, we build a general framework for analyzing the dual space models in order to find the inter and intra-relationship between the two spaces i.e. latent space and concept space. The dissertation conducted an extensive analysis of dual space model [1], probing into the interplay and complementarity between latent and concept spaces within the realm of cross-modal video-text retrieval. Three fundamental research questions were formulated and subsequently addressed, revealing insights into the optimal dimensions of these spaces, their potential complementarity, and the impact of ensemble learning. It revealed the optimal dimensions for these spaces, and their complementarity, and demonstrated that both spaces share similar capabilities, providing insights into model design.

6.1.3 Causal Inference in Video-Text Retrieval

Building upon the dual encoding model again proposed by *Dong et al.* [1], this contribution focuses on proving the causal explanation for dual encoding models' retrieval results. In this contribution, we have proposed an evaluation measure for quantifying the causality in ranking for retrieval of human-readable tags used in visual explanations. Then, we extended the dual encoding model [1] in a way to enforce a higher causality, without negatively impacting the performance of the system. Our proposal relies on a modification of the tag scores computation in order to increase the relative effect of the top tags. In such a case, the major part of the matching function (Jaccard or cosine) is supported by a few tens of dimensions in concept space, which is much more suitable for a causality-based explanation. We show quantitatively and visually that our proposal increases our causality measure by up to an order of magnitude without losing significantly on the accuracy while employing fewer training parameters. For instance, the best causal model (256-dimensional Jaccard Improved model in Table 5.3) utilizes only 37.74 million parameters, significantly fewer than the original dual encoding model (1536d+512d hybrid model) with 69.17 million parameters. This reduction in complexity is a cost-effective approach without sacrificing performance. This reduction in parameter can also lead to saving energy, and it is very important when training and using retrieval model on large scale data. It supports the wider aim of making large-scale computing applications more sustainable and efficient in their use of resources.

This study has been conducted for the dual encoding model, but both the observations and the improvements should be generalizable to other interpretable systems for multimedia retrieval that similarly rely on similarity in a conceptual space for instance [2]. This

preliminary work shows that, though it is possible to significantly improve the causality in visual explanations without sacrificing performance, a 100% causality in such visual justifications/explanations is still far away. Other experiments that we conducted show that it is possible to strictly enforce a 100% causality but with a very significant penalty on the accuracy. Any compromise in between is also likely to be achievable but, in general, users will not want to trade away accuracy for causality.

6.2 Insights and Implications

The findings and contributions of this dissertation bear significant implications for the field of cross-modal video-text retrieval and related domains:

- **Leveraging PoS Tags:** The integration of PoS tags into the dual encoding model represents a novel approach to enhancing cross-modal retrieval. It demonstrates the potential of linguistic information in improving the understanding and retrieval of multimedia content, opening new possibilities for more context-aware and accurate retrieval systems.

- **Understanding Complementarity:** The analysis of dual space models has illuminated the relationship between latent and concept spaces. The discovery that these spaces share similar optimal dimensions challenges previous assumptions and opens avenues for more efficient and effective model design. Understanding complementarity is crucial for improving retrieval system performance based on hybrid models and designing robust and interpretable models.

- **Enhancing Explanations:** The exploration of causal inference techniques has the potential to revolutionize the way explanations are generated in video-text retrieval systems. By focusing on the causal contribution of concept detection scores, it becomes possible to provide users with more interpretable and relevant explanations. This enhancement not only aids in improving the user experience but also fosters trust in AI-driven retrieval systems.

- **Balancing Causality and Accuracy:** The joint optimization strategies developed in this dissertation offer a practical approach to balance causality and accuracy. As users often prioritize retrieval performance, having the ability to fine-tune models to enhance explanations without compromising accuracy is a significant advancement. This balance is essential for real-world applications where both aspects are crucial.

6.3 Perspectives for future research

While this dissertation has made significant strides in leveraging PoS tags in video-text retrieval, understanding complementarity, and improving causality, there are several promising avenues for future research.

- **Expanding to Multimodal Data**

  By including the additional modalities such as audio could further enhance the richness of visual information in cross-modal retrieval. Investigating how causality and complementarity manifest in multimodal settings would be an exciting direction.

- **To explore ways to inspect approaches that may enforce stronger complementary of these spaces leading to new hybrid approaches**

  Another direction could concentrate on frameworks that support the study of spaces complementarity, for hybrid spaces in other contexts. Such a framework could help the community to detail the behaviors of any hybrid spaces. The usage of the nonlinear decomposition for the analysis of latent and concept space and correlation between the two can also be considered, considering the complexity of the inputs.

- **To improve the quality and accuracy of classifiers**

  One of the key challenges of tag-based visual explanations is the performance of the tag detectors. As discussed in the Section 5.6.2, some detectors learn poorly and others not always correspond to what humans would intuitively understand by the tag label. Several research directions that can be pursued to improve the quality of detectors are:

  - By utilizing more powerful feature representations and learning strategy. More sophisticated learning algorithms can be used to train the tag detectors. For example, ensemble learning methods can be used to combine the predictions of multiple classifiers, which can improve the overall accuracy of the predictions.
  - By enhancing the quality of training data. The training data used for training of classifiers was noisy or inaccurate, the detectors will learn to make mistakes. One way to improve the quality of the training data is to manually annotate a large corpus of video data. However, this can be a time-consuming and expensive process. An alternative approach is to use data augmentation techniques to generate more training data from a smaller set of annotated videos.
  - By leveraging natural language processing (NLP) techniques to improve the understanding of the concept behind each tag. For example, NLP can be used to identify synonyms and hyponyms for each tag, which can help the tag detectors

to learn more accurate and consistent representations. Additionally, NLP can be used to analyze the annotations and identify the semantic relationships between tags. This information can be used to develop a more comprehensive understanding of the meaning represented by the tags.

- **To incorporate the user feedback into the optimization process**

  This could involve mechanisms for users to indicate their preferences for explanations, allowing models to adapt and provide more personalized explanations. Moreover, user feedback can also help to evaluate the quality of explanation provided by the system.

- **To consider ethical implication**

  As AI-driven explanations become more prevalent, it is essential to consider the ethical implications of causality, complementarity, and the use of linguistic information in retrieval systems. Research on fairness, bias, and transparency in these models is crucial to ensure responsible AI deployment.

Overall, this dissertation has contributed significantly to the evolving landscape of cross-modal video-text retrieval. It has advanced our understanding of complementarity in dual space models, enhanced causal inference techniques, and showcased the potential of leveraging linguistic information through PoS tags. These insights offer a brighter future for AI-driven content recommendations, search engines, and more, emphasizing both accuracy and user understanding. As the field continues to evolve, the research presented here serves as a foundation for future innovations that prioritize both accuracy and user-centric explanations.

References

[1] Jianfeng Dong, Xirong Li, Chaoxi Xu, Xun Yang, Gang Yang, Xun Wang, and Meng Wang. Dual encoding for video retrieval by text. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[2] Jiaxin Wu and Chong-Wah Ngo. Interpretable embedding for ad-hoc video search. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 3357–3366, 2020.

[3] Qiubin Lin, Wenming Cao, and Zhiquan He. Level-wise aligned dual networks for text–video retrieval. *EURASIP Journal on Advances in Signal Processing*, 2022(1):58, 2022.

[4] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*, pages 818–833. Springer, 2014.

[5] Antoine Miech, Ivan Laptev, and Josef Sivic. Learning a text-video embedding from incomplete and heterogeneous data. *arXiv preprint arXiv:1804.02516*, 2018.

[6] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26, 2013.

[7] Yi-Jie Lu, Hao Zhang, Maaike de Boer, and Chong-Wah Ngo. Event detection with zero example: Select the right and suppress the wrong concepts. In *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*, pages 127–134, 2016.

[8] Jianfeng Dong, Xirong Li, Chaoxi Xu, Shouling Ji, Yuan He, Gang Yang, and Xun Wang. Dual encoding for zero-example video retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9346–9355, 2019.

[9] Plamen P Angelov, Eduardo A Soares, Richard Jiang, Nicholas I Arnold, and Peter M Atkinson. Explainable artificial intelligence: an analytical review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 11(5):e1424, 2021.

[10] Lizi Liao, Xiangnan He, Bo Zhao, Chong-Wah Ngo, and Tat-Seng Chua. Interpretable multimodal retrieval for fashion products. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 1571–1579, 2018.

[11] Christian Krupitzer, Tanja Noack, and Christine Borsum. Digital food twins combining data science and food science: System model, applications, and challenges. *Processes*, 10(9):1781, 2022.

[12] Waddah Saeed and Christian Omlin. Explainable ai (xai): A systematic meta-survey of current challenges and future opportunities. *Knowledge-Based Systems*, page 110273, 2023.

[13] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. MSR-VTT: A large video description dataset for bridging video and language. In *CVPR*, pages 5288–5296, 2016.

[14] Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid. Multi-modal transformer for video retrieval. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 214–229. Springer, 2020.

[15] Shizhe Chen, Yida Zhao, Qin Jin, and Qi Wu. Fine-grained video-text retrieval with hierarchical graph reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10638–10647, 2020.

[16] Maksim Dzabraev, Maksim Kalashnikov, Stepan Komkov, and Aleksandr Petiushko. Mdmmt: Multidomain multimodal transformer for video retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3354–3363, 2021.

[17] Jorge E Camargo and Fabio A Gonzalez. Multimodal visualization based on latent topic analysis. In *2014 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, pages 1–6. IEEE, 2014.

[18] Xirong Li, Chaoxi Xu, Gang Yang, Zhineng Chen, and Jianfeng Dong. W2vv++ fully deep learning for ad-hoc video search. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 1786–1794, 2019.

[19] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[20] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[21] Xirong Li, Fangming Zhou, Chaoxi Xu, Jiaqi Ji, and Gang Yang. Sea: Sentence encoder assembly for video retrieval by textual queries. *IEEE Transactions on Multimedia*, 2020.

[22] Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid. Cvpr 2020 video pentathlon challenge: Multi-modal transformer for video retrieval. In *CVPR Video Pentathlon Workshop*, 2020.

