Explainable Artificial Intelligence approaches for Image Captioning

Sofiane Elguendouze

UNIVERSITÉ D'ORLÉANS

ÉCOLE DOCTORALE Mathématiques, Informatique, Physique Théorique et Ingénierie des Systèmes

EA 4022 - LIFO

THÈSE présentée par :

Sofiane ELGUENDOUZE

soutenue publiquement le : **11 Janvier 2024**

pour obtenir le grade de : **Docteur de l'Université d'Orléans**

Discipline/ Spécialité : **Informatique**

Explainable Artificial Intelligence approaches for Image Captioning

Approches d'Intelligence Artificielle eXplicables pour le Sous-titrage d'Images

THÈSE dirigée par :

M. PEREIRA DE SOUTO Marcílio	Professeur des Universités, Université d'Orléans
M. HAFIANE Adel	Maître de conférences - HDR, INSA-CVL
Mme. LEFEUVRE-HALFTERMEYER Anaïs	Maître de conférences, Université d'Orléans

RAPPORTEURS :

M. LAGNIEZ Jean-Marie	Professeur des Universités, Université d'Artois
M. HERBIN Stéphane	Directeur de recherche, ONERA

EXAMINATEUR :

M. MAUDET Nicolas (Président du jury)	Professeur des universités, Sorbonne Université

UNIVERSITY OF ORLEANS

DOCTORAL SCHOOL Mathematics, Computer Science, Theoretical Physics, and Systems Engineering

Fundamental Informatics Laboratory of Orléans - LIFO

DISSERTATION prepared by

Sofiane Elguendouze

defended publicly on **January 11th, 2024**

for the purpose of obtaining the degree of

Doctor of Philosophy at the University of Orleans in **Computer Science**

Explainable Artificial Intelligence approaches for Image Captioning

Doctoral committee:

Mr. PEREIRA DE SOUTO Marcílio	Full Professor, University of Orleans
Mr. HAFIANE Adel	Associate Professor, INSA-CVL
Mrs. LEFEUVRE-HALFTERMEYER Anaïs	Assistant Professor, University of Orleans

REPORTERS:

Mr. LAGNIEZ Jean-Marie	Full Professor, University of Artois
Mr. HERBIN Stéphane	Senior Research Fellow, ONERA

EXAMINER:

Mr. MAUDET Nicolas (President of the jury)	Full Professor, Sorbonne University

Declaration

« In the presence of my peers, with the completion of my doctorate in Computer Science, in my quest for knowledge, I have carried out demanding research, demonstrated intellectual rigour, ethical reflection, and respect for the principles of research integrity. As I pursue my professional career, whatever my chosen field, I pledge, to the greatest of my ability, to continue to maintain integrity in my relationship to knowledge, in my methods, and in my results. »

En français :
« En présence de mes pairs, parvenu à l'issue de mon doctorat en Informatique, et ayant ainsi pratiqué, dans ma quête du savoir, l'exercice d'une recherche scientifique exigeante, en cultivant la rigueur intellectuelle, la réflexivité éthique et dans le respect des principes de l'intégrité scientifique, je m'engage, pour ce qui dépendra de moi, dans la suite de ma carrière professionnelle quel qu'en soit le secteur ou le domaine d'activité, à maintenir une conduite intègre dans mon rapport au savoir, mes méthodes et mes résultats. »

Acknowledgments

I express my deepest gratitude to my family, especially my parents, siblings, and my wife, for their unwavering support. I dedicate this achievement to my late grandfather, who passed away shortly before my defense, and had eagerly awaited this moment. Special appreciation goes to my friend Iheb and all my friends who have been a source of encouragement.

I would like to thank my supervisors Dr. Lefeuvre-Halftermeyer Anaïs, Prof. Pereira De Souto Marcílio, and Dr. Hafiane Adel for their invaluable guidance and continuous support throughout my doctoral journey. I am grateful to the jury members for their willingness to review and evaluate this work.

I am deeply thankful for the connections forged within the walls of LIFO, especially to the doctoral students and teaching colleagues whose engaging discussions and social gatherings added immense value to my academic journey.

Lastly, my gratitude extends to the French National Research Agency (ANR) and the University of Orléans for providing the grant that made this thesis possible. I wish also to thank the CaScIModOT federation for making available the computing resources of the regional parallel computing center.

Acronyms

AI Artificial Intelligence. 2

BU Bottom-Up. 8

CNNs Convolutional Neural Networks. 1

DL Deep Learning. 1

DNN Deep Neural Network. 18

IC Image Captioning. 1

LIME Local Interpretable Model-Agnostic Explanations. 6

LRP Layer-wise Relevance Propagation. 6

LSTM Long Short-Term Memory. 2

ML Machine Learning. 1

NLP Natural Language Processing. 1

RCNN Region-based Convolutional Neural Network. 8

RNN Recurrent Neural Network. 9

XAI Explainable Artificial Intelligence. 2

XIC Explainable Image Captioning. 35

Contents

List of Figures

List of Tables

List of Algorithms

1 Introduction

Over the past few years, Machine Learning (ML) architectures have witnessed an expansion in terms of usage and performance, resulting in a remarkable increase in complexity, particularly in Deep Learning (DL) approaches such as Convolutional Neural Networks (CNNs) and Transformers. The high complexity of these models has raised interpretability problems that prevent humans from understanding the reasoning behind the decisions they make, whether correct or erroneous. This problem is most acute in sensitive domains, where the cost of a wrong prediction or lack of understanding can amount to a human life, or put its safety at risk. Here, we are talking about defense, healthcare, autonomous driving, and so on.

Inspired by recent advances in the field of machine translation, researchers have increasingly explored the combination of Computer Vision and Natural Language Processing (NLP), giving rise to a thriving field of study at the intersection of these two fields, known as "Vision-to-Language". A specific facet of this field, known as Image Captioning (IC), has received particular attention. IC aims to generate accurate and representative textual descriptions of image content. For instance, in the context of biomedical imaging, IC has gained considerable momentum recently due to its remarkable capacities in assisting and accelerating the diagnosis processes [Pavlopoulos et al., 2019]. It has found applications in radiology reporting, where it aids in the interpretation of medical images, facilitating diagnostic and treatment decisions. Additionally, IC also enables improved communication among medical professionals and between professionals and patients, enhancing the exchange of medical information. Another compelling domain where IC demonstrates its utility is in natural disaster management [Klerings et al., 2019]. By augmenting disaster-scene images with descriptive captions, IC systems help structuralize the imagery content. This enrichment significantly enhances the effectiveness of image search and query, enabling emergency services to react rapidly in critical situations. Moreover, autonomous driving is one of the areas in which IC plays a decisive role [Mori et al., 2021]. It equips both autonomous vehicles and drivers with a deeper understanding of their surroundings. By

generating descriptive captions for images captured by on-board cameras, IC enables these vehicles to identify and describe objects including obstacles, pedestrians, traffic signs, etc. It also provides information on road and weather conditions, enhancing decision-making for safe navigation. IC is not limited to the vehicle alone, it also extends to driver monitoring in semi-autonomous (Level 2) vehicles, where the driver may be asked to take control of the vehicle when necessary, namely in ambiguous or dangerous situations.

Most of these IC architectures adhere to the Encoder-Decoder framework, leveraging deep learning modules like CNNs and Long Short-Term Memory (LSTM). Whatever the particularities of each model used in IC, the inner workings of the encoding and decoding components are often treated as black boxes, as their operating mechanisms are hidden from the end user and do not provide any explanations as to how their decisions were reached. In the context of the above-mentioned critical and sensitive domains, the need for an explanation becomes more obvious, as an erroneous prediction or ambiguity, however small, can have far-reaching consequences. Detecting a pathology from a machine-generated X-ray description without knowing the explicit reasons for its identification will necessarily lack public adoption and trust, and may therefore expose the patient's life to high risks in the event of errors. The actions of autonomous vehicles must be accompanied by explanations in order to identify safety-critical situations and mitigate risks, a measure through which accidents like Uber's fatal crash[1] in Tempe, Arizona (2018) [Daisuke, 2018] could have been avoided.

Explainable Artificial Intelligence (XAI) is an evolving field aimed at making Artificial Intelligence (AI) systems more transparent and understandable to humans [Adadi and Berrada, 2018], enabling users to comprehend how and why these systems reach specific results. Despite the growing interest in XAI, the realm of explainability for IC models remains relatively underexplored. This is not due to underestimating its importance but rather stems from the inherent complexity that characterizes this field. Only a few studies have therefore been proposed in the literature, most of which primarily seek to identify causal relations that link model outputs and inputs. For instance, in their pursuit of multimodal explanations[2], [Sun et al., 2022] have attempted to trace the captioning model decisions back to both vision-level and

[1] In March 2018, an Uber self-driving car struck and killed a pedestrian in Arizona who was walking her bicycle across a multi-lane road at night outside a designated crosswalk. The vehicle's system detected the pedestrian about six seconds before impact, classified her first as an unknown object, then as a vehicle, and finally as a bicycle with an unknown trajectory.

[2] By multi-modal explanations we refer to those involving different modalities of the original black box model.

language-level explanations. In the same fashion, in the context of medical image captioning, [Beddiar and Oussalah, 2023] have designed an explainable module that relates generated words to image regions. Additionally, it emphasizes word importance, highlighting the words that significantly influenced the encoder when calculating the word embedding of the target/current word. However, when it comes to multi-modal DL models, these explanation methods tend not to explicitly outline the roles of individual modalities in predicting the target outcome, such as a word in the caption for IC models, or even the weight of the explanations relative to the modality from which they result. Many of these methods share the same limitations as conventional XAI approaches, notably lacking to consider the influence of architecture components/modalities on the decision process and, therefore, on the explanatory information subsequently obtained. Merely establishing a relevance-based connection between model output and input data or intermediate representations, without emphasizing the relevance of each component, may inappropriately suggest uniform explanatory power across all modalities. Such a theory may prove insufficient to achieve a comprehensive understanding of the model's behavior. Moreover, the challenge also lies in the intricate nature of multimodal systems compared to other deep learning-based methods. This complexity stems mainly from the need to accurately integrate and interpret heterogeneous data from various modalities while carefully managing cross-modal dependencies.

If we assume that a specific modality should play a crucial role in generating explanations for IC models, it opens up a wide range of possibilities for exploring different approaches to extract and represent explanation elements from that modality. However, an important question that has often been overlooked is whether the choice of a specific approach can have a significant impact on the quality of the explanations. While there are numerous techniques available for generating explanations, ranging from simpler surrogate methods (e.g. LIME[3][Ribeiro et al., 2016]) to more complex and time-consuming approaches like back-propagation-based methods (e.g. LRP[4] [Bach et al., 2015]), it is not clear whether the use of more computationally intensive methods necessarily leads to better quality explanations. Moreover, in the general context of XAI, it is essential to consider the trade-off between the computational cost of generating explanations and the resulting explanation quality. While time-consuming methods can, in certain scenarios, offer a more detailed and fine-grained understanding of the model's behavior, it is worth noting that their granularity and depth of view might occasionally align with that of simpler techniques.

[3]LIME (Local Interpretable Model-Agnostic Explanations): a Post-hoc Model-agnostic technique.
[4]LRP (Layer-wise Relevance Propagation): a Post-hoc Model-agnostic technique.

It is also essential to recognize that such elaborate methods often require considerable computational resources and may be impractical for real-time applications or use with large-scale datasets. To determine the impact of various approaches on the quality of explanations, it becomes imperative to engage in comprehensive empirical evaluations and comparative studies that can systematically assess the relationship between the complexity of the method and the enhancement it brings to the overall understanding.

There are also ongoing concerns among deep learning researchers regarding the lack of visibility and comprehensive understanding of the representation/latent space. This space is considered crucial as it embodies essential information within the model [Rudin et al., 2022]. The data undergoes complex transformations and is ultimately transformed into a latent representation that captures key features and patterns. However, the nature and organization of this representation space remain somewhat elusive. Despite numerous efforts to explore this space, including the use of probing techniques [Ilinykh and Dobnik, 2021, Yu et al., 2022], we have yet to reach a definitive understanding of its underlying structure and the specific meaning associated with each element within the representation space embeddings. This concern is particularly more pronounced in the field of computer vision, where comprehending how models perceive and interpret visual information is a persistent challenge. For instance, it is essential to comprehend how models assign importance to salient objects, foregrounding them to exert more influence on predictions. This issue arises in various computer vision tasks, such as image classification, image captioning, and visual question answering.

1.1 Research questions and contributions

Within this specific context of explainability in IC, we have identified several key research questions that will guide our investigation. These research questions aim to address the aforementioned problems and could be listed as follows:

➔ **RQ 1: How different components in the representation space of IC architecture contribute to the overall understanding of the model's decision process?**

By analyzing the influence of different components or modalities within the representation/latent space **(Chapter 3)**, we can gain insights into their relative importance to the prediction. This analysis involves investigating how variations in specific components or modalities impact the subsequently generated captions. By considering

the varying strengths and objectives of different modalities, we would be able to uncover the interplay between visual and linguistic information and gain a deeper understanding of how the model combines and integrates these components to generate captions. Such knowledge is crucial for developing more interpretable and reliable image captioning models. Furthermore, it aids in pinpointing the pivotal components within the architecture from which subsequent explanations should be derived.

➔ **RQ 2: Does the scope/functioning of an explanation method for IC models have any effect on the quality of the explanations it provides? Furthermore, is there a more effective manner to evaluate the quality of explanations, that is closely aligned with the representation space?**
The scope of an explanation method refers to the extent to which the explanation method explores and captures the underlying factors. It encompasses considerations such as the range of elements covered by the explanation process (local instances or global behavior etc.). When comparing explanations generated by different methods, considering the scope and the functioning (surrogate vs back-propagation) may help identify potential causal relationships between these elements and the quality of the explanations provided **(Chapter 4)**. Additionally, the evaluation technique used to assess explanation quality plays a crucial role, and leveraging the inherent strengths of the representation space may offer opportunities for enhancing this evaluation process. This approach, combined with the confrontation of methods with different scopes and functioning, allows for a more comprehensive comparison and analysis and can significantly enhance our understanding of the strengths and limitations of explainability techniques.

➔ **RQ 3: Does the latent space have a discernible manner in which visual concepts, such as object quantity, are represented and structured? do these concepts have an explicit impact on object saliency in the final prediction?**
The lack of a clear structure and interpretability in the representation space poses significant obstacles to fully comprehending the inner workings of deep learning models. By manipulating the latent representations **(Chapter 5)**, we attempted to uncover the model's sensitivity, in both latent space and outputs, to changes in the quantitative visual concept. This provides insights into how it leverages this information to make decisions and whether these concepts hold any impact on objects' saliency within a scene.

1.2 Thesis overview

The remaining content of this thesis is structured and presented in the following manner.

- **Chapter 2:** This chapter begins by presenting an overview of image captioning (IC), highlighting recent advancements in the field. It then proceeds to review the existing research on XAI while providing a brief history, discussing the existing terminology, presenting a taxonomic overview of methods, categorizing different methods summarizing recent progress, and posing the problems and challenges emanating from this domain. Finally, it offers a comprehensive review of explainability in IC, including the latest approaches and evaluation techniques. It also discusses the existing gaps in this field and emphasizes the need for further exploration, while providing more detailed insights into our contributions.

- **Chapter 3:** This chapter highlights the use of two fundamental concepts: latent space and the perturbation paradigm. It embarks on comprehensive discussions and provides essential definitions. It then introduces the standard captioning architecture that has been adopted throughout this research work. The focal point is dedicated to the detailed presentation of the component influence identification approach, accompanied by the presentation of the experimental protocol. Finally, it wraps up with an in-depth analysis and discussion of the results obtained, offering valuable insights and implications.

- **Chapter 4:** This chapter provides a formal exposition of two novel attribution explanation approaches, BU-LRP and BU-LIME. These approaches draw inspiration from Layer-wise Relevance Propagation (LRP) and Local Interpretable Model-Agnostic Explanations (LIME) approaches while leveraging the representation space. Additionally, it introduces our new evaluation framework, known as Latent Ablation. Subsequently, it conducts an exhaustive comparative analysis between the two proposed techniques and engages in a thorough discussion of the results.

- **Chapter 5:** This chapter presents a new explanatory approach enabling a deeper exploration of latent space. It introduces a probing methodology tailored for image captioning models, with a specific emphasis on the quantitative dimension of visual information and its impact on visual saliency. This methodology relies on input editing techniques to reveal hidden facets of model behavior. The chapter presents the formal definition of the approach, followed by a series of experiments and in-depth discussions of the results.

2 Literature review

Contents

This chapter first embarks on an exploration of the realm of image captioning (IC), starting with a thorough overview of the field. Recent developments and advancements in IC are discussed, shedding light on the current state of the art. Subsequently, it delves into a survey of existing research on Explainable Artificial Intelligence (XAI). Recent advances in the field are summarized, while concurrently categorizing the various approaches that have been employed thus far. Then it provides a broad review of explainability in IC. This includes an analysis of the latest explanation methods and evaluation techniques employed in this area. Notably, it goes beyond a surface-level review by engaging in a more explicit discussion of the gaps present in current research.

2.1 Image Captioning

2.1.1 General framework

Image captioning (IC) is the process of generating textual descriptions (captions) for images that faithfully reflect their content, such as biomedical image captioning or autonomous driving scene descriptions. This is one of the well-known Vision to Language tasks, in which computer vision techniques are used to understand and encode the visual information contained in the image, and natural language processing (NLP) techniques to translate the encoded information into a coherent caption. The general captioning framework could be summarized in the following key steps:

1. **Data collection and processing:** Gathering a large dataset is the first step toward designing any learning-based model. Images are collected, resized to fixed dimensions, and aligned with human-annotated reference captions. Captions are tokenized and processed to create a vocabulary of words, which defines the lexicon from which the words forming the output caption will be selected during inference.

2. **Feature extraction:** The input images are encoded into visual features (image embeddings) that encapsulate the visual content. This process is often referred to as "feature extraction" and, depending on the image encoder, this step results in either a single feature map when using a CNN-based encoder, or several Bottom-Up (BU) visual features when opting for a Region-based Convolutional Neural Network (RCNN)-based one. Feature extraction is often decoupled from the captioning model which enables the reuse of pre-trained CNN models (such as ResNet, VGG, Inception, and Faster-RCNN), which are trained on large-scale image classification tasks, without the need to train the entire image captioning model from scratch. These pre-trained models have demonstrated their ability to learn rich visual representations thanks to the large amount of labeled data on hand, which enables captioning models to benefit from knowledge transfer through transfer learning. Furthermore, as feature extraction is computationally expensive, the pre-extracted features significantly reduce the computational burden, as the image features are computed offline and stored separately. The pre-computed features enable generalization to many image-related downstream tasks, such as image classification and visual question answering. Although feature extraction is often considered a separate step, it is an integral part of the image captioning pipeline.

3. **Captioning model design and training:** Most of the captioning architectures are

designed under the Encoder-Decoder [Karpathy and Fei-Fei, 2015] framework, with typically a CNN as encoder and Recurrent Neural Network (RNN) as decoder, such as LSTM or GRU, or more recently, transformer [Vaswani et al., 2017]-based models like the Vision Transformer (ViT) [Dosovitskiy et al., 2021] or the Caption Transformer (CPTR) [Liu et al., 2021]. The pre-extracted visual features are transformed/translated at each time step by the decoder into textual data (words) constituting the output caption. An attention mechanism is often incorporated between the two previous components resulting in a more sophisticated architecture (Encoder-Attention-Decoder) [Xu et al., 2015]. Its main role is to guide the model at each decoding step to focus only on relevant information (region) within the image and selectively weigh different visual features based on the current context, which results in more accurate and context-aware captions. During training, the model takes in aligned (Visual Features-Caption) pairs among the train data split, generates a system caption, compares it to the ground-truth[1] caption using a suitable loss function, and updates the model parameters through backpropagation and optimization techniques.

4. **Caption inference and evaluation:** Once the captioning model is trained, it is used to infer the system caption for new images (e.g. test split), word by word until a maximum caption length is reached. The evaluation of the output caption quality consists of measuring the discrepancy between the predicted caption and the target (reference) caption. Several metrics have been proposed in the literature including BLEU [Papineni et al., 2002], CIDEr [Vedantam et al., 2015], SPICE [Anderson et al., 2016], ROUGE [Lin, 2004], and METEOR [Banerjee and Lavie, 2005]. Further details are given in Section 2.1.3.

2.1.2 Captioning architectures

Image captioning (IC), which draws inspiration from machine translation, initially adopted the Encoder-Decoder architecture, popularly used in translation tasks. Nevertheless, the rapid advancements in IC have led to the development of more sophisticated architectures. These architectures have evolved from incorporating attention mechanisms within the Encoder-Decoder framework to embracing transformer-based and reinforcement-based approaches.

[1]The terms "ground-truth" and "gold standard", both referring to reference data used for evaluation, are often used interchangeably, but they can have slightly different connotations depending on the specific task or context.

Encoder-Decoder: The Encoder-Decoder architecture is the most widely adopted framework in IC (Section 2.1.1). It consists of two main components: an encoder and a decoder. The encoder (typically a CNN-like architecture) processes the input image and generates intermediate latent representations, capturing visual features. These representations are then passed to the decoder (typically an RNN-like architecture), which transforms them into a sequence of words that form the output caption describing the image [Kiros et al., 2014].

Encoder-Attention-Decoder: This framework extends the Encoder-Decoder architecture by incorporating an attention mechanism that allows the model to focus on the most important parts of the image while generating captions and enabling visual and textual information to be better aligned. To this end, many attention mechanisms were proposed in the literature. These include: semantic attention [You et al., 2016], adaptive attention [Lu et al., 2017], X-linear attention [Pan et al., 2020] and attention on attention [Huang et al., 2019]. [Liu et al., 2022] designed a specialized captioning method for Chinese based on visual attention and topic modeling. By leveraging the strengths of visual attention to understand the details of the image on the one hand, and the non-negative matrix factorization (NMF) topic model to include topic textual information to guide the caption generation on the other hand, their method has proven to be effective in generating more diverse and accurate sentences. Driven by the lack of image information and the deviation of the generated captions from the main content of the image, [Liu et al., 2020] took captioning a step further by incorporating image labels from the convolutional network into the decoding model and proceeding with a dual attention mechanism consisting of visual attention on image features, and textual attention on the aforementioned textual labels, whose role is to increase information integrity in the generated caption.

Transformer and semi-Transformer: The transformer [Herdade et al., 2019] architecture, initially proposed for NLP, has revolutionized various natural language processing tasks and has recently gained attention in the field of image captioning (IC). In some recent works, there has been a shift towards incorporating transformer-based decoders in the traditional frameworks, replacing the conventional RNN-based decoders. These approaches, which we can qualify as "semi-transformer-based architectures," aim to take advantage of the benefits offered by transformers, such as their ability to capture long-range dependencies and enable efficient parallel processing. While early attempts at utilizing these architectures in IC did not yield significant improvements in caption quality,

more recent works, such as [Liu et al., 2021] have shown considerable enhancements by extending the transformer to both the vision and language levels.

Reinforcement Learning: It has been utilized in IC to optimize the caption generation process. Self-critical Sequence Training (SCST) is a prominent method proposed by [Rennie et al., 2017] that employs reinforcement learning to optimize the evaluation metric (such as CIDEr). The key idea is to use the predicted captions obtained during inference as a reference to normalize the rewards. By doing so, SCST helps mitigate the exposure to bias problem, where models are trained using ground-truth captions during training but exposed to their imperfect predictions during inference. This enables the model to learn from its mistakes and improve the quality of the generated captions. Reinforcement Learning with Monte Carlo Search (RL-MCTS), introduced by [Yao et al., 2017], is an approach that combines reinforcement learning and Monte Carlo tree search to enhance the captioning process. RL-MCTS iteratively builds a tree structure that represents different captioning actions (trajectories) and their corresponding rewards. By simulating multiple trajectories and back-propagating the rewards, RL-MCTS selects the most promising actions to generate high-quality captions.

Scene-Graph: Focusing on the relationships between objects detected in the image constitutes a different branch of research in IC, resulting in more detailed and fine-grained captions. The scene-graph representation, introduced by [Zellers et al., 2018], builds a structured semantic graph where nodes represent objects and edges represent predicates (relationships between pairs of objects or attributes like colors and adjectives, etc.). The model relies on a Graph Convolutional Network (GCN) that aggregates neighborhood information within the graph and enriches the representation with contextual information that binds objects and predicates. During training, the model is typically optimized using a combination of language generation loss and scene graph alignment loss. In the context of scene-graph captioning, [Zhong et al., 2020] proposed a method that utilizes semantic representation through scene-graph decomposition. This approach allows for a more comprehensive understanding of the relationships within the scene, leading to improved caption generation.

2.1.3 Evaluation of caption quality

Many works in the literature have focused on developing methods for evaluating texts produced by text generation tasks, such as machine translation (text-to-text), image

captioning (image-to-text), visual question answering ("text+image"-to-text), and the examples are even more numerous. However, the challenge for tasks that include a visual aspect is much greater, since the visual content plays a key role in determining the quality of the corresponding output, which is not the case in some NLP tasks where the evaluation is uni-modal (text-text), requiring only the reference text to determine the quality of the output. Some works [Hessel et al., 2021] have already discussed the existing approaches for IC evaluation and how different metrics differ in their functioning. From our perspective, the existing paradigms in evaluating captioning models could be categorized under the three following classes:

Reference-based evaluation: This approach involves comparing the system (generated) caption with one or more reference captions that are considered to be ground-truth. Various evaluation methods exist for this purpose, including n-gram matching and overlapping metrics like BLEU [Papineni et al., 2002] and ROUGE [Lin, 2004], n-gram weighting metrics like CIDEr [Vedantam et al., 2015], word-level alignment metrics like METEOR [Banerjee and Lavie, 2005], token embedding similarity evaluation using BERTScore [Zhang et al., 2020b], and text semantic-graph comparison like SPICE [Anderson et al., 2016]. However, some authors [Jiang et al., 2019] argue that these metrics may not fully capture the image content and advocate for the inclusion of image content in the evaluation process, considering the inherent ambiguity of natural language.

Reference+Image-based evaluation: To address the mentioned problem of considering image content in the evaluation process of image captioning, several metrics have been proposed. One such approach is TIGEr [Jiang et al., 2019], which learns a model to ground and match text captions (reference and candidates) to the image content. VIFIDEL [Madhyastha et al., 2019] measures the Word Mover's Distance (WMD) [Kusner et al., 2015] between the candidate caption and object labels (from the object detection) previously weighted based on the reference captions. LEIC [Cui et al., 2018] on the other hand, is a learning-based discriminative metric that distinguishes between human and machine-generated captions using a trained neural network model. It utilizes this trained model to evaluate the quality of candidate captions within their context, which is composed of the input image and the reference caption. [Wang et al., 2021] introduced FAIEr, a metric that utilizes scene graph embeddings for both reference captions and input images. By fusing and comparing these embeddings with the candidate scene graph, FAIEr incorporates image content into

the evaluation process. However, like other metrics in this category, the effectiveness of FAIEr relies on the availability of large-scale datasets with aligned pairs of images and reference captions, which may not always be readily accessible or straightforward to obtain.

Image-based evaluation: This type of metric, commonly known as "reference-free", evaluates the quality of generated captions without relying on reference captions for comparison. Instead, it assesses the quality of the generated caption solely based on the input image content. One example is CLIPScore, introduced by [Hessel et al., 2021], which computes the cosine similarity between the embeddings of the image and the candidate caption using the cross-modal retrieval model CLIP [Radford et al., 2021]. Another metric is UMIC [Lee et al., 2021], which utilizes contrastive learning to compare and discriminate ground-truth captions from synthetic negative ones. However, these metrics may face challenges such as biases inherited from pre-training data, poor correlation with human judgments, lack of interpretability, and difficulty in detecting subtle errors in captions [Sai et al., 2022, Ahmadi and Agrawal, 2023].

While each category of metrics brings its strengths and weaknesses, certain metrics have garnered broader adoption in IC. Below, we provide detailed insights into key evaluation metrics, each discussed separately:

1. BLEU (Bilingual evaluation understudy) [Papineni et al., 2002]: is a precision-based metric used for evaluating the quality of generated captions. It compares the n-grams (segments) of the predicted caption with those of the reference high-quality caption. The scores are then averaged to obtain a similarity score ranging from 0 (low similarity) to 1 (high similarity). While BLEU is computationally efficient, it has limitations when applied to long texts, and an increase in the BLEU score does not always indicate good overall text quality. BLEU as defined in the original paper is given by Equation 2.1, where c and r are the lengths of the candidate caption and reference captions (closest value to c), respectively. BP denotes the Brevity Penalty which penalizes overly short candidates. p_n refers to the precision of n-grams matching between the candidate and the reference, with a modification to handle cases where a candidate n-gram appears more times in the candidate text than in any of the reference texts. The global BLEU

score combines the precision scores for multiple n-gram lengths ranging from 1 to N.

$$BP = \{ \begin{array}{l} 1 \quad if \quad c > r, \\ \exp(1 - r/c) \quad if \quad c <= r \end{array} \} \tag{2.1}$$

$$BLEU = BP \ * \ exp(\sum_{n=1}^{N} log(p_n)) \qquad ; \qquad BLEU \in [0, 1] \tag{2.2}$$

2. ROUGE (Recall-Oriented Understudy for Gisting Evaluation) [Lin, 2004]: is a set of metrics commonly used for evaluating text summarization and machine translation. It measures the overlap of sequences, such as words and n-grams. ROUGE-N, for example, measures the n-gram overlap at various orders (unigram, bigram, trigram, etc.). ROUGE-L measures the longest matching sequence of words. Unlike BLEU, ROUGE focuses on the ability to retrieve important information from the reference caption. The formula of ROUGE-L is given by Equation 2.3. LCS refers to the longest common sequence between the candidate and the reference.

$$\begin{array}{c} prec = (LCS/c) \\ rec = (LCS/r) \end{array} \tag{2.3}$$

$$ROUGE - L = \frac{2 \ * \ prec \ * \ rec}{prec \ + \ rec} \quad ; \quad ROUGE - L \in [0, 1] \tag{2.4}$$

3. METEOR (Metric for Evaluation of Translation with Explicit Ordering) [Banerjee and Lavie, 2005]: is a metric that combines precision and recall by calculating their harmonic mean, with recall weighted more than precision. It incorporates explicit word-to-word alignment and takes into account stemming and synonymy matching using WordNet[2] [Miller, 1995], making it more effective than previous metrics. It puts a penalty if the word order is not respected, using the concept of chunks which represent groups of adjacent unigrams. METEOR is designed to correlate well with human judgments at the sentence level, unlike BLEU and ROUGE, which operate at the corpus level. The METEOR formula is given in Equation 2.5. METEOR's precision and recall involve matching and counting words that are common between the

[2]WordNet: is a large semantics-based lexical corpus for the English language grouping words into sets of synonyms

candidate and reference sentences, in the candidate and the reference, respectively. *pen* represents the penalty for word order differences, and α is a parameter that balances precision and recall, typically set to 0.5 for equal weighting.

$$METEOR = (1 - \alpha) * prec + \alpha * rec * (1 - Pen) \quad ; \quad METEOR \in [0, 1] \tag{2.5}$$

4. CIDEr (Consensus-based Image Description Evaluation) [Vedantam et al., 2015]: is a metric specifically proposed for image captioning. It measures the similarity between a candidate caption and a set of ground-truth captions. CIDEr first transforms the words in candidate and reference sentences to their stem or root form and represents them using sets of n-grams. It applies TF-IDF weighting to penalize frequently occurring n-grams, which are likely to be less informative. This metric aims to capture the visual relevance of the captions. Considering a set of M reference sentences C and a candidate sentence $\hat{C}$, $g^n(C_i)$ stands for the vector formed by the TF-IDF weighting for all n-grams of length n in the i^{th} reference sentence, and similarly for $g^n(\hat{C})$. N is the maximum length of the n-grams (typically $N = 4$). A variation of this metric with several modifications, known as CIDEr-D [Vedantam et al., 2015], has also been proposed to capture caption diversity, and a factor of 10 is added to make the CIDEr-D scores numerically similar to other metrics.

$$CIDEr_n(\hat{C}, C) = \frac{1}{M} \sum_{i=1}^{M} \frac{g^n(\hat{C}) \cdot g^n(C_i)}{\|g^n(\hat{C})\| \|g^n(C_i)\|} \tag{2.6}$$

$$CIDEr(\hat{C}, C) = \sum_{n=1}^{N} \frac{1}{N} CIDEr_n(\hat{C}, C) \quad ; \quad CIDEr \in [0, 1] \tag{2.7}$$

5. SPICE (Semantic Propositional Image Caption Evaluation) [Anderson et al., 2016]: is a novel approach that is based on comparing the semantic graph representations of the candidate and reference captions. It utilizes a scene-graph parser to extract semantic graphs and computes the F1 score by matching tuples, which represent semantic relations in the scene graph. By analyzing the components of scene graphs and the semantic propositions extracted from captions, SPICE can capture semantic similarity, including handling synonyms or related terms based on its internal knowledge. While SPICE correlates well with human judgments, it only considers semantic similarity and disregards the syntactic structure of the sentence. The exact formula for SPICE is quite complex and involves multiple steps, including parsing captions into semantic

propositions and extracting semantic triplets (subject-predicate-object). Here is a simplified representation of the SPICE metric equation:

$$SPICE = \frac{SemanticContentMatches}{TotalNumberofSemanticTriplets} \quad ; \quad SPICE \in [0,1] \tag{2.8}$$

6. LEIC (Learning to Evaluate Image Captioning) [Cui et al., 2018]: addresses the limitations of SPICE by proposing a learning-based discriminative evaluation metric. It trains a binary classifier to distinguish between human-written and machine-generated captions. The classifier is trained with high-level features extracted from the train-set images using ResNet pre-trained on ImageNet, and their language context vectors, i.e. the candidate and reference caption encodings generated using an LSTM encoder. During evaluation, the learned classifier is used to measure the quality (to classify) of the candidate caption, while also taking into account its context (the image feature). LEIC outperforms existing metrics and correlates better with human judgments, adapting to both long and short sentences and capturing syntactic quality.

7. BERTscore [Zhang et al., 2020b]: is a recent word-to-word matching metric that leverages contextual word embedding representations. It extracts embeddings from both the candidate and reference captions using a BERT pre-trained language model, such as the original BERT model [Devlin et al., 2019] and computes cosine similarity between them. The metric employs a greedy pick strategy to select the maximum similarity values. It also offers the option to apply weighting by inverse document frequency (IDF). In the context of IC, the IDF weighting operates on the corpus composed of all images' captions. BERTscore excels at matching paraphrases and handles order differences and long distances well. However, it runs the risk of under-penalizing reference words that are not accompanied by matching candidate words.

8. CLIPScore [Hessel et al., 2021]: is an evaluation metric that utilizes a pre-trained cross-modal model called CLIP [Radford et al., 2021], trained on a large dataset of image-caption pairs. It calculates the cosine similarity between the latent vectors obtained from the input image and the candidate caption. CLIPScore has demonstrated a stronger correlation with human judgments compared to traditional reference-based and reference+image-based metrics like CIDEr and LEIC. However, a potential drawback of this metric is that the CLIP model is trained on heterogeneous images, which may pose challenges in generalizing the embeddings to specialized or domain-specific datasets.

It is crucial to acknowledge that ongoing debates surround the validity of learning-based explanation generation methods and evaluation metrics, as is the case with the last three metrics mentioned above. In situations where explanations are obtained and/or assessed through AI methods, a fundamental question arises: to what extent do the results of these methods themselves require further explanation? This scenario can give rise to a complex cycle of "explainability on explainability". Moreover, metrics such as CLIPScore and BERTScore rely on pre-trained models like CLIP and BERT to evaluate other models. Consequently, any biases present in the original models may lead to biased evaluations. These metrics not only tend to favor the outputs of their underlying models but also exhibit biases towards outputs from models that resemble their own. While it is acknowledged that any system dependent on third-party resources may introduce a certain bias, the issue becomes more pronounced when evaluations are based on resources integrated into a model's architecture. For example, CLIPScore evaluation of a CLIP embeddings-based captioning model can introduce a significant bias. Several studies have addressed the limitations of learning-based evaluation methods, including works like [Deutsch et al., 2022, Otani et al., 2023].