[23] Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. Hero: Hierarchical encoder for video+ language omni-representation pre-training. *arXiv preprint arXiv:2005.00200*, 2020.

[24] Antoine Miech, Jean-Baptiste Alayrac, Ivan Laptev, Josef Sivic, and Andrew Zisserman. Thinking fast and slow: Efficient text-to-visual retrieval with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9826–9836, 2021.

[25] Peng Wu, Xiangteng He, Mingqian Tang, Yiliang Lv, and Jing Liu. Hanet: Hierarchical alignment networks for video-text retrieval. In *Proceedings of the 29th ACM international conference on Multimedia*, pages 3518–3527, 2021.

[26] Jiaxin Wu, Chong-Wah Ngo, Wing-Kwong Chan, and Zhijian Hou. (un) likelihood training for interpretable embedding. *arXiv preprint arXiv:2207.00282*, 2022.

[27] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 1, pages 886–893. Ieee, 2005.

[28] David G Lowe. Object recognition from local scale-invariant features. In *Proceedings of the seventh IEEE international conference on computer vision*, volume 2, pages 1150–1157. Ieee, 1999.

[29] Thomas Cover and Peter Hart. Nearest neighbor pattern classification. *IEEE transactions on information theory*, 13(1):21–27, 1967.

[30] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20:273–297, 1995.

[31] Frank Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386, 1958.

[32] George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4):303–314, 1989.

[33] Henry J Kelley. Gradient theory of optimal flight paths. *Ars Journal*, 30(10):947–954, 1960.

[34] Kunihiko Fukushima. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological cybernetics*, 36(4):193–202, 1980.

[35] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[36] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[37] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.

[38] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015.

[39] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[40] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[41] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[42] Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In *International conference on machine learning*, pages 4651–4664. PMLR, 2021.

[43] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60:91–110, 2004.

[44] Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning. *Neurocomputing*, 508:293–304, 2022.

[45] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1728–1738, 2021.

[46] Simon Ging, Mohammadreza Zolfaghari, Hamed Pirsiavash, and Thomas Brox. Coot: Cooperative hierarchical transformer for video-text representation learning. *Advances in neural information processing systems*, 33:22605–22618, 2020.

[47] Yuying Ge, Yixiao Ge, Xihui Liu, Jinpeng Wang, Jianping Wu, Ying Shan, Xiaohu Qie, and Ping Luo. Miles: visual bert pre-training with injected language semantics for video-text retrieval. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXV*, pages 691–708. Springer, 2022.

[48] Xing Cheng, Hezheng Lin, Xiangyu Wu, Fan Yang, and Dong Shen. Improving video-text retrieval by multi-stream corpus alignment and dual softmax loss. *arXiv preprint arXiv:2109.04290*, 2021.

[49] Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L Berg, Mohit Bansal, and Jingjing Liu. Less is more: Clipbert for video-and-language learning via sparse sampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7331–7341, 2021.

[50] Jesús Andrés Portillo-Quintero, José Carlos Ortiz-Bayliss, and Hugo Terashima-Marín. A straightforward framework for video retrieval using clip. In *Pattern Recognition: 13th Mexican Conference, MCPR 2021, Mexico City, Mexico, June 23–26, 2021, Proceedings*, pages 3–12. Springer, 2021.

[51] Youngjae Yu, Jongseok Kim, and Gunhee Kim. A joint sequence fusion model for video question answering and retrieval. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 471–487, 2018.

[52] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[53] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[54] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017.

[55] Pascal Mettes, Dennis C Koelma, and Cees GM Snoek. Shuffled imagenet banks for video event detection and search. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 16(2):1–21, 2020.

[56] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[57] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

[58] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.

[59] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*, 2018.

[60] Jonathan Masci, Ueli Meier, Dan Cireşan, and Jürgen Schmidhuber. Stacked convolutional auto-encoders for hierarchical feature extraction. In *Artificial Neural Networks and Machine Learning–ICANN 2011: 21st International Conference on Artificial Neural Networks, Espoo, Finland, June 14-17, 2011, Proceedings, Part I 21*, pages 52–59. Springer, 2011.

[61] Xiao Sun, Xiang Long, Dongliang He, Shilei Wen, and Zhouhui Lian. Vsrnet: End-to-end video segment retrieval with text query. *Pattern Recognition*, 119:108027, 2021.

[62] Ning Han, Jingjing Chen, Chuhao Shi, Yawen Zeng, Guangyi Xiao, and Hao Chen. Bicnet: Learning efficient spatio-temporal relation for text-video retrieval. *arXiv preprint arXiv:2110.15609*, 2021.

[63] Ameen Ali, Idan Schwartz, Tamir Hazan, and Lior Wolf. Video and text matching with conditioned embeddings. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1565–1574, 2022.

[64] Xun Yang, Jianfeng Dong, Yixin Cao, Xun Wang, Meng Wang, and Tat-Seng Chua. Tree-augmented cross-modal encoding for complex-query video retrieval. In *Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval*, pages 1339–1348, 2020.

[65] Ji Lin, Chuang Gan, and Song Han. Tsm: Temporal shift module for efficient video understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7083–7093, 2019.

[66] Xudong Guo, Xun Guo, and Yan Lu. Ssan: Separable self-attention network for video representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12618–12627, 2021.

[67] Du Tran, Jamie Ray, Zheng Shou, Shih-Fu Chang, and Manohar Paluri. Convnet architecture search for spatiotemporal feature learning. *arXiv preprint arXiv:1708.05038*, 2017.

[68] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211, 2019.

[69] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017.

[70] Lin Sun, Kui Jia, Dit-Yan Yeung, and Bertram E Shi. Human action recognition using factorized spatio-temporal convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4597–4605, 2015.

[71] Zhaofan Qiu, Ting Yao, and Tao Mei. Learning spatio-temporal representation with pseudo-3d residual networks. In *proceedings of the IEEE International Conference on Computer Vision*, pages 5533–5541, 2017.

[72] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018.

[73] Han Fang, Pengfei Xiong, Luhui Xu, and Yu Chen. Clip2video: Mastering video-text retrieval via image clip. *arXiv preprint arXiv:2106.11097*, 2021.

[74] Yiwei Ma, Guohai Xu, Xiaoshuai Sun, Ming Yan, Ji Zhang, and Rongrong Ji. X-clip: End-to-end multi-grained contrastive learning for video-text retrieval. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 638–647, 2022.

[75] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[76] Jiaxin Wu, Phuong Anh Nguyen, and Chong-Wah Ngo. Vireo@ trecvid 2021 ad-hoc video search. 2021.

[77] Shaobo Min, Weijie Kong, Rong-Cheng Tu, Dihong Gong, Chengfei Cai, Wenzhe Zhao, Chenyang Liu, Sixiao Zheng, Hongfa Wang, Zhifeng Li, et al. Hunyuan_tvr for text-video retrivial. *arXiv preprint arXiv:2204.03382*, 2022.

[78] Qiang Wang, Yanhao Zhang, Yun Zheng, Pan Pan, and Xian-Sheng Hua. Disentangled representation learning for text-video retrieval. *arXiv preprint arXiv:2203.07111*, 2022.

[79] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, volume 2, page 4, 2021.

[80] Yang Liu, Samuel Albanie, Arsha Nagrani, and Andrew Zisserman. Use what you have: Video retrieval using representations from collaborative experts. *arXiv preprint arXiv:1907.13487*, 2019.

[81] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 21–37. Springer, 2016.

[82] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *Proceedings of the European conference on computer vision (ECCV)*, pages 305–321, 2018.

[83] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. Cnn architectures for large-scale audio classification. In *2017 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 131–135. IEEE, 2017.

[84] Dan Deng, Haifeng Liu, Xuelong Li, and Deng Cai. Pixellink: Detecting scene text via instance segmentation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[85] Amirhossein Habibian, Thomas Mensink, and Cees GM Snoek. Video2vec embeddings recognize events when examples are scarce. *IEEE transactions on pattern analysis and machine intelligence*, 39(10):2089–2103, 2016.

[86] Miroslav Kratochvíl, Patrik Veselỳ, František Mejzlík, and Jakub Lokoč. Som-hunter: video browsing with relevance-to-som feedback loop. In *MultiMedia Modeling: 26th International Conference, MMM 2020, Daejeon, South Korea, January 5–8, 2020, Proceedings, Part II 26*, pages 790–795. Springer, 2020.

[87] Amirhossein Habibian, Thomas Mensink, and Cees GM Snoek. Videostory: A new multimedia embedding for few-example recognition and translation of events. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 17–26, 2014.

[88] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[89] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

[90] Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne Van Den Berg, Ivan Titov, and Max Welling. Modeling relational data with graph convolutional networks. In *The Semantic Web: 15th International Conference, ESWC 2018, Heraklion, Crete, Greece, June 3–7, 2018, Proceedings 15*, pages 593–607. Springer, 2018.

[91] Kai Sheng Tai, Richard Socher, and Christopher D Manning. Improved semantic representations from tree-structured long short-term memory networks. *arXiv preprint arXiv:1503.00075*, 2015.

[92] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[93] Xiaohan Wang, Linchao Zhu, and Yi Yang. T2vlad: Global-local sequence alignment for text-video retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5079–5088, June 2021.

[94] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[95] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*, 2019.

[96] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

[97] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137, 2015.

[98] Subhashini Venugopalan, Marcus Rohrbach, Jeffrey Donahue, Raymond Mooney, Trevor Darrell, and Kate Saenko. Sequence to sequence-video to text. In *Proceedings of the IEEE international conference on computer vision*, pages 4534–4542, 2015.

[99] Li Yao, Atousa Torabi, Kyunghyun Cho, Nicolas Ballas, Christopher Pal, Hugo Larochelle, and Aaron Courville. Describing videos by exploiting temporal structure. In *Proceedings of the IEEE international conference on computer vision*, pages 4507–4515, 2015.