2.2 Explainable Artificial Intelligence (XAI)

The increasing complexity of AI systems has given rise to a crucial challenge: the lack of interpretability, hindering our understanding of the reasons behind these systems' decisions. Since then, the AI community has been actively working on methods and techniques to strike a trade-off between the robustness of AI-based systems and their interpretability. This pursuit of providing explanations that enhance our understanding of decision-making is nowadays called Explainable Artificial Intelligence (XAI).

2.2.1 Brief history

The XAI paradigm traces back to many years ago. While it does not have a definitive starting point associated with a specific date or event, the need for explainability was commonly recognized as intelligent systems experienced a resurgence, following a period marked by relative stagnation such was the case for expert systems (the 1970s) and recommendation systems (the 2000s). Early intelligent systems attempted to mimic human expertise but were often unable to provide comprehensible explanations for their decisions although the concept of explanation has already emerged [Buchanan and Smith, 1988,

Moore and Swartout, 1988]. The growing sophistication of AI-based systems, particularly with the rise of complex models such as Deep Neural Network (DNN), has allowed for more predictive power. Yet this has not come without a cost; in fact, it has often been at the expense of the effectiveness of the model's explanation, leading to much sharper challenges and greater needs in terms of XAI.

The concept of explainability has been a longstanding concern in various fields, including AI, social sciences (philosophy, social and cognitive psychology, etc.), and human-computer interaction, as indicated by [Miller, 2019]. In his view, the scope of explainable AI is circumscribed to human-agent interaction problems, which lie at the intersection of the aforementioned fields.

2.2.2 Terminology

Before embarking on the definition of XAI, it is first necessary to circumscribe the terminology that will be used throughout this manuscript. The field of XAI has seen the introduction of numerous terms and related concepts, such as "Interpretability" and "Transparency" which often leads to their interchangeable usage. However, the lack of clear distinctions between these terms has caused confusion and poses challenges in establishing common ground. Although these terms seem semantically close to each other, there are some important differences. In [Guidotti et al., 2018], the authors define Interpretability as the extent to which the ML model and/or its predictions are understandable to humans. Similarly, in [Doshi-Velez and Kim, 2017], Interpretability is described as the ability to explain or to present in understandable terms to a human. On the other hand, authors in [Barredo Arrieta et al., 2020] suggest that Interpretability is a passive characteristic of a model, while Explainability is an active characteristic that involves any action taken by the model in order to reveal its internal functioning. In line with this, some studies such as [Adadi and Berrada, 2018] note that explainability and interpretability are closely related, with interpretability being more commonly used within the machine learning community. They also define transparency as the need to describe, inspect, and reproduce the mechanisms used by AI systems to make their decisions. A system is thus considered transparent if it has good interpretability. In an effort to clear up confusion, researchers such as [Clinciu and Hastie, 2019] conducted an analysis of the existing literature to gain a deeper understanding of the terminology used in XAI and ensure their consistent and accurate use within the field.

The term XAI was first employed in 2004 by [Van Lent et al., 2004] to address explainability issues for AI systems in military simulations and computer games. They defined XAI as the extension of an AI system so that it can elucidate its behavior to a given user, either during execution or afterward, and can take the form of a chain of reasoning that leads to the result obtained. However, due to the diverse backgrounds and disciplines of researchers involved in XAI research, there is currently no universally agreed-upon definition of explainability or XAI.

Drawing from our extensive research and stance on this matter, we put forth the subsequent definitions that will serve for the remainder of this manuscript. ***Explainability*** refers to the extent to which a decisional system can elucidate its decisions, behavior, and internal workings, either through self-explanations or with the help of an auxiliary module or system. ***Interpretability***, on the other hand, refers to the inherent self-explanatory power of a decisional system, and an interpretable model can be attributed to a model whose behavior and decisions are understandable. ***Explainable Artificial Intelligence (XAI)*** encompasses the set of methods and techniques aimed at providing insights into the inner workings of AI systems, facilitating a deeper understanding of their decisional processes. In the context of XAI, an ***explanation*** is a representation provided by an explanation technique or a self-interpretable model, allowing users to comprehend the factors contributing to a specific decision. An ***explanation element***, as the most elementary unit of an explanation, refers to a specific piece of information or evidence that, when combined, constitutes a coherent explanation. It can take various forms such as a logical rule or feature importance. Lastly, the term ***black-box*** designates a specific type of ML model characterized by a lack of interpretability and an inability to provide clear insights into their inner workings or mechanisms.

2.2.3 The need of XAI, motivations and application domains

Obviously, risks induced by AI-based systems such as biased decision-making and security vulnerabilities were among the earliest factors that paved the way for the recognition of the necessity/importance of explainability and opened the door to many leads of research. According to [Van Lent et al., 2004], the main challenge was the difficulty encountered by users in understanding the actions taken by AI systems. As these systems have become more complex and used across a wider spectrum, the challenges have become even more acute. Nowadays, the need for XAI arises from several considerations, such as but not limited to:

➔ **Trust:** Making AI-generated decisions trustworthy is essential for their widespread adoption.

➔ **Accountability:** In the realm of AI, ensuring accountability for decisions can be a complex endeavor. However, it remains imperative to establish clear responsibility especially when it comes to critical domains, such as healthcare and finance. XAI can provide insights that help attribute responsibility for the outcomes of AI systems, which fosters a framework where individuals or entities can be held accountable for the actions and impacts of AI systems.

➔ **Regulatory requirements:** Regulations such as the General Data Protection Regulation (GDPR)[3] emphasize the need to explain and the right to explanation. They provide regulators with powerful means to ensure compliance, carry out audits or inspections, and mitigate risks.

➔ **Fairness:** Bias and discriminatory patterns ingrained in AI systems due to biased data are a pressing concern. XAI methods may offer potent means to identify and mitigate biases and foster equity in those systems.

➔ **Ethical Considerations:** It enables the detection of potential ethical violations and facilitates the design of AI systems that adhere to ethical principles and respect societal values.

According to [Adadi and Berrada, 2018], there are four principal reasons that encompass all motivations for XAI that can be listed as follows:

➔ **Justification:** XAI aims to provide explanations to justify and elucidate reasons for specific outcomes, especially when they are unexpected. By having an auditable and provable way to understand the model's decisions, trust in the system can be established.

➔ **Control:** It helps identify errors and correct them before they occur, similar to operating in a debug mode. This is particularly relevant in scenarios like simulations for autonomous cars, where understanding the decision-making process can help improve safety and performance.

➔ **Improvement:** Explainable models are known to be easily improved, because the inner workings and logic are well understood by the user. With transparency into

[3]GDPR: adopted by the European Union in 2018 (https://gdpr.eu/)

the model's operations, it becomes easier to identify areas for enhancement and make informed modifications.

➔ **Discovery:** Making a model explainable facilitates the collection of new information and the acquisition of knowledge. For instance, in games like chess, an explainable model can provide insights into its decision process, enabling humans to learn new strategies and movements.

Over time, the need for XAI has been emphasized in numerous domains where the aforementioned factors (human understanding, accountability, fairness, etc.) play a critical role. In healthcare, XAI is gaining traction in critical applications such as medical diagnoses, treatment recommendations, and clinical research. It helps doctors understand pathologies and the rationale behind recommending specific treatments based on their relationship with patient history and characteristics. Another domain where XAI holds immense promise is autonomous driving. Here it can contribute by providing explanations for traffic accidents, helping to improve autonomous systems and ensure safety. In the legal domain, certain AI-based systems, such as COMPAS [Tan et al., 2017], have demonstrated biases in criminal risk assessment. Anti-discriminatory tools were therefore needed under the guise of explainable systems, aimed at promoting fairness and accountability in legal decision-making. Cybersecurity, finance, and military also represent areas where the need for XAI has been widely recognized.

Authors in [Doshi-Velez and Kim, 2017] have pointed out two situations where explainability may not be necessary. First, when incorrect decisions have minimal impact or consequences. Second, when the problem at hand has been sufficiently studied and validated in the real world, and the system's decisions inspire great trust despite its imperfections. It follows that, in scenarios where incorrect predictions lead to high-stakes consequences, interpretability becomes crucial, shifting the focus accordingly.

In light of all these considerations, it can be argued that the development of XAI methods plays a central role in building trust, reinforcing accountability, and mitigating potential biases. Also, taking into account the particularities of each consideration and depending on the overreaching goal of the explanation, the latter should be adapted to the corresponding target audience.

2.2.4 Reasoning in XAI

Advances in XAI research have led to the emergence of a wide range of reasoning types that cover almost all existing explainable approaches and techniques. While the exact categorization may vary, some of these types take common logical reasoning forms [Cau et al., 2020] such as Inductive, Abductive, and Deductive reasoning. Abductive reasoning is concerned with generating plausible explanations or hypotheses for observed outcomes. It operates by abducting a cause from an effect and a rule. Inductive reasoning, on the other hand, focuses on extracting general patterns or rules from specific data or observations. It involves inducing a rule from a cause and an effect. Lastly, deductive reasoning involves drawing specific conclusions from established rules, knowledge, or logical principles. It entails deducing an effect from a cause and a rule. Counterfactual and contrastive reasoning, although originating several decades ago in domains like philosophy and cognitive science, have gained significant momentum due to the growing need for explainability [Miller, 2019]. Both types of reasoning share the causal side of reasoning in seeking explanations but differ in their approach and purpose. Counterfactual explanations focus on understanding the impact of changing inputs on a particular decision or outcome, by providing alternative scenarios, while contrastive explanations focus on comparing different instances or situations to highlight the distinguishing factors that lead to different outcomes.

While there is no standard rule to privilege among those types of reasoning, the choice of a reasoning method may depend on the specific characteristics of the problem domain, the explainee (user) requirements, and the goals of the explanation. The question of how explainability can be achieved in AI goes beyond discussing existing reasoning types. Further details on the various existing approaches in XAI will be given in section 2.2.

2.2.5 Taxonomy of XAI methods

In response to the increasing need for explainability, especially for deep learning models, a wide range of explainability methods has been developed to provide clues explaining the decision-making process. While many works have surveyed the literature on XAI, such as [Tjoa and Guan, 2021, Ras et al., 2022, Burkart and Huber, 2021, Adadi and Berrada, 2018, Carvalho et al., 2019, Došilović et al., 2018], there exist various criteria by which, XAI methods are grouped and categorized, as depicted in figure 2.1 which highlights the general landscape of XAI techniques.

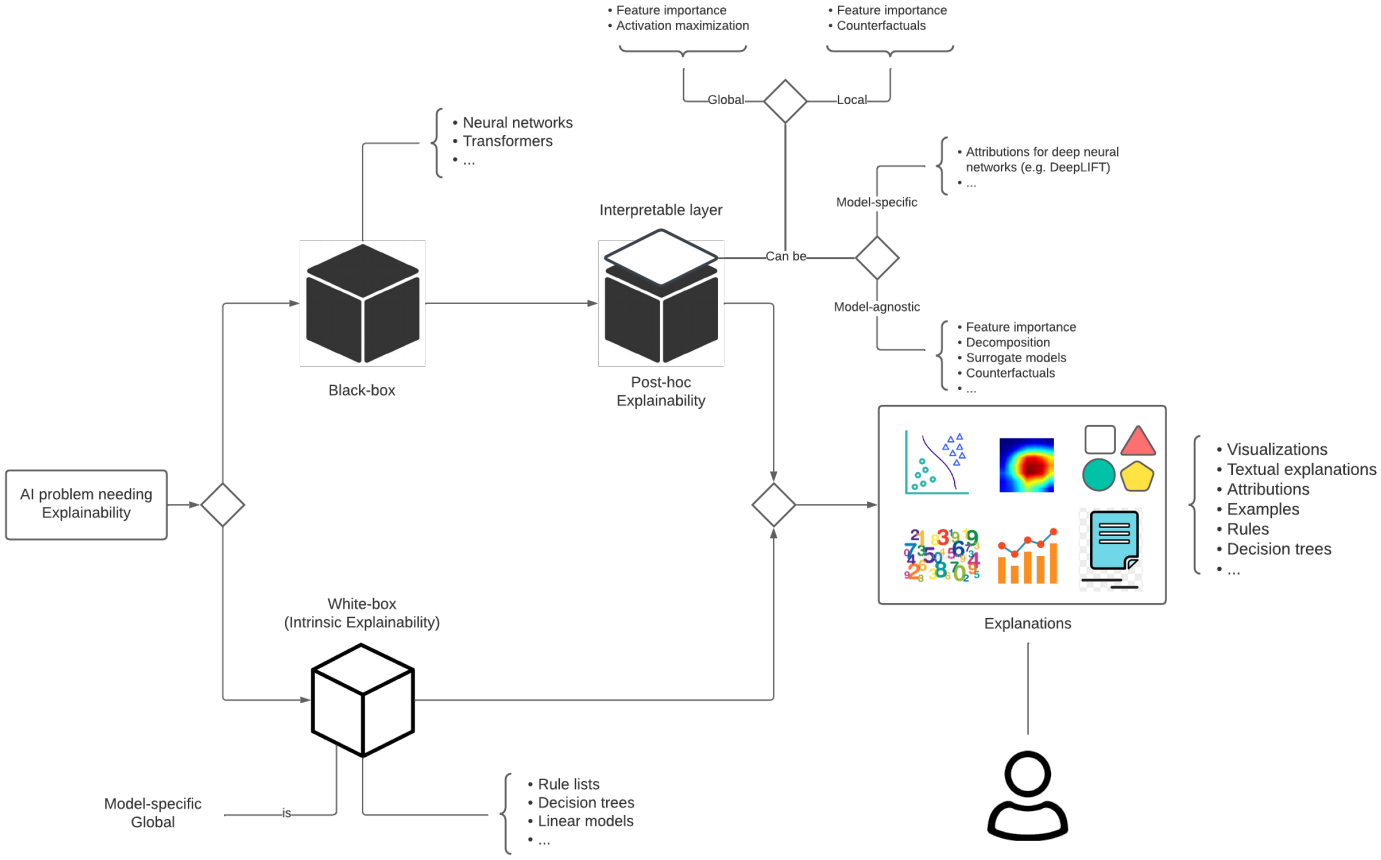


Figure 2.1: The XAI general landscape

Complexity: Explainability in AI can be approached in two main ways, depending on the complexity of the task and the nature of the data. Firstly, for less complex tasks or when the data is well-handled and annotated, ***interpretability by design*** or more commonly ***Intrinsic interpretability*** is a suitable strategy. This involves developing inherently interpretable models [Adadi and Berrada, 2018], also known as white boxes such as decision trees [Letham et al., 2015]. These methods provide explainability while maintaining good accuracy and relevance of results. As long as the model remains accurate for the task and utilizes a reasonably limited number of internal components, intrinsic interpretable models are sufficient. On the other hand, for more complex tasks and lightly processed data [Burkart and Huber, 2021], an alternative approach is to construct a black-box model with high accuracy, without considering any complexity constraints. Subsequently, separate techniques often referred to as ***Post-hoc explainability*** methods

can be employed, such as surrogate models [Ribeiro et al., 2016] and attributions [Sundararajan et al., 2017]. These techniques act as additional interpretable layers that overlay the original complex model. It should be noted that the choice between intrinsic and post-hoc methods may be influenced by factors other than the complexity of the task, such as the level of explainability we are seeking to achieve. Post-hoc methods can offer richer and more refined explanations which, in some cases, may make them prevail over intrinsic methods.

Scope: In terms of scope, explainability techniques fall into two main categories: Local methods and Global methods. However, the precise definition of each category within this scope-based perspective requires further discussion. According to [Adadi and Berrada, 2018], ***Local*** explainability refers to the set of methods that generate explanations for a specific single prediction [Lundberg and Lee, 2017]. Conversely, ***Global*** methods aim to unveil the entire logic of a black-box model, encompassing all possible outcomes [Nguyen et al., 2016]. It is worth noting that some techniques are capable of operating at both local and global levels [Bastani et al., 2017, Bach et al., 2015]. The distinction between local and global methods extends beyond the mere number of decisions for which explanations are sought. From our perspective, *Local* explainability consists of explaining individual decisions and targeted outputs using local information around the target instance, (see LIME [Ribeiro et al., 2016] in Section 2.2.6), offering insights into the immediate context and factors that influenced the model's decision. This approach aims to reveal only the logic associated with that particular decision. On the other hand, by adopting a *Global* perspective, the focus shifts to comprehending the overall functioning of the model, understanding the underlying patterns and relationships within all the data, and elucidating the model's entire behavior, including the knowledge acquired during training and utilized during inference to explain individual or multiple predictions, as offered by decision trees.

Dependency: We find the term "dependency" to be the most representative in distinguishing between two classes of explainability methods commonly referred to in the literature: ***Model-specific*** and ***Model-agnostic*** explainability. The first class includes methods that are specifically designed to explain the inner workings of a particular model architecture, making the intrinsic interpretable models inherently Model-specific. However, a drawback of such methods [Letham et al., 2015] is their limited applicability, as they can only be used with models for which they are specifically designed. In contrast, Model-agnostic methods are designed to provide explanations for any type of black-box

model, regardless of their internal structure or architecture [Casalicchio et al., 2019]. They offer a more flexible and generalizable approach to explainability.

The classification of explainability methods proposed by [Carvalho et al., 2019] extends the framework established by [Adadi and Berrada, 2018] by introducing two additional criteria:

Pre-Model vs. In-Model vs. Post-Model explainability (we propose the term **"chronological criterion"**). It reflects the point at which the explanation method is applied, distinguishing between three main stages. The Pre-Model explainability focuses on explaining the data used to train the model, while In-Model explainability aims to discern the internal workings of the model during its construction. Post-Model explainability involves explaining the model's outputs during inference.

Results (of Explanation Methods): This criterion groups explanation methods based on the nature of their outcomes. ***Feature summaries*** [Sundararajan et al., 2017] provide statistical information about the model's input features, such as their importance or relevance. ***Model internals*** are the results of intrinsic methods, such as the conditions of a decision tree [Letham et al., 2015] or the weights of a linear model, which provide insights into the inner workings of the model. Some techniques generate ***Data points***, often representative instances from the training dataset, to describe and interpret the model's decision process [Gurumoorthy et al., 2019]. Finally ***Surrogate intrinsically interpretable models (SIIM)*** are simpler models serving as interpretable representations or an approximation of the original black-box model [Bastani et al., 2017].

In table 2.1, we aim to extend the associations identified by [Carvalho et al., 2019] between the first three criteria to include the later ones. This extension enables a more complete and nuanced understanding of the relationships between different aspects of explainability methods.

Chronology	Dependency	Scope	Complexity	Results
Pre	NA	NA	L/G	Data Points
In	Intrinsic	Specific	L/G	Model internals
Post	Post-hoc	L/G	Agnostic	Feature summary, Data Points, SIIM

Table 2.1: Associations between classes of XAI techniques according to the five above-mentioned criteria. L/G denotes Local/Global methods.

From our perspective, complexity itself should not be considered as a criterion for categorizing explanation methods into Intrinsic and Post-hoc, as the very nature of an explanation method implies its operation at a post-hoc stage. Intrinsic interpretability, on the other hand, refers to the inherent explanatory power of an AI system, where the system itself is designed to be interpretable (Section 2.2.2). While complexity may influence the choice and implementation of explanation methods, it should not be used as a defining criterion for distinguishing between post-hoc explainability and interpretability by design. Therefore, it is important to differentiate between the categorization criteria of explanation methods and the nature of explainability.

The criterion based on the nature of explanations primarily serves to distinguish the different forms of explanation results (see Figure 2.1), often referred to as communication types of explanations (e.g. visualizations, textual descriptions, etc.) as discussed in [Burkart and Huber, 2021]. This criterion is also considered in other works, such as [Adadi and Berrada, 2018], to categorize explanation methods, particularly those belonging to the Model-agnostic class. While not comprehensive, the following provides an outline of these categories.

Visualizations: In this category, explanations take the specific form of visual representations that provide a visual understanding of how different parts of the input contribute to the model's decision, such as relevance/importance heatmaps [Sun et al., 2022, Selvaraju et al., 2017]. They are particularly effective in computer vision domains such as image classification, image captioning, and object detection. While some works categorize visualizations as a distinct form of explanation, they can also be considered as part of the broader category of attributions, where they offer an intuitive way to interpret and communicate the attribution values.

Attributions: Attribution values refer to importance/relevance scores assigned to input features or model's components. These scores indicate the degree to which each feature/component influences the model's decision. These include scene-graphs [Zellers et al., 2018], surrogate model weights [Ribeiro et al., 2016], attention and relevance heatmaps [Guo and Farrell, 2021], etc.

Textual explanations: Some explanation methods provide written descriptions that aim to convey the reasoning processes of the explained models in a human-understandable manner [Huk Park et al., 2017].

Examples: These stand apart from the other forms of explanations as they do not create the explanation in itself, but rather select instances from the dataset to elucidate the behavior of the ML model. These instances can serve different purposes: alternative scenarios that, if applied or observed, would have resulted in different model predictions (counterfactuals) [Jeanneret et al., 2023], or representative instances that capture the key characteristics of a specific decision and help us understand the factors driving the model's generalizations (prototypes) [Tan et al., 2020, Li et al., 2018].

Rule-based explanations: Refer to the use of symbolic descriptions capturing the knowledge learned by the ML model. Rule-based explanations can take various forms, including logical expressions, decision trees and rule lists [Letham et al., 2015, Zilke et al., 2016], etc. This form provides a human-readable way in which explanations can easily be understood and aligned with human reasoning, but may not capture the complexity of certain compound ML models.

2.2.6 Common explainability methods for DNNs

Prompted by trends in machine learning, the explainability of deep learning models has in turn taken the lead in XAI. The community has rapidly embarked on the development of explanatory methods for various deep learning-based models, and the number of such approaches continues to grow. According to [Ras et al., 2022], foundational explainability methods for DNN models can be grouped into three main classes, Visualization, Distillation, and Intrinsic. Visualization methods attempt to highlight features of the input that significantly affect the output, such as LRP [Bach et al., 2015] and Integrated Gradients [Sundararajan et al., 2017]. Distillation is a class of methods often known as "white-box" methods, which are supposed to approximate the functioning of the black-box model by a simpler one, such as LIME [Ribeiro et al., 2016] and Anchors [Ribeiro et al., 2018]. The last so-called Intrinsic class is similar to the one presented above in the previous section but has a more general scope. In addition to intrinsically interpretable models, there are also models that are capable of generating explanations along with the inference itself [Tavanaei, 2020]. Here are some of the most common XAI methods found in the literature.

LIME [Ribeiro et al., 2016]: is a perturbation-based approach that approximates the behavior of the black-box model for individual predictions by a simpler linear

model (surrogate) within a local neighborhood around the input instance. In the context of image classification [Schallner et al., 2020], LIME employs image segmentation to divide the image into super-pixels/regions. Perturbations are then applied to these regions to generate neighbor instances. The black-box model is queried with these perturbed instances, and the resulting predictions are used to train a linear model. The weight assigned to each region by the linear model indicates its relevance for a specific class decision. In terms of explanation results, LIME falls into the category of attribution-based methods, as it generates importance scores for input features. From the method-dependency perspective, LIME is regarded as a model-agnostic method.

LRP [Bach et al., 2015]: is a back-propagation-based method that, instead of gradients, back-propagates relevance scores to provide an explanation of how each input feature or neuron contributes to the model's prediction. It works by distributing relevance scores from the output layer to the input layer back along the architecture of the black-box model, following a set of rules that adhere to the conservation property. The conservation principle ensures that the sum of relevance scores at each layer is equal to the relevance score at the output layer. There are multiple variants of LRP each employing a different propagation rule such as ϵ-rule and β-rule, etc. LRP can be categorized as a global model-agnostic explanation method in the sense that it uses global information learned by the model during training and the one saved during inference to compute relevance scores. However, it can also be viewed as a local method since it provides explanations at the level of individual predictions.

SHAP: Originally proposed by [Lundberg and Lee, 2017], SHAP is an attribution-based method that assigns importance scores to input features. It is based on cooperative game theory that seeks to measure the contribution of each feature (the player) to the model's prediction (the payoff), taking into account the interaction between features and considering all their possible combinations (effect of including or excluding each feature from the input features). SHAP can be applied to various ML models, including complex architectures to explain individual predictions which makes it a local model-agnostic explanation method. One of the key strengths of SHAP is its ability to provide detailed explanations for individual predictions. However, it is important to note that recent research [Huang and Marques-Silva, 2023] has highlighted the potential limitations of SHAP. Specifically, it has been shown that SHAP can assign non-zero importance scores to features that may be irrelevant to the prediction. This raises concerns about the potential for misleading explanations when relying solely on SHAP values.

2.2.7 XAI problems and challenges

The major problems with XAI are deeply rooted in the subjective nature of the domain. This subjectivity mainly arises from the reliance on human interpretation and judgment, as humans play a crucial role in understanding and evaluating AI explanations. The same explanation can be perceived differently by different individuals, based on their cognitive biases, knowledge, and social expectations. Different users may have different expectations and requirements for what constitutes a good/satisfactory explanation [Miller, 2019], leading to variations in interpretation and trust. Addressing the problems of subjectivity in XAI requires careful consideration of the human factor, along with the development of simple and robust explanation methods. Efforts should be made to engage users in the design and evaluation of XAI systems to account for their diverse perspectives and requirements.

The definition of XAI and how it should be assessed has been the subject of various points of view. At present, a consensus on the precise definition of XAI is yet to be reached, which may pose further problems of subjectivity and hinder the development of XAI methods. However, ongoing debates and discussions within the research community are actively addressing this issue and striving to establish a more unified understanding of XAI. [Carvalho et al., 2019] in their review, highlight the problem of the lack of explainability/interpretability evaluation, particularly in the ML community where most efforts focus on improving interpretability in prediction tasks and maintaining the trade-off between explainability and predictive accuracy without developing measures and approaches for XAI evaluation. They also note that existing work in this domain primarily focuses on defining properties for explanations (e.g. accuracy, fidelity, consistency, stability, comprehensibility, Certainty, Novelty, Representativeness, etc. [Robnik-Šikonja and Bohanec, 2018]). Various indicators have also been proposed to measure these interpretability properties [Sundararajan et al., 2017, Honegger, 2018]. However, the practical application of these properties and measures in assessing the quality of explanations and the effectiveness of explanation methods remains largely unclear. The challenge lies in establishing concrete guidelines or frameworks that utilize these properties and measures to provide meaningful assessments of explanations in practice.

One fundamental issue that arises in many explanation methods for ML models is the lack of consideration for determining which components of the model are most suitable for generating explanations and how to represent them effectively. These methods often assume equal explanatory power for all model components [Sun et al., 2022] as long as the

explanation meets user expectations. However, this overlooks the possibility of imbalances in the importance of different explanation modalities. It is crucial to recognize that certain explanation modalities may inherently carry more weight or be more effective in conveying information than others. Neglecting the varying strengths and capabilities of different modalities can lead to overlooking valuable sources of explanation or placing excessive emphasis on the wrong modality. To address this issue, it is imperative to ensure that the selection of explanation modalities aligns with the objectives of the explanation and is tailored according to the design of the ML model. In cases where prioritizing one modality over another is not feasible, different modalities can excel at conveying different aspects of the model's behavior and, thus, be complementary.

Although some work [Wu and Song, 2019, Rudin et al., 2022] has pointed to the importance of latent space in increasing the interpretability of ML models, the exploration of this space remains largely untapped. Questions such as how a neural network-based model gradually learns concepts over layers could be addressed by disentangling the latent structures [Chen et al., 2020]. [Wu and Song, 2019] take a notable stride by defining model interpretability as the capability to methodically explore the latent space corresponding to a given task, while capturing the fundamental statistical insights within this latent space in a coherent and structured manner. It is evident that there exists considerable room for further investment of effort in this particular direction. Exploring the latent space, comprehending its dynamics, and devising methodologies to harness its interpretability potential hold much promise for advancing our understanding of ML models. All the issues and challenges mentioned above provide fertile ground for ongoing research, which continues to shape the landscape of XAI.

2.3 Explainability in Image Captioning (XIC)

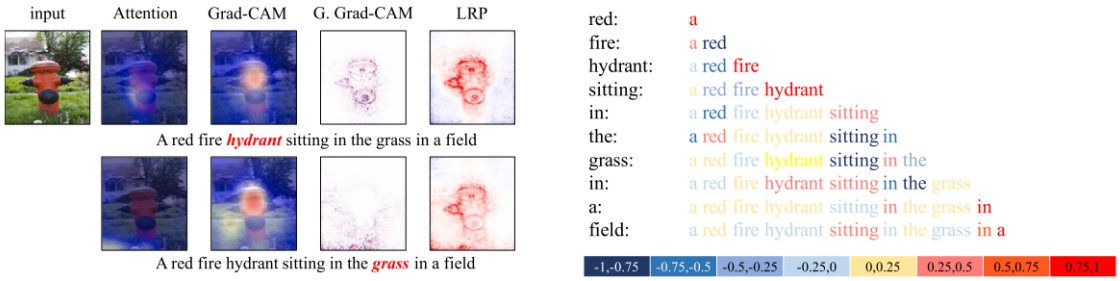
The increased awareness of the need to develop more accurate captioning models, mainly based on deep learning, has led to exponential progress in terms of complexity. However, the lack of interpretability remains a challenge, as it is difficult to understand why these models generate specific captions. The field of explainability in image captioning (XIC) aims to address this issue by providing insights into the inner functioning of captioning models. Most of the very few existing approaches in XIC adhere to the post-hoc explanation paradigm. These techniques mainly focus on answering questions such as "Which parts of the image contribute the most to the generated caption?" or even "How does each word in the caption relate to specific image regions?". In the following, we discuss the leading

works in this field of XIC.

The paper by [Han and Choi, 2018] presents an explainable approach to IC, providing a visual connection between a specific image region/object and the corresponding word in the generated caption. This connection is achieved through the utilization of an explanation module with a dual role. First, during the training phase, it serves as a loss function that evaluates the correlation between the words generated in the caption and the objects detected in the image. It then back-propagates the error, thereby enhancing the performance of the captioning model. Subsequently, during the inference step, the same module generates a weight matrix that assigns attention coefficients (importance scores) to regions within the input image according to their involvement in predicting each word of the output caption.

The authors in [Sahay et al., 2021] employed the LIME [Ribeiro et al., 2016] surrogate technique to approximate the behavior of the black-box captioning model by a simpler linear model for each data instance (image). The core idea behind this approach is to create a set of perturbed instances locally around a given image. These perturbed instances are generated using operations like blurring and blacking on the original input (the image pixels). The goal is to assess the impact of these perturbations on predictions, which are then integrated into the cost function of the linear model. During the training of this surrogate model, weights are assigned to various regions/patches that were perturbed in the input image. These weights are determined based on how effectively each region, through its presence or absence in the perturbed instances, influences the prediction of specific words in the caption. This approach essentially allows for the interpretation of the contribution of different parts of the input image to the captioning process.

In their study, the authors of [Sun et al., 2022] proposed several attribution-based explanation methods, including an adapted version of Layer-wise Relevance Propagation (LRP) [Bach et al., 2015], as well as Grad-CAM and Guided Grad-CAM [Selvaraju et al., 2017] which are based on Gradient-weighted Class Activation Mapping. The new LRP-based method proposed in their research generates pixel-wise visual explanations in the form of heatmaps, highlighting both supporting and opposing pixels for predicting a target word in the output caption (Figure 2.2a). Additionally, linguistic explanations are provided, indicating the contribution of each word in the generated caption to the prediction of the target word in the same caption (Figure 2.2b). By obtaining explanations at both the visual and linguistic levels, the authors claim that the dependencies between words and image regions become more explicit.

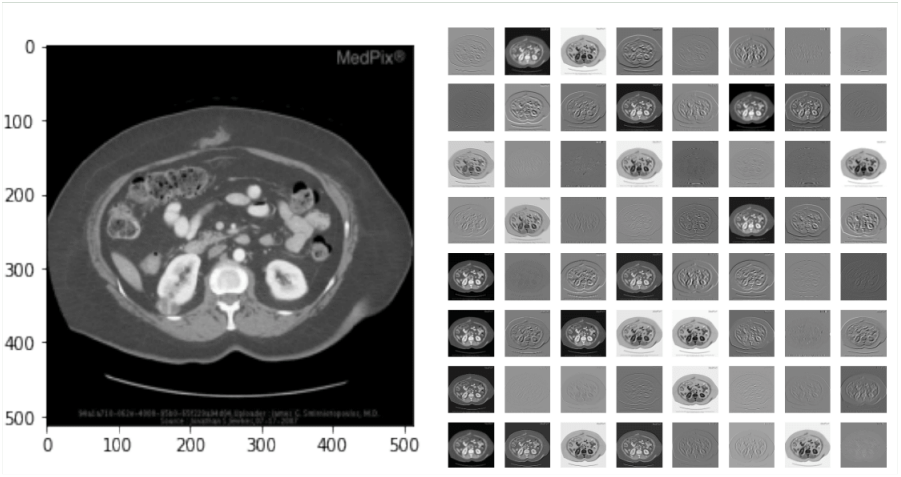


(a) image explanations (b) linguistic explanations of LRP

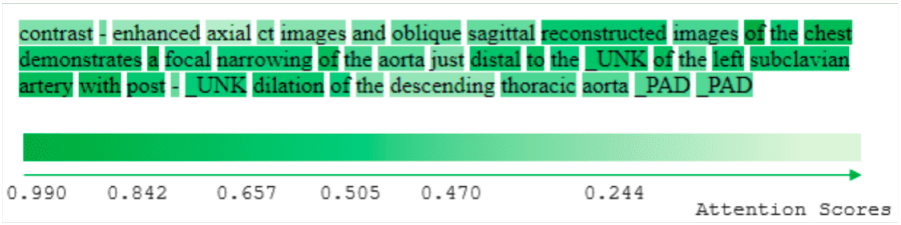
Figure 2.2: Image and linguistic explanations from [Sun et al., 2022].

In the medical application, [Beddiar and Oussalah, 2023] have designed an explanatory module within their captioning model by exploiting the attention scores generated by a self-attention mechanism, the encoder-decoder attention weights, and the convolutional layer feature maps from a pre-trained CNN. Visual features were extracted from images using a pre-trained Resnet model. The feature maps from the convolutional layers were visualized and exploited as perspective explanations, showcasing features that correspond to a specific prediction (Figure 2.3a). Additionally, the explanation module incorporated word importance visualizations by leveraging the weights of the self-attention module, which served to highlight the most important words in the generated caption to the computation of the word embedding (i.e. the prediction) of other words within the same output caption. These self-attention weights were computed intrinsically to the generated caption and were independent of the input image (Figure 2.3c). Furthermore, attention maps were also computed through the attention-based encoder-decoder. These maps were used to illustrate the regions of the image considered most important by the decoder when generating captions (Figure 2.3b).

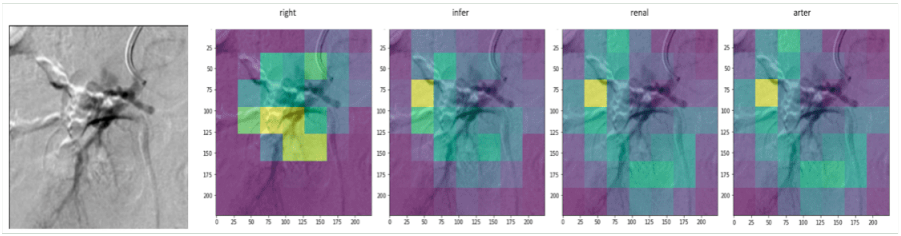
Continuing in the medical application, a study by [Wu et al., 2023] presents an approach using expert-defined keywords as interpretability boosters for retinal IC models. The main objective of their approach is to align human-comprehensible keywords with specific local image patches and guide the captioning model to accurately predict these keywords within the output caption. To achieve this, they first generate image features from a classical CNN, which are then fed into a multi-label classifier to predict relevant keywords (labels). These keywords are used in conjunction with expert-defined keywords, which may not always be available. The embeddings of these keywords serve as semantic features and are fused with the image content by an image-keyword attention-based encoder. The keyword-centric strategy employed in their approach proved to be effective in advanced



(a) Feature maps explanations



(b) Word importance **explanations**



(c) Attention maps **explanations**

Figure 2.3: Qualitative explanations from [Beddiar and Oussalah, 2023].

word sampling and the generation of precise and contextually meaningful captions for retinal images. However, a notable limitation of their approach is its dependence on the availability of expert-defined image keywords. Consequently, for datasets lacking such keywords, the performance of their method may be compromised. Additionally, it is essential to note that their approach primarily focuses on enhancing caption quality rather

than generating detailed explanations.

Regarding the utilization of Shapley values in XIC, there are limited works where these values have been explored. In a recent study by [DEWI et al., 2023], a Shapley-based approach was introduced. This approach involves segmenting images into superpixels along specific axes and subsequently attributing importance scores to these superpixels based on their relevance to predicting each word in the generated caption. Figure 2.4 represents the heatmaps involving image patches, generated through Shapley value explanations, corresponding to each output word in the predicted caption.

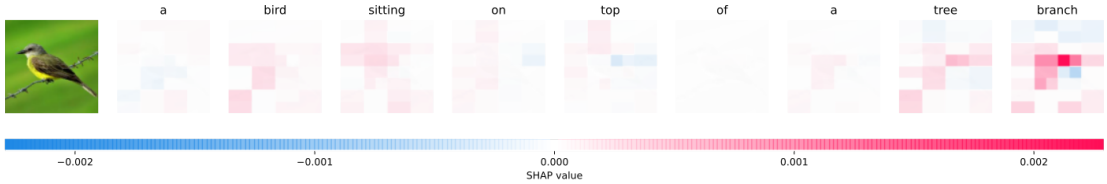


Figure 2.4: Qualitative explanations from [DEWI et al., 2023].

Although other recent explanation approaches for IC have been proposed, such as the PIC-XIC and iPIC-XIC frameworks in [Al-Shouha and Szűcs, 2023], these methods almost all point in the same direction. In their work, [Al-Shouha and Szűcs, 2023] express the target words in the output caption needing explainability as queries, and the generated explanations as answers to those queries. Their approach mainly relies on segmented patches (pixels) from the original input image to generate explanation proposals and measure the proposal's relevance by combining these patches to various uniform backgrounds constituting new test images that will further be passed to the captioning model to generate captions. They consider that an explanation proposal is good if the query (word to be explained) appears in any of the generated captions. Qualitative examples of the explanations generated by their method are presented in Figure 2.5, showcasing from left to right: the original input image and the answers/explanations from PIC-XAI and iPIC-XAI explanation methods, respectively.

Several techniques have been proposed to improve the explainability of image classification models. For instance, [Hendricks et al., 2016] introduced a method that generates natural language explanations for classification decisions by combining textual class definitions, such as a bird species, and image descriptions (captions). Their method was found to produce relevant explanations that take into account both class-discriminative image features and class definitions. These explanations are intended to explain why a specific

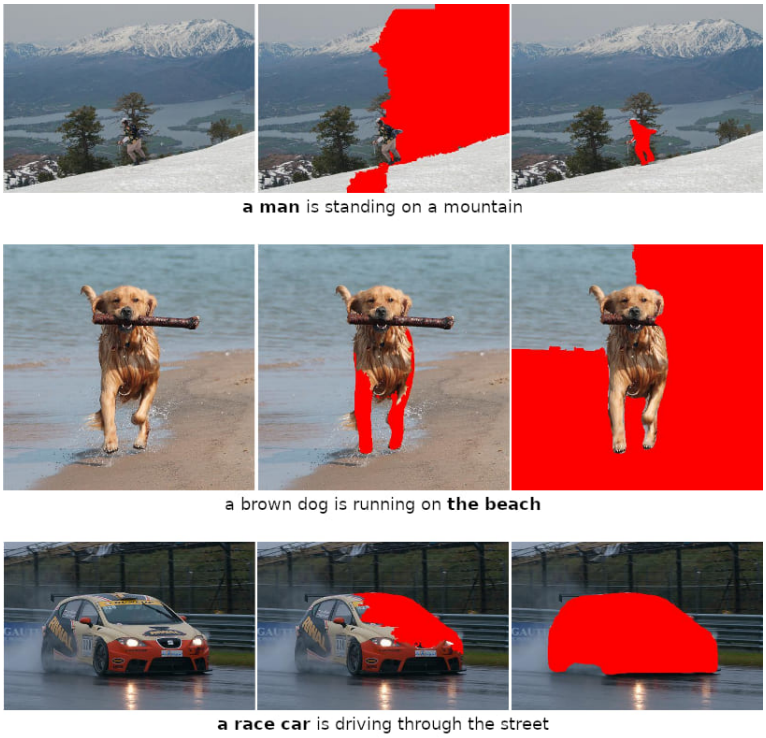


Figure 2.5: Qualitative explanations from [Al-Shouha and Szűcs, 2023].

class label has been assigned to a given image. To condition sentence generation, they also incorporated a reinforcement module that learns to generate sentences that embody a global property of the sentence, such as class specificity. [Fong and Vedaldi, 2017] used a perturbation paradigm in the original input space (the image) to generate meaningful explanations. In their model, they employed an optimization technique that learns perturbation masks that most or least affect the model's output. In another approach, [Yang et al., 2021] employed morphological fragmentation to divide the input image into multi-scale fragments. These fragments were then subjected to masking via perturbation to generate heatmap explanations.

The landscape of Explainable Image Captioning (XIC) research reveals some prevailing trends and characteristics, as summarized in Table 2.2. Most notably, many existing XIC approaches operate in a post-hoc fashion, primarily offering local explanations. While IC primarily deals with the linguistic modality on the output side, some research endeavors venture into seeking explanations within this modality. This can occur either in conjunction with the visual modality or independently. The mechanisms used in these approaches offer a perceptible means of representing explanations. Furthermore, it is worth mentioning

that XIC research predominantly relies on well-established large-scale datasets, such as MSCOCO and Flickr, while domain-specific applications may lack open-source data. This particular challenge raises several fundamental questions, which will be examined in the next section.

Reference	Technique	Modalities of explanations	Nature of explanations	Complexity	Scope	Dataset
[Han and Choi, 2018]	Attention	Vision	Weight matrix	PH	Local	MSCOCO/Flickr30K
[Sahay et al., 2021]	LIME	Vision	Image patches	PH	Local	Flickr30K
[Sun et al., 2022]	LRP	Vision+Language	Heatmaps	PH	Global/Local	MSCOCO/Flickr30K
[Beddiar and Oussalah, 2023]	Attention	Vision+Language	Feature maps+Heatmaps	I+PH	Local	ImageCLEF[4]
[Al-Shouha and Szűcs, 2023]	Region proposals	Vision	Image patches	PH	Local	MSCOCO/Flickr8K
[Wu et al., 2023]	Keywords	Language	Image patches	I	Local	DeepEyeNet[5]
[DEWI et al., 2023]	SHAP	Vision	Heatmaps	PH	Local	–

Table 2.2: Overview of key contributions in Explainable Image Captioning. I/PH denotes intrinsic/post-hoc, respectively.

2.4 Discussion and conclusion

Our literature review, covering both the broader field of XAI and the specific field of XIC, alongside the challenges in XAI previously explored in Section 2.2.7, has brought to light several problems and shortcomings within the realm of IC explainability. These insights have allowed us to identify several ongoing challenges in this field, as outlined in the introduction to this study (Chapter 1), and we reiterate them below.

Building upon the insights gleaned from previous research endeavors, particularly those that delve into both visual and linguistic explanations [Sun et al., 2022, Beddiar and Oussalah, 2023], where a certain degree of parity between these two modalities is implicitly assumed, our first contribution was directed towards addressing a foundational question concerning the reliability of explanations and the sources from which explanations should be derived. This investigation aimed to discern whether these modalities play particular roles in the decision process, a facet that has often been underexplored in the broader context of XAI. The core concept revolved around

[4]ImageCLEF: a medical captioning dataset [Pelka et al., 2021].
[5]DeepEyeNet: a retinal captioning dataset [Wu et al., 2023].

identifying and isolating the components of the architecture that exhibit the most significant involvement in the captioning process. The aim was to determine whether decisions are based exclusively on the input data or are influenced by other elements such as imbalances in the degree of influence exerted by various components/modalities within the captioning architecture. This enables the behavior of the internal components to be implicated in the subsequent explanation process.

Leveraging the results of our former investigations, we have embarked on an in-depth study taking into account the specificities of the architecture components, highlighting the crucial issue of explainability-complexity trade-off. While certain approaches opt for more complex explanation methods [Sun et al., 2022, DEWI et al., 2023] such as back-propagation two-pass (forward/backward) techniques and others, we find it essential to examine whether these time-consuming methods truly provide added value compared to simpler alternatives [Al-Shouha and Szűcs, 2023, Sahay et al., 2021]. Drawing from the advantages we observed within the latent space, our second contribution extends its application in both explanation generation and evaluation. This holistic approach allows us to extract more concrete and tangible insights to elucidate the decision processes in IC. In our pursuit, we have designed variants of two popular explanation methods, LIME and LRP, adhering them to the paradigm of latent space. The choice of these two methods was certainly far from being arbitrary. It aimed to investigate the potential causal relationships between the quality of the explanations, the scope of the explanation methods (local vs. global), and their intricacies. Notably, BU-LRP stands out as a new contribution in the field, as it exploits the BU features, exploring their eventual potential in providing improved explanations in the latent space. Regarding the evaluation of explanation quality, our new method, Latent Ablation, leverages the strengths of the latent space. This manifests in the manipulation of low-level features considered to be very close to internal model theory enabling us to characterize its functioning, as well as the ways in which data is encoded. Depending on the granularity of the concepts manipulated within the latent space, whether they are partial such as individual features, or complete as entire objects, we have designed two versions of the LIME-based method and Latent Ablation. Our objective is to study the impact of this granularity on the construction and evaluation of explanation methods.

Constructing complex data dependencies to help improve overall performance has become relatively straightforward for many machine learning models. However, disentangling the complex structures that these models tend to form within their internal

components proves to be a much more challenging endeavor. As highlighted by numerous researchers in the literature (Section 2.2.7), exploring the latent space presents a valuable means for investigating how ML models structure information, as well as their encoding and decoding processes. Surprisingly, only a few studies have ventured down this path, probably due to its inherent difficulty.

In our third contribution, we focus on investigating semantic concepts such as object position, size, and quantity, through latent space manipulation. This approach helps bridge the gap between the low-level encoding and the high-level semantic understanding of captioning models. To this end, we aim to delve into the intricate process of how the visual concept of quantity is manipulated within the captioning architecture, its impact on object saliency and how it is processed by the captioning model components before it can be captured in the output text description.

It is crucial to highlight a notable distinctive facet of the whole present work: our exclusive focus on operating within the latent space for both seeking and deriving explanations. This approach represents an ambitious endeavor within the realm of eXplainable Image Captioning (XIC), setting it apart from established methodologies. This approach not only underscores the potential of exploiting the latent space but also opens up exciting new leads for advancing the field of explainability of captioning systems in particular, and multimodal systems more generally.

3 Component Influence Identification

Contents

The problem of opacity in AI systems, especially those based on deep learning, has prompted extensive research in the field of XAI. However, although XAI tools represent important strides toward explainability, they may present some incompleteness [Gilpin et al., 2018]. By incorporating insights from the humanities and social sciences, XAI can benefit from a more holistic understanding of explanations and how they might be attained [Borch and Min, 2022]. *Decisiveness*, as defined in psychology, refers to "the ability of the individual to engage in the decision-making process" [Weissman, 1976]. While

decisiveness shares a close relationship with explainability, as far as we know, it has not been the subject of any previous research in the field of XAI. In this specific context of XAI, we define the concept of *decisiveness* as:

> ***"A measure of the active engagement exhibited by individual components of a decision model in producing the final outcome, how each of these components interacts with the others and how it influences the decision process."***

The analogy to the psychological definition can be illustrated as follows: Consider a team of experts working collaboratively towards a shared objective, symbolizing a complex system characterized by a collective decision process. In this analogy, each team member represents a distinct component within the decision model, contributing their unique expertise and knowledge. Similar to how each team member contributes to the team's success, the *decisiveness* of each component determines its influence/impact on the overall decision outcomes. Some components may possess specialized skills and assume leadership roles in making pivotal decisions, while others provide support and complementary contributions. *Decisiveness* exhibited by each component has a fundamental impact on the collective decision process and ultimately shapes the system's ability to achieve optimal outcomes.

Carefully considering the possible limitations of delving into fine-grained explanations without fully incorporating the internal behavior of components can help prevent potential issues related to explainability, such as the risk of providing irrelevant and unjustified explanations or the presence of biases. This becomes particularly relevant when it comes to providing multi-modal explanations [Sun et al., 2022, Beddiar and Oussalah, 2023], notably in sensitive areas such as biomedical image description. In that case, the reasoning behind the decisions made by the model may not be straightforward to decipher and we cannot state that these decisions are exclusively made based on inputs or whether they are affected by other elements within the captioning architecture. In addition, explanations from different modalities may not have the same significance and should therefore not be interpreted and considered in the same way. Studying the influence of each component that belongs to the IC architecture, individually and comparatively to each other, allows us to identify the components that are more or less decisive and impactful in the captioning process and highlight their respective roles in making decisions.

In this chapter, we present a meta-explainability approach for common IC architectures.

Our proposed approach incorporates a formal perturbation protocol that allows us to systematically investigate the influence of model components at both the vision and language levels. Through this protocol, we aim to address fundamental questions regarding the decision process in IC. Specifically, we seek to understand which components play the most influential roles in shaping the generated captions. Additionally, we explore the idea of component *decisiveness*, questioning whether all components have an equal impact on the captioning process or if certain components are more decisive than others. Through this meta-explainability approach, we gain valuable insights into the inner workings of IC models, shedding light on the importance and contribution of individual components. Consequently, the insights gained from our study have significant implications for subsequent explainability choices in IC. For example, this can inform regarding the modality from which the most relevant explanations should be derived, or even more, the possibility of assigning different levels of importance to the components based on their contribution to the captioning process and their respective roles.

3.1 Latent space

Latent space refers to a fundamental concept that has emerged through the development of various deep learning (DL) methods. It has been widely used and discussed in the machine learning literature [Bengio et al., 2013] including their representation learning capabilities and applications such as in self-supervised learning. During training, this space is capable of capturing representations of a lower dimension than the original inputs. Formally, a latent space $\tilde{\mathcal{I}}$ can be defined as the codomain of a function f, which maps each input I from a given input space $\mathcal{I}$ to a compact representation $\tilde{I}$ in the latent space $\tilde{\mathcal{I}}$. This can be expressed as:

$$\begin{aligned} f : \mathcal{I} &\longmapsto \tilde{\mathcal{I}} \\ I &\longmapsto \tilde{I} = f(I) \end{aligned} \tag{3.1}$$

The function f can be as simple as a linear projection or can involve more complex transformations often implemented using DNN models. Those transformations can be learned through various techniques such as auto-encoders and convolutions. The learning process involves training the model until the latent vectors reach an optimal representation that best encodes the input.

The strengths of the latent space stem from its ability to underlie patterns of the data and the extent to which factors can be disentangled, for example, specific dimensions might

correspond to object attributes including the size, color, and position, or other meaningful factors (Chapter 5). The way in which information is compressed and processed, or even structured, often allows irrelevant or redundant aspects to be discarded and the focus to be placed on the most salient and informative aspects. This can be clearly discerned in certain image-related tasks such as IC and image classification, where priority during the decision process is given to foreground and most conspicuous objects in the image rather than less salient objects in the background scene for example.

Although the usage and exploration of latent space have evolved over time in the DL community, notably to develop more sophisticated learning techniques and capture richer representations, this concept remains less exploited for explainability despite its remarkable ability to derive insights about the data and the inner workings of the model. In the present work, we have focused exclusively on the latent space to develop explainable approaches for IC systems. This endeavor represents a first step, but also a breakthrough in our ongoing efforts toward gaining more explainability within the confines of our research scope.

3.2 The perturbation paradigm

In the context of XAI, the perturbation paradigm refers to a methodology that aims to understand the behavior of a decisional system by modifying/altering specific information. Perturbations have been widely employed in post-hoc explainers, such as in building surrogate models like LIME that approximate the functioning of black-box models by means of perturbed instances generated locally around a given input (Section 2.2.6). Perturbations have also been used in sensitivity analysis, which aims to measure how intentional changes in the model's inputs impact its output/decision. [Fong and Vedaldi, 2017] in their work, focused on employing meaningful perturbations explicitly editing the input image, in a way that preserves their semantic meaning while inducing changes in the model's predictions. Their method helps identify the image regions that significantly influence the model's output scores when perturbed.

It is worth mentioning that some research has used the perturbation paradigm for adversarial attacks and/or training. [Liu et al., 2023] proposed a score-based attack model for the textual attack task. Their method was based on selecting important words using a self-attention mechanism and generating a correlation degree of words inside the texts, which is to some extent related to the aspect of explainability. In [Xian et al., 2021], the

authors sought to identify the vulnerability of link prediction methods for graph-structured data using a deep architecture-based adversarial attack method. They also investigated other adversarial attack methods such as heuristic and evolutionary perturbation methods. Other methods have also been proposed for various adversarial attack-based tasks such as Graph Neural Networks (GNNs) structure enhancement [Wu et al., 2022b] and time-series prediction [Wu et al., 2022a]. Regarding the image captioning domain, we are only aware of the methods proposed in [Xu et al., 2018, Zhang et al., 2020a] where the authors focus only on improving the robustness and stability of captioning models, but do not address the explainability issues in the domain.

More generally, we define the perturbation paradigm as the process that involves introducing changes, denoted as η, to a specific component, denoted as ω_i, among the set of all components $\Omega = \{\omega_1, ..., \omega_i, ...\}$ within an AI system and observing the resulting output changes. In this context, a component refers to any unit of the system that represents data or a module that can be reused in other architectures. This includes raw input data, data encodings, attention mechanisms, language decoders, and more. The perturbation η is applied to a predefined set of elements of interest within the selected component. These elements could be regions of an image, extracted features, latent vectors, model weights, and so on. Perturbations can take various forms, such as random perturbations involving random increases or decreases in element values, adding random noise, or patterned perturbations that introduce specific variations or meaningful changes.

By leveraging perturbations, one can gain a deeper understanding of the contributions and sensitivities of different components, and identify critical factors that determine system decisions, thus facilitating the explainability of complex models.

3.3 Selected captioning architecture

We employ the standard captioning architecture Ada-LSTM from [Sun et al., 2022] which is designed upon the common topology of Encoder-Attention-Decoder (Figure 3.1) and rely on the use of Bottom-Up (BU) image features rather than classical global CNN features. The main difference between these two types of features lies in the level of granularity and spatial information they capture. The CNN global features are obtained from the last layer of a classical CNN architecture when encoding the input image. They represent a high-level overview of the image content and provide low-granularity representations encoding the global context. Bottom-up features, on the other hand, are obtained following

an object detection operation that provides region-based encodings. These encodings represent target regions or objects in the image and capture fine-grained information such as objects and people. Bottom-up features also show good object localization capabilities and allow captioning models to attend to specific regions of the image using attention mechanisms, resulting in more contextualized and detailed captions.

Several key factors influenced the selection of this architecture for our work. First and foremost, from a captioning standpoint, the use of BU features proved more effective in generating more accurate and descriptive captions, as stated in [Anderson et al., 2018]. Many recent captioning models have embraced the use of BU features, as they enable the capture of more fine-grained information compared to traditional global CNN features. Additionally, the Encoder-Attention-Decoder framework has emerged as a widely adopted architecture in the IC community. By leveraging this framework, our findings can be generalized to various IC architectures that follow this structure, enhancing the applicability of our approach.

Secondly, from an explainability standpoint, it is crucial to emphasize that the original black box model must inherently possess a high level of accuracy. Indeed, if the original model fails to generate accurate outcomes, any subsequent attempts at employing explainability methods would inherently lack the capacity to provide relevant explanations. Therefore, to assess the effectiveness of our explainability method, it becomes essential, even at this preliminary stage, to possess a captioning model and system references of a certain level of relevance. Without such baseline quality, the assessment of provided explanations remains impractical. In fact, explanations derived from an inaccurate model, then also compared with low-quality references derived from the same inaccurate model, can potentially introduce biased or irrelevant information, ultimately skewing the entire explainability process. Finally, it is noteworthy to highlight that the work of [Sun et al., 2022] represents one of the pioneering efforts in addressing the challenge of explainability, specifically within the captioning task, thus representing a benchmark for our work.

Although more sophisticated architectures exist, such as Transformer-based ones (Sec. 2.1) which have gained significant attention, when it comes to explainability, they can pose challenges due to their inherent characteristics. The high complexity of these architectures relative to their ordinary performances for the IC domain in particular, makes them less appropriate for explainability issues, where a trade-off must be preserved [Adadi and Berrada, 2018]. For example, Transformers operate on distributed

representations and perform complex computations, making it challenging to trace the influence of individual input features or tokens on the model's predictions. Furthermore, those models are known by their large number of parameters and it is often very complicated to disassociate the components of the architecture from each other since they have a lot of interactions and dependencies. Given all these arguments, we are convinced that for applications that prioritize explainability, alternative architectures or approaches that are inherently easier to manipulate, such as the Encoder-Attention-Decoder, may be more suitable options.

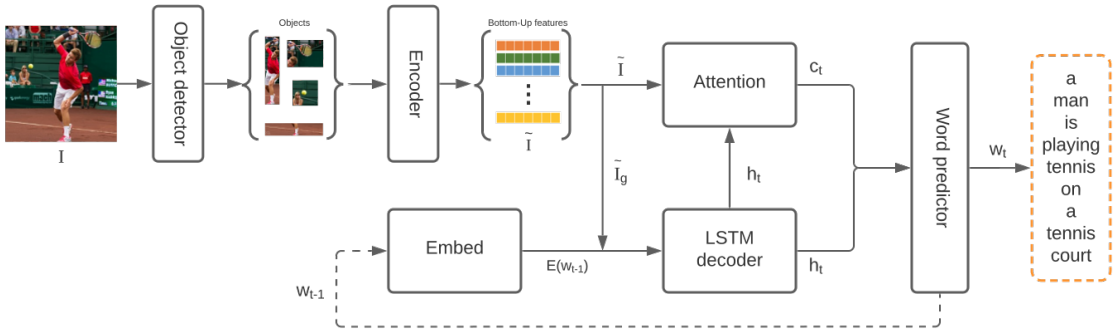


Figure 3.1: Bottom-Up captioning architecture overview.

3.3.1 Image encoding:

In their work, [Sun et al., 2022] used Faster-RCNN [Ren et al., 2015] to generate the BU features. Faster-RCNN is an object detection algorithm designed to localize and delimit objects in an image using bounding boxes. It comprises three key components: a backbone network, a region proposal network (RPN), and a region-CNN. The backbone network, typically a pre-trained CNN like VGG16 [Simonyan and Zisserman, 2015] or ResNet [He et al., 2016], extracts high-level features from the input image. These features are fed into the RPN, which generates region proposals representing regions likely to contain objects. The RPN consists of a classifier and a regressor. By sliding a window (called anchor) over the image at various scales, the classifier predicts an objectness score, indicating the likelihood of an object's presence. The regressor iteratively refines the anchor coordinates based on the classifier's predictions. Finally, the region-CNN takes the proposed regions from the RPN and performs feature extraction (encoding) and classification to generate feature vectors and assign an object label/class to each region along with a confidence score/probability. The backbone network is typically pre-trained on a large-scale classification dataset, while the RPN and region-CNN are systematically

fine-tuned on the data at hand based on the bounding box (anchor) regression loss and the region classification loss. The number of BU features to generate with Faster-RCNN may vary depending on several factors such as the complexity of the images within the dataset, the diversity of objects present, and the requirements of the downstream tasks. In the present work, we choose to utilize the Faster-RCNN encoder as our image encoder. The Faster-RCNN architecture has gained significant popularity in the field of image captioning due to its widespread adoption and robust performance compared to other state-of-the-art CNN architectures.

Given an aligned pair (I, S) consisting of an image instance I and its corresponding ground-truth captions S, the Faster-RCNN algorithm extracts a set of bottom-up (BU) visual features (latent representations), denoted as $\tilde{I}$ (Equation 3.2). These features are obtained by applying the region-CNN encoder to the set of regions detected within the image using the RPN. The resulting BU features $\tilde{I}$ form a set of latent vectors, where each vector v_i represents an individual region (ideally an object). The number of BU features is denoted as V, and each feature has a dimensionality of d_v.

$$\tilde{I} = (v_i)_{i=1}^{V} \quad , \quad v_i \in \mathbb{R}^{d_v} \tag{3.2}$$

3.3.2 Attention and decoding:

The process of combining visual features, the previous hidden state, and the word embedding is a crucial step in the caption generation process. It involves creating a global visual feature $\tilde{I}g$ by averaging the set of visual features. This global visual feature is then combined with the previous hidden state $ht - 1$ generated by the LSTM decoder and the word embedding of the previous word in the sequence $E(w_{t-1})$. The combination of these components is fed into the LSTM decoder, which generates the current hidden state h_t using the specified equation (Equation 3.3). The dimensions d_w and d_h refer to the dimensions of the word embedding and hidden state vectors, respectively. The hidden state vector, represented by h_t, captures the linguistic context preceding the current word w_t in the caption generation process, providing essential information for generating accurate and contextually appropriate captions.

$$h_t = LSTM(\tilde{I}_g, h_{t-1}, E(w_{t-1})) \quad , \quad h_t \ \in \mathbb{R}^{d_h} \tag{3.3}$$

The adaptive attention mechanism [Lu et al., 2017], as employed in this architecture, plays a crucial role in selectively attending to specific regions of the image during the caption

decoding process. By dynamically assigning attention weights to the BU visual features based on the information contained in the hidden state, the model generates a context representation c_t of dimension d_c (as shown in Equation 3.4). This context representation captures the most informative and relevant visual information for word prediction at each decoding step, resulting in more accurate and contextually rich captions.

$$c_t = ATT(h_t, \tilde{I}) \ , \ \ c_t \in \mathbb{R}^{d_c} \tag{3.4}$$

Finally, both the hidden state and context vector are used by the language LSTM to predict a score over vocabulary for the current word w_t (Equation 3.5) of the output sequence C (Equation 3.6). The word with the highest score is returned. L is the caption length.

$$w_t = WordPredict(h_t, c_t) \tag{3.5}$$

$$\hat{C} = (w_l)_{l=1}^{L} \tag{3.6}$$

During training, the IC model is optimized with cross-entropy loss measuring the discrepancy between the predicted caption $\hat{C}$ and the ground-truth one C. To calculate the cross-entropy loss, these captions are compared word by word. Each word in the predicted caption is treated as a probability distribution over the vocabulary, representing the model's confidence in generating each word. Similarly, the ground-truth caption is represented as a one-hot encoded vector, where the position corresponding to the correct word is assigned a value of 1, and all other positions are assigned a value of 0. For a single training instance, the cross-entropy loss can be calculated as follows:

$$\mathcal{L}_{ce}(p^{\hat{c}}, p^{c}) = -\frac{1}{L}\sum_{l=1}^{L}\sum_{q=1}^{Q} \ p_{l,q}^{c} \ \ \log(p_{l,q}^{\hat{c}}) \tag{3.7}$$

where Q is the vocabulary size, p^c is the ground truth one-hot vector, where $p_{l,q}^c = 1$ if the l^{th} token is the q^{th} word in the vocabulary, and $p_{l,q}^c = 0$ otherwise. $p_{l,q}^{\hat{c}}$ is the probability of the l^{th} word being the q^{th} word in the vocabulary. The inner sum computes the cross entropy loss for each word in the caption, and the outer sum averages the losses across all the words in the caption. The loss formula can therefore be simplified as given in Equation 3.8.

$$\mathcal{L}_{ce} = -\sum_{l=1}^{L} \ \log(p_l^{\hat{c}}) \tag{3.8}$$

3.4 Our approach

In this section, we introduce a novel perturbation approach that aims to capture the *decisiveness* and sensitivity of the components in IC architectures, as outlined in our work [Elguendouze et al., 2022]. Our approach stands out by operating entirely in the representation space, rather than the original space (images, text), to achieve explainability [Sun et al., 2022, Han and Choi, 2018]. Specifically, we focus on two distinct latent levels within the architecture: Vision and Language as shown in Figure 3.2. r_V, r_C, r_W, r_H in red boxes represent perturbations on VF, CT, WE, and HT components respectively. $\oplus$ denotes the direct sum.

➔ Vision: we introduce perturbations to the BU visual features (VF) and context representations (CT), which result from the image encoder and attention mechanism, respectively. By perturbing these two components, we aim to understand the impact of the visual part of the IC model on the overall captioning process.

➔ Language: the perturbation concerns the word representations (WE) and hidden states (HT) generated by the word embedding encoder and the language decoder, respectively. This allows us to explore the influence of these language-related components on the captioning process.

The intuition behind targeting these specific levels (Vision, Language) with our perturbation approach is mainly guided by their mutual complementarity. By analyzing the effects of perturbation on both levels, we gain insights into their individual behaviors and can compare them to each other, thereby covering all crucial parts of the architecture. Each level consists of two essential components that need to be carefully examined and compared to assess their involvement in the captioning process. In addition, these components embody important sub-tasks such as encoding, attention, and decoding, further elucidating their respective roles in the overall captioning task. Gaussian perturbations have often been applied to original images in computer vision (CV) to mimic common real-life distortions, such as changes in lighting conditions, focus, and contrast [Borkar and Karam, 2019, Laugros et al., 2019]. Building upon this understanding, we extend the application of perturbations in a different context, specifically in the latent space rather than the original input space (i.e., images) and for explainability purposes. Our objective is to uncover patterns and dependencies that may not be immediately apparent in the original input space. By applying Gaussian perturbations in the latent space, we get the possibility to explore how the model behaves and responds towards these

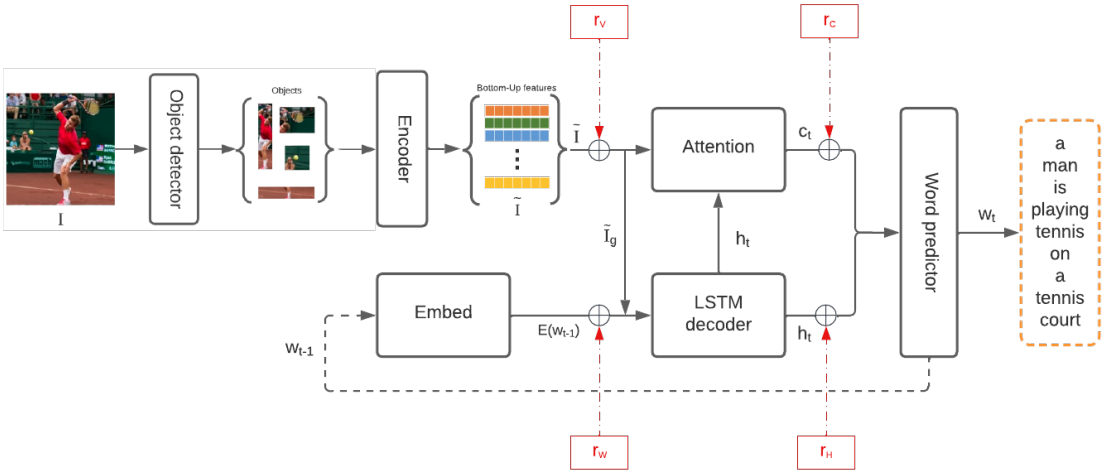


Figure 3.2: Overview of the latent perturbation protocol.

changes, which can provide valuable insights about the inner workings of deep models in general and captioning models in particular, at different stages of the decision process.

To introduce our perturbation approach, we focus on the four key components involved in the captioning process: visual features (VF), word embeddings (WE), context vectors (CT), and hidden states (HT). We adopt a perturbation function, denoted as $r(\phi)$, which adds random noise $\eta = (\eta_j)_{j=1}^{d_\eta}$ element-wise to a representation vector $\phi \in \mathbb{R}^{d_\phi}$. Intuitively, the dimensions of the perturbation vectors (d_η) match the dimensions of the corresponding component latent vectors (d_ϕ). The random noise η_j is sampled from a Gaussian distribution $\mathcal{N}$ (Equation 3.9), where μ represents the mean and σ represents the standard deviation. By utilizing this Gaussian distribution, we enable the perturbation to cover a range of possible variations in the component representations.

$$\eta_j \sim \mathcal{N}(\mu, \sigma^2) \tag{3.9}$$

To ensure consistent perturbations within the definition domain of each component, we impose a condition on η_j using Equation 3.10. This condition guarantees that each random value lies within the boundaries defined by min_G and max_G, representing the global minimum and maximum values for the component, respectively.

$$min_G \leq \eta_j \leq max_G \tag{3.10}$$

Below, we present comprehensive details regarding the perturbation applied to both the visual and linguistic levels of the captioning architecture, which is systematically carried out on each module within these architectures.

3.4.1 Perturbation of the visual level

Visual features (VF): This perturbation is applied to the BU visual features at the output of the Faster-RCNN. For each visual feature v_i (Section 3.3.1), we add random noises ηj to all the dimensions of the feature vector, as shown in Equation 3.11. To ensure that the perturbed feature remains within the valid range, a second condition is placed on each perturbed dimension as given in Equation 3.12. In this equation, vij represents the j^{th} dimension of the original visual feature v_i, and min_j and max_j represent the global minimum and maximum values for the j^{th} dimension across the entire latent space of the visual feature component.

$$r(v_i) = (v_{ij} + \eta_j)_{j=1}^{d_v} \tag{3.11}$$

$$min_j \leq v_{ij} + \eta_j \leq max_j \tag{3.12}$$

Context representations (CT): The context vector c_t, generated by the attention module, indicates the regions of the image that are most relevant for generating the current word in the output sequence, i.e. the ones that should acquire more attention/focus by the decoding module afterward. In this perturbation, random values η are added to c_t at each time step t, as shown in Equation 3.13. As the values of the context vectors may vary across different executions, the second condition above is slightly adjusted for this perturbation to consider the global minimum and maximum values instead of per dimension, as shown in Equation 3.14.

$$r(c_t) = (c_{tj} + \eta_j)_{j=1}^{d_c} \tag{3.13}$$

$$min_G \leq c_{tj} + \eta_j \leq max_G \tag{3.14}$$

The perturbation of the visual component is a central aspect of our investigation, as it specifically targets the encoding stage within the captioning process. Subsequently, the next phase of the perturbation methodology addresses the linguistic part responsible for decoding the caption, another essential aspect.

3.4.2 Perturbation of the language level

Word embeddings (WE): Similarly, we define the perturbation of the word embeddings at the output of the embedding module. At each time step t, a random perturbation η is added to the embeddings of the predecessor word w_{t-1}, as shown in Equation 3.15. The same condition as for the context representation perturbation applies to ensure the perturbed embeddings remain within the valid range.

$$r(E(w_{t-1})) = (E(w_{(t-1)j}) + \eta_j)_{j=1}^{d_w} \tag{3.15}$$

Hidden states (HT): The perturbation of hidden states involves the hidden representations generated by the LSTM decoder. These representations encapsulate the textual information that has been generated up to the previous word, w_{t-1}. The perturbation is applied to each dimension of the hidden states h_t, as illustrated in Equation 3.16. The same condition as for context representations must be complied with.

$$r(h_t) = (h_{tj} + \eta_j)_{j=1}^{d_h} \tag{3.16}$$

Once the overall methodology has been carefully outlined, we now delve into the intricacies of the experimental protocol that was followed. This protocol, meticulously designed to validate the proposed concepts, deserves in-depth exploration.