[100] Subhashini Venugopalan, Huijuan Xu, Jeff Donahue, Marcus Rohrbach, Raymond Mooney, and Kate Saenko. Translating videos to natural language using deep recurrent neural networks. *arXiv preprint arXiv:1412.4729*, 2014.

[101] Kazuya Ueki, Koji Hirakawa, Kotaro Kikuchi, Tetsuji Ogawa, and Tetsunori Kobayashi. Waseda_meisei at trecvid 2017: Ad-hoc video search. In *TRECVID*, 2017.

[102] George Awad, Jonathan Fiscus, David Joy, Martial Michel, Alan F Smeaton, Wessel Kraaij, Maria Eskevich, Robin Aly, Roeland Ordelman, Marc Ritter, et al. Trecvid 2016: Evaluating video search, video event detection, localization, and hyperlinking. In *TREC Video Retrieval Evaluation (TRECVID)*, 2016.

[103] Kazuya Ueki, Takayuki Hori, and Tetsunori Kobayashi. Waseda_meisei_softbank at trecvid 2019: Ad-hoc video search. In *TRECVID*, 2019.

[104] Kazuya Ueki, Yu Nakagome, Koji Hirakawa, Kotaro Kikuchi, Yoshihiko Hayashi, Tetsuji Ogawa, and Tetsunori Kobayashi. Waseda meisei at trecvid 2018: Ad-hoc video search.

[105] Foteini Markatopoulou, Anastasia Moumtzidou, D Galanopoulos, T Mironidis, V Kaltsa, A Ioannidou, S Symeonidis, K Avgerinakis, S Andreadis, I Gialampoukidis, S Vrochidis, Alexia Briassouli, V Mezaris, I Kompatsiaris, and I Patras. Iti-certh participation in trecvid 2016. In *Proceedings TRECVID 2016 Workshop*, 2016.

[106] Foteini Markatopoulou, Damianos Galanopoulos, Vasileios Mezaris, and Ioannis Patras. Query and keyframe representations for ad-hoc video search. In *Proceedings of the 2017 ACM on international conference on multimedia retrieval*, pages 407–411, 2017.

[107] Cees GM Snoek, Xirong Li, Chaoxi Xu, and Dennis C Koelma. University of amsterdam and renmin university at trecvid 2017: Searching video, detecting events and describing video. In *TRECVID*, 2017.

[108] Amirhossein Habibian, Thomas Mensink, and Cees GM Snoek. Videostory embeddings recognize events when examples are scarce. *arXiv preprint arXiv:1511.02492*, 2015.

[109] Ruoyue Shen, Nakamasa Inoue, and Koichi Shinoda. Text-guided object detector for multimodal video question answering. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1032–1042, 2023.

[110] Lu Jiang, Shoou-I Yu, Deyu Meng, Teruko Mitamura, and Alexander G Hauptmann. Bridging the ultimate semantic gap: A semantic search engine for internet videos. In *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*, pages 27–34, 2015.

[111] Amirhossein Habibian, Thomas Mensink, and Cees GM Snoek. Composite concept discovery for zero-shot video event detection. In *Proceedings of International Conference on Multimedia Retrieval*, pages 17–24, 2014.

[112] Niluthpol Chowdhury Mithun, Juncheng Li, Florian Metze, and Amit K Roy-Chowdhury. Learning joint embedding with multimodal cues for cross-modal video-text retrieval. In *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval*, pages 19–27, 2018.

[113] Atousa Torabi, Niket Tandon, and Leonid Sigal. Learning language-visual embedding for movie understanding with natural-language. *arXiv preprint arXiv:1609.08124*, 2016.

[114] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. Vse++: Improving visual-semantic embeddings with hard negatives. *arXiv preprint arXiv:1707.05612*, 2017.

[115] Jianfeng Dong, Xirong Li, and Cees GM Snoek. Predicting visual features from text for image and video caption retrieval. *IEEE Transactions on Multimedia*, 20(12):3377–3388, 2018.

[116] Mayu Otani, Yuta Nakashima, Esa Rahtu, Janne Heikkilä, and Naokazu Yokoya. Learning joint representations of videos and sentences with web image search. In *European Conference on Computer Vision*, pages 651–667. Springer, 2016.

[117] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information fusion*, 58:82–115, 2020.

[118] Aniek F Markus, Jan A Kors, and Peter R Rijnbeek. The role of explainability in creating trustworthy artificial intelligence for health care: a comprehensive survey of the terminology, design choices, and evaluation strategies. *Journal of Biomedical Informatics*, 113:103655, 2021.

[119] Zeynep Akata, Dan Balliet, Maarten De Rijke, Frank Dignum, Virginia Dignum, Guszti Eiben, Antske Fokkens, Davide Grossi, Koen Hindriks, Holger Hoos, et al. A research agenda for hybrid intelligence: augmenting human intellect with collaborative, adaptive, responsible, and explainable artificial intelligence. *Computer*, 53(08):18–28, 2020.

[120] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *International conference on machine learning*, pages 3145–3153. PMLR, 2017.

[121] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[122] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?" explaining the predictions of any classifier". In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

[123] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Anchors: High-precision model-agnostic explanations. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[124] Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harv. JL & Tech.*, 31:841, 2017.

[125] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016.

[126] Ashwin Bhat, Adou Sangbone Assoa, and Arijit Raychowdhury. Gradient backpropagation based feature attribution to enable explainable-ai on the edge. In *2022 IFIP/IEEE 30th International Conference on Very Large Scale Integration (VLSI-SoC)*, pages 1–6. IEEE, 2022.

[127] Zhongang Qi, Saeed Khorram, and Fuxin Li. Visualizing deep networks by optimizing with integrated gradients. In *CVPR Workshops*, volume 2, pages 1–4, 2019.

[128] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017.

[129] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

[130] Yu-Liang Chou, Catarina Moreira, Peter Bruza, Chun Ouyang, and Joaquim Jorge. Counterfactuals and causability in explainable artificial intelligence: Theory, algorithms, and applications. *Information Fusion*, 81:59–83, 2022.

[131] Mukund Sundararajan and Amir Najmi. The many shapley values for model explanation. In *International conference on machine learning*, pages 9269–9278. PMLR, 2020.

[132] Scott M Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. From local explanations to global understanding with explainable ai for trees. *Nature machine intelligence*, 2(1):56–67, 2020.

[133] María Vega García and José L Aznarte. Shapley additive explanations for no2 forecasting. *Ecological Informatics*, 56:101039, 2020.

[134] Danny Francis, Phuong Anh Nguyen, Benoit Huet, and Chong-Wah Ngo. Eurecom at trecvid avs 2019. In *TRECVID*, 2019.

[135] Po-Yao Huang, Junwei Liang, Vaibhav Vaibhav, Xiaojun Chang, and Alexander Hauptmann. Informedia@ trecvid 2018: Ad-hoc video search with discrete and continuous representations. In *TRECVID Proceedings*, volume 70, 2018.

[136] Phuong Anh Nguyen, Qing Li, Zhi-Qi Cheng, Yi-Jie Lu, Hao Zhang, Xiao Wu, and Chong-Wah Ngo. Vireo@ trecvid 2017: Video-to-text, ad-hoc video search, and video hyperlinking. In *TRECVID*, 2017.

[137] Supriyo Chakraborty, Richard Tomsett, Ramya Raghavendra, Daniel Harborne, Moustafa Alzantot, Federico Cerutti, Mani Srivastava, Alun Preece, Simon Julier, Raghuveer M Rao, et al. Interpretability of deep learning models: A survey of results. In *2017 IEEE smartworld, ubiquitous intelligence & computing, advanced & trusted computed, scalable computing & communications, cloud & big data computing, Internet of people and smart city innovation (smartworld/SCALCOM/UIC/ATC/CBDcom/IOP/SCI)*, pages 1–6. IEEE, 2017.

[138] Yu Zhang, Peter Tiňo, Aleš Leonardis, and Ke Tang. A survey on neural network interpretability. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 5(5):726–742, 2021.

[139] Angelos Chatzimparmpas, Rafael M Martins, Ilir Jusufi, and Andreas Kerren. A survey of surveys on the use of visualization for interpreting machine learning models. *Information Visualization*, 19(3):207–233, 2020.

[140] Jonathan G Richens, Ciarán M Lee, and Saurabh Johri. Improving the accuracy of medical diagnosis with causal machine learning. *Nature communications*, 11(1):3923, 2020.

[141] Andreas Holzinger, Georg Langs, Helmut Denk, Kurt Zatloukal, and Heimo Müller. Causability and explainability of artificial intelligence in medicine. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 9(4):e1312, 2019.

[142] Judea Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan kaufmann, 1988.

[143] Andreas Holzinger. Explainable ai and multi-modal causability in medicine. *i-com*, 19(3):171–179, 2021.

[144] Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. Building machines that learn and think like people. *Behavioral and brain sciences*, 40:e253, 2017.

[145] Judea Pearl. Causality: models, reasoning, and inference, 1980.

[146] Zhichao Hu and Marilyn A Walker. Inferring narrative causality between event pairs in films. *arXiv preprint arXiv:1708.09496*, 2017.

[147] Chikara Hashimoto, Kentaro Torisawa, Julien Kloetzer, Motoki Sano, István Varga, Jong-Hoon Oh, and Yutaka Kidawara. Toward future scenario generation: Extracting event causality exploiting semantic relation, context, and association features. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 987–997, 2014.

[148] Xun Yang, Fuli Feng, Wei Ji, Meng Wang, and Tat-Seng Chua. Deconfounded video moment retrieval with causal intervention. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1–10, 2021.

[149] Shangwen Li, Sanjay Purushotham, Chen Chen, Yuzhuo Ren, and C-C Jay Kuo. Measuring and predicting tag importance for image retrieval. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2423–2436, 2017.

[150] Ran Xu, Caiming Xiong, Wei Chen, and Jason Corso. Jointly modeling deep video and compositional text to bridge vision and language in a unified framework. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015.

[151] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

[152] Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.", 2009.