3.5 Experimental protocol

In this section, we present the detailed experimental protocol that we followed in this work. We begin by introducing the datasets utilized and describing the steps taken to prepare them for our specific study. Next, we provide a technical overview of the captioning model employed, including hyper-parameters values, the encoder implementation, and the vector dimensions. We then delve into the experimental settings of Gaussian perturbation, explaining the specific parameters and choices. Finally, we outline the evaluation metrics that were used to assess the results of our explanation method.

3.5.1 Dataset selection

In this work, we utilized two popular benchmarks for sentence-based image descriptions: MSCOCO2017 [Lin et al., 2014] and Flickr30k [Young et al., 2014]. MSCOCO2017, short for Microsoft Common Objects in Context, is a large-scale dataset consisting of images paired with human-annotated captions. It comprises approximately 123,000 images, each accompanied by 5 captions, resulting in a total of around 616,000 captions. The images in MSCOCO are collected from professional photographers and cover a wide range of object categories in various everyday scenes, thereby offering greater diversity in terms of content. In contrast, the Flickr30k dataset includes 31,783 images and approximately 158,915 captions. It consists of images from the Flickr platform focusing on people involved in everyday human-centric activities and events. Furthermore, there is a distinction in caption style between the two datasets. MSCOCO captions are in general shorter and more concise, focusing on describing the main objects and their relationships in the image. On the other hand, the captions in Flickr30k tend to be longer and more detailed, providing additional contextual information and describing various aspects of the scene. Table 3.1 presents a comparative summary highlighting the key differences between the MSCOCO and Flickr30k datasets, regarding their inherent characteristics and pre-processing results for our specific context of image captioning.

Dataset	**MSCOCO2017**	**Flickr30k**
Size	123k images, +600k captions	32k images, +158k captions
Image Source	Professional photographers	User-contributed images on Flickr
Caption Style	Short, concise	Longer, more detailed
Content composition	Wide range of subjects, objects and styles	Human-centric scenes and activities
Caption Diversity	Multiple (5) captions per image	Multiple (5) caption per image
Vocabulary size	11023	9582
Karpathy split	110k/5k/8k	29k/1k/1k

Table 3.1: Summary of comparison between MSCOCO2017 and Flickr30k datasets.

Although other datasets such as "Conceptual Captions" and "nocaps" are available for IC, we were unable to use them in our work for several reasons. In the case of the Google "Conceptual Captions" dataset, we conducted a preliminary evaluation using a sample from this dataset to assess the quality of the captions. However, the evaluation scores of common metrics on a test set of 6,500 images for the Ada-LSTM model trained on a subset of 150,000 images, were significantly lower compared to those obtained for MSCOCO2017 and Flickr30k (Table 3.2). This discrepancy in performance would introduce biased results in our perturbation experiments and subsequent explanation methods in the next chapters.

Dataset	Train/Test	BLEU-4	CIDEr-D	SPICE	ROUGE	METEOR	MSICE
MSCOCO2017	110k/5k	0.3403	1.0774	0.2007	0.5522	0.2692	0.5747
Flickr30k	29k/1k	0.2543	0.5258	0.1512	0.4680	0.2081	0.5033
Concept. Cap. (subset)	150k/6.5k	0.0417	0.4154	0.0991	0.1874	0.0685	0.1231

Table 3.2: Evaluation results of Ada-LSTM captioning model on MSCOCO2017, Flickr30k, and Conceptual Captions.

The low quality of the alternative texts in the Conceptual Captions dataset may constitute the major factor contributing to these results. These captions, which serve as alternative texts for web images, do not necessarily reflect the most salient content of the images and often contain errors, as we have manually verified. Our hypothesis is that these alternative texts may emphasize a salient target object in the scene with regard to the author's purpose, given that the images are labeled according to the concept they are meant to represent. In contrast, datasets like MSCOCO2017 allow for descriptions of the entire image, without being biased towards a specific purpose. Additionally, the Conceptual Captions dataset comprises highly diverse images in terms of context and quality, and it often includes synthetic images, which can further affect the performance of captioning models. The scores obtained on this dataset for some common captioning architectures are also not as impressive as those achieved on MSCOCO2017 and Flickr30k, as we can notice when comparing the results reported in the original papers presenting these models with those on the dataset platform.

As for the "nocaps: novel object captioning at scale" dataset, it is important to note its unique focus on capturing a wider variety of visual concepts. However, when used in conjunction with our explanation methods, the dataset's emphasis on novel objects may introduce challenges in interpreting the absence of specific objects in the captions. Specifically, our perturbation method relies on the model's ability to predict objects in the output caption following the Gaussian perturbation. However, since the nocaps dataset is designed to evaluate the model's capacity to recognize novel objects, it becomes difficult to distinguish whether the absence of an object is due to perturbation or simply due to the inherent composition of the image from the nocaps dataset, which intentionally includes novel objects that the model has never encountered before.

Considering these factors, we opted to focus on well-established datasets like MSCOCO2017 and Flickr30k for our experiments, as they provide more consistent results, aligning with the common practices in the captioning community. While the Conceptual Captions dataset may not be widely employed in the captioning community for

experimental purposes, it still holds potential for other tasks such as vision-language model pre-training and representation learning [Radford et al., 2021, Mokady et al., 2021].

To evaluate the performance of our perturbation method with these two datasets, we follow the widely adopted "Karpathy" split [Karpathy and Fei-Fei, 2015] for reporting results. This split divides a dataset into three subsets: train, validation, and test. The Karpathy split method provides a standardized evaluation setup, allowing for consistent comparison of different models and algorithms. By adhering to this split, we ensure compatibility and enable meaningful comparisons with prior work in the field. Following a similar approach to [Sun et al., 2022], we construct the vocabularies based on the captions from the training sets of each dataset. Words that appear less than 3 times in MSCOCO2017 and 4 times in Flickr30k are encoded using the unknown token "<unk>". We believe that the intuition behind the choice of these thresholds is that words that appear with such occurrences are so rare that they do not even appear simultaneously on all five ground-truth captions associated with the same caption, implying that they are either outliers or unfamiliar terms, and therefore necessarily have a substitute in the vocabulary. The results presented in Section 3.7 are reported on the test partition of MSCOCO2017 and Flickr30k, consisting of 5,000 and 1,000 instances, respectively.

3.5.2 Captioning model preparation

As indicated earlier in Section 3.3, we utilized the standard captioning architecture based on BU features from [Sun et al., 2022] for our experiments. It is worth recalling that this architecture consists of a Faster-RCNN encoder, an adaptive attention module, an LSTM decoder, and an LSTM word predictor. The model was retrained on the train partitions of the MSCOCO2017 [Lin et al., 2014] dataset (110k instances) and the Flickr30k [Young et al., 2014] dataset (29k instances). The training process was stopped automatically when the best CIDEr score was achieved, which occurred at the 21^{st} epoch for MSCOCO2017 and the 8^{th} epoch for Flickr30k.

For the image encoder, we utilized the Detectron2 [Wu et al., 2019] implementation of Faster-RCNN. It selects the top $V = 36$ detections (regions) per image, and each detection was represented by a BU visual feature of dimension $d_v = 2048$. This choice of 36 features is a common practice in image captioning as it captures a spectrum of diverse and informative visual information. In order to reduce their sparsity and condense them into more compact vectors, these BU features underwent a further transformation via a simple linear layer which reduced them to 1024 dimensions. The dimensions of the context vectors,

word embeddings, and hidden state vectors were set to $d_c = 512, d_w = 512, d_h = 1024$, respectively. The maximum length of the output caption sequence was capped at $T = 20$.

3.5.3 Perturbation settings

We conducted experiments using Gaussian perturbations with mean $\mu = 0$ and various standard deviation σ values. To determine the appropriate σ values for each perturbed component, we empirically computed the upper and lower bounds of their representation space. These interval bounds served to determine the range of σ values as shown in Table 3.3. Intuitively, the σ controls the magnitude of the perturbations applied to the values of each component.

	Dataset	VF	WE	CT	HT
$[min_G, max_G]$	MSCOCO2017	[0, 16.5]	[-5, 5]	[-1, 9.5]	[-1, 1]
	Flickr30k	[0, 16.5]	[-5, 5]	[-1, 14.6]	[-1, 1]
$\sigma \leq$		1.5	1.5	1.5	0.375

Table 3.3: Global range bounds and standard deviation values per perturbed component.

To find the optimal range of σ values, we started by initializing σ with the highest possible value that gives the maximum perturbation. We then iteratively halved the σ and applied the perturbation until we observed stationarity in the quality of the captions. Stationarity was reached when the difference in caption scores became negligible, with a difference less than a predefined threshold ($\epsilon = 10^{-2}$), indicating that the component became insensitive to perturbations. At this point, we stopped the process and recorded the previous σ value as the lower boundary for the perturbation. This process was repeated for all the components, ultimately retaining only the intervals of sigma values that were most representative of the results. Figure 3.3 illustrates the evolution of evaluation scores for a VF perturbation across varying sigma (perturbation magnitude) values. The results reveal that the informative range, where the most representative results are observed, lies within [0.1875, 1.5]. The evaluation scores were obtained by comparing the captions generated after perturbation with those before perturbation, using several metrics introduced in Section 2.1.3.

The sigma values selected for the perturbation experiments were (1.5, 0.75, 0.375, 0.1875) for VF, CT, and WE, and (0.375, 0.1875, 0.0938, 0.0469) for HT. It is worth noting that for the HT component, the values of σ were adjusted by shifting the interval to

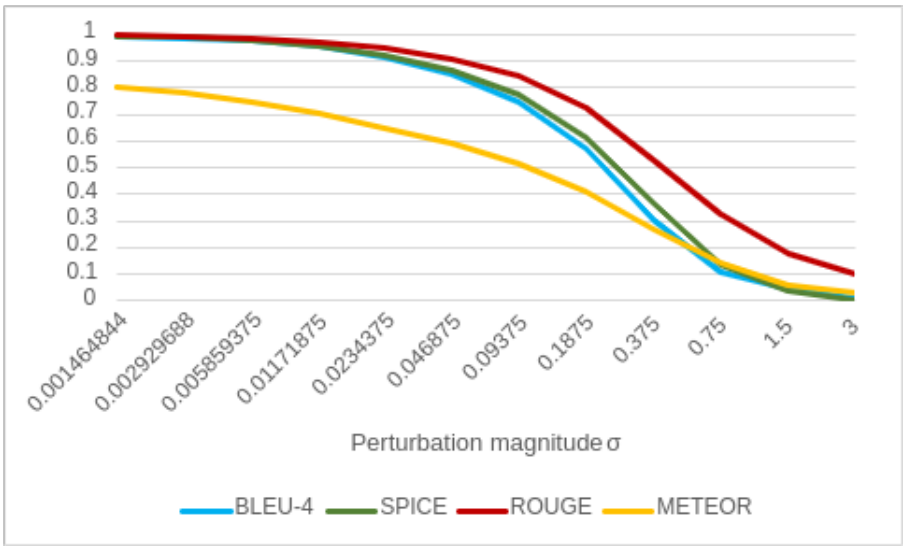


Figure 3.3: Perturbation magnitude retrieval for the Visual Features component.

accommodate its smaller definition range compared to the other components. This adjustment was necessary to ensure consistency across all perturbed components and maintain a balanced level of perturbation. By tailoring the sigma values to each component's specific range, we ensured that the perturbation process remained within the appropriate bounds and allowed for meaningful comparisons between the components.

To ensure the reliability and stability of our findings, we repeated the perturbation process 30 times for each sigma value and each component. By repeating the experiments, we accounted for the inherent randomness in the perturbation process. We evaluated and reported the results on the test partitions of both the MSCOCO2017 and Flickr30k datasets, allowing us to analyze the robustness and sensitivity of the captioning model under different levels of perturbation.

3.6 Explanation evaluation

3.6.1 Evaluation measures

We evaluate the quality of captions for different perturbation configurations using well-established evaluation measures commonly employed in image captioning (Section 2.1.3). These metrics include BLEU [Papineni et al., 2002], METEOR [Banerjee and Lavie, 2005], CIDEr-D [Vedantam et al., 2015],

SPICE [Anderson et al., 2016], and ROUGE [Lin, 2004], which have gained widespread adoption in IC evaluation due to their long-standing presence in the NLP domain. Our selection of these metrics was driven by multiple factors, including their suitability for the particularities and requirements of our specific task, which focuses on the explainability of IC models rather than the pure comparison of different IC systems. While we prioritize explainability throughout this work, the selected metrics appear most appropriate for our analyses as there is an ongoing debate about the use of learning-based approaches for either explanation generation or evaluation, such as ClipScore and BERTScore, as we previously discussed in Section 2.1.3. Nonetheless, it is crucial to acknowledge that conventional metrics might also have inherent limitations, paving the way for developing new evaluation measures that better cater to our specific requirements of explainability.

To address this need, we introduce a new evaluation metric called MSICE (Morphology-semantic-based Image Caption Evaluation). Unlike some existing metrics, such as BLEU, ROUGE-L, and CIDEr, which primarily rely on n-gram overlap, MSICE incorporates two crucial linguistic aspects: morphology, concerning the surface form of the captions, and semantics. Existing measures often overlook lemmatization and synonymy when computing scores, leading to unfair penalization of the output captions, especially in the context of our goal of improving explicability. Specifically, in terms of synonymy, lexically different yet semantically similar words (e.g., *car* and *vehicle*) in different sentences (e.g., "*a car is driving down a street*"; "*a vehicle is driving down a street*") should be considered equivalent. In addition, from an explainability perspective, the presence of desired objects holds greater importance than quantifiers. Therefore, under MSICE, sentences like "*Four chairs around a table*" and "*A chair and a chair and a chair and a chair around a table*" would be considered highly similar, capturing the essence of the explanation task more effectively, especially as we successfully manage to rebuild visual sequentiality on the language side. These two features of MSICE enable us to refine the evaluation scores, aligning with our specific objectives of explainability.

To further define our metric in this section, it is essential to understand some key concepts in the domain of NLP. A parser is an algorithm that analyzes the grammatical structure of a sentence and assigns a tree structure to represent the internal organization of a text. This process, known as "parsing", enables us to determine the relationships between words in a sentence, such as subject-verb-object relationships, and to represent the sentence's structure in the form of a syntactic tree. Part-of-speech (POS) tagging involves assigning a specific category to each word in a sentence, indicating its syntactic

role within the text (noun, verb, adjective, etc.). Lemmatization is the process that enables reducing words to their base form, called lemma, for example: the lemma of the word "explainability" is "explain". The preferred forms of lemmas are typically singular for nouns and infinitive for verbs, as these forms often serve as dictionary entries. Spacy[1] and NLTK[2] are open-source libraries that provide a vast collection of corpora, lexical resources (such as Wordnet[3]), and pre-trained models to assist in developing methods for various NLP tasks including lemmatization, parsing, and part-of-speech tagging.

MSICE employs three key functions to process the text: f_O extracts objects from a sequence of words using a POS-tagger and a parser, from Spacy and NLTK, respectively. The POS-tagger identifies words representing nouns (subjects and objects), while the parser handles polylexical expressions, also known as noun chunks, which are nouns composed of two or multiple words (e.g., "*fire hydrant*" corresponds to "*firehydrant*"). f_L converts sets of objects into lemmas using the lemmatizer from Spacy and f_S retrieves synonyms for all lemmas from the Wordnet corpus. For a given evaluation instance $(C, \hat{C})$, composed of one or multiple reference captions C and a candidate caption $\hat{C}$, we first apply f_O to extract the object sets for both references and the candidate. Next, we transform these object sets into lemmas using f_L and then find synonyms for the candidate object lemmas with f_S. Finally, we compute the proportion of common words between the reference object lemmas and the synonyms of the candidate object lemmas. Finally, the scores obtained for individual evaluation instances are averaged to obtain the overall MSICE score. To optimize finding synonyms and avoid redundancy, we only search for synonyms for words in candidate captions and not in references, as the search would be slower in the presence of multiple reference captions. The formal definition of MSICE for a single evaluation instance is given in Equations 3.17 through 3.21, where $\circ$ is the function composition, $|.|$ signifies the cardinality and L represents the caption length.

$$MSICE = |A| \tag{3.17}$$

$$A = \{f_L \circ f_O(C)\} \cap \{f_S \circ f_L \circ f_O(\hat{C})\} \tag{3.18}$$

$$f_O : x \mapsto f_O(x) = \{o_i\} \ , \ 1 \leq i \leq L \tag{3.19}$$

$$f_L : y \mapsto f_L(y) = \{l_i\} \tag{3.20}$$

$$f_S : z \mapsto f_S(z) = \{s_j\} \ , \ i \leq j \tag{3.21}$$

[1]Spacy: `https://spacy.io/`
[2]NLTK (Natural Language Toolkit): `https://www.nltk.org/`
[3]WordNet: `https://wordnet.princeton.edu/`

3.6.2 Evaluation protocol

We adopt a tripartite evaluation protocol to comprehensively assess the effects of perturbation. This protocol builds upon the classical IC evaluation, where the predicted captions are conventionally compared to the ground-truth captions. Then, we go beyond this evaluation by further investigating the impact of perturbation related to each of these references separately, as depicted in Figure 3.4.

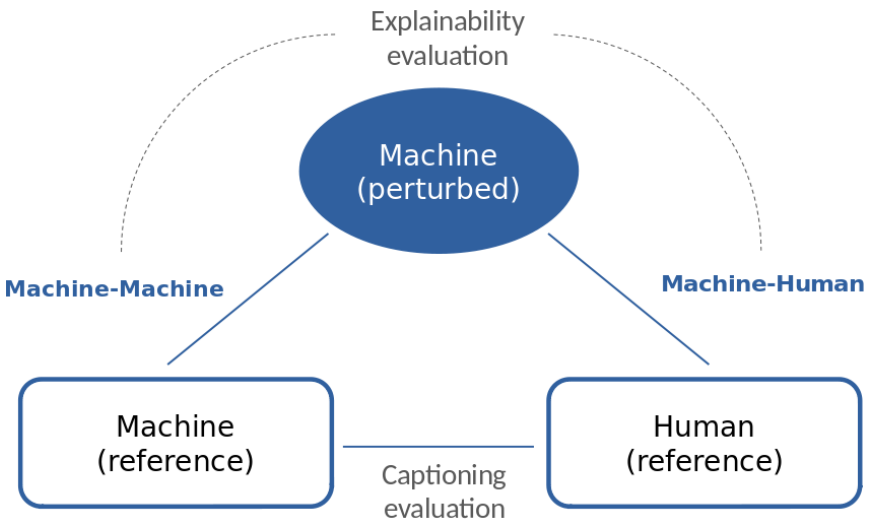


Figure 3.4: The tripartite perturbation evaluation protocol.

The first evaluation, referred to as *Machine-Machine (M-M)*, involves comparing the perturbed system captions (candidates) with the original system captions before perturbation (references). This evaluation focuses solely on the system's performance without considering human influence. By comparing the perturbed system captions with their unperturbed counterparts, we gain valuable insights into the perturbation's effects on the system's output independent of human judgments.

In addition, we conduct a second evaluation called *Machine-Human (M-H)*, where we compare the perturbed system captions (candidates) against the ground-truth human captions (references). This evaluation allows us to gauge how well the system performs in relation to human-generated captions, providing insights into the system's alignment with human language understanding.
By employing this tripartite evaluation, it is possible to investigate and identify potential artifacts that may arise from human involvement represented by human-annotated captions, in the evaluation process.

3.7 Results and discussion

Figure 3.5 provides a visual illustration of the qualitative outcomes achieved through perturbation on samples from the MSCOCO2017 test set, using the maximum perturbation magnitudes, i.e. $\sigma = \{1.5, 1.5, 1.5, 0.375\}$ for {VF, WE, CT, HT} respectively. Captions contained within purple boxes denote the system references. The captions presented in the lower boxes correspond to the respective perturbations indicated alongside. The results highlight distinct trends in the various perturbation types. Notably, the VF perturbation displays the highest divergence from the reference captions, leading to captions that do not effectively reflect the image's content. In contrast, the CT perturbation manages to capture some fundamental concepts such as the color "white" of the mug (depicted in the left example) and the "room" (depicted in the right example), though these captions remain incomplete and incoherent. Remarkably, both the WE and HT perturbations exhibit positive results, successfully generating captions that encompass nearly all the essential elements needed to describe the image content.

Figure 3.5: Qualitative examples of perturbation on MSCOCO2017 test set.

The qualitative analysis of the perturbation results on the Flickr30k dataset reveals consistent trends in the perturbation results compared to the MSCOCO dataset. The model does, however, exhibit a heightened ability to correctly identify the subject in various contexts, such as the man or the baseball player depicted in Figure 3.6. This increased robustness in subject identification could be attributed to the distinctive content composition of the Flickr30k dataset. The nature of the Flickr30k images predominantly centers around people engaged in diverse everyday scenarios. As a consequence, the model has been extensively exposed to these particular entities enabling it to preserve better performance in identifying subjects even when subjected to perturbations.

Tables 3.4a, 3.4b, 3.4c, 3.4d and Tables 3.5a, 3.5b, 3.5c, 3.5d provide comprehensive summaries of the *Machine-Machine (M-M)* and *Machine-Human (M-H)* evaluations following the perturbation experiments conducted on VF, CT, WE, and HT components,

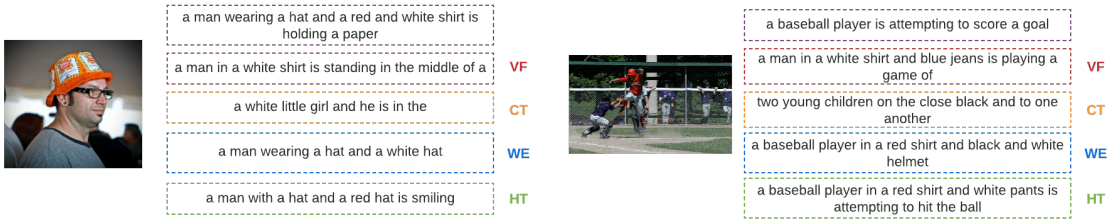


Figure 3.6: Qualitative examples of perturbation on Flickr30k test set.

respectively, using the MSCOCO2017 test set. The presentation of results follows a systematic approach to highlight the impact of perturbation magnitudes, denoted by σ, on each component. These tables encapsulate valuable insights into the performance of the model under various perturbation conditions.

For each σ value, which signifies the extent of perturbation, the final results are presented as the average scores across the 30 iterations in the first row. Additionally, to gauge the influence of randomness, the standard deviations for each metric are reported in the second row. As a point of reference, it is important to reiterate the evaluation metrics employed and their respective ranges of definition: BLEU-4 [0,1], CIDEr-D [0,10][4], SPICE [0,1], ROUGE-L [0,1], METEOR [0,1], and MSICE [0,1].

Controlling for randomness:
In line with experimental procedures, the reliability of our measurements can be influenced by inherent stochasticity. To address this, we employ an approach to quantify the extent of dispersion within our measurement outcomes, encompassing each σ value and component configuration. A noteworthy observation, as can be observed in Tables 3.4a, 3.4c, 3.4b, 3.4d and Tables 3.5a, 3.5c, 3.5b, 3.5d, is the remarkably low standard deviations displayed across our measurements. This underlines the fact that each iteration consistently exhibited similar performance behaviors for both evaluation configurations (M-M/M-H), thereby minimizing the potential influence of random factors and substantiating the validity and reproducibility of our experimental results.

Comparing Scores: Understanding Sensitivity and Influence
An insightful lens through which we examine our perturbation experiments is by analyzing their comparative impact across different components, as elaborated in Tables 3.4a, 3.4c,

[4]In the literature, CIDEr-D is multiplied by a factor of 10 to make CIDEr-D scores numerically similar to other measures.

σ	BLEU-4	CIDEr-D	SPICE	ROUGE-L	METEOR	MSICE
0.1875	0.5684	5.3097	0.6150	0.7253	0.4093	0.6442
	0.0021	0.0212	0.0017	0.0013	0.0010	0.0013
0.3750	0.3000	2.6216	0.3610	0.5241	0.2657	0.4447
	0.0019	0.0166	0.0017	0.0013	0.0008	0.0018
0.7500	0.1127	0.7211	0.1350	0.3261	0.1432	0.1861
	0.0014	0.0102	0.0013	0.0015	0.0008	0.0019
1.5000	0.0485	0.1897	0.0402	0.1787	0.0633	0.0574
	0.0013	0.0058	0.0011	0.0016	0.0009	0.0013

(a) Visual Features (VF)

σ	BLEU-4	CIDEr-D	SPICE	ROUGE-L	METEOR	MSICE
0.1875	0.7706	7.4036	0.7945	0.8599	0.5279	0.7540
	0.0043	0.0464	0.0040	0.0030	0.0028	0.0030
0.3750	0.6275	5.8390	0.6718	0.7714	0.4415	0.6802
	0.0047	0.0524	0.0043	0.0034	0.0026	0.0034
0.7500	0.3728	3.3176	0.4679	0.6149	0.3194	0.5521
	0.0027	0.0230	0.0026	0.0020	0.0013	0.0029
1.5000	0.0633	0.4643	0.1467	0.2917	0.1116	0.1985
	0.0015	0.0086	0.0020	0.0015	0.0009	0.0027

(b) Context Vectors (CT)

σ	BLEU-4	CIDEr-D	SPICE	ROUGE-L	METEOR	MSICE
0.1875	0.7785	7.4913	0.8016	0.8654	0.5337	0.7569
	0.0047	0.0534	0.0044	0.0031	0.0034	0.0027
0.3750	0.6396	5.9709	0.6819	0.78	0.4489	0.6815
	0.0041	0.0427	0.0038	0.0023	0.0022	0.0028
0.7500	0.4337	3.8493	0.5138	0.6536	0.3486	0.567
	0.004	0.0418	0.0043	0.0028	0.0019	0.0033
1.5000	0.1785	1.579	0.3182	0.4812	0.2453	0.3962
	0.0024	0.0176	0.0025	0.0021	0.0012	0.0027

(c) Word Embeddings (WE)

σ	BLEU-4	CIDEr-D	SPICE	ROUGE-L	METEOR	MSICE
0.0469	0.8577	8.3793	0.8708	0.9134	0.5940	0.7989
	0.0029	0.0325	0.0028	0.0018	0.0026	0.0021
0.0938	0.7620	7.3058	0.7862	0.8551	0.5225	0.7472
	0.0038	0.0425	0.0036	0.0024	0.0025	0.0024
0.1875	0.6201	5.7628	0.6651	0.7689	0.4390	0.6706
	0.0038	0.0374	0.0034	0.0028	0.0022	0.0031
0.3750	0.3956	3.5411	0.4888	0.6385	0.3419	0.5540
	0.0031	0.0296	0.0034	0.0024	0.0015	0.0026

(d) Hidden States (HT)

Table 3.4: M-M evaluation of latent components perturbation

σ	BLEU-4	CIDEr-D	SPICE	ROUGE-L	METEOR	MSICE
0.1875	0.3059	0.9787	0.1868	0.5292	0.2556	0.5561
	0.0013	0.0027	0.0005	0.0008	0.0004	0.0013
0.3750	0.2193	0.7225	0.1457	0.4604	0.2135	0.4651
	0.0016	0.0034	0.0006	0.0011	0.0006	0.0018
0.7500	0.1128	0.2643	0.0708	0.3397	0.1448	0.2434
	0.0012	0.0025	0.0006	0.0011	0.0006	0.0022
1.5000	0.0541	0.0639	0.0237	0.2150	0.0750	0.0862
	0.0013	0.0014	0.0004	0.0012	0.0007	0.0015

(a) Visual Features (VF)

σ	BLEU-4	CIDEr-D	SPICE	ROUGE-L	METEOR	MSICE
0.1875	0.3365	1.0685	0.1986	0.5503	0.2675	0.5714
	0.0013	0.0035	0.0006	0.0008	0.0005	0.0016
0.3750	0.3267	1.0449	0.1959	0.5450	0.2637	0.5653
	0.0024	0.0057	0.0009	0.0013	0.0008	0.0019
0.7500	0.2732	0.9207	0.1829	0.5154	0.2455	0.5388
	0.0023	0.0055	0.0011	0.0012	0.0008	0.0024
1.5000	0.07330	0.2359	0.0722	0.3169	0.1227	0.2559
	0.0016	0.0038	0.0009	0.0016	0.0008	0.0030

(b) Context Vectors (CT)

σ	BLEU-4	CIDEr-D	SPICE	ROUGE-L	METEOR	MSICE
0.1875	0.3371	1.0700	0.1990	0.5506	0.2679	0.5716
	0.0015	0.0032	0.0006	0.0010	0.0005	0.0016
0.3750	0.3301	1.0555	0.1971	0.5468	0.2654	0.5661
	0.0023	0.0056	0.0008	0.0010	0.0006	0.0021
0.7500	0.2982	0.9830	0.1894	0.5290	0.2550	0.5454
	0.0025	0.0061	0.0009	0.0017	0.0009	0.0020
1.5000	0.1814	0.6824	0.1633	0.4595	0.2225	0.4647
	0.0021	0.0053	0.0010	0.0015	0.0008	0.0035

(c) Word Embeddings (WE)

σ	BLEU-4	CIDEr-D	SPICE	ROUGE-L	METEOR	MSICE
0.0469	0.3389	1.0744	0.1993	0.5516	0.2686	0.5728
	0.0016	0.0042	0.0006	0.0009	0.0005	0.0013
0.0938	0.3362	1.0687	0.1986	0.5508	0.2679	0.5708
	0.0016	0.0037	0.0007	0.0009	0.0005	0.0019
0.1875	0.3257	1.0463	0.1963	0.5452	0.2644	0.5654
	0.0025	0.0050	0.0008	0.0014	0.0007	0.0020
0.3750	0.283	0.953	0.1894	0.5233	0.2554	0.5444
	0.0031	0.0062	0.0010	0.0015	0.0010	0.0024

(d) Hidden States (HT)

Table 3.5: M-H evaluation of latent components perturbation

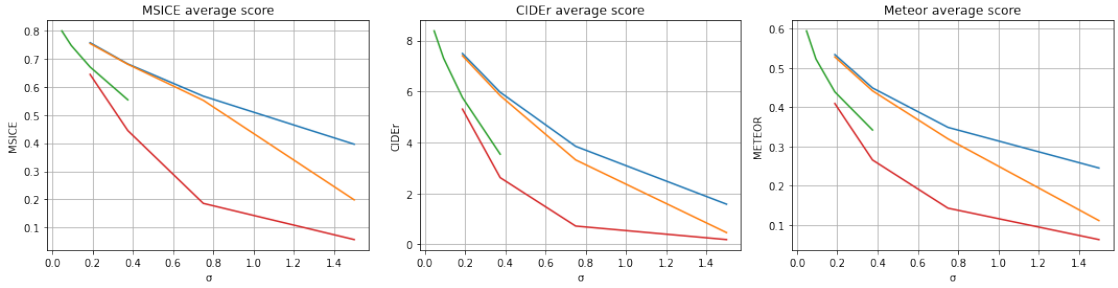
3.4b, 3.4d and Tables 3.5a, 3.5c, 3.5b, and 3.5d. Evidently, discernible patterns emerge, highlighting distinct behaviors among the various components in response to perturbations. Specifically, for the Visual Features (VF) and Context Vectors (CT) components, there is a discernible trend of significantly diminishing scores as the perturbation magnitude escalates. This suggests that these components are more susceptible to perturbations, resulting in reduced performance. In contrast, the Word Embeddings (WE) and Hidden

States (HT) components demonstrate robust stability in their respective scores despite varying degrees of perturbation. This observation suggests that the captioning model exhibits a heightened resilience against the latter perturbations.

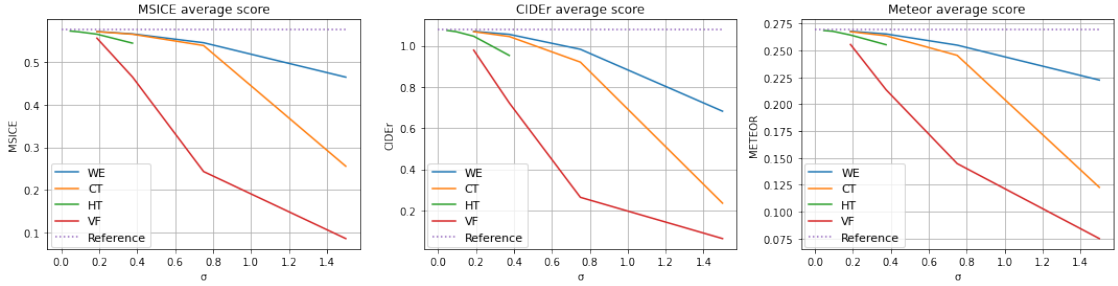
To visually depict these trends, Figure 3.7a portrays curves that capture the impact of perturbation for each component, assessed through the *M-M* evaluation. These graphics display the average perturbation scores of MSICE, CIDEr-D, and METEOR on MSCOCO2017's test set. Higher scores mean lower sensitivity, hence lower influence, and vice-versa. straight lines in purple represent the reference scores (i.e. the captions generated without perturbation compared to ground-truth). This graphical representation highlights a parallel behavior between VF and CT components exhibiting a steeper descent in their curves, as well as between WE and HT components that demonstrate a more gradual decline. MSICE emerges as a cohesive thread that aligns harmoniously with other metrics, evincing a smoother penalty curve in response to perturbations.

Transposing our analysis to the *M-H* evaluation mode, Figure 3.7b broadly mirrors the patterns observed in Figure 3.7a on the various components. This synchronization of trends between *M-M* and *M-H* evaluations effectively neutralizes potential biases stemming from human-generated references in the context of our explanation approach. However, some subtle variations in the values obtained between the two evaluation modes appear, with the *M-M* evaluation tending to produce higher values compared to the *M-H* evaluation. This discrepancy was expected, as the reference captions in the *M-M* evaluation originate from the captioning model before perturbation, making them closer in nature and structure to the captions generated by the same model after perturbation than those curated by human annotators.

Figure 3.8 depicts the curves obtained for the two evaluation modes on the Flickr30k dataset, mirroring the trend observed in the MSCOCO dataset with some distinctions. Notably, the decay rate of the curve representing the CT component perturbation appears swifter than that of the VF component. This reversal of behavior prompts an inquiry into the reasons underlying this shift when compared to the MSCOCO dataset. We attribute this phenomenon to a confluence of factors linked to the intrinsic characteristics of the Flickr30k dataset. This dataset is renowned for its emphasis on human-centric scenes, featuring complex human interactions and detailed scenarios. In such datasets, subjects (humans) and their interactions often carry paramount semantic significance. Given the pivotal role of the attention mechanism in accurately highlighting these salient entities and their relationships within the scene, any perturbations affecting the recognition of these



(a) *M-M* evaluation



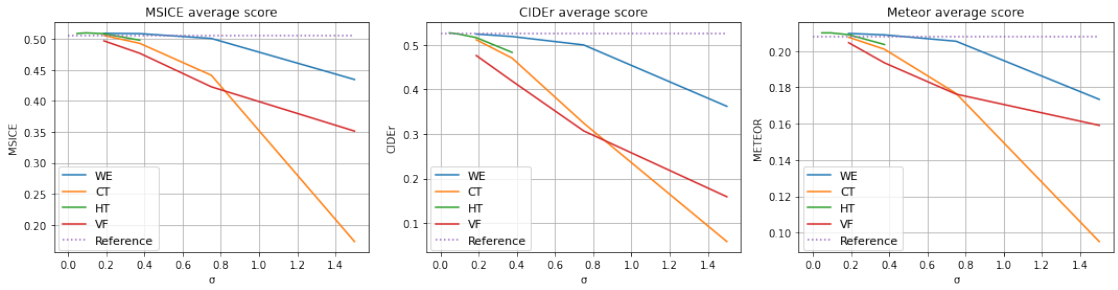
(b) *M-H* evaluation

Figure 3.7: Graphical representations of perturbation scores on MSCOCO2017 test set.

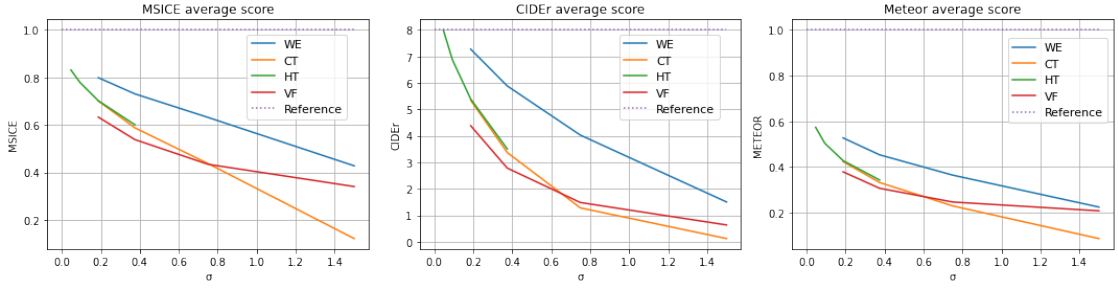
interactions can lead to greater distortion in attention weights, consequently amplifying the impact of perturbation relative to other components.

Irrespective of the dataset considered, a key rationale behind these findings can be attributed to the varying dominance levels of the visual and linguistic modalities, along with their distinct roles in vision-to-language tasks. The imbalance in the potency of these modalities fosters a deeper involvement of the first one in the captioning process, rendering it more susceptible to alterations and changes. This nuanced observation forms a pivotal cornerstone for further explanatory endeavors.

Our experiment unveils that the visual part, composed of the image encoder and attention mechanism, stands as the central element in IC architectures. The greater sensitivity of visual features and context vectors to perturbations translates into a more pronounced influence on the decision process compared to the linguistic part, encompassing word embeddings and the LSTM decoder. An illuminating study by [Sun et al., 2022] conducted a word ablation experiment. In this experiment, they attempted to generate a target word by dropping the three preceding words in the generated captions. The outcome displayed a nearly 50% decline in the quality of the newly generated word following random



(a) *M-H* evaluation



(b) *M-M* evaluation

Figure 3.8: Graphical representations of perturbation scores on Flickr30k test set.

ablation and an 85% decline when the three words with the highest relevance scores were deleted (words that significantly contributed to predictions as indicated by LRP scores). These findings suggested that the ablated words held crucial importance for the model in predicting target words. However, it is crucial to note that the ablation process introduces potential gaps within the sequence. This raises ambiguity as to whether the decline in scores can be attributed to the words' significance for generating the target word, their sequential positioning, or the inconsistency in the structure of the previous sequence due to ablation. This can also shed light on the internal nature of vectors. Specifically, it may suggest that visual vectors tend to be more compositional, while language vectors exhibit greater density and complexity. Further research in this area could make a significant contribution to our understanding of the inner workings of these vectors.

In our WE perturbation, we sidestep sequence gaps by altering all words within the sequence rather than removing them. This approach ensures that captions maintain structural coherence, and counteracts positional bias. Here, the only variable is the magnitude of perturbation, which signifies the extent of changes made to preceding words before generating the current one. Under such assumptions, we can infer that the primary reason behind captions' resistance to WE perturbation lies in the relatively weak influence

of this component (WE). In contrast, VF perturbation leads to a marked drop in caption quality. Analogous to WE, this perturbation strategy averts introducing gaps in the image, unlike ablation, by modifying all extracted features simultaneously. CT perturbation yields comparable effects owing to the pivotal role of attention mechanisms in the model. In fact, context features encapsulate the degree of attention required by the word predictor for each visual feature during word generation. Such perturbation disrupts the focus of the word predictor, resulting in a random emphasis on the image content during decoding. We deduce that both the VF and CT components exert a higher influence on captioning decisions than the WE component.

The final experiment, focusing on the hidden states (HT) of the LSTM decoder supports our findings. It reveals significant stability in caption quality post perturbation, implying that the preceding sequence encoded by HT vectors has minimal impact on word prediction. Notably, the word predictor takes as input both context vectors, encompassing the visual features and their attention weights, and the hidden states. This elucidates that the word predictor prioritizes visual information over linguistic information, accentuating the role of the visual modality. It may also be plausible that the reduced sensitivity of the language component to perturbations emanates from its lower robustness compared to the visual component. Recent endeavors in IC have ventured into enhancing captioning architectures by intervening at the language decoder level [Herdade et al., 2019, Li et al., 2019], particularly through the substitution of the RNN decoder with a transformer module [Vaswani et al., 2017]. Surprisingly, these methods have not yielded significant progress and have, at times, even yielded lower scores compared to classical CNN+RNN models. On the other hand, the study by [Liu et al., 2021] shows that the integration of transformers in the visual part markedly improves caption quality. This underscores the notion that the language component's insensitivity -or low sensitivity- to perturbations is not indicative of its robustness but rather its modest influence. Our current findings reinforce the argument that the linguistic modality exerts a relatively weak impact on captioning decisions when juxtaposed with the powerful influence of the visual modality. This comparison is facilitated by the operation of perturbation on a shared intermediate latent space. Moreover, our conclusions go beyond the faster R-CNN architecture and also encompass CNN encoder-based architectures, as both models converge on a common latent space, namely the image latent space, and share a common step when encoding the image input.

The nuanced dynamics within the component influence landscape seem to steer vision-to-language tasks, such as IC, towards a greater emphasis on enhancing and explaining the visual modality, encompassing encoding and attention mechanism. However, this emphasis on the visual part does not completely nullify the significance of the language component in the overall captioning process. Our study operates at a post-hoc stage, facilitating the evaluation of architecture components and offering insights that pave the way for more comprehensive investigations into the specific roles of each element. Notably, the language component appears to primarily play a stylistic role within the captioning process, while the visual component shoulders the responsibility of generating the lexical items in the output. This division underscores the distinct contributions of these two modalities in the creation of coherent and informative captions. We believe that understanding the behavior of the model's components serves as an inaugural but fundamental step towards achieving more explainability in black-box models. Such understanding can provide invaluable insights into their inner functioning. The identification and assessment of the influence exerted by each component can significantly shape subsequent choices for enhancing explainability. A component exhibiting a higher influence is likely to hold more potential for insightful explanations than its counterparts. The relevance and subtlety of these explanations will intrinsically depend on the levels of influence and sensitivity specific to each component.

Pioneering this quest for linguistic explanations, [Sun et al., 2022] laid the foundation. In this study, we go a step further by conducting comprehensive experiments that underscore the role of the linguistic modality in the realm of IC. Our findings illuminate that, in the pursuit of explanations, the linguistic modality may not wield as much significance as the visual aspect. This realization resonates particularly strongly in several fine-tuned IC domains, more acutely in scenarios marked by high risks and contexts where the cost of an erroneous decision or explanation is exceptionally steep. Take, for instance, biomedical IC, such as describing X-rays. In such contexts, the utmost precision is crucial in pinpointing the reasons behind specific words within a caption, which often depict pathologies. In such a case, it is necessary to explain to the end-users that a detected pathology in a radiology examination emanates from a particular region in the image, and this explanation must specifically exclude any ambiguity caused by previously generated words in the caption, which could potentially be misleading. In line with our first definition of decisiveness presented at the outset of this chapter, a clear demarcation can be made between the contribution of the visual modality manifested through visual regions and that of the linguistic modality materialized by the words in the generated caption. The former

guarantees the accurate prediction of target pathologies, while the latter constructs the scientific description of the medical image, both cooperating to deliver relevant captions. A fair conclusion is that explanations of specific diseases should be sought primarily in the visual component, rather than in the linguistic side.

3.8 Conclusion

In this chapter, we introduced an explanation approach for deep learning models in image captioning (IC), exploiting the representation space of information. We demonstrated that explainability in deep models extends beyond establishing links between outputs and inputs, and does not necessarily have to be perceptive. To do so, we proposed a perturbation paradigm, where information within deep models is modified rather than truncated. Perturbations proved to be a more coherent and insightful means of dissecting model behavior, transcending mere ablation. Central to our endeavor is the novel MSICE evaluation protocol, which stands as a new advancement in the evaluation of IC and explainability. This protocol, rooted in morphology and semantics, addresses the intrinsic challenge of evaluating sentences with similar semantics but different surface forms. Our results illuminate that this approach enhances the evaluation process and offers a finer-grained assessment of captions. The most pivotal revelation of this study is the discernment of the substantial influence exerted by the visual component within captioning architectures, overshadowing the impact of the language element. Consequently, we expect the visual part to constitute a key point for subsequent model's explanations. Beyond its impact on the field of IC, our findings bear significance for a broader array of vision-language tasks. The current results hold the potential to catalyze the development of accurate, subtle, and trustworthy explanations across various domains, ensuring that the synergy between vision and language yields insights that are both complementary and dependable.

4 Attribution-based explanations

Contents

In our prior research [Elguendouze et al., 2022], we established that the visual components of IC models stand as key elements for subsequent explanatory endeavors. Building upon this insight and on the distinctive strengths gleaned from the representation space, in this chapter we embark on a more intricate exploration of the visual modality. Our goal is to extract more overt insights shedding light on the complex workings of captioning models and latent space.

To achieve this, we extend the work of [Sun et al., 2022] by adapting the LRP (Layer-wise Relevance Propagation) explanation technique to IC models that leverage bottom-up

(BU) visual features extracted via Faster-RCNN. This adaptation marks a departure from the conventional use of global CNN features. Simultaneously, we exploit latent space to develop a variant of the popular attribution-based explanation method called LIME. It is worth reiterating that attribution methods are those that operate in post-hoc explainability mode, assigning relevance scores, or attributions (Section 2.2), to inputs based on their importance for output predictions. While LRP and LIME are conventionally deployed on the original input space, we propose a shift to the latent space, thereby emphasizing its role in generating finer-grained explanations.

It is worth noting that BU features diverge from CNN features by being anchored in local image pixels. Each BU feature encapsulates a specific region of the image, unlike CNN features that encompass the entire image as a global entity. The above distinction, outlined earlier in Section 3.3, aligns with our choice of a standard architecture founded on BU features. This choice retains its validity, as contemporary captioning models widely adopt BU features. Moreover, the incorporation of these features augments our capacity to unveil further insights about the captioning pipeline, the way in which the input image is analyzed, as well as the intricate encoding and decoding of data within the latent space.

Operating within the latent space while subjecting attribution-based explanation methods, LRP and LIME, to a meticulous comparative analysis based on their distinct operational scopes (Section 2.2.6), offers us a deeper grasp of how the scope and mechanisms of these methods influence the quality of explanations. It is important to distinguish the term "scope" in its contextual usage here, referring to the operational range, not the scope of explanation outcomes. LRP and LIME, while both producing attributions as explanations, diverge in their approach, LRP following a backpropagation paradigm and LIME adopting a perturbation-based paradigm in their explanation generation.

4.1 Bottom-up Layer-wise Relevance Propagation Explanations

4.1.1 Background of LRP

Layer-wise Relevance Propagation (LRP) [Bach et al., 2015] is an explainability technique aimed at uncovering the contribution of input features to output prediction. This method, as introduced earlier in Section 2.2.6, operates through a backward propagation process. Given a neural network, LRP back-propagates the final prediction (output) along the

network by recursively assigning a relevance score to each neuron within the network, ultimately reaching the input layer. Specifically, this approach dissects the relevance of individual neurons by iteratively re-distributing the scores to each of the neurons in the preceding layer. This decomposition is guided by various propagation rules defined by [Bach et al., 2015], including ϵ-rule (presented hereafter) and β-rule. LRP unfolds in two main stages, illustrated in Figure 4.1: on the left, the forward pass, where predictions are made by the neural network model during inference, and on the right, the backward pass, where the relevance is retrogressively disseminated.

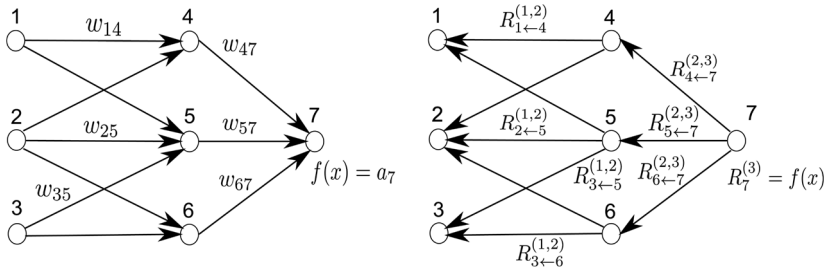


Figure 4.1: Forward inference Vs. relevance propagation [Bach et al., 2015].

Consider a neural network composed of a sequence of interconnected neural layers, following a layer-wise architecture. A neuron denoted as x_j within a given layer (l) is characterized by a linear transformation of all neurons from the preceding layer (l - 1), followed by an activation function $g(\cdot)$ (as depicted in Equation 4.1). Here, x_i represents an input neuron, y_j denotes the linear output, x_j is the output following activation, and w_{ij} corresponds to the network weights.

$$x_j = g(y_j) \;\; ; \;\; y_j = \sum_i x_i w_{ij} + b_j \tag{4.1}$$

$$R^{(l\text{-}1,l)}_{i\leftarrow j} = R^{(l)}_j \;\cdot\; \frac{x_i w_{ij}}{y_j \;+\; \epsilon} \tag{4.2}$$

$$R^{(l\text{-}1)}_i = \sum_j R_{i\leftarrow j} \tag{4.3}$$

Given a neuron x_j in a specific layer (l) with a relevance score $R^{(l)}_j$, LRP employs a decomposition rule, such as $\epsilon - rule$ (Equation 4.2), to assign a contribution score $R^{(l\text{-}1,l)}_{i\leftarrow j}$ to each input neuron x_i from the preceding layer (l - 1). Regardless of the propagation rule used to compute the relevance scores, this essentially quantifies the proportional influence of neuron x_i among all neurons within the same layer that collectively contribute to the

computation of neuron x_j. In the case of $\epsilon - rule$, the parameter ϵ works as a stabilizer, preventing the relevance scores $R_{i \leftarrow j}$ from becoming unbounded, particularly when the linear output y_j attains small values. Finally, to derive the overall relevance score for neuron x_i, the contributions stemming from all incoming connections within layer (l) to neuron x_i are summed, as illustrated in Equation 4.3.

4.1.2 LRP for Bottom-up captioning architectures

Adapting Layer-wise Relevance Propagation (LRP) to the bottom-up (BU) captioning architecture employed in the Ada-LSTM model outlined by [Sun et al., 2022] involves customizing the method to align with the distinctive components of this BU-based architecture. The Ada-LSTM captioning model encompasses a Faster-RCNN encoder, an adaptive attention module, and an LSTM decoder. The authors in [Sun et al., 2022] stated that, as far as the entire captioning architecture is concerned, the LRP rules adhere the same topological flow as backpropagation (component by component). In this regard and in line with the architecture's sequence, we initialized the relevance scores for the words within the output caption using the logits generated by the last layer of the word predictor sub-module. We back-propagate each word's relevance score $R(w_t)$ across the architecture. This process traverses through the Word predictor, LSTM decoder, and Attention components, halting at the encoder's output, as illustrated in Figure 4.2. The feature referring to a tennis ball attained the highest importance score for predicting the word "tennis" in the caption. Of particular note is our focus on the BU visual features that were originally employed for prediction during the forward pass. These visual features, totaling V in number, have their individual relevances computed through the LRP mechanism, tailored to the specifics of the BU architecture.

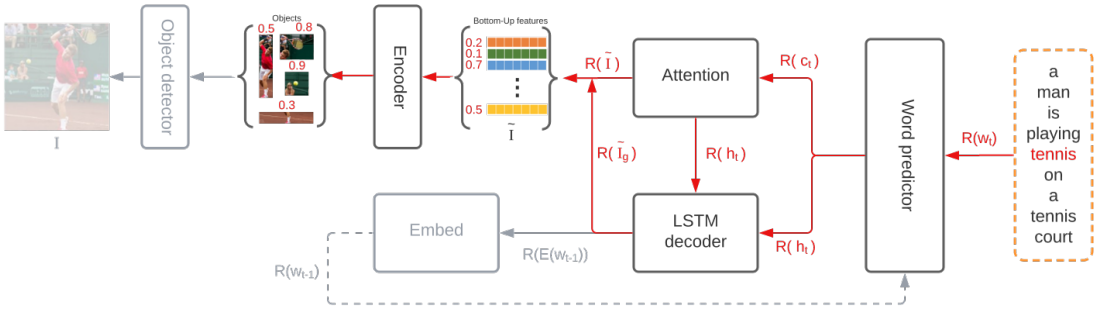


Figure 4.2: LRP flow through the BU-based captioning architecture.

At each component's output within the backpropagation pathway, we derive the relevance scores corresponding to the respective data flow. Specifically, denoted as $R(c_t)$, $R(h_t)$, $R(\tilde{I}_g)$, and $R(\tilde{I})$, these relevance scores respectively pertain to the context vectors, hidden states, global image feature, and BU image features, respectively. Our primary focus lies in investigating the relevance of the BU features concerning the final word prediction. This emphasis on BU features stems from our prior findings (as highlighted in Section 3.8) that, during this phase of explanation, the linguistic modality exhibits comparatively reduced significance.

The relevance scores vector $R(v_i)$ for each individual BU feature v_i obviously aligns with the shape of the feature vector, i.e. $R(v_i) \in \mathbb{R}^{d_v}$. Consequently, the global relevance score for a specific BU feature is the summation of all the elements $R(v_{ij})$ within the relevance vector. This process yields a coefficient vector $\hat{\alpha} \in \mathbb{R}^V$ as the final outcome of the explanation (Equation 4.4), encompassing the importance of all BU features to the prediction of a given word within the caption.

$$\hat{\alpha} = R(\tilde{I}) = \{R(v_i)\}_{i=1}^{V} \quad , \quad R(v_i) = \sum_{j=1}^{d_v} R(v_{ij}) \tag{4.4}$$

4.2 Bottom-up Local Interpretable Model-Agnostic Explanations

4.2.1 Background of LIME

LIME, an acronym for Local Interpretable Model-Agnostic Explanations, is a perturbative method that can be used to explain any black-box model. This methodology becomes particularly valuable in scenarios where the complexity of the original model prevents it from being explained globally. LIME's central idea lies in its ability to transform complex models into locally interpretable explanations around each input instance. It achieves that by introducing perturbations to the original input, generating similar yet slightly altered neighbor instances. These perturbed instances are then utilized to predict new outcomes and train a linear model whose inputs and labels are the perturbed instances and the associated outputs. The resulting coefficients of the linear model provide local explanations in the form of importance attributions assigned to individual input features based on their role in predicting a specific output. It is crucial to clarify the terminology here. Throughout the subsequent sections of this chapter, the term "Intrinsic perturbations"

will refer to the perturbations intrinsic to the LIME technique itself. This should be differentiated from the Gaussian perturbations introduced in Chapter 3, which served a distinct purpose in the previous context.

The core procedure of a typical LIME approach can be distilled into five key steps, as illustrated in Figure 4.3. For a given input instance requiring explainability, LIME initiates by generating a set of P perturbed samples in the vicinity of the input. These modifications might involve conventional operations like darkening or blurring, particularly applicable to image data. Subsequently, both the original input and these perturbed samples are passed to the original black-box model to predict their respective labels (e.g. classes for image classification). This set of instances, coupled with their predicted labels, serves as the training dataset for constructing a simple linear model designed to mimic the intricate behavior of the complex original model. The coefficients learned by the surrogate model then offer insights into the importance of input features in making the prediction. In this manner, LIME is capable of approximating the behavior of a black-box model locally, without necessitating any knowledge about the model's internal mechanisms.

4.2.2 LIME for Bottom-up captioning architectures

In this section, we introduce our BU-LIME method by adapting the principles of Local Interpretable Model-Agnostic Explanations (LIME) [Ribeiro et al., 2016] to the context of IC in architectures that employ BU visual features. To develop BU-LIME, we adhere to the same underlying rationale as in the original LIME framework, albeit with a distinct approach to perturbation. Specifically, we adopt gradual intrinsic Gaussian perturbations in latent space (i.e. visual features). These perturbations show good performances and better consistency compared to the alternative of altering the original image through techniques like blurring and blackening [Sahay et al., 2021]. Such conventional operations can often lead to issues like data truncation or information loss.

The process of BU-LIME can be outlined as follows. Initially, we generate a set of P neighboring instances $\Gamma = \{\acute{\tilde{I}}^{(p)}\}_{p=1}^{P}$ around a given image I. Each perturbed instance is obtained by introducing random perturbations to a subset of its visual features. The selection of this feature subset can be approached in various ways, which will be detailed in the experimental section. However, it is crucial to ensure a degree of randomness in feature selection while maintaining consistency in the number of selected features. A binary vector $X^{(p)} \in \{0,1\}^V$ is associated with each perturbed instance $\acute{\tilde{I}}^{(p)}$ through an indexing mechanism, where '1' denotes perturbed features and '0' represents unperturbed

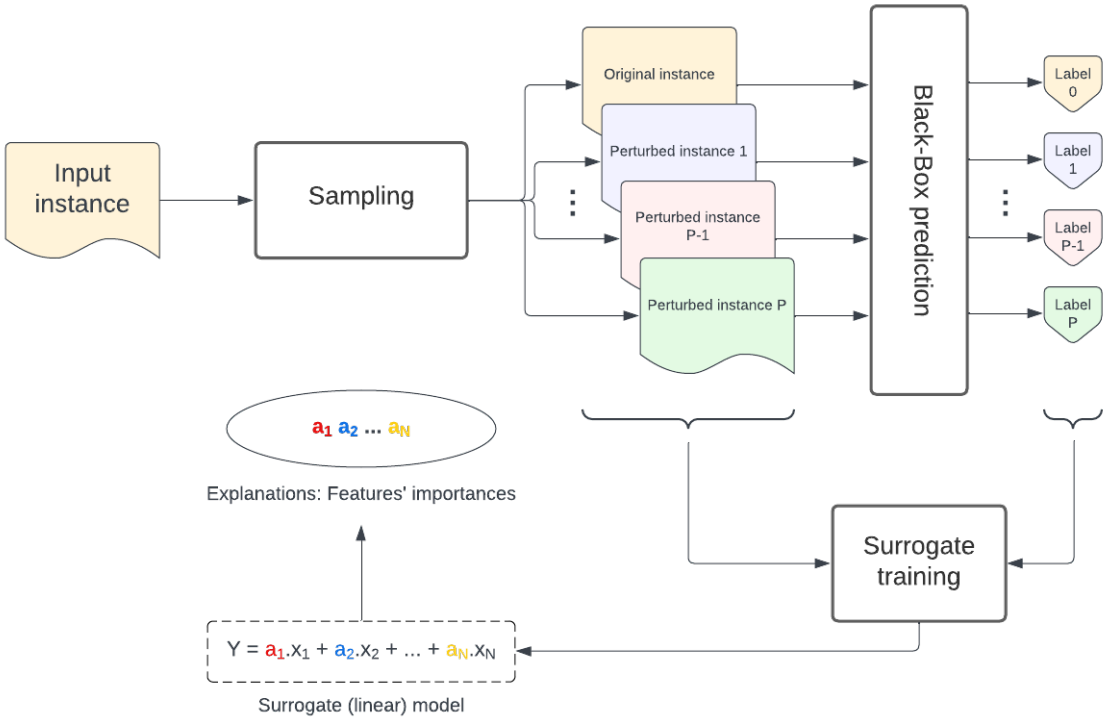


Figure 4.3: Overview of LIME.

features. Here, V represents the number of BU features, and the set of all binary vectors $X = \{X^{(p)}\}_{p=1}^{P}$ is denoted as the binary matrix $X \in \{0,1\}^{P \times V}$. Subsequently, each perturbed instance is fed into the captioning model. While the generated captions themselves lack coherence due to the applied perturbations, our focus is on the generated weights (logits from the last layer of the LSTM) associated with every word in the vocabulary during caption prediction.

Since the probability of each word is computed during every decoding step, we aggregate the maximum probabilities from all decoding steps to derive the final probability. For instance, if we consider the word 'man' and obtain weights of 0.5, 0.3, 0.33, 0.4, and 0.7 across five decoding steps for the first perturbed instance, the final weight for that word is 0.7. This process is repeated for all words in the vocabulary for each perturbed instance. Upon this stage, we obtain a weight matrix $\Delta = (\delta_{pq}) \in \mathbb{R}^{P \times Q}$, where Q represents the size of the vocabulary.

To obtain explanations for a specific word w_q from the vocabulary, we build a linear regression model (Equation 4.5) using the paired data (Binary matrix X, Weight vector y) as the training set, where $y = \delta_{.q} \in \mathbb{R}^P$ represents the q^{th} column of Δ. Each training instance is composed of a pair (X_p, y_p). The cost and objective functions of the regression model are given in Equations 4.6 and 4.7 respectively. $\beta \in \mathbb{R}^V$ denotes the vector of coefficients to be estimated, while $\hat{\beta}$ is the estimator, and $\gamma \in \mathbb{R}^V$ is a noise vector.

$$Y = X \cdot \beta + \gamma \tag{4.5}$$

$$cost = \frac{1}{P}\|Y - X\beta\|_2^2 + \lambda\|\beta\|_2^2 = \frac{1}{P}\sum_{p=1}^{P}(\ y_p - \sum_{v=1}^{V}(X_{pv} * \beta_v)\ )^2 + \lambda\sum_{v=1}^{V}{\beta_v}^2 \tag{4.6}$$

$$\hat{\beta} \in argmin(cost) \tag{4.7}$$

The entire approach is illustrated in Figure 4.4. In this visualization, the coefficients of the linear model reflect the importance of individual visual features in predicting the word 'man', with the highest coefficient corresponding to the feature representing a man. A concise and formal summary of this approach is provided in Algorithm 1. In this algorithm, w_q denotes the word for which we seek explanations, $\hat{\beta}$ represents the explanation vector (Equation 4.7), Γ represents the set of all perturbed instances, X is the set of all perturbation indexes, and y is the vector of predicted captioning weight. The functions *perturb*, *capt*, and *linear* denote the perturbation, captioning, and linear regression functions, respectively.

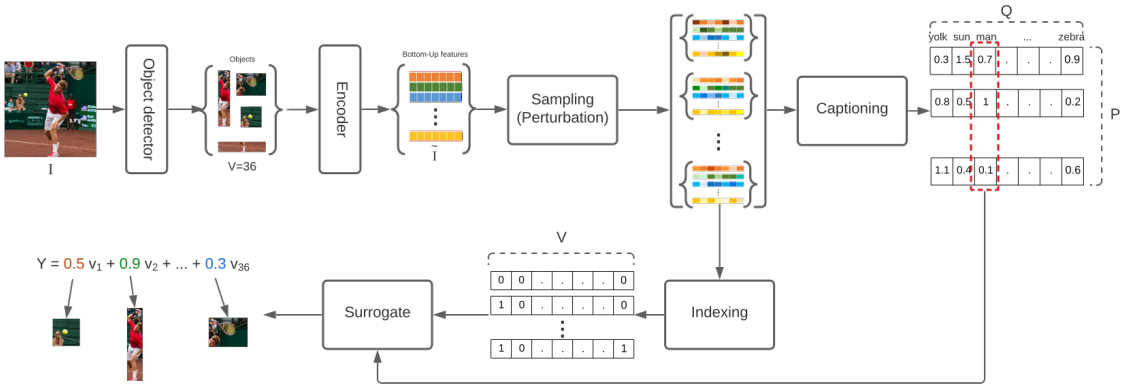


Figure 4.4: BU-LIME method overview.

Algorithm 1 BU-LIME algorithm.

Inputs: $\tilde{I}, q$
Outputs: $\hat{\beta}$
$\Gamma \leftarrow \{\varnothing\}$
$X \leftarrow \{\varnothing\}$
for $p \in \{1, ..., P\}$ **do**
  $\acute{\tilde{I}}^{(p)}$, $X^{(p)} \leftarrow perturb(\tilde{I}^{(p)})$
  $\Gamma \leftarrow \Gamma \cup \{\acute{\tilde{I}}^{(p)}\}$
  $X \leftarrow X \cup \{X^{(p)}\}$
  $\delta_{p.} \leftarrow capt(\acute{\tilde{I}}^{(p)})$
end for
$y \leftarrow \delta_{.q}$
$\hat{\beta} \leftarrow linear(X, y)$
return $\hat{\beta}$

Knowing that images often encode regions or objects through multiple visual features, which can sometimes introduce redundancy or ambiguity, we propose in the next section a more complete investigation. This entails exploring the effect of manipulating such features through an extended version of the LIME-based explanation method, wherein we shift our focus from individual visual features to entire objects within the image. The key difference lies in the intrinsic perturbation process of LIME.

4.2.3 Object aware BU-LIME

Referred to as BU-LIME-N-OBJ, this new method involves perturbing for up to N objects within the image. The number of perturbed instances per image, in this case, is determined by $P = \sum_{i=0}^{N} C_n^i$, with n representing the number of objects detected in the image. Obviously, we only retain images where $n \geq N$. We experiment with two values of N (5 and 8). Our choice to avoid lower values for N is because this could result in an insufficient number of instances to train the LIME linear model. For instance, with $N = 2$ and $n = 7$, we would have $P = 29$ instances. Perturbing entire objects amounts to manipulating all the visual features associated with that particular object, irrespective of whether the corresponding region precisely delimits the object or only represents a partial detection. To this end, we leverage the class labels generated by the object detector to select all the features that correspond to each individual object, ensuring a comprehensive perturbation of the object's representations.

Interestingly, through empirical exploration, we've found that the maximum number of unique objects within an image rarely exceeds 15, not accounting for overlapping or redundant regions. Consequently, we adopt the practice of perturbing at most half of the detected objects, aiming to mitigate the risk of excessive perturbation. It is worth noting that from experimental observations, employing a larger number of perturbed objects tends to penalize the learning of the linear model. For a clearer grasp of the concept of redundant visual features, please refer to the illustrative example in Figure 4.7.

4.3 Evaluation of explanation quality

The existing research on IC explainability assessment predominantly revolves around a qualitative [Xu et al., 2015, Huang et al., 2019, Han and Choi, 2018, Sahay et al., 2021] evaluation (e.g. saliency maps) and quantitative [Sun et al., 2022] evaluation (e.g.ablation measures), focusing on the fidelity aspect of generated attributions. These approaches entail either a visual validation of the correctness of the explanation with regard to the explained item or quantifying the disparity between a new prediction, obtained by the omission of the explanation element from the input, and the original prediction. However, the visual validation of the explanation's quality might not capture subtle or complex patterns in the data, and it can be time-consuming and resource-intensive when applied to large datasets or complex models. The traditional masking of the explanation element in the input, also called ablation, to evaluate the prediction changes, inherits the same limitations as conventional perturbation techniques, like blackening and blur, as already discussed in the previous section.

In this section, we introduce two novel evaluation techniques. The first method gauges the correlation between the explanations and the object detection scores, while the second relies on the Latent Ablation principle. These two evaluation metrics hold the potential to not only advance explanation evaluation but also establish comparative benchmarks. The current dearth of evaluation metrics in the realm of explainability poses a challenge to this evolving research domain. The ensuing sections provide detailed descriptions of both evaluation methods.

4.3.1 Correlation measure

This metric measures the correlation between the explanations and the classification scores of the regions (objects) jointly generated by the object detector during image encoding.

However, this evaluation does not involve an absolute comparison between the explanations generated by the attribution methods and the object detector probabilities (Faster-RCNN). Rather, we have developed an evaluation strategy that projects the compared items into a unified ranking space. Building upon our earlier findings regarding the role of the visual modality, our initial motivation in employing this evaluation method is to gain insights into how well the explanations align with the object detection scores, a component integral to the visual modality. This evaluation serves as a complementary perspective to other evaluation methods that will be discussed subsequently.

Consider the candidate (predicted) caption $\hat{C}$ for a given image I. Let $O_{\hat{C}}$ represent the object words that appear in $\hat{C}$, and let O_V denote the labels corresponding to the visual objects detected by the Faster-RCNN. The evaluation process focuses exclusively on the explanations for the words that jointly appear within $O_{\hat{C}}$ and O_V. The core concept involves quantifying the disparity between the importance of a given region v_i in the prediction of its label $l \in \{O_{\hat{C}} \cap O_V\}$ within the output caption, and the class (label) probability as anticipated by the object detector. This could be achieved by using the explanation vectors $\hat{\alpha}$ and $\hat{\beta}$ derived from the BU-LRP and BU-LIME methods respectively when applied to the word l. These vectors contain the attribution (importance) scores assigned to the BU features. The elements of each vector are sorted and indexed in descending order using the *sort_index* function resulting in $E_{\hat{\alpha}}$ and $E_{\hat{\beta}}$, respectively (Equation 4.8). Therefore, the features (regions) considered most relevant for predicting the word l will occupy the initial positions in these vectors, with less important ones following suit. The evaluation consists in measuring the distance that separates the position of the region v_i in both $E_{\hat{\alpha}}$ and $E_{\hat{\beta}}$ as determined by *position* function, and its corresponding ranking from Faster-RCNN's output (i.e. the index i). Note that the Faster-RCNN ranks the regions in descending order according to their class probabilities, bringing those with the most confident prediction at the top of the list and the least certain further down the list.

$$E_{\hat{\alpha}} = sort\_index(\hat{\alpha}) \;\; ; \;\; E_{\hat{\beta}} = sort\_index(\hat{\beta}) \tag{4.8}$$

$$d_i^{\hat{\alpha}} = max(\ 0\ ,\ position(E_{\hat{\alpha}}, v_i) - i\ ) \tag{4.9}$$

$$s1_i^{\hat{\alpha}} = \frac{1}{1\ +\ d_i^{\hat{\alpha}}} \in [0, 1/(V+1)] \tag{4.10}$$

$$s2_i^{\hat{\alpha}} = 1 - \frac{d_i^{\hat{\alpha}}}{V} \in [0, 1] \tag{4.11}$$

Nevertheless, if the region v_i attains a position in the sorted vectors $E_{\hat{\alpha}}$ and $E_{\hat{\beta}}$ that is equivalent to or surpasses its position in the ranking provided by the object detector (in terms of importance), the resulting distance is reduced to the minimum value '0' (as outlined in Equation 4.9). In such situations, this would mean that the explanation produced by the concerned method precisely correlates with the classification score associated with the predicted label for a given BU feature. The computed distance is subsequently transformed into a similarity score [Segaran, 2007], further quantifying the alignment between the explanation and the object detection score, as expressed by Equation 4.10. However, we propose a new similarity score in Equation4.11, which seems to more accurately capture this correlation concept by providing a fairer penalization for medium and long distances. In fact, consider the maximum possible distances that can separate the position of the feature index in the explanation vector and the one in the classification scores vector, which, in this case, is equal to V (let's take V=36 as an example). According to Equation4.10, the similarity score reaches a minimum value of $s1 = 1/37$ in this scenario, whereas it attains $s2 = 0$ according to Equation4.11. Similarly, if we take a distance equal to one-third of the maximum distance, i.e., 12, the similarity scores would be $s1 = 1/13$ based on the first similarity formula, whereas it would reach $s2 = 1/2$ for the second one. This demonstrates that the second similarity formula better matches our quest for a fair score in such situations.

4.3.2 Latent Ablation measure

In addition to the correlation metric defined above, we also propose Latent Ablation. The original concept of ablation, as depicted in Figure 4.5, consists of removing specific explanation elements from the input data to assess the resulting impact on prediction. For instance, in their study, [Sun et al., 2022] evaluated the quality of captioning explanations, both visual and linguistic, by masking the top 20 high-relevant image patches (parts) for predicting a specific word in the output caption, as well as the top 3 high-relevant words in a separate experiment. The objective is to eliminate their contribution to the prediction and measure the change in the output caption's quality compared to the original one.