[153] Fabian Berns, Luca Rossetto, Klaus Schoeffmann, Christian Beecks, and George Awad. V3c1 dataset: an evaluation of content characteristics. In *Proceedings of the 2019 on International Conference on Multimedia Retrieval*, pages 334–338, 2019.

[154] Luca Rossetto, Heiko Schuldt, George Awad, and Asad A Butt. V3c–a research video collection. In *MultiMedia Modeling: 25th International Conference, MMM 2019, Thessaloniki, Greece, January 8–11, 2019, Proceedings, Part I 25*, pages 349–360. Springer, 2019.

[155] Alexey Dosovitskiy and Thomas Brox. Generating images with perceptual similarity metrics based on deep networks. *Advances in neural information processing systems*, 29, 2016.

[156] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[157] Ian T Nabney, Yi Sun, Peter Tino, and Ata Kabán. Semisupervised learning of hierarchical latent trait models for data visualization. *IEEE transactions on knowledge and data engineering*, 17(3):384–400, 2005.

[158] Xiaonan Ji, Han-Wei Shen, Alan Ritter, Raghu Machiraju, and Po-Yin Yen. Visual exploration of neural document embedding in information retrieval: Semantics and feature selection. *IEEE transactions on visualization and computer graphics*, 25(6):2181–2192, 2019.

[159] Bartosz Krawczyk, Leandro L Minku, João Gama, Jerzy Stefanowski, and Michał Woźniak. Ensemble learning for data stream analysis: A survey. *Information Fusion*, 37:132–156, 2017.

[160] Xiao-Lei Zhang and DeLiang Wang. A deep ensemble learning method for monaural speech separation. *IEEE/ACM transactions on audio, speech, and language processing*, 24(5):967–977, 2016.

[161] Thomas G Dietterich et al. Ensemble learning. *The handbook of brain theory and neural networks*, 2(1):110–125, 2002.

[162] Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7), 2015.

[163] Anuvabh Dutt, Denis Pellerin, and Georges Quénot. Coupled ensembles of neural networks. *Neurocomputing*, 396:346–357, 2020.

[164] Duc Hau Nguyen, Guillaume Gravier, and Pascale Sébillot. A study of the plausibility of attention between rnn encoders in natural language inference. In *2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 1623–1629. IEEE, 2021.

[165] Duc Hau Nguyen, Guillaume Gravier, and Pascale Sébillot. Filtrage et régularisation pour améliorer la plausibilité des poids d'attention dans la tâche d'inférence en langue naturelle. In *TALN 2022-Traitement Automatique des Langues Naturelles*, pages 95–103. ATALA, 2022.

[166] Renjie Liao, Alex Schwing, Richard Zemel, and Raquel Urtasun. Learning deep parsimonious representations. *Advances in neural information processing systems*, 29, 2016.

[167] Liting Zhou, Jianquan Liu, Shoji Nishimura, Joseph Antony, and Cathal Gurrin. Causality inspired retrieval of human-object interactions from video. In *2019 International Conference on Content-Based Multimedia Indexing (CBMI)*, pages 1–6. IEEE, 2019.

[168] Patrick Schwab and Walter Karlen. Cxplain: Causal explanations for model interpretation under uncertainty. *Advances in Neural Information Processing Systems*, 32, 2019.

[169] Sumedha Singla, Stephen Wallace, Sofia Triantafillou, and Kayhan Batmanghelich. Using causal analysis for conceptual deep learning explanation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24*, pages 519–528. Springer, 2021.

[170] Susanne Dandl and A Christian. General pitfalls of model-agnostic interpretation methods for machine learning models. *stat*, 1050:17, 2021.

[171] Mathias Kraus, Daniel Tschernutter, Sven Weinzierl, and Patrick Zschech. Interpretable generalized additive neural networks. *European Journal of Operational Research*, 2023.

[172] George A. Miller. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *The Psychological Review*, 63(2):81–97, March 1956.

[173] J. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in Large Margin Classifiers*, 10(3), 1999.

[174] Michael Wray, Hazel Doughty, and Dima Damen. On semantic similarity in video retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3650–3660, 2021.

[175] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15750–15758, 2021.

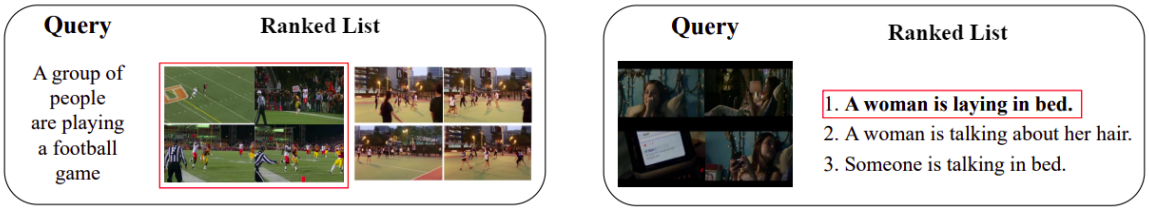
Appendix A

Résumé en Français

A.1 Introduction

À l'ère actuelle de la génération et de la collecte de données à grande échelle sur Internet, l'indexation et la recherche efficaces sont difficils. Avec des millions de vidéos et de textes téléchargés quotidiennement sur Internet, le défi consiste à localiser les informations pertinentes. Ce défi met en évidence l'importance du domaine de la recherche *cross modale* où, pour une requête donnée dans n'importe quelle modalité (vidéo, image ou texte), la tâche consiste à trouver des informations pertinentes dans la même modalité ou dans une autre modalité. Le travail de recherche de cette thèse est basé en particulier sur la recherche vidéo-textuelle (VTR) [175, 29, 27, 14, 16, 22, 23, 43], qui permet aux utilisateurs d'exprimer leurs besoins d'information en langage naturel et de retrouver des vidéos pertinentes ou vice versa.

Depuis une dizaine d'années, les réseaux neuronaux profonds ont gagné en popularité dans le domaine de l'indexation et de la recherche multimédia, ce qui en fait un choix populaire pour le développement et l'évaluation de tels systèmes. La figure A.1 illustre les deux principales sous-tâches de VTR : la recherche de texte à vidéo (TTV) et la recherche de vidéo à texte (VTT).



(a) Text-to-Video Retrieval **(b)** Video-to-Text Retrieval

Figure A.1: Recherche cross modale vidéo-texte

La recherche TTV consiste à utiliser une requête textuelle pour retrouver une liste de vidéos pertinentes, tandis que la recherche VTT fait l'inverse, en récupérant les légendes textuelles pertinentes pour une requête vidéo donnée en requête. Ces tâches sont difficiles à réaliser en raison de l'hétérogénéité inhérente et de la haute dimensionnalité des données multimédias. Pour s'attaquer à la VTR, les chercheurs ont exploré trois approches principales : les méthodes basées sur des concepts, les méthodes sans concepts et les méthodes hybrides. Les approches basées sur les concepts utilisent des concepts ou des catégories prédéfinis pour représenter les informations sémantiques dans les vidéos et les textes, et les font correspondre sur la base de la similarité des concepts dans l'espace conceptuel. Les approches sans concept extraient directement les enchâssements des vidéos et des textes, en faisant correspondre les caractéristiques dans un espace latent commun. Les approches hybrides combinent les méthodes basées sur les concepts et les méthodes sans concept.

La recherche TTV A.1a présente un exemple de recherche "texte-vidéo (TTV)". Ici,

une requête textuelle donnée est utilisée pour récupérer une liste de vidéos et les classer en fonction de leur score de pertinence. Alors que dans VTT A.1b, la requête est sous forme de vidéo, et il récupère les légendes pertinentes et les classe de la même manière. Les vidéos et les légendes pertinentes sont encadrées en rouge. Répondre à ces tâches nécessite des techniques sophistiquées telles que les approches basées sur l'apprentissage profond pour obtenir des performances élevées. La recherche multimodale (en particulier la VTR) implique principalement trois types d'approches : *approche basée sur les concepts* [7, 103, 104, 111] – méthode qui utilise des concepts ou des catégories prédéfinis pour représenter les informations sémantiques des vidéos et des textes, puis les fait correspondre sur la base de leur similarité conceptuelle dans **espace conceptuel**, *approche sans concept* [8, 21, 25] – une méthode qui apprend et extrait l'intégration directement de la vidéo et du texte et fait correspondre les caractéristiques en commun **espace latent**, et *approche hybride* – une méthode qui combine à la fois les méthodes basées sur les concepts (espace conceptuel) et les méthodes sans concept (espace latent).

Une extension importante de la VTR est la recherche de l'explicabilité, permettant aux utilisateurs de comprendre pourquoi des résultats de recherche spécifiques ont été obtenus. Avec l'augmentation de la complexité des modèles et de la quantité de données, il est devenu crucial de fournir aux utilisateurs non techniques ou techniques des informations sur le processus de prise de décision des systèmes de recherche de vidéotextes. La recherche explicative de vidéos et de textes s'efforce d'être transparente et interprétable, permettant aux utilisateurs de comprendre quels aspects de la vidéo ou du texte ont influencé le résultat de la recherche. Cela permet non seulement de renforcer la confiance des utilisateurs, mais aussi de soutenir les applications dans lesquelles la responsabilité et la compréhension du comportement du système sont essentielles. Les explications fournies peuvent prendre la forme d'un texte en langage naturel, de cartes thermiques, ou bien d'un nuage de tags (comme le montre la figure A.2) .

Figure A.2: Recherche texte à vidéo. Les nuages de tags (sous la requête) et à gauche des vidéos justifient les résultats results de la requête [1]

Nous concentrons notre travail de thèse sur la création d'un modèle hybride explicable composé d'un espace double composé d'un espace latent (non explicable) et d'un espace conceptuel (explicable), et capable d'effectuer une double tâche : la recherche et la classifica-

tion. L'espace conceptuel est supposé non seulement supporter l'interprétabilité des résultats récupérés mais également compléter l'espace latent, améliorant la performance globale de l'approche hybride [1]. Etayer une telle affirmation nécessite d'étudier de manière détaillée ***les espaces d'intégration appris ainsi que leurs caractéristiques***. Cette analyse détaillée serait également utile pour comprendre comment le modèle hybride, très efficace, représente et utilise les informations multimodales.