Our *Latent Ablation* differs from conventional ablation methods by operating in the latent space (features) rather than the original input space (i.e. images, as in [Sun et al., 2022]), aiming to explore the potential benefits offered by this space with regard to explainability evaluation. Analogous to our earlier proposal in Section 4.2 where we discussed two versions of the LIME-based explanation approach (visual features

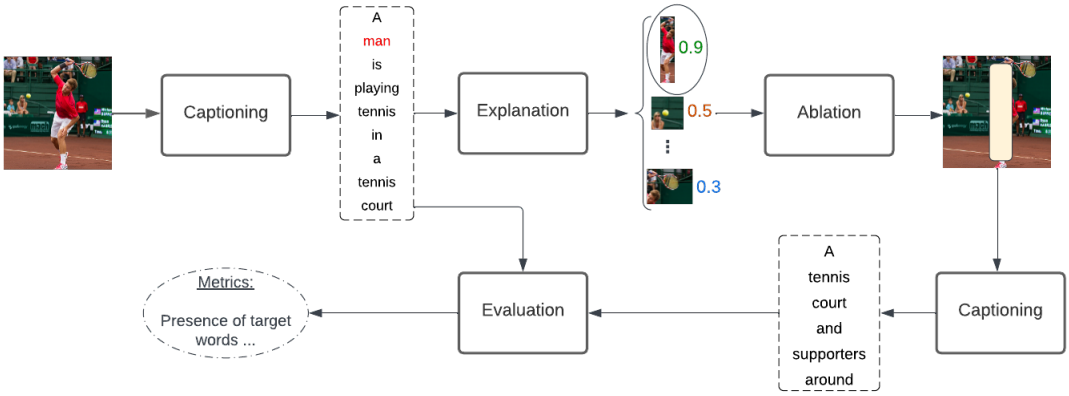


Figure 4.5: Standard ablation applied to XIC evaluation.

versus objects), we extend this concept to Latent Ablation. Here, we suggest investigating two variants of Latent Ablation: the first involves masking the k most important visual features, while the second masks the top k entire objects, both within the latent space. This dual approach enables us to delimit the impact of handling individual visual features versus whole objects when assessing the explanatory capacity of the proposed methods.

The process of masking the top k visual features, as depicted in Figure 4.6, amounts to neutralizing (ablating) those features that are deemed most influential for the final prediction while keeping the other features unaltered. We conduct this procedure with various ablation magnitudes. The first level of intensity operates by applying Gaussian perturbations to the visual features. It is essential not to confuse this perturbation with the Gaussian perturbation discussed in the first chapter or the intrinsic perturbation employed in the BU-LIME explanation method. The second one involves saturating the visual features by setting them to either the minimum or maximum values of the VF component (refer to Section 3.5.3). Following this, we re-generate the image caption and examine whether the word associated with the explanation remains present in the new caption. Our evaluation includes all object words, excluding stop words and predicates. The results are reported as the percentage of missing words, which reflects the fidelity of the explanation elements in relation to their corresponding predictions.

Algorithm 2 provides a formal synthesis of the ablation process for BU-LRP explanations applied to a single explanation instance (word). In this regard, $\tilde{I}$ represents the set of visual features, C is the original caption, w^* is the target word for explanation, $\hat{\alpha}$ denotes the BU-LRP explanation vector, k stands for the number of features for ablation,

mode captures the ablation configuration, and *eval* encapsulates the ablation evaluation function. It is important to note that the same process applies to BU-LIME explanations, and the instances/words considered for evaluation are identical.

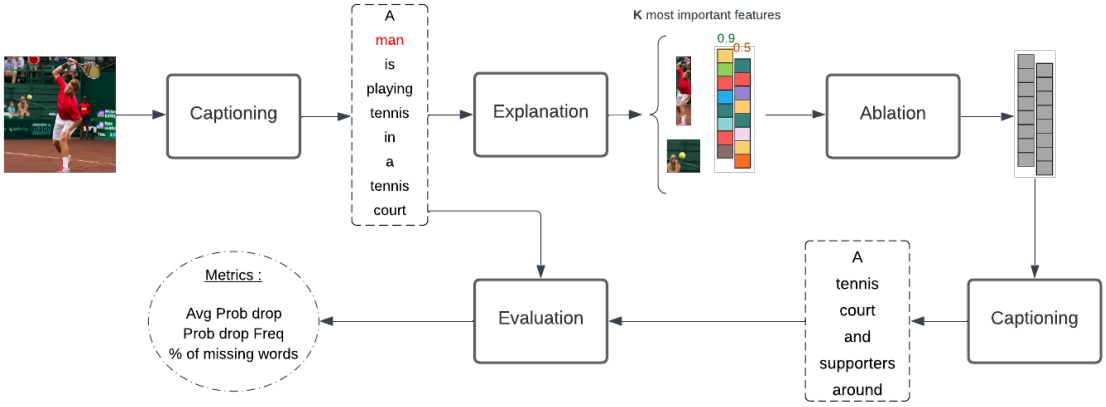


Figure 4.6: Latent ablation for XIC evaluation.

Algorithm 2 Latent feature ablation algorithm.

Inputs: $\tilde{I}, C, w^*, \hat{\alpha}, k, mode$
Outputs: $\hat{C}, \hat{\mathcal{P}}$
Global variables: $Min^{VF}, Max^{VF}, \sigma$
for $i \in \{1, .., k\}$ **do**
  if $mode =$ Normal **then**
    $\tilde{I}_{\hat{\alpha}_i} \leftarrow \tilde{I}_{\hat{\alpha}_i} + \mathcal{N}(0, \sigma)$
  else if $mode =$ Min **then**
    $\tilde{I}_{\hat{\alpha}_i} \leftarrow Min^{VF}$
  else if $mode =$ Max **then**
    $\tilde{I}_{\hat{\alpha}_i} \leftarrow Max^{VF}$
  end if
end for
$\hat{C}, \hat{\mathcal{P}} \leftarrow capt(\tilde{I})$
return $\hat{C}, \hat{\mathcal{P}}$

The second version of ablation involves masking the top k objects rather than individual features. An object is represented by a set of visual features, and each visual feature corresponds to a single object within the image. For a specific object, there may be multiple visual features generated, often detected by Faster-RCNN as partial portions of the object. These partial areas typically possess lower label probabilities than the entire object as the object detector is less confident in assigning the correct object category to the partial detection. For instance, Figure 4.7 illustrates the object detection outcomes

on an image representing a man playing tennis. In this case, the main object "man" is enclosed by several bounding boxes, each assigned a different probability (man 83%, man 80%, man 35%...). Consequently, the latent space accommodates multiple visual features that correspond to various segments of the main object "man". These features often represent fragments of the entire object "man", which is assigned the highest probability.

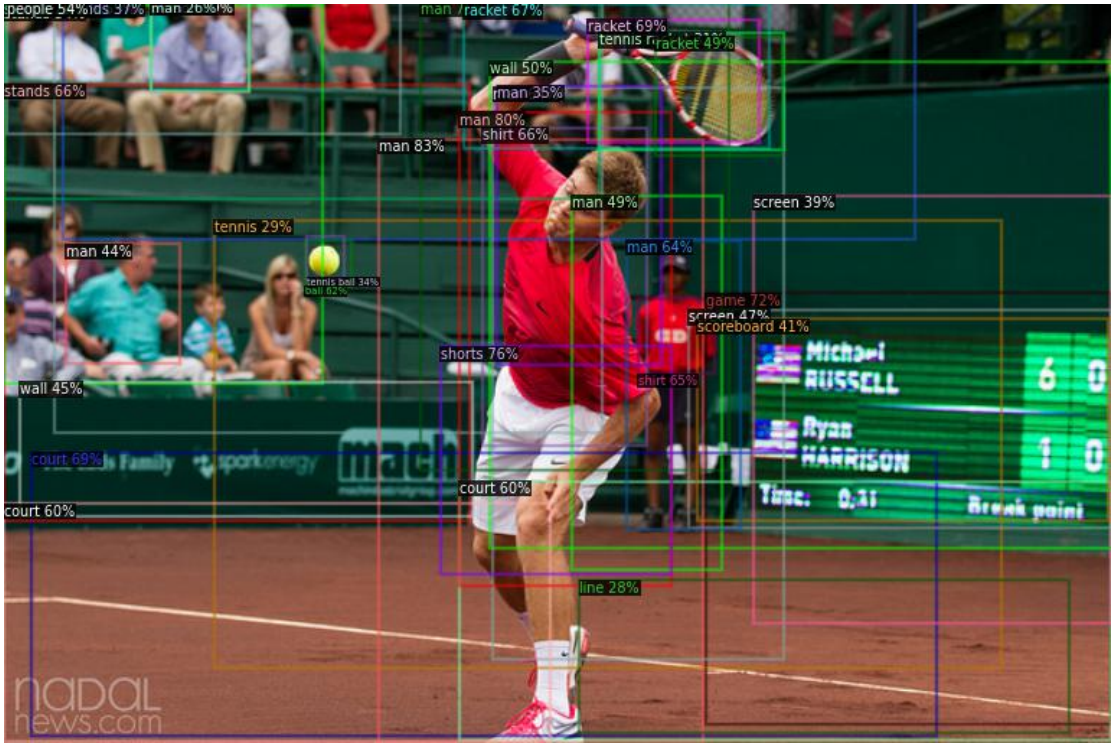


Figure 4.7: Illustrative example of object detection results using Faster-RCNN.

The selection of the top k objects is determined by the ranking of their first occurrence in the explanation vector, which is composed of indices representing the 36 extracted features, ranked by importance. This well-ranked position corresponds to the most probable feature among all those associated with the same object.

As discussed earlier in this section, conventional ablation experiments are commonly evaluated by measuring changes in new captions following ablation and comparing them to captions before ablation. This assessment often involves quantifying the presence or absence of target words or assessing caption quality through established metrics such as CIDEr and ROUGE. In our latent ablation method, besides quantifying the *absence of target words*, we introduce two additional evaluation measures: *average probability drop* and *probability drop frequency*.

The average probability drop reflects the decrease in a word's probability during the prediction of the new caption. Given that words' probabilities are generated for the entire vocabulary in each decoding step, only the maximum value is retained, indicating the highest level of confidence the decoder attains for that particular word during caption decoding. Subsequently, probability drop frequency is defined as the number of instances where such a decrease in word probability occurs. An instance here refers to a word considered for explanation. Algorithm 3 defines the evaluation function, which takes two inputs: the reference prediction (prior to ablation) consisting of the vocabulary probabilities $\mathcal{P}$, and the new prediction post ablation, comprising the new caption and new vocabulary probabilities $\hat{\mathcal{P}}$. The f_O function stands for the object words extractor from a caption text.

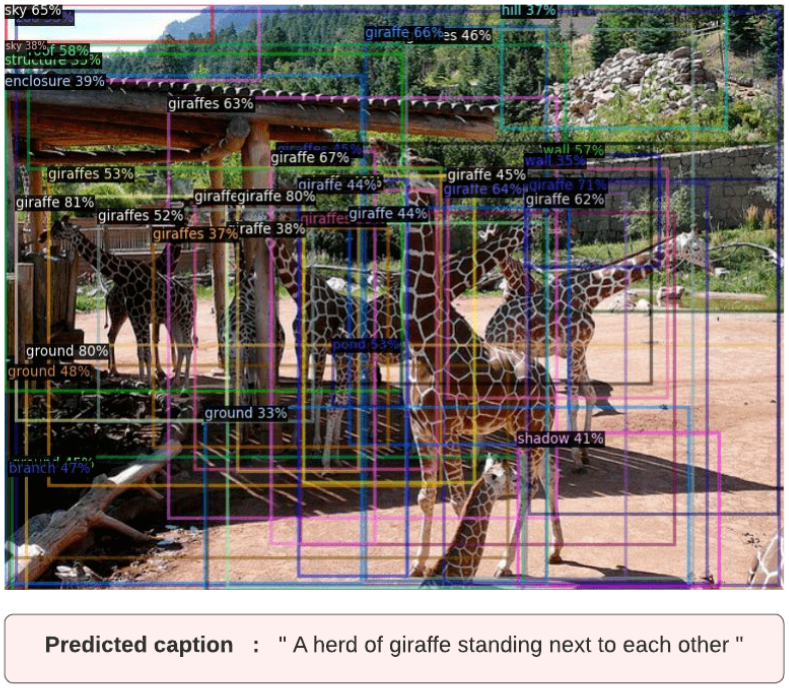
Algorithm 3 Explanation fidelity assessment algorithm.

Inputs: w^*, $\hat{C}$, $\mathcal{P}$, $\hat{\mathcal{P}}$
Outputs: $missing\_word$, $prob\_drop$, $drop\_freq$
$\lambda \leftarrow f_O(\hat{C})$
if $w^* \notin \lambda$ **then**
    $missing\_word \leftarrow 1$
else
    $missing\_word \leftarrow 0$
end if
$prob\_drop \leftarrow max(0, \mathcal{P}_{w^*} - \hat{\mathcal{P}}_{w^*})$
if $prob\_drop \neq 0$ **then**
    $drop\_freq \leftarrow 1$
else
    $drop\_freq \leftarrow 0$
end if
return $missing\_word$, $prob\_drop$, $drop\_freq$

While quantifying the *absence of the target words* relies only on the captions, the other two metrics, *average probability drop* and *probability drop frequency*, employ word probabilities. All three metrics are applied across all evaluation instances, considering only object terms within the captions as target words for explanation. The final fidelity scores are obtained by averaging individual scores across the entire test set.

4.4 Results and discussion

In this section, we conduct an evaluation and comparative analysis of the explanations provided by the two distinct explanation methods, BU-LRP and BU-LIME. Our evaluation protocol employs the two evaluation methods presented in Section 4.3: correlation with object detection scores and Latent Ablation, specifically tailored for evaluating the explainability of BU-based IC models. Both BU-LRP and BU-LIME operate within the representation space and rely on the visual component to derive explanation elements. However, they differ in the nature of their scope, with a global scope for BU-LRP and local explanations for BU-LIME. Given this distinction, it is interesting to assess the precision of their outcomes in relation to this criterion.



0	giraffe	81%
1	giraffe	80%
2	ground	80%
3	giraffe	71%
4	giraffe	67%
5	giraffe	66%
6	sky	65%
7	giraffe	64%
8	giraffes	63%
9	giraffe	62%
10	roof	58%
11	wall	57%
12	giraffes	55%
13	pond	53%
14	giraffes	53%
15	giraffes	52%
16	giraffe	50%
17	giraffe	49%
18	ground	48%
19	branch	47%
20	giraffes	46%
21	giraffe	45%
22	ground	45%
23	giraffes	45%
24	giraffe	44%
25	giraffe	44%
26	shadow	41%
27	enclosure	39%
28	giraffe	38%
29	sky	38%
30	hill	37%
31	giraffes	37%
32	structure	35%
33	zoo	35%
34	wall	35%
35	ground	33%

Figure 4.8: Object detection applied to an image featuring a herd of giraffes.

Figure 4.8 illustrates object detection applied to an image featuring a herd of giraffes. Detected regions are encoded as BU features and assigned category labels through classification, accompanied by classification scores/probabilities. Figure 4.9 displays BU-LRP's explanation outcomes at the level of individual visual features. The two heatmaps/matrices indicate the relevance scores attributed to each element of the first and third BU feature

vectors (v_0 and v_2) in the prediction of the word "giraffe" belonging to the image caption in Figure 4.8. These two selected features correspond to the labels "giraffe" and "ground" with probabilities of 81% and 80%, respectively. Cumulative importance scores for these two visual features ($R(v_0) = 0.00051$ and $R(v_0) = -0.000045$) are reported above the heatmaps. These scores are derived by averaging the elements of each feature's explanation vector, as indicated earlier in Equation4.4).

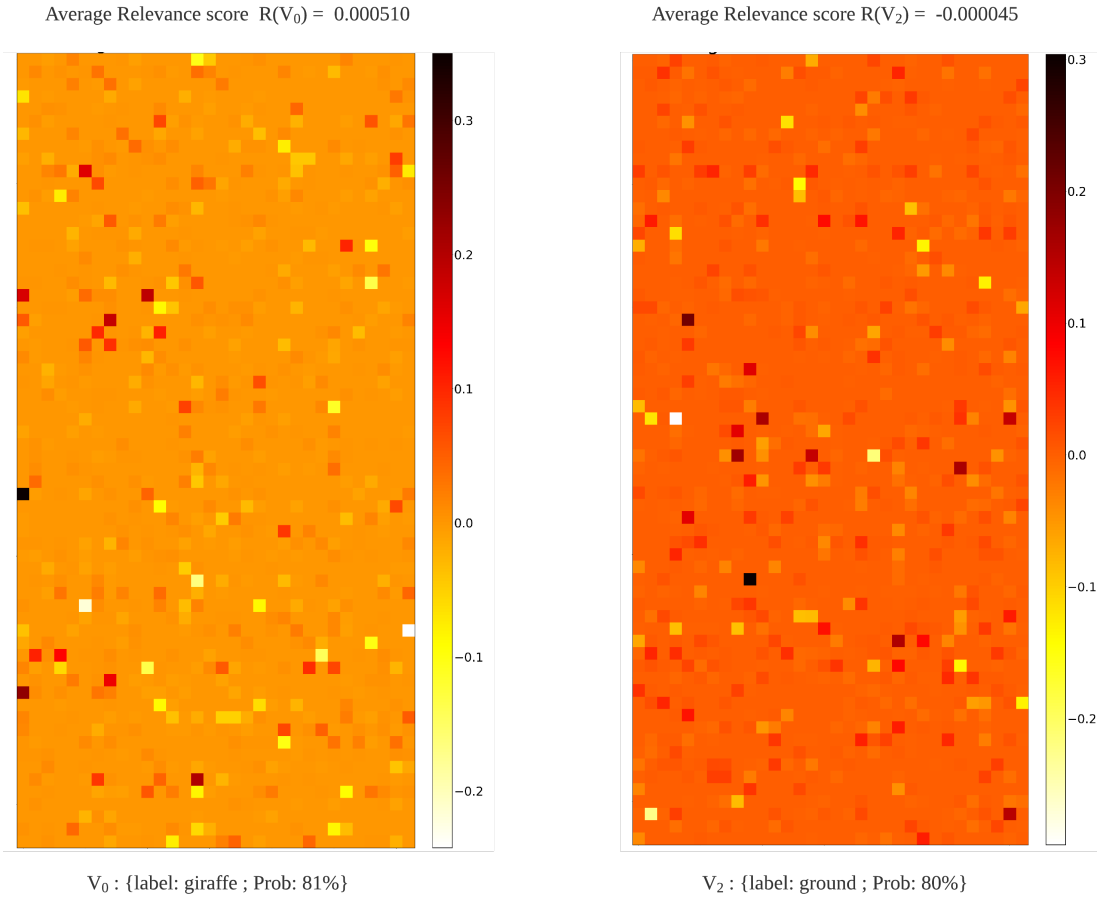


Figure 4.9: Visualization of importance distribution explanations for individual visual features.

It is important to note that, for the sake of simplicity of visualization, the importance vectors, each comprising 2048 elements, are displayed here as compact heatmaps of dimension (64*32). In these heatmaps, warmer colors correspond to higher importance values, while cooler colors indicate lower values. One notable observation is that explanation values within the same vector tend to exhibit a remarkable level of homogeneity. This is

demonstrated by the consistency of heatmap colors within the same feature's importance vector, with a few exceptions characterized by notable high or low importance values. This pattern hints at the possibility that certain dimensions of a specific feature could be disproportionately influential, either negatively or positively, in predicting a particular word within the caption.

In cases where the contribution is extremely high, these dimensions may encapsulate the most crucial information within the corresponding feature's region, such as the patterns representing the core concept present within that particular region of the image. Conversely, when confronted with very low contribution values, these dimensions could potentially impede the prediction of the desired word, i.e. "giraffe" in this example. This may arise due to the presence of information or patterns that either counteract the essence of the studied concept (in other words, they cannot coexist simultaneously in the output) or, intriguingly, contain information that, when viewed in a negating context, actually aligns with the target concept. Nonetheless, at this juncture, it remains challenging to definitively assert whether negative and positive (or high and low) relevance scores act in direct competition, as there is also merit in considering them as complementary contributions, mutually supporting to highlight different facets of the explanation.

At a broader scale, Figure 4.10 presents a good example of heatmap explanations generated for the word "giraffe" within the image caption showcased in Figure 4.8. The heatmaps in Figure 4.10a indicate the relevance distribution across the 36 visual features, while Figure 4.10b portrays the explanation vector derived from BU-LRP. For both heatmaps, warm colors indicate higher importance, and vice versa. Remember that these explanatory visualizations are obtained by applying the BU-LRP explanation method, reflecting the back-propagated contribution throughout the captioning architecture. The back-ward process starts with the end outcome, i.e. probability of the predicted word in the output caption, and reiterates back to obtain the importance of each element of the BU visual feature vector, and finally, the overall importance of each visual feature when averaged. For each word in the caption, we get 36 importance vectors displayed as heatmaps, each corresponding to a visual feature (Figure 4.10b).

In Figure 4.10, it becomes clear that some visual features exert a greater influence on the prediction of the word "giraffe". These features manifest as warmer-colored heatmaps, specifically those corresponding to the 9^{th}, 10^{th}, 18^{th}, 21^{st}, 27^{th}, and 28^{th} features, matching respectively with regions labeled as giraffe, roof, ground, giraffe, enclosure, and giraffe, respectively. For a clearer perspective, Figure 4.10b presents the ranking of visual

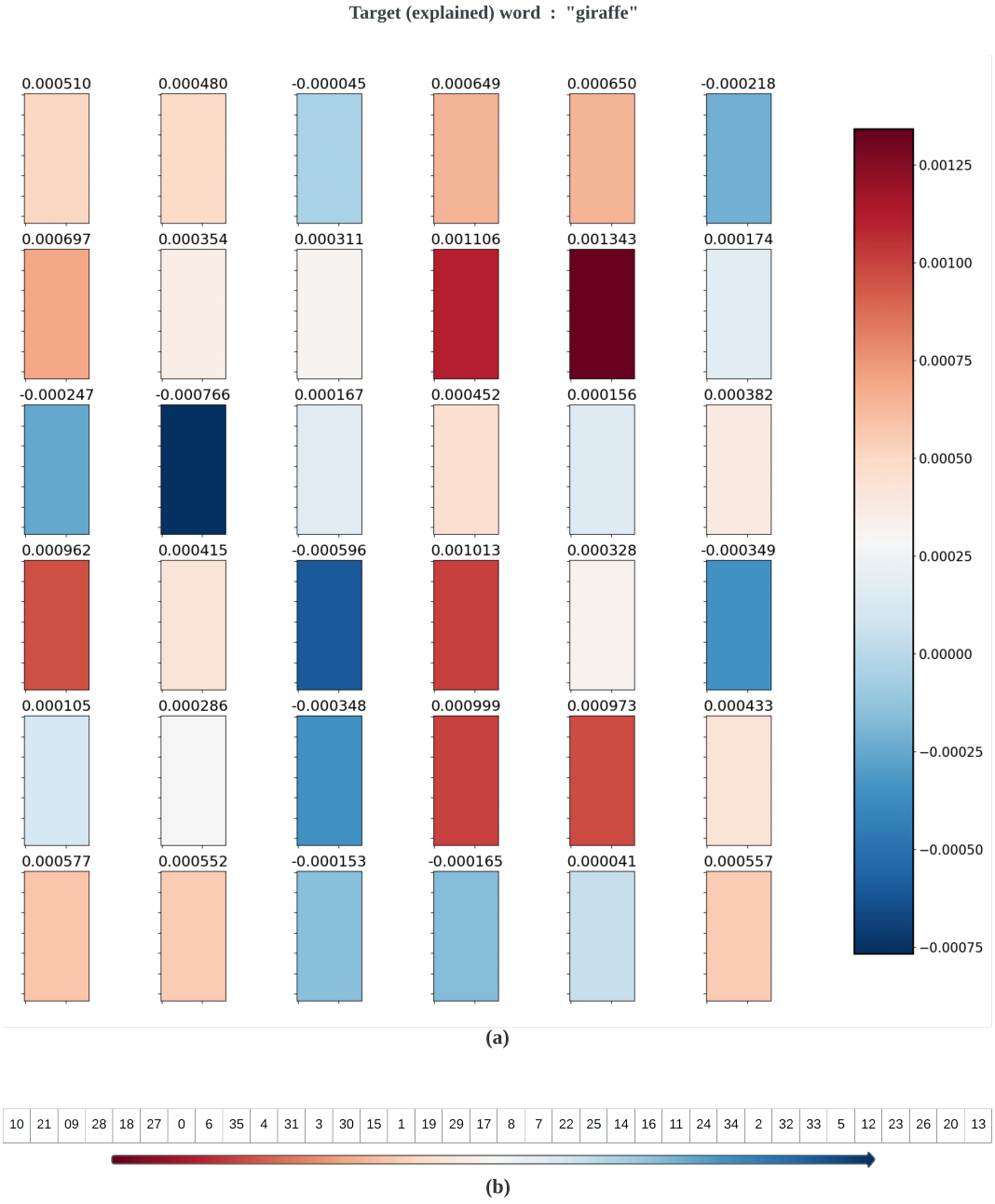


Figure 4.10: Larger-scale explanatory visualizations.

features based on BU-LRP relevance scores, ranging from the most to the least important. In contrast, certain features exhibit a negative contribution to the prediction of this same word "giraffe", notably the 12^{th} and 13^{th} features, corresponding to labels "giraffes" and

"pond" with probabilities of 55% and 53%, respectively. It may appear odd to observe a negative contribution for a feature representing the same entity as indicated by the word being explained (i.e., giraffe). This apparent paradox can be rationalized by considering that the captioning model, during its inference process, tends to focus on the most salient features, even if other features comprising similar content are also present. This could apply, for example, when the object detector generates multiple partial detections, leading the captioning model to focus primarily on the feature encapsulating the entire object.

However, it is important to note that not all instances adhere to this behavior. In some cases, features representing partial objects gain higher relevance scores than those representing entire objects. Given the complex interactions between visual features within the captioning model, their fusion, and aggregation, we opt to further discuss this concept in the third chapter of this manuscript. In the current chapter, however, our focus remains confined to studying and comparing the outcomes of attribution explanations.

4.4.1 The correlation of explanations to object detection scores

Table 4.1 reports the evaluation results on the MSCOCO2017's test set, presenting the total average distance $\overline{d}$ separating the position of a feature in the explanation vector from the rank assigned by the Faster-RCNN, along with the average correlation scores expressed in terms of two similarity scores $\overline{s1}$ and $\overline{s2}$, as described in Section 4.3.1. The *common_ words* and *non_ common_ words* represent the number of instances where the word to be explained either appears or does not appear (i.e. word to be ignored while computing scores) in the list of detected objects.

Our experimentation involved two versions of the BU-LIME model, each with distinct strategies for generating perturbed instances. In the case of BU-LIME-1-2, we performed a comprehensive perturbation approach, encompassing all combinations, for up to two visual features simultaneously. This led to a total of 676 instances per image, one of which remains unperturbed. On the other hand, BU-LIME-5 is built based on a random selection strategy, perturbing five features simultaneously. Given the large number of possible combinations (C_{36}^{5}) and the proportional increase in execution time, we constrained the number of perturbed instances to 50 per image. The insights provided by the results in Table 4.1 will be discussed in Section 4.4.3.

Illustrated in figure 4.11, the distribution shows how the elements of the explanation vectors stack up in terms of position against the object rankings identified by Faster-RCNN.

	common_words	*non_common_words*	$\overline{d}$	$\overline{s1}$	$\overline{s2}$
BU-LRP	8256	6422	10.2633	0.3754	0.7149
BU-LIME-1-2			13.9510	0.3492	0.6125
BU-LIME-5			13.7025	0.3418	0.6194

Table 4.1: Global correlation scores for BU-LRP and BU-LIME explanations.

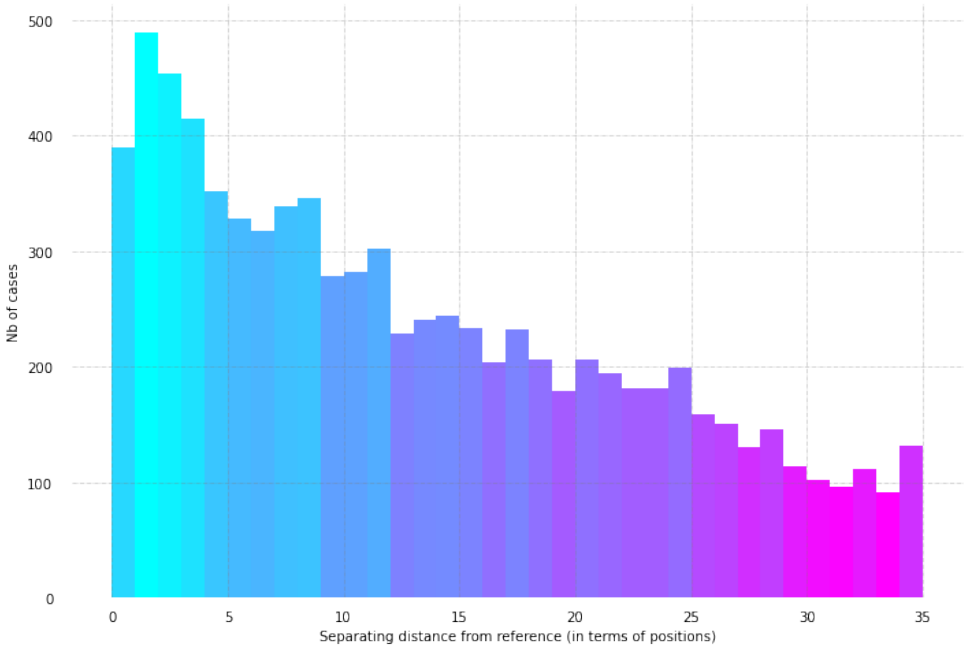
These distributions reveal any patterns followed by the explanation techniques, helping to determine whether the scores listed in Table 4.1 are influenced by specific peaks at particular positions, or if they show a more global trend. It also enables us to better assess the extent of the conclusions drawn from the aforementioned table.

4.4.2 Latent Ablation

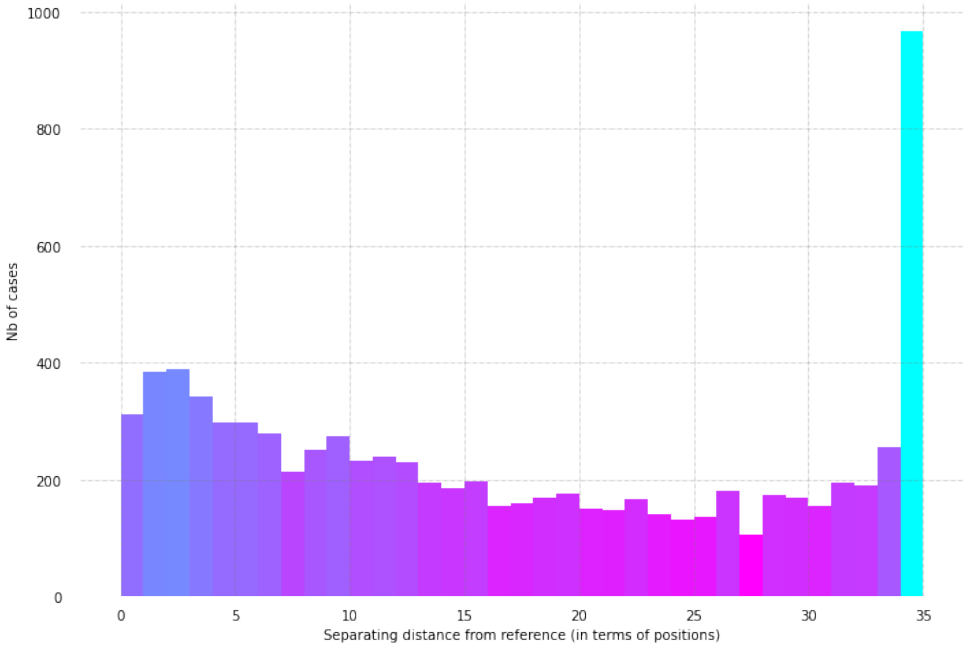
Latent Ablation experiments were conducted using various experimental configurations, as outlined in Section 4.3.2. These configurations involved latent ablation of either the individual visual features most influential in predicting object words in output captions, or of whole objects, based on their corresponding visual features. The main objective was to discern the potential impact of manipulating these two concepts. Experimental results are reported on the MSCOCO2017 test set.

Individual features manipulation

Table 4.2 shows the results of the **visual features ablation** for all explanation methods, expressed as the percentage of missing words in new captions following ablation. k denotes the total number of visual features masked simultaneously. It is important to remember that our visual feature ablation employs three distinct levels of intensity: Gaussian/Normal perturbations and saturation of features using the minimum (min) and maximum (max) values within the VF component (BU features). Specifically, the magnitude of normal ablation involves perturbing the visual features by adding Gaussian values with $\sigma = 1.5$, the maximum standard deviation achievable within the VF component. The min and the max magnitudes correspond to substituting the feature values with the minimum and maximum values of the respective vector dimensions. In addition to BU-LRP, and BU-LIME along with its two versions, we also include a random ablation baseline (explanations generated randomly) for comparison purposes.



(a) BU-LRP positions



(b) BU-LIME positions

Figure 4.11: Distribution of the distances separating the explanation elements from object detection rankings.

k	Explanation method	Ablation magnitude		
		normal	min	max
1	Random	0.1159	0.2823	0.2735
	BU-LRP	**0.1211**	**0.2840**	**0.2756**
	BU-LIME-1-2	0.1057	0.2797	0.2619
	BU-LIME-5	0.1089	0.2807	0.2648
3	Random	0.1403	0.5116	0.6094
	BU-LRP	**0.1445**	0.5170	**0.6125**
	BU-LIME-1-2	0.1160	**0.5189**	0.6015
	BU-LIME-5	0.1224	0.5154	0.5997
6	Random	0.1779	0.6881	0.8452
	BU-LRP	**0.1948**	0.6892	**0.8456**
	BU-LIME-1-2	0.1616	**0.6944**	0.8364
	BU-LIME-5	0.1559	0.6847	0.8340
9	Random	0.2338	0.7700	**0.9019**
	BU-LRP	**0.2418**	**0.7744**	0.9012
	BU-LIME-1-2	0.2029	0.7725	0.8945
	BU-LIME-5	0.1985	0.7703	0.8910

Table 4.2: Explanation coherence scores for individual features ablation experiment.

Object features manipulation

The second ablation experiment, manipulating whole objects instead of individual features, yields different results, as reported in Table 4.3. Evaluation here is expressed by the percentage of missing words, as well as two additional metrics: the average probability drop and the frequency of probability drops. The average probability drop quantifies the decrease in the prediction probability of a target word in the new caption after ablation. This decrease is measured in relation to the probability of the same word before ablation. The probability drop frequency, on the other hand, tracks the frequency of such a probability drop. It should be noted that, for brevity, only latent object ablation with the minimum value of the visual component is presented and discussed hereafter. This is based on the observation that the results showed similar trends for the three perturbation magnitude settings: *min*, *max*, and *normal*.

k	explanation method	Avg Prob drop	Prob drop Freq	% of missing words
1	Random	1.3521	0.9028	0.4460
	BU-LRP	2.1259	**0.9311**	0.5931
	BU-LIME-1-2	**2.1548**	0.9242	0.5922
	BU-LIME-5	2.1322	0.9224	**0.5960**
	BU-LIME-5-OBJ	1.3781	0.9011	0.4568
	BU-LIME-8-OBJ	1.1813	0.8908	0.4191
3	Random	2.8782	0.9507	0.7309
	BU-LRP	3.4346	0.9595	0.8033
	BU-LIME-1-2	**3.4776**	**0.9600**	0.8068
	BU-LIME-5	3.4396	0.9589	**0.8091**
	BU-LIME-5-OBJ	2.4333	0.9417	0.6598
	BU-LIME-8-OBJ	2.1437	0.9374	0.6284
5	Random	3.4810	0.9332	0.8061
	BU-LRP	3.8915	0.9623	0.8617
	BU-LIME-1-2	**3.9280**	**0.9640**	**0.8677**
	BU-LIME-5	3.9019	0.963 3	0.8654
	BU-LIME-5-OBJ	2.8814	0.9486	0.7414
	BU-LIME-8-OBJ	2.5787	0.9458	0.7097

Table 4.3: Explanation coherence scores for object features ablation experiment.