Dans cette thèse, nous visons non seulement à augmenter la précision et l'efficacité du modèle de recherche, mais aussi à améliorer la fiabilité et la satisfaction des explications basées sur un nuage de tags lorsque des vidéos pertinentes sont récupérées pour une requête textuelle ou vice versa. Cependant, la partie explicable du modèle hybride, basée sur les concepts, ne règle pas les problèmes d'ambiguïté dans le vocabulaire de ses concepts. Nous cherchons à minimiser en intégrant les informations d'extraction de groupes morphosyntaxiques dans le vocabulaire des concepts. Notre objectif est d'améliorer la précision et l'explicabilité du modèle hybride en affinant l'espace conceptuel afin de garantir la sélection de classificateurs de concepts visuels pertinents.

En outre, ce travail vise également à intégrer la causalité dans les explications basées sur le nuage de tags afin de fournir une compréhension plus profonde des décisions d'extraction. Globalement, notre recherche aborde les questions suivantes :

1. Pouvons-nous améliorer le vocabulaire conceptuel pour minimiser l'ambiguïté des concepts visuels ?

2. L'idée de complémentarité entre l'espace conceptuel et l'espace latent est-elle vraiment valable ?

3. Quel sera l'effet de l'intégration de la causalité dans l'explication basée sur le nuage de balises pour la recherche vidéotextuelle ?

Ces questions portent sur des lacunes importantes dans la littérature de recherche actuele, et notre thèse vise à fournir des idées et des solutions pour améliorer la recherche de texte vidéo et son explicabilité.

A.2 Modèle de double encodage basé sur les balises PoS

Comme première contribution à cette thèse, nous étendons le modèle de codage double [1] en explorant l'impact de l'incorporation des balises morphosyntaxiques, aussi appelées Part-of-Speech (PoS), dans le pipeline de codage de texte pour l'entraînement du modèle de codage double. L'étiquetage PoS est un processus crucial qui attribue des étiquettes spécifiques aux mots, représentant leurs catégories syntaxiques. En incorporant les balises PoS, nous visons à réduire l'ambiguïté des concepts visuels présents dans le vocabulaire conceptuel et

à exploiter les informations syntaxiques et grammaticales fournies par les balises PoS afin d'améliorer les performances, la pertinence et l'explicabilité de la recherche vidéotextuelle.

A.2.1 Motivation

Considérons une requête textuelle *"Un homme mesure la taille d'un serpent à tête de cuivre avec un mètre ruban"*, le concept visuel "mesure" est présent à la fois en tant que verbe et en tant que nom. La présence de "mesure" en tant que verbe et nom dans la requête oblige le système de recherche à se concentrer sur les vidéos où l'action de mesurer et les objets à mesurer sont tous deux présents, ce qui est susceptible d'afficher des vidéos pertinentes. Les balises PoS clarifient l'action prévue et mettent en évidence la pertinence des vidéos qui décrivent cette activité spécifique.

De même, dans d'autres cas, par exemple "une personne arrose ses fleurs tandis que des gens marchent sous l'eau" et "un homme et une femme cuisinant dans une émission culinaire", la présence de balises PoS permet de désambiguïser le verbe ou le nom, comme l'eau dans le premier cas et la cuisine dans le second, et d'identifier les noms singuliers ou pluriels, par exemple les gens ou la personne. Cela permet d'obtenir des résultats de recherche vidéo plus précis et plus pertinents, conformes au sens de la requête.

Les exemples ci-dessus montrent clairement que l'inclusion de balises PoS offre plusieurs avantages dans le contexte de la recherche de textes vidéo. Tout d'abord, elle permet d'aborder ***ambiguïté*** dans les concepts au sein du vocabulaire de l'espace conceptuel. Par exemple, si le concept de "mesure" n'est pas distingué en fonction de son étiquette PoS (verbe et nom), il est probable que le système de recherche se concentrera davantage sur les objets de mesure que sur les activités de mesure lors de la recherche des vidéos. En tenant compte des catégories syntaxiques des mots, nous pouvons désambiguïser leur signification et améliorer la précision de la classification. Cela est particulièrement utile dans les scénarios où plusieurs interprétations ou sens sont possibles. Deuxièmement, l'étiquetage du PoS contribue à la ***interprétabilité*** des résultats récupérés, actuellement réservés aux utilisateurs techniques. En incorporant les balises PoS, nous pouvons analyser et expliquer l'influence des mots avec des balises PoS dans le processus de recherche.

En outre, l'intégration des balises PoS permet une analyse plus approfondie du contenu textuel. En examinant la structure syntaxique des phrases, nous obtenons des informations sur les ***relations entre les mots*** et leurs rôles dans la phrase. Ces informations supplémentaires peuvent aider à saisir les nuances et le contexte du texte, améliorant ainsi les performances de la recherche vidéotextuelle.

A.2.2 Formulation

Formellement, étant donné un ensemble de vidéos $\mathbb{V} = \{v_1, v_2, ..., v_n\}$ et un ensemble correspondant de légendes $\mathbb{S} = \{s_1, s_2, ..., s_m\}$, où n et m représentent le nombre total de vidéos et de légendes dans l'ensemble de données, notre modèle vise à atteindre deux objectifs principaux :

1. deux fonctions de mise en correspondance, $f()$ et $g()$, pour les encodages visuels et textuels dans deux espaces : les encodages de l'espace latent $(f(v_i), f(s_i))$ et les encodages de l'espace conceptuel $(g(s_i), g(v_i))$.

2. Apprendre deux fonctions de similarité pour calculer la similarité entre la vidéo v_i et la légende s_j : $sim_{lat}(v_i, s_j)$ pour la similarité dans l'espace latent et $sim_{con}(v_i, s_j)$ pour la similarité dans l'espace conceptuel PoS-tag.

Dans cette section, nous nous concentrons sur l'intégration des balises PoS dans le pipeline d'encodage de texte du modèle d'encodage double et sur l'apprentissage des fonctions de similarité $sim(v, s)$ pour déterminer la similarité entre le texte s et la vidéo v à la fois dans l'espace latent et dans l'espace conceptuel.

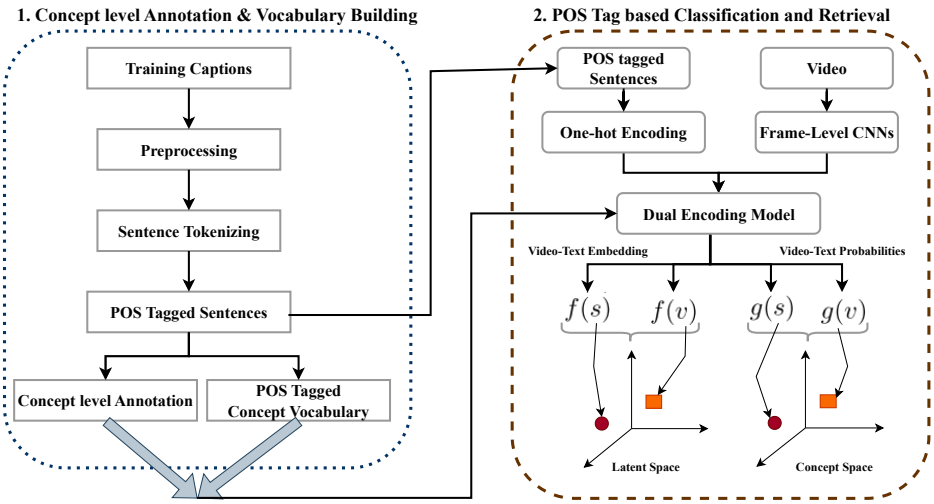


Figure A.3: Architecture de codage double avec balises PoS pour la classification et la recherche vidéotextuelle

Notre approche globale s'articule autour de deux étapes, comme le montre la figure A.3, i) l'annotation au niveau du concept basée sur les balises PoS & la construction du vocabulaire, et ii) la classification et l'extraction basées sur les balises PoS. Dans un premier temps, nous annotons les vidéos avec des concepts ou des balises, ainsi qu'avec les balises PoS correspondantes. Nous construisons ensuite un vocabulaire de concepts basé sur ces annotations de balises PoS. Dans un deuxième temps, nous utilisons les annotations et le vocabulaire au niveau des concepts pour former l'espace conceptuel de notre modèle hybride. Lors de la

recherche, nous classons les vidéos ou les légendes en fonction de leur similarité pondérée combinée dans l'espace latent et l'espace conceptuel.

A.2.3 Résultats

Dans notre évaluation, nous avons comparé notre modèle de double encodage marqué par le PoS avec le modèle de double encodage [1] sur cinq ensembles de données vidéo et texte, couvrant les tâches de recherche de texte à vidéo et de recherche de vidéo à texte. Pour garantir une analyse complète, nous avons soigneusement sélectionné divers ensembles de données et utilisé des mesures d'évaluation rigoureuses. En outre, nous nous sommes concentrés sur l'évaluation de l'impact du modèle étiqueté PoS sur la précision de la recherche et l'explicabilité. Nous avons également mené des expériences comparables en utilisant différents PoS-taggers, à savoir TreeTagger (TT) [1], WordNet (WN) [152], et Spacy [2]. En étudiant ces facteurs, nous avons cherché à obtenir des informations précieuses sur les performances et les avantages potentiels de notre modèle de double encodage étiqueté PoS pour la recherche vidéotextuelle.

Les performances d'un modèle de codage double avec étiquette PoS pour les tâches de recherche texte-vidéo (TTV) et vidéo-texte (VTT) ont été évaluées à l'aide de diverses mesures. Les mesures comprenaient des paramètres tels que R@K, le rang médian (Med R), la précision moyenne (mAP) et SumR pour différentes configurations du modèle, en mettant l'accent sur la dimensionnalité et les méthodes d'étiquetage PoS.

Méthode	**Recherche Texte vers Vidéo**					**Recherche Vidéo vers Texte**					**SumR**
	R@1	R@5	R@10	Med r	mAP	R@1	R@5	R@10	Med r	mAP	
Apprentissage Hybride											
Dual Encoding TPAMI [1]	11.6	30.3	41.3	17	21.2	22.5	47.1	58.9	7	10.5	211.7
Dual Encoding GitHub [1]	11.8	30.6	41.8	17	21.4	21.6	45.9	58.5	7	10.3	210.2
Dual Enc. (Conf. Ver. 2048-d)[1])	11.0	29.2	39.8	19	20.2	18.8	42.7	56.2	8	9.3	197.7
Dual Enc. (Conf. Ver. 1536-d)[1])	11.0	29.3	39.9	19	20.3	19.7	43.6	55.6	8	9.3	199.0
(1536+512)-d Dual Enc. Re-Run	11.78	31.00	42.08	16.10	21.52	20.70	45.10	57.74	7.00	10.13	**208.40**
(1536+512)-d Dual Enc. (TT)	12.09	31.39	42.50	16.10	21.87	20.68	45.32	58.12	6.90	10.32	210.10
(1536+512)-d Dual Enc. (WN)	12.02	31.44	42.59	16.00	21.84	20.56	45.60	58.28	6.90	10.36	**210.50**
(1536+512)-d Dual Enc. (Spacy)	11.84	31.07	42.08	16.40	21.59	20.29	45.00	57.78	7.00	10.22	208.05
512-d Dual Enc. (Concept Space)	9.9	26.8	37.4	23	18.7	17.9	41.5	53.9	8	9.0	**187.4**
512-d Dual Enc. (Concept Space - TT)	10.10	27.09	37.31	22.65	18.86	18.64	41.83	54.79	8.50	9.14	189.74
512-d Dual Enc. (Concept Space - WN)	10.14	27.23	37.53	22.40	18.94	18.54	42.25	55.19	8.00	9.11	**190.87**
512-d Dual Enc. (Concept Space - Spacy)	9.85	26.71	36.76	23.55	18.54	18.15	41.45	54.22	8.30	9.02	187.15
Apprentissage concept											
512-d Dual Enc. (Concept Space)	9.84	26.79	37.02	23.30	18.57	18.57	41.85	54.22	8.40	8.88	**188.28**
512-d Dual Enc. (Concept Space - TT)	10.15	27.33	37.65	22.10	18.98	18.82	42.32	55.22	8.15	9.17	**191.49**
512-d Dual Enc. (Concept Space - WN)	10.14	27.30	37.58	22.40	18.96	18.69	42.13	54.99	8.00	9.15	190.82
512-d Dual Enc. (Concept Space - Spacy)	9.96	26.92	37.18	23.00	18.71	18.39	41.59	54.52	8.50	9.04	188.55

Table A.1: Résultats sur MSR-VTT, sur l'ensemble de test de [13].

Les résultats, présentés dans le tableau A.1, évalués sur l'ensemble de données MSR-

[1] https://pypi.org/project/treetaggerwrapper/
[2] https://github.com/explosion/spaCy

VTT, montrent que le modèle de PoS-tagged proposé a légèrement amélioré la précision de la recherche, TreeTagger étant le plus performant parmi les PoS-taggers. Cependant, les gains n'étaient pas substantiels en raison de la présence limitée de mots ambigus dans l'ensemble de données. L'évaluation du modèle de double codage et balises PoS sur le sous-ensemble de requêtes sélectionnées avec plusieurs mots ambigus a montré que notre proposition améliore significativement la précision par rapport au système d'origine. Plus précisément, le score SumR est passé de 312,59 à 319, 17. Cette amélioration souligne l'efficacité de l'incorporation de l'étiquetage PoS pour la recherche vidéotextuelle.

A.3 Analyse de complémentarité dans les modèles à double espace

La deuxième contribution à cette thèse est l'étude analytique des modèles à double espace. L'étude analytique menée dans cette section visait à approfondir la relation et la complémentarité entre les espaces latents et les espaces conceptuels dans le contexte de la recherche multimodale de textes et de vidéos. L'analyse complète et les résultats expérimentaux mettent en lumière les dimensions optimales de ces espaces, leur complémentarité potentielle et l'impact de l'apprentissage d'ensemble afin de répondre aux trois questions de recherche suivantes.

> *R1 : Le nombre de **dimensions optimales** est-il le même dans les espaces conceptuels et latents pour deux sous-tâches de la recherche vidéotextuelle multimodale, Texte vers Vidéo (TTV) et Vidéo vers Texte (VTT)?*
>
> *R2 : Les espaces latents et conceptuels représentent-ils des informations **complémentaires**?*
>
> *R3 : Est-ce que **l'apprentissage par ensembles** présente une complémentarité dans les espaces latents et conceptuels ?*

La discussion ci-dessous résume les principaux résultats et leurs implications.

A.3.1 Dimensions optimales et analyse ACP (R1)

La première question de recherche (R1) vise à déterminer si les espaces conceptuels et latents ont des dimensions optimales similaires lorsqu'ils sont appris indépendamment. En explorant les dimensions optimales de chaque espace séparément, nous cherchons à savoir si ces espaces possèdent des capacités de représentation similaires. Cette question est importante pour comprendre si la complémentarité entre ces espaces provient de différences inhérentes

à leur représentation, ou s'ils partagent des caractéristiques communes. Nous avons utilisé deux techniques, l'une sans ACP (analyse en composantes principales) et l'autre avec ACP. Sans ACP, nous avons entraîné l'espace latent avec un nombre variable de dimensions afin de trouver l'espace optimal. Dans le cas de l'ACP, nous avons utilisé l'espace latent et conceptuel de haute dimension appris initialement et nous avons utilisé l'analyse en composantes principales pour explorer les dimensions linéaires saillantes et leur variance dans l'espace de plus faible dimension.

L'étude de la dimensionnalité dans l'espace latent et l'espace conceptuel a révélé que le nombre de dimensions optimales pour les deux espaces est le même, ce qui est nettement inférieur aux dimensions utilisées par *Dong et al.* dans des travaux antérieurs (par exemple, 1536-D et 512-D) [1]. L'espace conceptuel, bien que légèrement moins performant, présente des valeurs optimales identiques à celles de l'espace latent. La performance inférieure peut s'expliquer par le fait que la tâche de classification associée à l'espace conceptuel impose de fortes contraintes à ce dernier. L'asymétrie entre les comportements TTV et VTT peut s'expliquer par l'asymétrie du rapport entre les légendes et les vidéos. Des fluctuations peuvent être observées sur toutes les courbes. Elles sont dues à l'effet de l'initialisation aléatoire dans l'entraînement et sont du même niveau que ce qui est observé lorsque les mêmes expériences sont réalisées plusieurs fois. L'analyse ACP menée a confirmé le nombre optimal de dimensions dans l'espace latent.

A.3.2 Corrélation et complémentarité (R2)

La question de recherche R2 vise à déterminer si les espaces latents et conceptuels capturent des informations distinctes ou similaires à partir des données. Pour répondre à cette question, l'analyse de corrélation canonique (CCA) est utilisée pour évaluer la corrélation et la complémentarité entre ces deux espaces de caractéristiques à haute dimension. La principale distinction entre les deux espaces réside dans le fait que l'espace conceptuel est associé à une tâche de classification, alors que l'espace latent ne l'est pas. Si la tâche de classification est supprimée, l'espace conceptuel devient simplement un deuxième espace latent avec des caractéristiques différentes (par exemple, en utilisant une similarité de Jaccard au lieu d'un cosinus). L'analyse considère quatre combinaisons possibles d'espaces latents et/ou conceptuels, y compris l'apprentissage indépendant (deux espaces identiques), l'homogène couplé (apprentissage conjoint de deux espaces latents identiques), l'hétérogène couplé (différents espaces latents avec des métriques de similarité variables, et un concept hétérogène couplé (système hybride incorporant une tâche de classification pour créer un espace conceptuel). Tous les espaces sont normalisés pour avoir une dimensionnalité de 512 pour des comparaisons significatives.

Les résultats obtenus ont révélé des corrélations élevées entre les mappages de concepts

hétérogènes et hétérogènes couplés dans toutes les configurations, avec de légères variations. Notamment, la formation indépendante a montré la corrélation la plus faible, tandis que la configuration homogène couplée a montré la corrélation la plus élevée. Il est intéressant de noter que la présence d'une tâche de classification dans la configuration de concepts hétérogènes couplés n'a pas eu d'effet significatif sur la corrélation, car les configurations de concepts hétérogènes couplés et hétérogènes couplés ont présenté des profils de corrélation intermédiaires et presque identiques. Cette analyse suggère que la tâche de classification a un impact minimal sur la corrélation entre ces espaces.

A.3.3 Apprentissage par ensembles et complémentarité (R3)

La troisième question de recherche (R3) explore la complémentarité des espaces latents et conceptuels par le biais de **l'apprentissage ensembliste**. Cette question vise à déterminer si l'utilisation de techniques d'apprentissage d'ensemble dans ces espaces permet d'améliorer les performances et de démontrer leur complémentarité. L'apprentissage d'ensemble permet de combiner plusieurs modèles ou représentations, et si la performance du modèle d'ensemble est similaire à l'utilisation indépendante de l'un ou l'autre espace, alors les espaces sont en effet similaires et ne présentent pas de forte complémentarité.