4.4.3 Discussion on attribution explanations

Based on the results presented in Table 4.1 within Section 4.4.1, the correlation scores observed across different explanation models appear promising, even better results may be expected. The average distance $\overline{d}$ separating the positions of explanations and the predictions from Faster-RCNN is close to one-third of the total number of positions (36), with a correlation score $\overline{s2}$ ranging between 61% and 71%. As can be noticed, the average similarity scores given by $\overline{s2}$ align better with the average distance scores $\overline{d}$, while $\overline{s1}$ shows a large mismatch, which confirms our hypothesis regarding these two metrics. The new metric provides better readability of results and better reflects method performance. These results indicate that, on average, explanations highlighting the most important features for the prediction of a specific word are displaced by around one-third of the expected position compared to Faster-RCNN predictions, which is fairly promising.

A closer look at the distribution of position distances, which reflect the gap between the explanations and the object detector's classification scores, reveals a clear outperformance of BU-LRP over BU-LIME in the context of this evaluation. In particular, the distribution is mostly concentrated towards the left side of the graph, implying smaller distances. As the distance widens, the number of matching explanation instances decreases. Remarkably, BU-LRP achieves its peak at distances (gaps) of one and two positions. On the other hand, BU-LIME's peak occurs at the maximum distance of 36 positions. The emergence of such a particular difference, while the rest of the distances show a comparable behavior to that of BU-LRP, can be attributed to several factors. The first one is the BU-LIME's intrinsic perturbations, which focus mainly on BU's characteristics. In essence, the detected regions within input images by Faster-RCNN are provided with their encoded features, together with their classification probabilities and labels. Manipulating these features by introducing considerable noise to build the linear model substantially alters the existing relationship between these features and their associated probabilities for certain "sensitive" instances. Consequently, this perturbation can introduce an artifact that surfaces during evaluation. This effect is particularly pronounced in the correlation evaluation, given that the reference criterion revolves around the probability scores of these perturbed features. It is important to note that the instances affected by this scenario constitute around one-fifth of the total instances evaluated, which remains within acceptable limits for the current context.

Elaborating on instances characterized by large position distances, this could be attributed to insufficient correlation between the probabilities generated by the object detection module (Faster-RCNN) and some specific instances of explanation. Indeed, the presence of several features designating the same object can lead to ambiguities during comparison. To illustrate this, let's consider the scenario described in Figure 4.7, which features a man playing tennis. In this example, the explanation technique highlights the importance of the feature *man 80%* in predicting the term "man" in the caption. However, this importance does not necessarily translate into a misalignment of the importance ranking, simply because *man 83%* is deemed to be the most important feature according to Faster-RCNN's rating, as both features represent a man and have a high probability of detection. Redundant detections by the Faster-RCNN may therefore be one of the main causes of such misalignment. Nevertheless, it is reasonable to expect that certain objects, whose probability is lower under the Faster-RCNN, could hold considerable importance under the explanation method. This tendency becomes particularly evident in instances of singular detections where redundancy is absent.

The results of latent ablation evaluation, involving the visual features masking, may not appear highly promising at this initial stage. The results are expressed in terms of the percentage of missing words, as shown in Table 4.2. Notably, the explanation methods do not seem to demonstrate a significant advantage over the random baseline, with BU-LRP explanations displaying a slight superiority over BU-LIME explanations. This observation could be attributed to several factors. One plausible explanation is that manipulating individual visual features during evaluation, especially ones that often represent partial objects rather than complete concepts could potentially introduce artifacts that negatively impact the evaluation process. Additionally, an essential aspect is that the information masked by individual feature ablation can be recovered using the rest of the features. This phenomenon may be underpinned by two key facts: firstly, predictions may be influenced by the contextual surroundings, represented by neighboring regions/visual features. Secondly, the presence of many redundant features for a single object, whether complete or partial, allows missing information from an ablated feature to be easily reconstructed using other overlapping features.

In our investigation of individual features ablation, we conducted experiments with four distinct ablation parameter values, specifically $k \in \{1, 3, 6, 9\}$. We have chosen this range of k values to explore both small-scale and moderate ablation scenarios. Notably, the upper limit of $k = 9$ corresponds to approximately a third of the total features detected within an image. Beyond this threshold, ablation could lose its interpretability due to the difficulty in distinguishing whether high percentages of missing words are primarily due to the relevance of the ablated features or over-ablation causing a total absence of information and loss of context. Impressively, with an ablation of only 9 features, we achieve relatively high scores, reaching up to 90% for the MAX ablation mode. An important observation emerges when comparing the fidelity scores (percentage of missing words) between 1-feature and 3-feature ablation, which consistently exhibits substantial improvement regardless of the explanation model or ablation magnitude. However, the increase in scores becomes less pronounced for the remaining values of the k parameter. This suggests a strong dependency of score augmentation on the k parameter. Specifically, the greatest sensitivity is observed when transitioning from a single feature to multiple features, followed by a less marked increase when progressing from multiple features to even more features. These results cast fresh insights into word prediction dynamics in captioning systems. Indeed, they highlight that the prediction of a specific word in a caption is influenced by several features of the input image, although some of these features do not correspond to the object designated by the word in the caption. This implies that

words attributed to objects in output captions do not depend exclusively on the associated object in the image, but extend to the image's contextual neighbors. These findings highlight the complex interplay between object representations during caption decoding, reinforcing our initial assertion that achieving explainability in captioning systems requires surpassing a simple one-to-one link between output and input. The internal workings of these systems are considerably less obvious than this straightforward connection would suggest.

To further investigate the possible undesirable impact of manipulating individual features rather than complete concepts in latent ablation and its consequences on the evaluation process, we proceeded to explore Object-based latent ablation. This approach takes into consideration the concern that manipulating individual features might introduce artifacts that adversely affect the evaluation process. Instead of masking individual features, Object-based ablation involves masking entire objects. The results, as reported in Table 4.3, show a marked increase in the percentage of missing words and the average probability drop across different explanation methods. Notably, the BU-LIME-5 model shows a clear shift in performance between object-based and visual feature-based ablation (0.5960 for object ablation vs. 0.2648 for visual feature ablation), along with a notable difference between all proposed methods and the random baseline.

The BU-LIME-N model performs best among all models in most scenarios. Conversely, the BU-LIME-N-OBJ models, which utilize intrinsic object perturbation to LIME instead of perturbing individual features, show lower performance than expected. This suggests that the manipulation of complete objects instead of isolated visual features holds more value when employed for Latent Ablation-based evaluation rather than being integrated intrinsically into the development of explanation models, such as the case with LIME. Therefore, this approach does not appear to enhance the efficacy of the explanation method itself. These observations lead us to believe that methods based on object manipulation when designing surrogate explanation models may introduce gaps within the data, potentially causing inconsistencies in the weights and activations of the linear model during the training phase. This may explain the lower scores often observed for these models, some of which even fall below the random baseline (BU-LIME-5-OBJ and BU-LIME-8-OBJ in Table 4.3). Thus, As a result, we argue that object manipulation may be more appropriate for post-hoc evaluations such as latent ablation, but is not recommended for intrinsic use when designing explanatory approaches, as it does not necessarily improve the quality of explanations.

As far as the "Prob drop freq" column is concerned, the advantage of one approach over the other is less expressive. This can be attributed to the introduction of noise into the visual features representing the object, which obviously leads to a change in predictions (a natural consequence). This is reflected in the large number of cases where such alterations occur (more than 90% of the cases), irrespective of whether the explanation elements (objects) in question are determined by the explanation methods or if they are chosen at random. Notably, this change tends to be negative, leading to a decrease in probability, which aligns with rational expectations. This decrease is reasonable since, if a word was already present in the caption before the ablation, it implies that its probability was sufficiently high to have appeared in the caption. Thus, after feature ablation, the model's confidence in predicting that specific word is likely to decrease due to the lack of information compared to the pre-ablation situation. In the specific case of the "Prob drop freq" measure, this rationale contributes to the minor variation between randomly generated explanations and other explanations produced by BU-LRP and BU-LIME.

Regarding the parameter k, in this experiment, we explored various values that determine the number of objects subject to latent ablation. The maximum number of objects to ablate was set to 5. This number corresponds to approximately one-third of the total number of objects within MSCOCO's images, as the total count of distinct objects in MSCOCO images does not exceed 15, a fact based on empirical observations. Analogous to our observations in the individual features ablation, the transition from ablation of a single object to ablation of multiple objects is characterized by significantly larger discrepancies across various scores, as opposed to the shift from multiple objects to the ablation of an even larger number of objects. This observation further reinforces our earlier findings, emphasizing that predicting object-related words in the caption's output does not depend exclusively on the corresponding labeled object in the image, but rather on several objects, which essentially encompass neighboring objects.

Overall, both LRP-based and LIME-based explanations demonstrate a good quality in comparison with the only existing (random) baseline. In the context of explainability within the latent space, these two techniques show remarkable similarity in the correctness of their explanations. It becomes apparent that the scope of the explanation technique does not significantly affect the quality of explanations for captioning models. Furthermore, it appears that the global method, which back-propagates the relevance of a given prediction back to the input, is not that instrumental for capturing the most important aspect of an image. In contrast, the local method achieves equivalent performance by opting for a more

direct shortcut, bypassing the intricacies of the captioning architecture' mechanisms. It does so by considering only the linkage of the output prediction to the input via a causal dependency.

This observation may also depend on the extent to which a detailed and nuanced explanation is desired. BU-LRP, by virtue of its precision-oriented mechanisms and customized adaptations to captioning architectures, is able to generate precise, fine-grained explanations that delve into a remarkably low level of granularity. This capability enables it to pinpoint the involvement of each element of the feature vectors in the overall prediction process. This characteristic holds considerable importance from an explanatory perspective, as it facilitates a more profound exploration of sub-components in deep neural networks, particularly within captioning models. This is particularly crucial in critical domains where even the minutest detail carries importance. Conversely, in scenarios where the granularity of explanations is not a central concern, priority is given to other factors such as computational efficiency, data volume, and complexity. These methods are often faster and more practical to realize in real-world applications than those of global explainability [Adadi and Berrada, 2018], the latter requiring more complex and time-consuming techniques.

4.4.4 Limitations of attribution explanations

Having explored the strengths of both explanation methods, it is now appropriate to examine their possible shortcomings and limitations. For an illustrative test image sourced from the MSCOCO2017 dataset, Figure 4.12 presents a scenario in which the accuracy of the explanations appears somewhat compromised. For instance, the prediction of the word "stop" in the output caption seems to be roughly related to features representing a "stop sign". However, the explanation method has prioritized other features, specifically those representing more common concepts such as "sky" and "building". These concepts are broader and less contextually specific. In our opinion, these explanation errors could be attributed to several factors. Firstly, the presence of redundant visual features introduces a certain bias into the captioning model, making it unclear whether decisions are based on a particular feature or its duplicates. Secondly, some common concepts such as "sky", "tree", or even "road" are frequent objects in a large part of the dataset. This widespread occurrence can lead to explanation biases that are not intrinsic to the explanation methods themselves but rather inherited from the biases present within the captioning models.

Moreover, the ablation experiments carried out in Section 4.4.2 reveal that masking

such objects based on their explanation importance significantly affects the prediction of the target word. This is particularly noteworthy given that, logically, there is no inherent correlation between these objects and the target word present in the caption ("stop" in this case). These results highlight a key insight further supporting our finding in the previous section: captioning models have the ability to establish inter-dependencies between different concepts during the learning process by recognizing the coexistence of specific objects in a scene. Then, this acquired coexistence knowledge is subsequently translated into predicting specific words from the vocabulary during the inference stage.

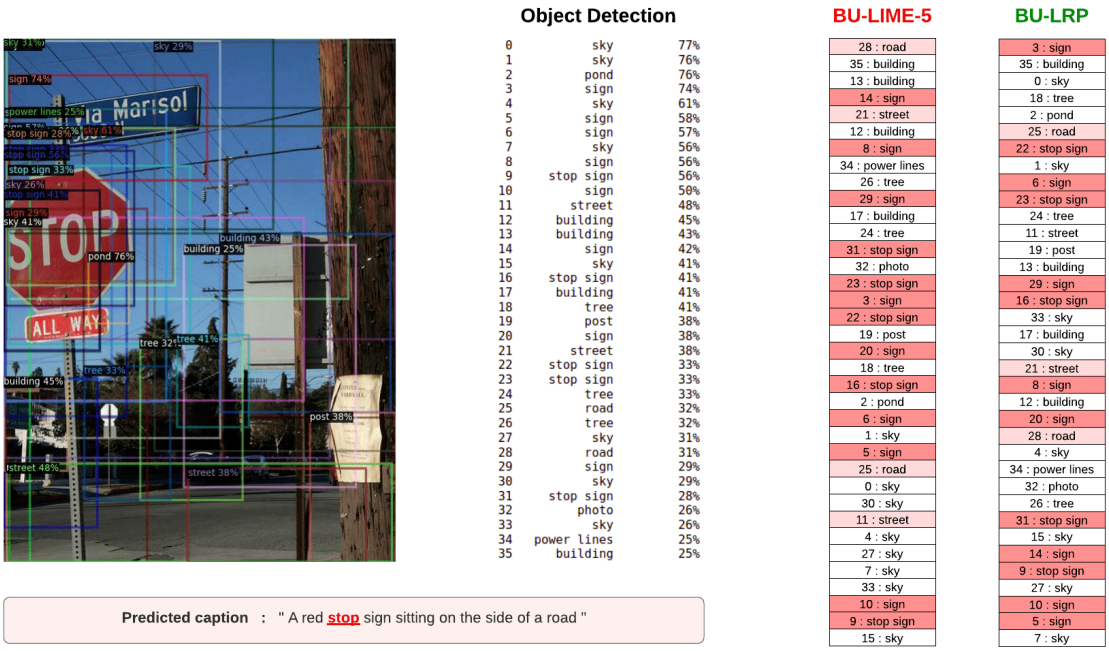


Figure 4.12: Comparative example of BU-LIME-5 and BU-LRP explanation outcomes.

4.5 Conclusion

This chapter dived deeper into the visual modality, which stands as the most decisive component of captioning architecture. Within this framework, we have tailored two attribution-based explanation methods that are distinct in their scope and functioning, LRP and LIME, to the context of Image Captioning (IC). Through a comparative study, we evaluated the effectiveness of these methods in providing accurate explanations. The two approaches yielded comparative results in terms of the quality of their explanations.

In addition, we have introduced the new concept of "Latent Ablation" to assess explanation quality. In contrast to classical ablation, Latent Ablation operates within the latent stages of the architecture. This allowed for a finer manipulation of image objects, offering several levels of alteration and enabling a more precise evaluation. We have also scrutinized the impact of concept completeness, i.e. entire object versus individual features, on the performance and evaluation of explainability. This has led us to design two versions of LIME and Latent Ablation, each focusing on either full object handling or individual feature-based approaches. Intriguingly, our results indicated that while object handling does not necessarily enhance the robustness of the surrogate explanation model, it remains a crucial element in assessing explanation quality, particularly in the case of Latent Ablation.

The scope of the explanation method proved to be less decisive in the quest for superior explanation quality, but instead plays a role in determining the granularity of the explanations produced. It is noteworthy that the choice between these two approaches is not dictated purely by their scope, but extends to the degree of subtlety of the expected results. In this context and building upon the subtle explanation provided by BU-LRP, the next chapter will showcase how explanations obtained at the level of individual feature vector elements can extend the horizons of explainability. This move will help to affirm established facts, uncover new insights, and pave the path for promising future endeavors in the field.

5 Discerning latent concepts

Contents

In the preceding chapter, we delved deeper into the visual modality, seeking to extract more tangible clues from the latent space and explore the possible correlation between the quality of the explanation and the scope/complexity of the explanation method. In addition to the results obtained within the scope of the approach, the endeavor yielded promising related findings, unveiling important facts about the structuring and patterning of information in this space. The heatmap explanations showcasing the contributions of bottom-up features suggested a highly structured representation in the latent space. Such structured encoding of information holds significant implications for understanding the inner workings of the captioning model.

Building on these foundations, this chapter embarks on an in-depth exploration of the latent representations seeking for a better understanding of how the latent space effectively handles specific visual concepts. Our focus in this work is confined to the concept of object appearance frequency in the original image. Our investigation delves into the process of encoding this visual concept in the latent space and how it is subsequently processed by the captioning model before leading to the output caption. Our primary objective is to unveil the underlying mechanisms governing crucial aspects such as the visual saliency of objects within a given scene. Consequently, this study is particularly interested in

examining potential correlations between the saliency aspect and the aforementioned quantity concept. Through this exploration, we also endeavor to illuminate the role of saliency in the complex interplay between vision and language during image captioning.

It is important to note that the primary aim of this chapter is to provide a stepping stone for further explorations of latent space that may lead to a deeper grasp of its inner workings and its role in the IC process.

5.1 Probing IC models via input editing

In the quest to comprehend the inner workings of ML models and gain insights into their decision-making processes, probing-based methods play a pivotal role. These methods encompass a range of specialized analyses aimed at exploring the internal states, inputs, and outputs of ML models. By exploring these components, probing-based methods not only help to unveil the nature of the information they encapsulate but also shed light on how they are utilized/produced during the decision process. Probing includes several techniques, such as sensitivity experiments where the model's reaction to specific linguistic or visual alterations (similar to perturbation) is evaluated, and neuron activations within the model's hidden layers are examined. This aims to detect potential biases within the data or the model itself, comprehend the model's behavior patterns, and even make comparisons between different ML architectures. Probing methods provide a practical approach to studying the inner workings of complex ML models and improving the understanding of how these models represent, encode, and decode information.

The concept of probing has been used in various forms in earlier research, namely in the field of linguistics, where researchers have employed probing techniques to investigate the linguistic knowledge encoded in the human brain. Recently, probing methods have gained widespread adoption within the NLP field. In the work of [Clark et al., 2019], the authors proposed a probing method to analyze the attention mechanisms of BERT. Another approach proposed by [van Aken et al., 2019] intended to inspect question-answering models by visualizing token representations that reveal information about the internal state of Transformer networks. [Tenney et al., 2019] probe the syntactic and semantic information captured by contextualized word representations obtained from pre-trained language models like BERT, ELMo, and GPT-1, and investigate whether these representations encode sentence-level structural information. Their study allows us to understand how contextualized word representations learn and encode higher-level linguistic features.

Several studies in the machine learning literature have adopted probing techniques to assess and compare the performance of different systems [Wexler et al., 2020, Yu et al., 2022, Ilinykh and Dobnik, 2021]. For instance, [Yu et al., 2022] introduced a meta-morphic testing approach for validating IC systems. Their approach involved evaluating changes in model predictions after inserting objects into images to identify and report any issues present in the models. To achieve this, they inserted objects into selected backgrounds using object resizing and location tuning algorithms and analyzed image pairs (background-synthetic) whose captions exhibited unexpected differences.

In the context of IC systems, probing techniques may involve analyzing hidden representations to understand how visual information is processed and linked to textual descriptions. Inspired by the concepts of object size and location explored in [Yu et al., 2022], we extend the probing techniques in this work to investigate the crucial concept of quantification in the context of image captioning explainability. We aim to explore and analyze the latent and output spaces to gain insights into how this concept is represented and processed by the IC architectures. Specifically, we make quantity-related changes to the input images and examine their impact on latent representations through an analysis of visual encoding on the one hand, and model output by analyzing predicted captions on the other. This study enables us to understand how IC models deal with the concept of quantification and how it is reflected in the various stages of captioning.

5.2 Quantification concept probing

The quantification concept refers to the representation of quantities related to entities present in the image. During the captioning process, the system needs to describe the (precise) count of certain objects within the scene. For instance, in an image containing a group of people, the encoder detects each person by assigning a bounding box, a label, and a feature vector. The attention mechanism and the decoding module then aggregate this information to predict the appropriate quantifiers in the output sequence, which can be generic quantifiers (e.g., "a group of") or exact quantifiers (e.g., "four," "five," etc.).

The ability of an IC system to effectively handle quantification is crucial for generating informative captions that accurately describe the visual content of the image. As part of the probing study, investigating the quantification concept helps to gain insights into how well the system captures quantity aspects of the scene and how it relates this information to the corresponding textual description. One of the aims was also to investigate the

concept of "saliency" in this context. Our working hypothesis posits a possible dependency relationship between an object's correct identification and two factors: its saliency and its appearance frequency within the image. This exploration could potentially shed light on the role of saliency in the interaction between vision and language during captioning.

5.2.1 Object corpus preparation

The idea of probing IC models using input editing consists of introducing meaningful changes to the input images and then observing the subsequent changes in the latent structures and the output. These meaningful changes are typically achieved by inserting new objects into original background images from the dataset. To avoid novel object biases that could result from inserting objects external to the dataset used throughout this study, we extract objects from the train partition of the same dataset (MSCOCO2017). This approach ensures the homogeneity of the inserted objects with the background images, as they share a common source.

To attain a satisfactory level of object diversity and quality, a semi-automated extraction approach is adopted. By semi-automated extraction, we mean an automated image segmentation algorithm followed by a manual selection of objects based on their classification scores and edge quality. Objects with high classification scores and well-defined contours are selected. In our implementation, we use Mask-RCNN, an extension of Faster R-CNN by [He et al., 2017] which produces object masks in addition to object class labels and bounding boxes. This mask guarantees a more precise extraction of an object's spatial layout.

Figure 5.1 illustrates the object extraction process for an instance from the MSCOCO2017 train set. As exemplified, the object segmentation module identifies multiple objects within the scene. Among these, two objects share the same label ("bus"), accompanied by the highest classification scores of 99.97% and 99.81%, respectively. In such cases where the objects of interest are not unique, the one with the highest score is selected, as can be seen in Figure 5.1d. This chosen object is then isolated from its background, and cropped to its bounding box. The object corpus obtained at the end of this process comprises an assortment of categories, including Humans (woman, man), Animals (dog, bird, giraffe), and Objects (car, chair, clock, bus).

Figure 5.1: Illustrative example for object extraction.

5.2.2 Quantification probing method

The quantification probing process covers several stages, including the generation of quantification data, the prediction of new captions, and evaluation. The main objective is to introduce quantity-related modifications to original inputs and then to examine the subsequently undergone changes in the information, particularly those related to

the quantification aspect. Figure 5.2 provides an overview of the quantification probing method and its various stages, using an illustrative example. In the following, we present a detailed breakdown of each of these stages.

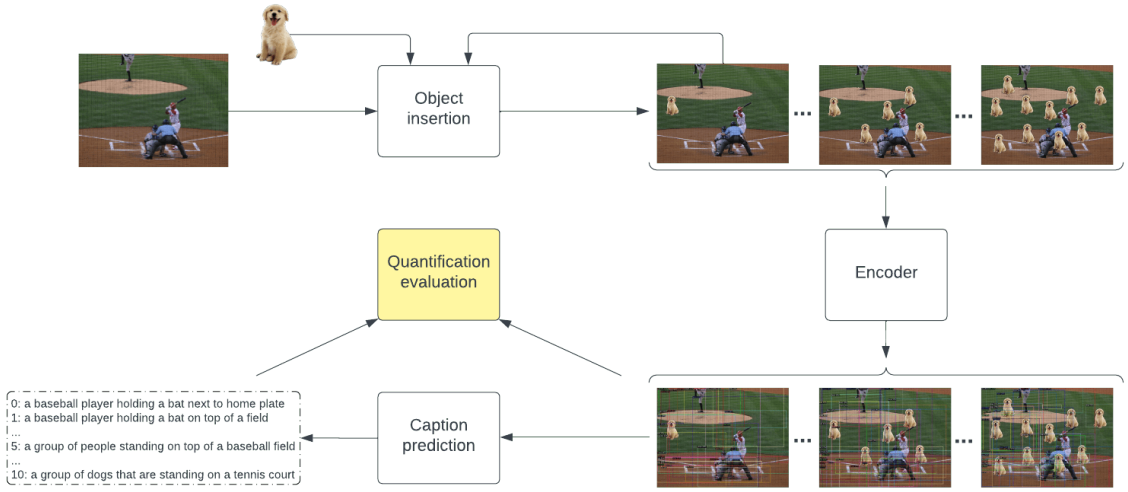


Figure 5.2: Quantification probing pipeline.

1. Synthetic data generation: Given a background image I_B and an object image I_O, generating a synthetic image I_S^k consists of inserting a number (k) of objects I_O into I_B at random coordinates, as indicated in Equation 5.1. At each step, the previously generated synthetic image becomes the background for the current insertion.

$$I_S^k = Combine(I_S^{k-1}, I_O, random\_coordinates) \;\; ; \;\; I_S^0 = I_B \tag{5.1}$$

During this stage, two challenges emerged concerning the size of inserted objects and their spatial location in the image. Certain studies attempted to address these issues, such as [Yu et al., 2022] where the authors, in their specific context of IC models testing, proposed object resizing and location tuning algorithms during the synthetic images generation phase. However, their approach presupposed that the inserted object in the background image should always be salient.

By taking a distinct perspective, we argue that the parameters governing the saliency of an object within a scene remain enigmatic and deserve further exploration. Our hypothesis is based on the premise that this saliency factor may depend on various concepts, such as

the position of the object, its size, and the frequency of appearance of similar objects. To elucidate these complex relationships, our methodology isolates the specific concept of quantity, while keeping the others invariant. This approach serves to mitigate the risk of the interplay between these concepts. It enables gaining a clearer understanding of the specific impact of the quantitative aspect first by examining it individually while laying the foundations for more comprehensive studies of the other aspects in further dedicated studies.

To this end, regarding the size of inserted objects, a fixed and shared rescaling parameter ($\lambda = object\_image\_size/background\_image\_size$) has been uniformly applied to all the images we experimented with. The new size of each object image is then given by ($object\_image\_size = \lambda \cdot background\_image\_size$). This ratio plays an essential role in maintaining a normalized object scale in the background image and prevents the inserted objects from appearing overly large or excessively small. Its value is derived from empirical observations of the average object size in the images. Concerning the positioning of each inserted object, we generate random coordinates within the range of the height and width of the background image and ensure a slight buffer equal to the object's height and width to avoid overflow. After each iteration, we record the coordinates of the inserted object, which we use as reference points to avoid overlaps when placing subsequent objects. Concerning the number of objects inserted in background images, we explore configurations with a maximum of K occurrences. By doing so, we examine the potential impact of this parameter on object saliency and assess how quantification information evolves across various settings. The process of generating synthetic quantification data for a single instance is summarized in Algorithm 4. The $Check\_overlap()$ function returns $True$ if no overlap is identified between objects previously inserted in the background image, $False$ otherwise.

2. Captioning: Once the quantification data is prepared, the set of synthetic images associated with each background image is fed into the captioning model. Image encodings, including object detections with their labels and newly generated captions, are saved for further examination in the evaluation stage.

3. Quantification evaluation: In this approach, we focus on inspecting quantification information at two main levels: input encodings and output decodings. The primary objective of this evaluation is to facilitate a thorough examination of quantification information from its appearance in the raw data to its manifestation in the final outcomes. This encompasses the visual and linguistic aspects, enabling a nuanced exploration of

Algorithm 4 Synthetic image generation for quantification probing.

Inputs: I_B, I_O, K, λ
Outputs: $\{I_S^k\}_{k=1}^K$

$I_O \leftarrow Rescale(I_O, I_B, \lambda)$
$list\_coordinates \leftarrow \{\varnothing\}$
for $k \in \{1, .., K\}$ **do**
  $coordinates_k \leftarrow Random()$
  while $Check\_overlap(coordinates_k, list\_coordinates)$ *is* $False$ **do**
    $coordinates_k \leftarrow Random()$
  end while
  $list\_coordinates \leftarrow list\_coordinates \cup \{coordinates_k\}$
  $I_S^k \leftarrow Combine(I_B, I_O, coordinates_k)$
  $I_B \leftarrow I_S^k$
end for

return $\{I_S^1, ..., I_S^K\}$

the unique properties present in both modalities with regard to the representation of quantification information.

On the visual part, the evaluation consists of identifying the successful detections of the inserted objects and their number. This is achieved by using the labels generated by the detection module. Concerning the language part, the evaluation is less straightforward to achieve. The linguistic structures, by their intrinsic complex forms, require several case-by-case studies. The evaluation consists of checking the presence of both object tags and quantifiers in the output caption. The quantifiers considered for this study are extracted from English dictionaries (e.g. Cambridge dictionary[1]), we then propose a classification into two categories: generic quantifiers, and specific quantifiers. These quantifiers are listed in Table 5.1. Some quantifiers that are not part of the dataset vocabulary are simply excluded. The evaluation principle in this context is to assess the change in quantification information between the reference and the synthetic image captions. This assessment is based on the presence (or not) of the object tag and the quantifier. To facilitate this evaluation, we have introduced an evaluation grid detailed in Table 5.2, where scores range from 0 to 1.

The evaluation considers a range of scenarios: the highest score is granted when both the object and the quantifier appear in the candidate caption after being absent in the

[1] `https://dictionary.cambridge.org/`

Generic quantifiers	Cardinal (precise) quantifiers
some, many, few, several, much, enough, a couple of, lot of, lots of, plenty of, tons of, ton of, loads of, bit of, deal of, number of, little of, bunch of, pile of, hundreds, thousands, amount, herd, group, crowd, flock, quantity, quantities, numerous	a, one, two, three, four, five, six, seven, eight, nine, ten

Table 5.1: Quantification terms in English.

reference (the first row). A high score is also assigned when only the quantifier appears in the candidate caption while the object remains absent, indicating it corresponds to the inserted objects (the second row). On the other hand, if only the object appears in the candidate caption and the quantification information remains unchanged, this suggests that the quantifier may not have been successfully detected in case of absence, or that it is ambiguous whether the quantifier refers to the inserted objects or one of the objects already present in the image in case of presence. For both cases, we award half of the maximum score (the third and fourth rows, resp). Another scenario is where the object appears but the quantifier disappears, possibly because the object originally associated with that quantifier has been hidden by newly inserted objects, in which case a low score should be assigned (the fifth row). Lastly, cases where there is no change between the reference and the candidate, in both object and quantifier or when the quantifier disappears while the object remains absent, represent the worst-case scenarios and are awarded a score of zero. Cases not falling into these categories (N/A) are excluded from the evaluation, as we do not take into account images where the inserted object was already present.

5.2.3 Experimental protocol

For our quantification experiment, we utilized background images from the test partition of MSCOCO2017 dataset (5,000 instances). For each instance, we systematically inserted objects ranging in quantity from 1 to 10, as illustrated in Figure 5.3. This procedure resulted in approximately a total of 50,000 synthetic images for each object, excluding instances where background images initially contained objects with the same label as the inserted one and those with technical problems during insertion. The various quantities of objects inserted enable us to study the evolution of object identification and how the quantification information is captured at both the visual and language levels.

Reference		Candidate		Score
Object	Quantifier	Object	Quantifier	
✗	✗	✓	✓	1
✗	✗	✗	✓	0.75
✗	✗	✓	✗	0.5
✗	✓	✓	✓	0.5
✗	✓	✓	✗	0.25
✗	✗	✗	✗	0
✗	✓	✗	✓	0
✗	✓	✗	✗	0
✓	✗	✗	✗	N/A
✓	✗	✗	✓	N/A
✓	✗	✓	✗	N/A
✓	✗	✓	✓	N/A
✓	✓	✗	✗	N/A
✓	✓	✗	✓	N/A
✓	✓	✓	✗	N/A
✓	✓	✓	✓	N/A

Table 5.2: Quantification evaluation grid. ✗and ✓denote Object/Quantifier absence and presence, respectively, N/A denotes the not supported cases.

The resizing ratio of objects with respect to the background image size is consistently set at 2%. This specific value was determined empirically, with the intention of creating synthetic images that minimize inserted object overlap with pre-existing objects. However, it is important to note that achieving entirely automated object insertion without any overlap with existing objects is practically unfeasible due to the complex nature of real-world images. Nevertheless, maintaining a fixed ratio of 2% ensures uniformity in object size throughout the entire dataset. This standardization of object size is essential to isolate and investigate the quantity aspect of objects without interference from variations in size. Consequently, for any given object category, the inserted objects share identical dimensions.

Remember that the evaluation of quantification information consists of two main stages: the latent level of the visual part which consists of bottom-up visual features detected by the Faster-RCNN, and the linguistic output level, which explores the captions generated from synthetic images. Through this dual-stage analysis, we aim to track and compare the progression of the quantification information in these two modalities and investigate its impact on the concept of saliency in both the visual and linguistic parts.

Figure 5.3: An illustrative example of object insertion.

5.2.4 Results and discussion

Quantification analysis at the vision level

In this experiment, we investigate the distribution of object occurrences in detections for synthetic images across different quantities ranging from 1 to 10. An object is considered "well detected" if it is absent in the original background image but becomes part of the set of detected regions identified by the Faster-RCNN in the new synthetic image. When the number of inserted objects exceeds 2, collective detections may occur. To automate the identification of individual and collective detections, we leverage the Faster-RCNN detection labels, by examining the number of singular labels corresponding to the inserted object or their plural forms. Results showing the evolution of quantification information detection at the vision level (Faster-RCNN encoding outputs) for the various objects we experimented with are presented in Figure 5.4.

On these histograms, there seems to be a consistent overall pattern for various objects, notably "clock", "bird" and "chair". In certain synthetic instances, these inserted objects are promptly and successfully detected upon their first appearance. It is natural to observe a gradual decrease in this number as we move from one insertion stage to the next since we exclusively account for first-time detections. However, for objects like "giraffe" and "car", detections are occasionally deferred until subsequent insertions, with the object being detected on the second insertion in roughly the same number of cases. In such scenarios, the object only reaches saliency at subsequent stages of insertion. This phenomenon strongly suggests that the concept of saliency may indeed be tied to the quantitative aspect of objects within a given scene. Additionally, one could consider the influence of an object's frequency within the dataset, which might lead to certain objects being consistently regarded as part of the background. However, analyzing data for "giraffe" and "car" classes reveals a substantial imbalance, with 2647 images containing "giraffe" versus 12786 for "car" in the MSCOCO dataset. This disparity refutes the hypothesis

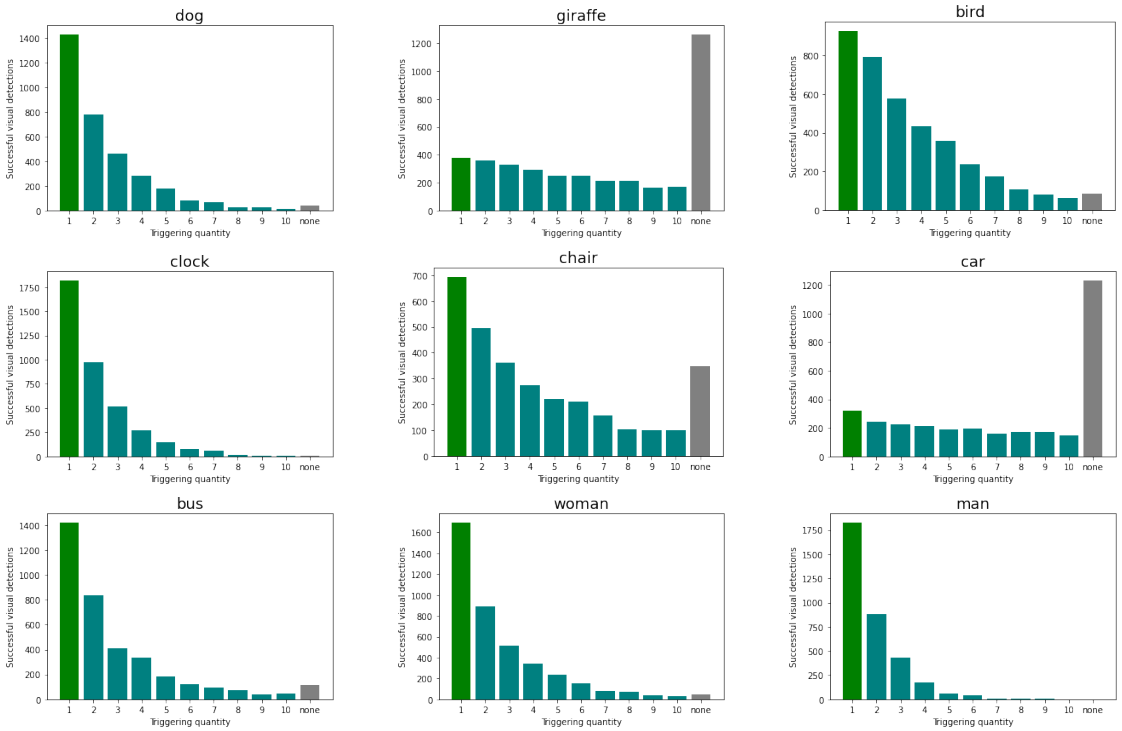


Figure 5.4: Distributions of detection results at visual level, in the function of the quantity at which the object is identified.

that an object's frequency within the dataset alone drives the observed behavior.