Méthode	**Recherche Texte vers Vidéo**					**Recherche Vidéo vers Texte**					**SumR**
	R@1	R@5	R@10	Med r	mAP	R@1	R@5	R@10	Med r	mAP	
Evaluation sur un espace											
Latent independent 1536 (10)	10.94	29.12	39.73	19.30	20.21	19.00	42.60	54.81	8.10	9.26	196.20
Latent independent 512 (20)	10.88	29.06	39.73	19.25	20.15	19.42	42.91	55.20	7.95	9.30	197.21
Latent-latent coupled homogeneous (10)	11.17	29.83	40.58	18.10	20.63	19.95	43.77	56.31	7.60	9.57	201.61
Latent-latent coupled heterogeneous (10)	11.37	30.25	41.11	17.80	20.93	19.85	43.70	56.26	7.60	9.63	202.54
Latent-concept coupled heterogeneous (10)	11.42	30.29	41.16	17.70	21.00	19.65	43.24	55.84	7.90	9.57	201.60
Evaluation sur deux espaces											
Latent-latent indep. homogeneous (10)	11.50	30.30	41.22	17.55	21.04	20.92	44.78	56.99	7.25	9.89	205.71
Latent-latent coupled homogeneous (10)	11.41	30.31	41.16	17.70	20.98	20.30	44.57	56.85	7.10	9.80	204.60
Latent-latent coupled heterogeneous (10)	11.78	31.12	42.28	16.20	21.60	21.23	45.65	58.08	7.00	10.36	210.14
Latent-concept coupled heterogeneous (10)	11.76	30.98	42.10	16.40	21.52	20.25	44.74	57.48	7.10	10.09	207.31

Table A.2: Résultats de l'apprentissage par ensembles sur MSR-VTT

En évaluant les quatre combinaisons pour l'apprentissage par ensembles, l'étude a mené des expériences quantitatives sur les espaces fusionnés en utilisant les métriques MSR-VTT. Dans le tableau A.2, les résultats moyennés sur 10 exécutions avec différentes initialisations aléatoires n'ont montré aucune différence statistiquement significative entre la formation indépendante et la formation couplée pour les espaces latents homogènes. Les meilleures performances ont été observées dans la configuration hétérogène couplée latent-latent, qui utilisait des espaces latents présentant des similarités différentes (cosinus et Jaccard). Les expériences avec des configurations homogènes couplées latent-latent ont été moins performantes, ce qui indique une diminution des performances lorsque la similarité cosinus est utilisée dans l'espace conceptuel. L'ajout d'une tâche de classification a entraîné une légère diminution des performances, bien que la signification statistique soit minime. La première

partie du tableau montre qu'il n'y a pas de différence significative entre un espace latent de 1536 D et un espace latent de 512 D lors de l'exécution de la tâche en utilisant uniquement le premier espace latent.

A.4 Inférence causale dans l'extraction de vidéotextes

Comme vu en section A.1, les explications fournies par le modèle hybride de recherche vidéotextuelle peuvent se présenter sous la forme d'un nuage de tags. L'un de ces systèmes est le modèle de codage double pour la recherche vidéotextuelle [1]. Ces systèmes mettent en œuvre simultanément une tâche de mise en correspondance texte-vidéo et une tâche connexe de classification des concepts. Il est essentiel que la mise en correspondance entre le texte et la vidéo soit effectuée en utilisant uniquement les résultats de la classification des concepts, ce qui impose une relation causale stricte entre la classification des concepts et l'extraction TTV/VTT [2]. En outre, le calcul du score de similarité pour la recherche est dérivé des scores de classification à l'aide d'une fonction simple, facile à comprendre et intuitivement logique, en pratique la similarité de Jaccard ou de cosinus. Le fonctionnement du système peut alors être interprété comme des décisions de recherche (basées sur des similarités) utilisant uniquement des scores de classification correspondant à des étiquettes significatives pour les humains et d'une manière qui l'est également pour eux.

La figure A.2 illustre une explication/justification fournie à un utilisateur à l'aide d'une approche hybride : les nuages de tags montrent les concepts jugés les plus pertinents (avec des tailles liées à leur importance estimée) pour la requête et pour les 4 documents les mieux classés. L'utilisateur peut apprécier dans quelle mesure ces nuages de tags sont réellement pertinents pour la requête et les documents, et dans quelle mesure ils correspondent. Cependant, il est important de noter que ces nuages de mots-clés ne fournissent pas d'informations sur leur contribution relative à la décision de recherche globale. Notre travail se situe à ce niveau, avant de tels affichages. Notre troisième contribution étudie comment mesurer la contribution causale des scores de détection de concepts dans les décisions de recherche, et propose des moyens d'améliorer cette causalité sur un système de l'état de l'art.

A.4.1 Analyse de la causalité

A.4.1.1 Quantification de la causalité

Nous nous concentrons sur la partie basée sur les concepts du modèle de codage double, dans ce cas les vidéos et les requêtes sont représentées uniquement par les scores de détection des concepts. Les vidéos sont alors classées par ordre décroissant de similarité de leurs représentations avec celles du contenu de la requête. *Dong et al.* dans le modèle de codage double [1] utilise par défaut la fonction de similarité de Jaccard pour les concepts entre v et

s respectivement un échantillon vidéo et un échantillon de texte.

La fonction de similarité par cosinus peut également être envisagée, car elle est déjà utilisée par défaut dans l'espace latent de [1]. Dans le cas de la similarité en cosinus, les scores de détection sont utilisés sans fonction sigmoïde.

Nous quantifions la causalité d'un groupe d'étiquettes dans une décision de recherche par la somme de leurs effets de caractéristiques, eux-mêmes pris comme leur contribution globale *relative* dans la mesure de similarité utilisée pour le classement des résultats. On obtient ainsi les contributions individuelles des balises Jaccard et cosinus :

$$w_i(v,s) = \frac{\min(g(v)_i, g(s)_i)}{\sum_{j=0}^{j=K} \min(g(v)_j, g(s)_j)} \quad \text{or} \quad \frac{|h(v)_i.h(s)_i|}{\sum_{j=0}^{j=K} |h(v)_j.h(s)_j|} \tag{A.1}$$

L'évaluation de la contribution relative d'un groupe d'étiquettes peut ensuite être effectuée par sommation, par exemple, les k les plus importants, comme ceux qui sont affichés dans les nuages d'étiquettes :

$$c_k(v,s) = \max_{G \subset [\![1,K]\!], |G|=k} \sum_{i \in G} w_i(v,s)$$

À partir de cette mesure, définie pour une unique paire (v,s), nous dérivons des mesures globales sur une collection multimodale entière en calculant des statistiques telles que la moyenne (équation (A.2)) et l'écart type de cette valeur sur un ensemble de paires P.

$$C_k(P) = \frac{1}{|P|} \sum_{(v,s) \in P} c_k(v,s) \tag{A.2}$$

P peut être l'ensemble de toutes les paires possibles dans la collection ou seulement l'ensemble des paires correspondantes. Nous pouvons également considérer l'ensemble des paires obtenues en utilisant toutes les requêtes textuelles et, pour chacune d'entre elles, les premières-n vidéos retrouvées, ou l'inverse en utilisant les requêtes vidéo et les textes retrouvés.

Dans notre cas, la causalité dans les explications/justifications repose uniquement sur les scores de détection des balises affichées. C'est le cas pour les dimensions d'un espace conceptuel, mais pas pour les dimensions d'un espace purement latent, car elles n'ont pas de sens pour les humains. Le poids causal de tout élément provenant de l'espace latent dans l'explication/justification visuelle basée sur les concepts doit donc être strictement nul. Dans l'approche fondée uniquement sur l'espace latent [8], aucun score de détection de concept n'est de toute façon disponible pour l'affichage des nuages de tags. Cependant, de tels scores sont disponibles dans les approches hybrides [1], car la décision est prise en partie sur les similarités $sim_{lat}(v,s)$ provenant de l'espace latent et en partie sur les similarités $sim_{con}(v,s)$

provenant de l'espace conceptuel. La similarité globale est une somme pondérée (après une normalisation globale de l'échelle) $sim(v,s) = \alpha.sim_{lat}(v,s) + (1-\alpha).sim_{con}(v,s)$. La causalité globale devrait logiquement être une somme pondérée basée uniquement sur les scores des concepts multipliés par le facteur $(1-\alpha)$ dans lequel, comme la causalité sur la partie latente devrait être nulle.

A.4.2 Évaluation de la causalité du système [1]

Nous avons évalué la causalité entre le score de détection des balises et la similarité en utilisant le modèle hybride pré-entraîné fourni par les auteurs de [1] sur l'ensemble de données MSR-VTT [13]. Dans ce modèle hybride, la similarité basée sur les concepts représente 40

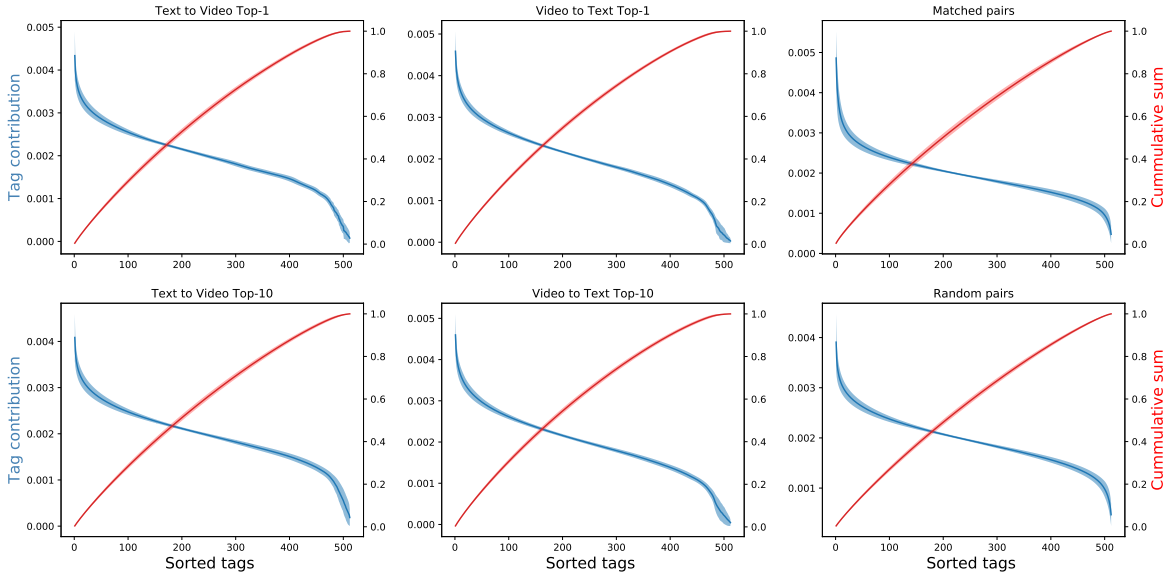


Figure A.4: Contribution individuelle et cumulative (moyenne $\pm$ écart-type), des étiquettes par valeur de contribution décroissante

La figure A.4 présente la moyenne et les écarts types des contributions individuelles et cumulatives des balises (représentées par $w_k(v,s)$ et $c_k(v,s)$) dans différentes paires de vidéos et de légendes (Top-1, Top-10, appariées et aléatoires) pour les tâches TTV et VTT. Les résultats montrent que pour des paires similaires, les premières balises ont des causalités individuelles et cumulatives minimales, ce qui indique que les explications visuelles reposent principalement sur une petite partie des balises, avec seulement environ 4% d'influence pour une similarité basée sur le concept uniquement et 1,6% si l'on considère la similarité dans une approche hybride. Alors que la plupart des balises ont un impact significatif sur les décisions de similarité et de classement, le fait de n'inclure que les quelques balises initiales dégrade considérablement les performances de recherche, soulignant l'importance des termes au-delà des premières dizaines dans la distance de Jaccard.

A.4.3 Amélioration de la causalité

Nous avons vu ci-dessus que la causalité dans les explications visuelles réelles est très faible (Figure A.4) parce que, au lieu d'avoir les poids causaux principalement distribués sur seulement quelques balises comme on pourrait s'y attendre si seules les balises pertinentes étaient détectées avec des scores significatifs, les poids causaux sont tellement répartis sur toutes les balises disponibles avec une "probabilité" moyenne d'environ 0,4. Cela signifie qu'en moyenne, environ 200 balises sur 512 sont détectées, ce qui n'est pas ce que l'on attend et est beaucoup plus important que la fréquence moyenne des balises dans les données d'apprentissage.

Afin d'améliorer la causalité des premières étiquettes, nous proposons de modifier les scores de détection en leur appliquant une fonction de sorte que le poids de la causalité soit plus concentré sur les premières étiquettes. Il y a plusieurs façons de procéder. Tout d'abord, en considérant les probabilités de balises utilisées dans la similarité de Jaccard (équation (A.1)), le simple fait d'appliquer une transformation de puissance avec un exposant p supérieur à 1 augmente automatiquement les poids relatifs des premiers termes. Deuxièmement, les probabilités de marquage $g(v)$ ou $g(s)$ sont obtenues en appliquant une fonction sigmoïde aux scores de détection "bruts" $h(v)$ ou $h(s)$; nous pouvons alors appliquer un biais b (décalage) et/ou un gain a (échelle) à ces scores bruts avant d'appliquer la fonction sigmoïde, en effectuant une sorte de normalisation de Platt [173], en corrigeant éventuellement l'influence de la perte d'extraction dans la calibration de la classification. Combinaison de transformations, nous remplaçons $g(x)_i = \sigma(h(x)_i)$ par:

$$\left(g_{(a,b,p)}(x)\right)_i = (\sigma(a(h(x)_i - b)))^p \tag{A.3}$$

avec σ étant la fonction sigmoïde (expit) et x étant soit un échantillon vidéo v soit un échantillon de texte s. La fonction originale correspond à $(a,b,p) = (1,0,1)$.

De même, afin d'améliorer la causalité des premières balises avec la similarité cosinus (equation A.1), nous remplaçons $h(x)_i$ par :

$$\left((h_{(a,b,p)}(x)\right)_i = (a(h(x)_i - b))^p \tag{A.4}$$

La principale différence avec Jaccard est que la transformée sigmoïde n'est pas utilisée pour la similarité en cosinus. Encore une fois, la fonction originale correspond à $(a,b,p) = (1,0,1)$ mais on peut noter que, en tant que facteur d'échelle, le paramètre a n'a pas d'effet dans la similarité en cosinus, qui est liée à un angle entre les vecteurs. Nous garderons donc $a = 1$ dans ce cas.

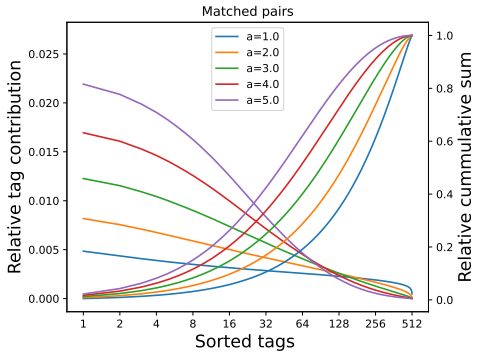


Figure A.5: Causalité par tag (courbe décroissante) et cumularive (courbe croissante) pour plusieurs valeurs d'échelle a

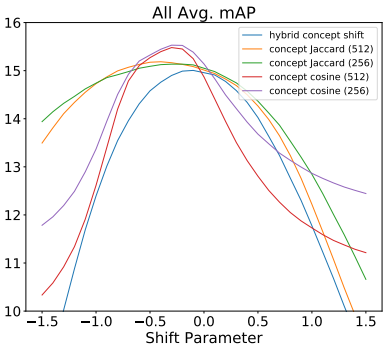


Figure A.6: Evolution globale de ma mAP suivant le paramètre de décalage (pour l'échelle optimale) pour cinq variantes de systèmes

A.4.3.1 Impact sur la causalité

L'impact des paramètres de causalité, notamment a, b et p, montre que des valeurs plus élevées de ces paramètres entraînent une causalité accrue, comme l'illustre la figure A.5. Comme nous nous intéressons à des valeurs aussi élevées que possible pour la causalité à quelques dizaines d'étiquettes, pour tous les systèmes, nous devrions utiliser des valeurs aussi élevées que possible pour les paramètres p, a (le cas échéant) et b.

A.4.3.2 Impact sur la précision moyenne

Le choix de valeurs aussi élevées que possible pour les paramètres p, a et b est susceptible d'avoir un impact négatif sur la précision de la recherche. La figure A.6 montre l'évolution de la mAP globale en fonction du paramètre b (shift) des équations A.3 et A.4. La valeur optimale donne généralement une légère amélioration des performances par rapport à la ligne de base, parfois significative. En ce qui concerne les paramètres p et a (le cas échéant), la valeur optimale est significativement plus élevée que la valeur de base, ce qui indique qu'il est possible d'avoir un gain simultanément sur la causalité et sur la précision. A l'inverse, la valeur optimale de la précision pour le paramètre b correspond à une valeur inférieure à la ligne de base, de sorte que l'on perd sur un critère si l'on optimise sur l'autre.

A.4.3.3 Optimisation conjointe

Les utilisateurs ne veulent généralement pas sacrifier la qualité du système au profit de la l'explicabilité. Pour trouver un compromis entre causalité et précision, une stratégie d'optimisation conjointe est employée. Cette approche prend en compte les paramètres p, a et b et leur impact sur la causalité et la précision. Nous n'optimisons pas conjointement les paramètres p et a car ils ont un effet similaire. Les valeurs optimales des paramètres sont sélectionnées sur la base de l'ensemble de validation, et les causalités et les précisions sont mesurées sur l'ensemble de test. Nous avons également vérifié que les valeurs optimales sont

Training	inference	*p*	*a*	*b*	C@10	C@30	mAP	SumR
1536d+512d hyb.	original				1.6	4.0	15.8	**210.2**
512-d (hyb. tr.)	original				3.9	10.0	13.7	191.1
	improved	1.00	2.7	0.0	10.9	25.5	15.0	**203.0**
512-d Jaccard	original				4.0	10.0	14.5	194.8
	improved	1.00	2.9	0.0	16.0	29.7	15.0	198.7
256-d Jaccard	original				8.2	19.6	14.7	193.2
	improved	1.00	1.8	0.0	32.0	51.8	15.3	**200.8**
512-d cosine	original				10.4	23.5	14.9	199.5
	improved	1.07	n/a	-0.25	15.6	31.3	15.5	207.0
256-d cosine	original				17.4	37.8	15.1	201.0
	improved	0.98	n/a	-0.24	22.3	44.1	15.5	**206.7**

Table A.3: Résultats sur MSR-VTT en moyenne. Ensemble de test tiré de [13]

assez proches sur l'ensemble de validation et sur l'ensemble de test. Les résultats montrent que cette stratégie d'optimisation conjointe peut aider les utilisateurs à affiner leurs modèles pour améliorer les explications visuelles tout en maintenant une précision de recherche compétitive, comme le montre le tableau A.3. Tous les modèles avec différentes dimensions d'espace conceptuel et différents paramètres (c'est-à-dire Jaccard et cosinus) dans le tableau A.3 peuvent conduire à une amélioration significative de la causalité sur les premières étiquettes ou dizaines d'étiquettes sans sacrifier la précision d'extraction ou avec même une légère augmentation de la précision (différence entre l'original et l'amélioré dans chaque ligne), sauf dans la première étape considérée qui est d'abandonner l'utilisation de l'espace purement latent dans l'étape d'extraction (les deux premières lignes).

A.5 Conclusion and Perspectives

Dans cette thèse, nous nous sommes attaqués aux défis des systèmes de recherche multimodale explicables, en particulier dans le contexte de la recherche vidéotextuelle. Notre recherche a contribué de manière significative à l'amélioration de la compréhension et de la performance de ces systèmes. Nous avons étendu le modèle de double encodage en incorporant des balises Part-of-Speech (PoS), améliorant ainsi sa capacité à interpréter les requêtes textuelles et à catégoriser le contenu en concepts visuels. En outre, nous avons effectué une analyse approfondie des modèles à double espace, révélant la complémentarité entre les espaces latents et conceptuels et ouvrant la voie à des modèles plus efficaces. En outre, nous avons exploré des techniques d'inférence causale afin de fournir des explications significatives pour les résultats de recherche, en obtenant une causalité plus élevée sans sacrifier la précision. Ces contributions ont de vastes implications, allant de l'amélioration de la précision du système et de la confiance des utilisateurs à l'équilibre entre causalité et précision dans les applications du monde réel.