The insightful patterns continue, particularly when we examine the behavior of the object "car". The number of successful detections increases steadily, as the number of objects inserted increases, meaning that objects that gain saliency in the later stages are more numerous than those in the earlier ones. However, for some objects, detection does not occur even after multiple insertions, extending up to the tenth insertion stage. This occurrence suggests that certain inserted objects are still not recognized by the object detector. Objects like "giraffe" and "car" fall into this category, representing the majority of such cases. The reasons behind this phenomenon can be attributed to the inherent nature of the object itself. It might indicate a kind of reluctance on the part of the object detector to manifest the object. It is plausible that this reluctance results from the detector's perception that the object does not seamlessly fit within the context of the overall image content, which could give rise to a form of contextual inconsistency.

Quantification analysis at the output level

From the language output perspective, it is essential to study the evolution of quantification information in synthetic image captions. This exploration helps us understand how information aggregates within the captioning model and how the language component selects what information to incorporate into the captions. The first key aspect to study is the evolution of the prediction of the object tag and quantification information (i.e. their saliency) in output captions when the number of inserted objects varies from one to ten. The objective here is not only to determine object saliency after a certain number of occurrences, but also to characterize how quantification information is mirrored in these outputs, either simultaneously with the presence of the target object or not.

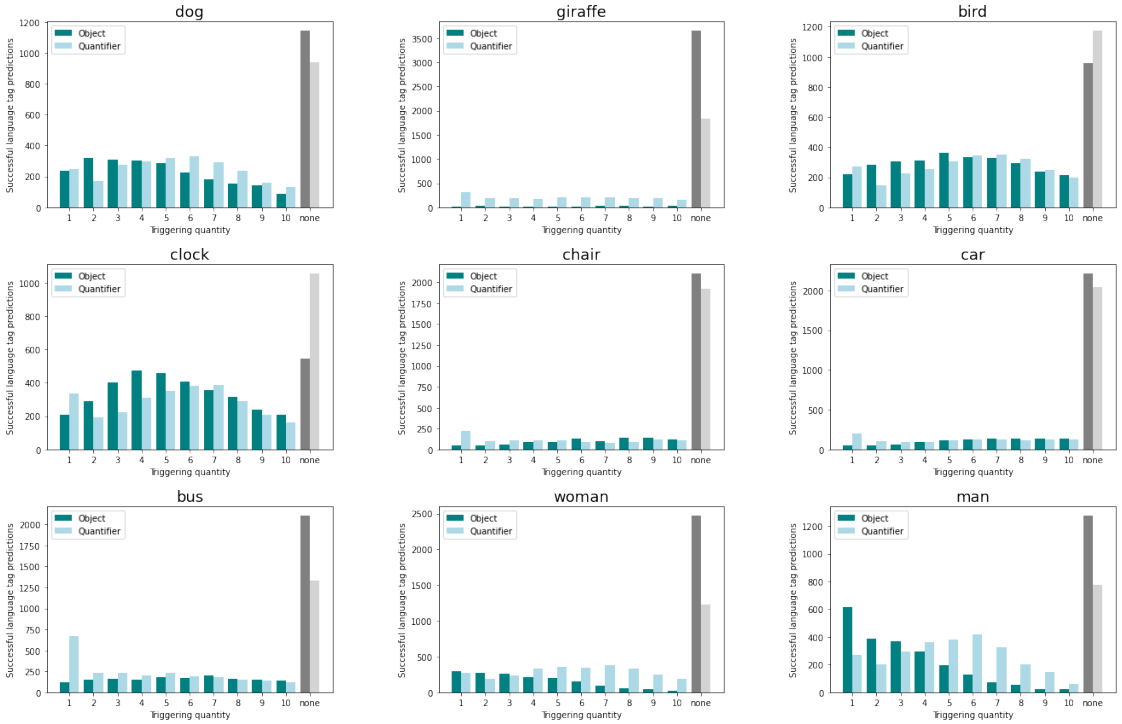


Figure 5.5: Distributions of object/quantifier tags prediction at the output level, in the function of the quantity from which they appear.

In Figure 5.5, the histograms display the frequency of successful object and quantification information identifications starting from a specific number of inserted objects. The "none" category represents cases where the object/quantifier has not been successfully predicted in any of the previous insertion steps. Our initial observation reveals a similar behavior in the identification of objects and quantifiers in output captions across a range of

object classes, including "dog", "bird", "clock", "chair", "car", and "bus". In these cases, both the object and quantifier predictions reach their peak at a specific insertion step, which varies depending on the object class. A qualitative example showcasing the evolution of quantity information at both the visual and language levels is presented in Table 5.3. In this example, the ✓symbol indicates a successful object detection at the visual level. It is evident that for this object class, the more objects we insert into the image, the more confident the model becomes, allowing it to accurately identify the inserted objects. Notably, in terms of vision-level detection results, the detector successfully located all objects from the second insertion stage onwards.

Image	Quantity	Vision	Language
	0	-	"a man on a surfboard in the water"
	1	-	"a man in a wet suit standing on a surfboard in the water"
	2	✓	"a man and a dog are on a boat in the water"
	3	✓	"a couple of people that are standing in the water"
	4	✓	"a couple of dogs that are standing in the water"
	5	✓	"a group of people that are standing in the water"
	6	✓	"a group of people that are standing in the water"
	7	✓	"a group of people that are standing in the water"
	8	✓	"a group of dogs that are standing in the water"
	9	✓	"a group of dogs that are standing in the water"
	10	✓	"a group of dogs are standing in the water"

Table 5.3: A qualitative illustration of quantification probing with object "dog".

The object "giraffe" however exhibits a distinctive behavior in which quantification information appears in the output captions independently of the specific object being inserted. This suggests that the quantity aspect, although inserted objects are not explicitly detected by the image encoder, can still influence the captioning outputs. In such scenarios, the quantification information may not necessarily correspond to the specific quantified

item in the scene. As illustrated in Table 5.4, this inconsistency between quantifier and object is evident, with the "giraffe" object remaining undetected until the ninth insertion. Notably, at the linguistic level, the object tag is not successfully predicted in any of the insertion steps, while the quantification information begins to appear from the third insertion. Moreover, the predicted quantifier is associated with one of the existing objects in the scene (in this case, the "man"). It is worth highlighting that the model pluralizes the most salient object, "man," within the predicted caption, even though the synthetic image contains a single occurrence of that object. These observations raise questions about the captioning model's ability to accurately relate quantitative information to the various elements of a scene and the potential hallucinations that may occur, its ability to withstand input perturbation attacks, and the relevance of the explanations provided in the case of objects with multiple occurrences in a scene.

Image	Quantity	Vision	Language
	0	-	"a man on a surfboard in the water"
	1	-	"a man riding a surfboard on top of a body of water"
	2	-	"a man riding a surfboard on top of a body of water"
	3	-	"a group of people that are standing in the water"
	4	-	"a group of people that are standing in the water"
	5	-	"a group of people that are standing in the water"
	6	-	"a group of people that are standing in the water"
	7	-	"a group of people that are standing in the water"
	8	-	"a group of people that are standing in the water"
	9	✓	"a group of people that are standing in the water"
	10	✓	"a group of people that are standing in the water"

Table 5.4: A qualitative illustration of quantification probing with object "giraffe".

The patterns observed in the qualitative examples presented in Table 5.4 and Table 5.3 diverge significantly due to the model's inability to consistently link the quantifier to

the object (giraffe), a connection successfully established in the first example (dog). This divergence can be attributed to the specific class of objects manipulated in both experiments. In the case of the "dog," it typically appears in scenes alongside humans or near water, leading the model to confidently integrate the "dog" into the context of the background image representing a man doing water surfing, both at the visual and language levels. However, when dealing with the "giraffe," the model exhibits hesitancy and only allows the quantification information to appear, while the object itself is not as readily incorporated. This distinct behavior appears to be unique to the "giraffe" among all the object classes considered in this experiment. It may be influenced by the nature of the images within the MSCOCO dataset, where giraffes are rarely found in conjunction with other object categories that typically feature in everyday human activities. As a result, the model struggles to identify giraffes in contexts where they are not commonly encountered.

Another notable pattern that surfaces from these graphics is the distinctive trend exhibited by some objects in relation to their saliency. Objects like "clock" and "bird" exhibit peaks in both object and quantification information prediction around the fourth or fifth object insertion. This observation lends further credence to our hypothesis regarding the correlation between object saliency and the quantity aspect. A stronger presence of an object in a scene, possibly beyond a certain threshold, implies a greater saliency and, consequently, a more explicit reflection in the captioning result.

Furthermore, a closer examination of the results concerning some object classes like "bird" reveals an intriguing phenomenon when contrasting object detection at the visual level with object prediction at the output level. In many cases, even when the image encoder accurately identifies an object, it fails to appear in the output caption, despite multiple object insertions. This hints at a deliberate dismissal of this information, exclusively by the language decoder which may have reservations about incorporating the object into the generated text. Such behavior probably stems from a contextual misalignment between the well-detected inserted object and other objects in the scene that might possess greater saliency and contextual appropriateness within the output caption. Table 5.5 provides a set of comparative statistics that shed light on this contrast between object detection on the visual part, and object and quantifier prediction on the language level. Each column in the table, corresponding to an insertion quantity/step, displays the count of instances where at least one of the occurrences of an inserted object was identified at that step, without prior identifications in earlier insertions. More specifically, out of a

total of 3,839 synthetic instances, the object "bird" was successfully detected by the image encoder at the first insertion step in 929 instances, and this number steadily declined until the tenth insertion, where it was detected in 63 new instances. In 84 instances, the object was not detected in any of the ten insertion steps. On the language level, the object tag was successfully predicted in the output caption in only 219 instances, while the quantifier was successfully predicted in 270 instances. This presents a notable disparity compared to the detection at the visual level. The visual saliency of the inserted object peaks at the first insertion step, i.e. even with small quantities, whereas the peak for both the object and the quantifier occurs at a later insertion step (the fifth) on the language level. It follows that the reluctance to consider the inserted object and the evaluation of object coherence within the context of the scene is predominantly handled by the language decoder. These findings are broadly applicable to all objects under investigation in this study, as the results across various objects demonstrate consistent patterns when comparing the two modalities.

Insertion Quantity	1	2	3	4	5	6	7	8	9	10	None	Total
Object detection (Vision)	929	793	579	433	360	239	173	108	78	63	84	3839
Object prediction (Language)	219	282	306	313	362	334	328	292	235	212	956	
Quantifier prediction (Language)	270	148	224	252	308	344	349	321	250	197	1176	

Table 5.5: Visual detection Vs. Language prediction evolution for object "bird".

As a second step in this analysis, we examine the overall distribution of quantification information predictions for each object within their respective synthetic images' captions. Figure 5.6 offers histograms that illustrate the frequency of occurrences for different quantification cases. These quantification cases are based on the evaluation grid presented in Table 5.2. For the sake of clarity and a more compact representation, we encode these diverse quantification information prediction cases in a binary format using four bits. In this encoding scheme, 0 indicates the absence of the object or quantifier, while 1 signifies their presence. For example, a code like "0011" represents the scenario in which the object tag and the quantifier were initially absent in the reference caption (the bits "00"), and they have subsequently been successfully predicted in the output caption (the bits "11").

These graphics yield several noteworthy observations. Firstly, the scenario "0000", in which both the objects and quantifiers remain absent, dominates the majority of cases for most objects. This observation suggests a pattern of information dismissal by the language decoder, which aligns with the findings in Figure 5.5 (the "none" category). For objects such as "bird" and "clock," there is a marked presence of cases where both the object

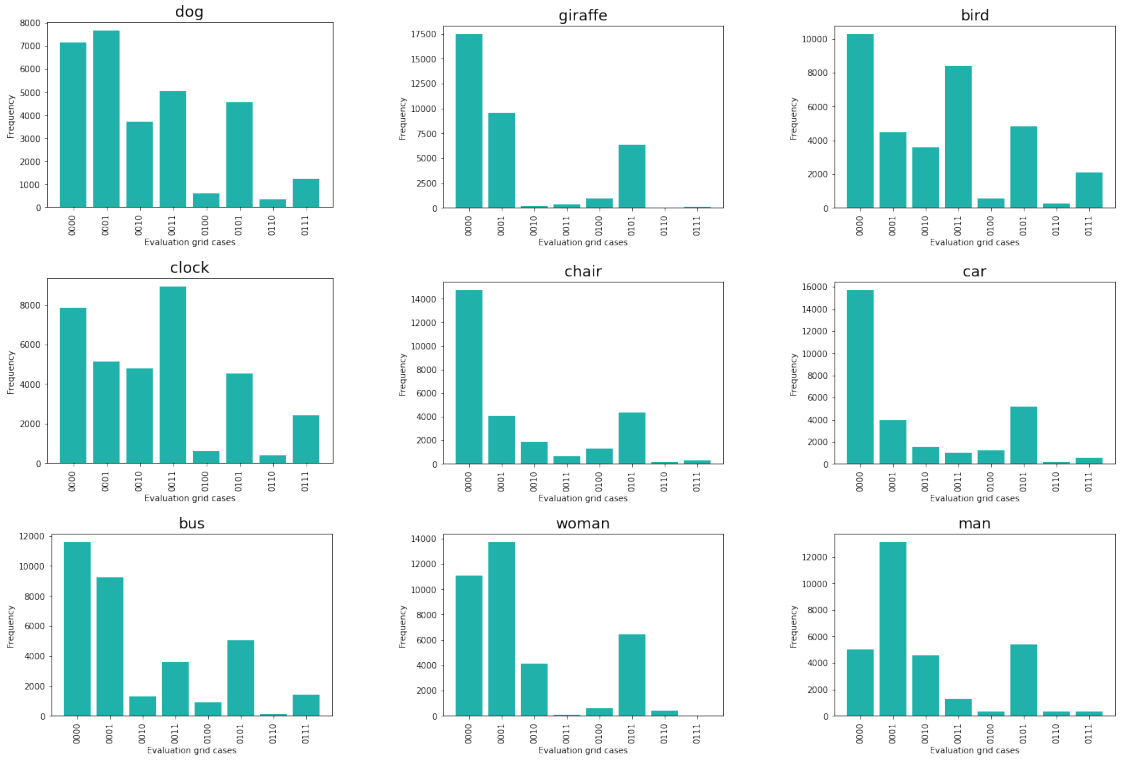


Figure 5.6: Distribution of the cases per quantification scenario at the output level.

and quantifier are accurately predicted. Additionally, a sort of correlation can be noticed between two scenarios: the stable case when the object is absent in the output caption but the quantifier is present (indicated as "0101"), and the case where the object is absent but the quantifier appears in the new caption (noted as "0001"). This suggests that in the latter scenario, the quantification information in the output caption may be related to the inserted objects, much like in the first scenario. As a result, cases in which the quantifier is reflected in the output caption without the presence of the associated object are frequent. This observation hints that information about quantity may be encoded within the image features, and it raises questions about the capacity of current captioning models to effectively link quantification information to the correct object.

In addition to these global statistics across all insertion steps, it is insightful to explore the average quantification scores per object quantity to gain more detailed insights into the relationship between object quantity and object saliency. Table 5.6 provides the quantification scores obtained based on the evaluation grid previously presented in Section 5.2.2. These scores, ranging from 0 to 1, are also summarized in Figure 5.7.

Insertion Quantity	1	2	3	4	5	6	7	8	9	10
Dog	0.0699	0.1478	0.2215	0.3036	0.3893	0.4635	0.5376	0.5945	0.6374	0.6683
Giraffe	0.0052	0.0899	0.1138	0.1406	0.1769	0.2202	0.2565	0.2886	0.3223	0.3567
Bird	0.0547	0.1235	0.1813	0.2433	0.3372	0.4286	0.5194	0.6010	0.6704	0.7250
Clock	0.0449	0.1328	0.2044	0.2982	0.3959	0.4924	0.5887	0.6626	0.7250	0.7716
Chair	0.0140	0.0761	0.0939	0.1201	0.1418	0.1698	0.1966	0.2262	0.2677	0.3006
Car	0.0101	0.0648	0.0821	0.1027	0.1298	0.1641	0.1972	0.2334	0.2678	0.3090
Bus	0.0314	0.1807	0.2272	0.2692	0.3212	0.3612	0.4091	0.4503	0.4893	0.5202
Woman	0.0668	0.1234	0.1763	0.2398	0.3010	0.3597	0.4152	0.4606	0.5002	0.5267
Man	0.1282	0.1899	0.2711	0.3582	0.4210	0.4947	0.5431	0.5721	0.5915	0.6015
Average	0.0258	0.0974	0.1351	0.181	0.2363	0.295	0.3517	0.4024	0.4506	0.4926

Table 5.6: Detailed quantification scores per object.

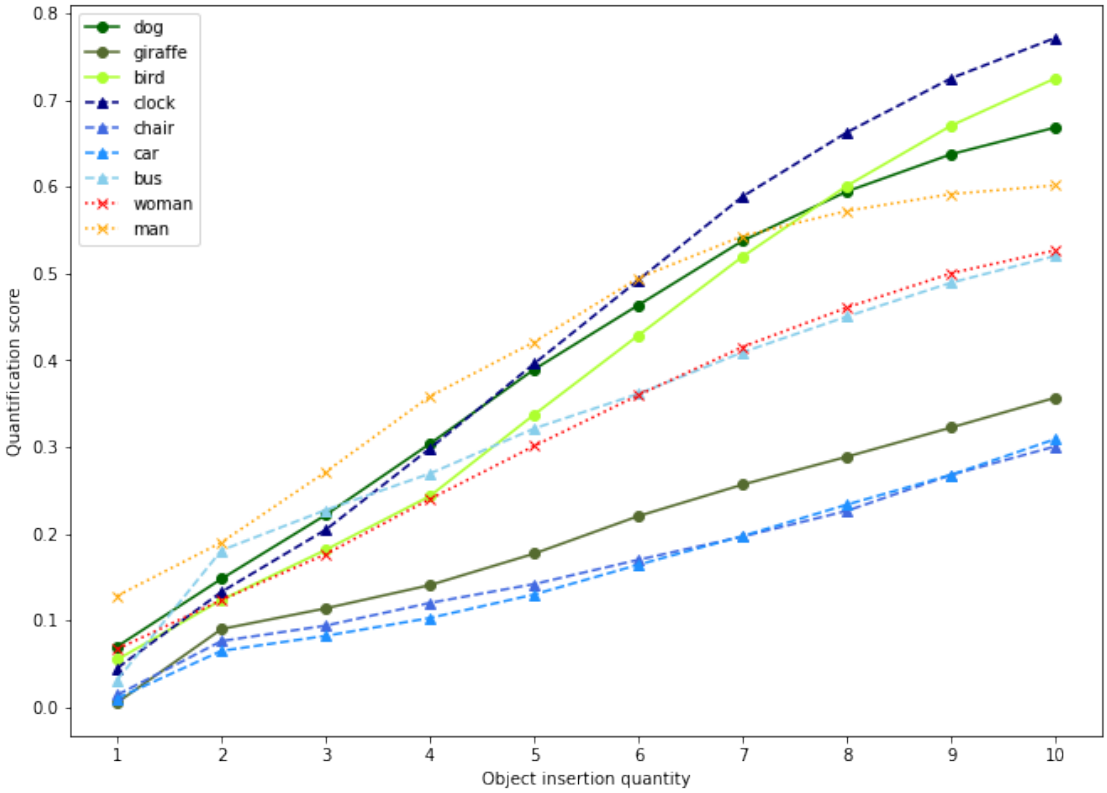


Figure 5.7: Quantification scores.

The overall trend observed in these results is that as the quantity of objects increases, more quantification information is retrieved, even in the later stages of insertion. This further supports our earlier findings regarding the correlation between object quantity and the saliency of objects and their quantifiers. It is also worth noting that certain

objects, such as "bird" and "clock," achieved higher scores compared to other objects, and demonstrated a better ability to predict both the object and its quantifiers as already discussed above. This can be attributed to the object encoder's superior performance in detecting these two objects compared to the others, which, in turn, translates into more accurate object/quantifier predictions in the language output.

5.3 Conclusion

In this chapter, we embarked on an exploration of explainability in captioning architectures, dealing with the representation of visual concepts and the saliency aspect. We introduced a probing approach that employs input editing to uncover hidden facets of model behavior, with a particular emphasis on the quantitative concept of visual information. This approach provided us with valuable insights into the role of this concept and its influence on model decisions at both input encoding and output decoding levels. Notably, our findings revealed a dynamic relationship between the number of object insertions and shaping the object's saliency in the visual detection and language prediction processes. It was also found that for certain objects, saliency may not manifest even after multiple insertions, potentially owing to contextual inconsistency or selective omission by the language decoder.

6 General conclusion

Throughout this work spanning multiple chapters, we tackled the challenge of improving explainability within image captioning systems. Each chapter addressed a distinct facet of this endeavor, and our investigation took us through various methodologies. In this research, we primarily investigated image captioning architectures based on the Encoder-Attention-Decoder framework, which incorporates deep learning components such as CNNs and LSTMs.

The third chapter dissected the core components of these captioning systems, examining their roles and influence in the caption prediction process. A perturbation approach operating on the latent space of the captioning models was introduced. This investigation, centered around an important aspect called decisiveness, highlighted the predominant importance of the visual component, demonstrating its central role in captioning models compared to the language component, resulting in more subtle explanations. The merits of these findings are much more pronounced in some domain-specific tasks like biomedical IC, where the need for precision and accurate explanations is paramount, and any ambiguity in explanations can make AI technologies unusable. In this context, understanding the interaction between different modalities to derive the appropriate explanations poses an important challenge.

Continuing our quest for deeper insights, the fourth chapter dived into the latent space of the visual modality, seeking to extract more tangible and concrete clues. Novel explainability methods were introduced, and their effectiveness was evaluated within this image captioning framework. Through comparisons of LIME and LRP latent-based variants, the chapter unveiled that the scope of an explanation method does not explicitly impact the quality of explanations, but instead plays a central role in determining the granularity of the explanations produced. The concept of "Latent Ablation", which operates within the latent space, provided a new lens to assess granularity and precision in explanations. The findings highlighted the complex interplay between visual and linguistic modalities, emphasizing the nuanced factors guiding the choice of explanation

methods. Furthermore, promising results, unveiling important facts about how information is structured and encoded within this space, have been revealed. The explanation elements showcasing the contributions of visual features suggested a highly structured representation in the latent space. Such structured encoding of information holds significant implications for understanding the inner workings of the captioning model.

In the fifth chapter, through the introduction of a probing-based method, an isolated analysis of the quantitative aspect of objects within visual scenes has been carried out, enabling a clearer understanding of its specific impact on both visual encoding and language decoding. This exploration yielded interesting insights, notably a robust correlation between object saliency and quantity, shedding light on how these models prioritize and generate information based on these factors. A second noteworthy finding was the phenomenon of "information dismissal" by the language decoder, where objects well-detected visually did not appear in the language output. This highlights the complexities of multimodal systems and their behavior, as well as the need for further investigation. Additionally, our research has shown that quantification information can appear in output captions even when the corresponding objects are absent. This raises questions about the models' capacity to effectively link quantification information to specific objects.

The diverse findings emerging from this thesis have revealed promising leads for further research and exploration. These research directions align with the continuation of our contributions and can be categorized into multiple dimensions. A compelling line to consider involves conducting an in-depth investigation into the potential sources of bias and hallucinations in captioning architectures. Particular attention should also be directed toward understanding how salient objects are identified and distinguished from the background scene within an image. These factors may play a pivotal role in influencing captioning decisions and warrant thorough investigation. Although a part of this work has focused primarily on the LRP and LIME explanation methods due to their widespread adoption in the field, the latent space approach developed in this thesis can be extended to encompass other explanation techniques. For instance, future studies could explore the applicability of Shapley's explanations and assess their compatibility with our proposed framework. Our exploration of the quantification aspect and its potential correlation with the saliency aspect has laid a solid foundation for further investigations. In-depth studies can delve into other visual attributes, such as the position and size of objects, and assess their potential relationships with object saliency. The findings within this thesis, in conjunction with the prospective research directions

mentioned above, hold promise for expanded applicability to complex architectures such as Transformer-based and Graph-based structures. Such an expansion would encompass a thorough assessment and comparative explanatory exploration across a diverse spectrum of existing image captioning architectures, thereby offering a more profound insight into their unique operational mechanisms and distinctive characteristics. At both fundamental and application levels, in this thesis, we have pursued the goal of making the captioning architectures more interpretable, which helped pave the way for further research enabling to address growing problems in captioning systems explainability.

Publications during this thesis

1. Explainability in Image Captioning based on the Latent Space. *Neurocomputing, 2023 (Impact Factor: 5.779, SJR: Q1).* [Elguendouze et al., 2023]

2. Towards Explainable Deep Learning for Image Captioning through Representation Space Perturbation. *IJCNN, 2022 (CORE B).* [Elguendouze et al., 2022]

7 Résumé substantiel de la thèse en français

" Approches d'Intelligence Artificielle eXplicables pour le Sous-titrage d'Images "

Introduction

Au cours des dernières années, les architectures d'apprentissage automatique ont connu une expansion en termes d'utilisation et de performance, entraînant une augmentation remarquable de leur complexité, en particulier dans les approches d'apprentissage profond telles que les réseaux de neurones convolutifs (CNNs) et les Transformers. La grande complexité de ces modèles a soulevé des problèmes d'interprétabilité qui empêchent les humains de comprendre le raisonnement qui sous-tend les décisions qu'ils prennent, qu'elles soient correctes ou erronées. Ce problème est particulièrement aigu dans les domaines sensibles, où le coût d'une prédiction erronée ou d'un manque de compréhension peut s'élever à une vie humaine, ou mettre sa sécurité en danger. Nous parlons ici de défense, de la santé, de conduite autonome, etc.

Inspirés par les récentes avancées dans le domaine de la traduction automatique, les chercheurs ont de plus en plus exploré la combinaison de la vision par ordinateur et du traitement automatique des langues, donnant naissance à un champ d'étude florissant à l'intersection de ces deux domaines, à savoir la "Vision au Langage". Une facette spécifique de ce domaine, connue sous le nom de sous-titrage d'images, a fait l'objet d'une attention particulière. Le sous-titrage vise à générer des descriptions textuelles précises et représentatives du contenu des images. Par exemple, dans le contexte de l'imagerie médicale, le sous-titrage d'images a récemment pris un essor considérable en raison de ses capacités remarquables à assister et à accélérer les processus de diagnostic et de traitement [Pavlopoulos et al., 2019]. En outre, la conduite autonome est l'un des domaines dans lesquels le sous-titrage d'images joue un rôle important [Mori et al., 2021]. Il permet aux véhicules autonomes et aux conducteurs de mieux comprendre ce qui les entoure, en

générant des descriptions pour les images capturées par les caméras embarquées, ce qui permet à ces véhicules d'identifier et de décrire des objets tels que des obstacles, des piétons, des panneaux de signalisation, etc. La gestion des catastrophes naturelles est un autre domaine dans lequel le sous-titrage d'images démontre son utilité [Klerings et al., 2019] en facilitant l'analyse des images provenant des satellites/réseaux sociaux, ce qui pourrait contribuer à accélérer les opérations de sauvetage.

La plupart des architectures de sous-titrage adhèrent au cadre Encodeur-Décodeur, en se fondant sur des modules d'apprentissage profond tels que les CNN et LSTM. Quelles que soient les particularités de chaque modèle utilisé dans le sous-titrage, le fonctionnement interne des composants d'encodage et de décodage est souvent considéré comme boîte noire, car leurs mécanismes de fonctionnement sont cachés à l'utilisateur final et ne fournissent aucune explication sur la manière dont leurs décisions ont été prises. Dans le contexte des domaines critiques et sensibles susmentionnés, la nécessité d'une explication devient plus évidente, car une prédiction erronée ou une ambiguïté, aussi minime soit-elle, peut avoir des conséquences considérables. Détecter une pathologie à partir d'une description radiographique générée par une machine sans connaître les raisons explicites de son identification ne sera pas nécessairement adopté par le public et ne suscitera pas sa confiance, et peut donc exposer la vie du patient à des risques élevés en cas d'erreur. Les actions des véhicules autonomes doivent être accompagnées d'explications afin d'identifier les situations critiques pour la sécurité et d'atténuer les risques.

L'intelligence artificielle explicable (XAI) est un domaine en pleine évolution visant à rendre les systèmes d'IA plus transparents et compréhensibles pour les humains [Adadi and Berrada, 2018], permettant aux utilisateurs de comprendre comment et pourquoi ces systèmes atteignent des résultats spécifiques. Malgré l'intérêt croissant pour la XAI, l'explicabilité des modèles de sous-titrage reste relativement peu exploré en raison de la complexité inhérente à ce domaine. Seules quelques études ont donc été proposées dans la littérature, la plupart d'entre elles cherchant principalement à identifier les relations causales qui lient les sorties et les entrées du modèle. Par exemple, dans leur recherche d'explications multimodales, [Sun et al., 2022] ont tenté de retracer les décisions du modèle de sous-titrage à partir d'explications au niveau de la vision et du langage. De la même manière, dans le contexte du sous-titrage d'images médicales, [Beddiar and Oussalah, 2023] ont conçu un module explicable qui relie les mots générés aux régions de l'image. En outre, il met l'accent sur l'importance des mots, en soulignant les mots qui ont influencé de manière significative l'encodeur lors du calcul

des plongements des mots cibles/courants.

Cependant, lorsqu'il s'agit de systèmes multimodaux, ces méthodes d'explication ont tendance à ne pas souligner explicitement les rôles des modalités individuelles dans la prédiction du résultat cible (mot de la légende dans le cadre du sous-titrage d'images). La plupart de ces méthodes présentent les mêmes limites que les approches XAI conventionnelles, notamment l'absence de prise en compte de l'influence des composantes/modalités de l'architecture sur le processus de décision et, par conséquent, sur les informations explicatives obtenues par la suite. Le simple fait d'établir un lien basé sur la pertinence entre les résultats du modèle et les données d'entrée ou les représentations intermédiaires, sans mettre l'accent sur la pertinence de chaque composant, peut suggérer de manière inappropriée un pouvoir explicatif uniforme pour toutes les modalités. Une telle théorie peut s'avérer insuffisante pour parvenir à une compréhension globale du comportement du modèle. En outre, le défi réside également dans la nature complexe des systèmes multimodaux par rapport à d'autres méthodes basées sur l'apprentissage profond. Cette complexité découle principalement de la nécessité d'intégrer et d'interpréter avec précision des données hétérogènes provenant de diverses modalités tout en gérant soigneusement les dépendances intermodales.

Étude de la littérature

Au vu des recherches menées dans ce domaine, il apparaît que certaines tendances et caractéristiques prévalent. En particulier, de nombreuses approches de sous-titrage d'images explicable (XIC) existantes fonctionnent de manière post-hoc, en proposant principalement des explications locales. Alors que le sous-titrage d'images traite la modalité linguistique du côté de la sortie, certains travaux de recherche se sont concentrés récemment à rechercher des explications au sein de cette modalité. Cela peut se faire en conjonction avec la modalité visuelle ou indépendamment. Les mécanismes utilisés dans ces approches offrent un moyen perceptible de représenter les explications. En outre, il convient de mentionner que la pluspart des travaux en XIC s'appuie principalement sur des ensembles de données à grande échelle bien établis, tels que MSCOCO et Flickr, alors que les applications spécifiques à un domaine peuvent manquer de données librement accessibles.

Identification de l'influence des composants

S'appuyant sur les idées tirées de ces travaux antérieurs, en particulier ceux qui portent sur les explications visuelles et linguistiques [Sun et al., 2022, Beddiar and Oussalah, 2023], où un degré de parité entre ces deux modalités est implicitement supposé, notre première contribution visait à répondre à une question fondamentale concernant la fiabilité des explications et les sources à partir desquelles celles-ci devraient être dérivées. Cette étude visait à déterminer si ces modalités jouent un rôle particulier dans le processus de décision, une facette qui a souvent été sous-explorée dans le contexte plus large de l'XAI. Le concept de base s'articule autour de l'identification et de l'isolement des composants de l'architecture qui jouent le rôle le plus important dans le processus de sous-titrage. L'objectif était de déterminer si les décisions sont basées exclusivement sur les données d'entrée, ou si elles sont influencées par d'autres éléments, tels que des déséquilibres dans le degré d'influence exercé par divers composants/modalités au sein de l'architecture de sous-titrage. Cela permet d'impliquer le comportement des composants internes dans le processus d'explication qui s'ensuit.

Nous avons proposé une approche explicative pour les modèles de sous-titrage d'images, en exploitant l'espace de représentation (latent) de l'information. Nous avons démontré que l'explicabilité dans les modèles profonds va au-delà de l'établissement de liens simples entre les sorties et les entrées, et qu'elle ne doit pas nécessairement être perceptive. Pour ce faire, nous avons proposé un paradigme de perturbation, dans lequel l'information au sein des modèles profonds est modifiée plutôt que tronquée. Les perturbations se sont avérées être un moyen plus cohérent et plus perspicace de disséquer le comportement du modèle, transcendant le simple mécanisme de masquage. Nous avons également proposé une nouvelle mesure d'évaluation "MSICE" ancrée dans la morphologie et la sémantique, relevant le défi intrinsèque de l'évaluation de phrases ayant une sémantique similaire mais des formes de surface différentes.

Nos résultats montrent que cette approche améliore le processus d'évaluation et permet une évaluation plus fine des légendes. La révélation la plus importante de cette étude est le discernement de l'influence substantielle exercée par la composante visuelle dans les architectures de sous-titrage, éclipsant l'impact de l'élément linguistique. Par conséquent, nous nous attendons à ce que la partie visuelle constitue un point clé pour les explications ultérieures du modèle. Au-delà de leur impact sur le domaine du sous-titrage d'images, nos résultats sont significatifs pour un éventail plus large de tâches vision-langage ou encore de systèmes multimodaux. Les résultats actuels ont le potentiel de catalyser le développement

d'explications précises, subtiles et fiables dans divers domaines, en s'assurant que la synergie entre la vision et le langage produit des idées qui sont à la fois complémentaires et fiables.

Explications basées sur les attributions

En nous appuyant sur les résultats de nos recherches antérieures, nous avons entrepris une étude approfondie tenant compte des spécificités des composants de l'architecture, en mettant en évidence la question cruciale du compromis entre la capacité d'explication d'un modèle et sa complexité. Alors que certaines approches optent pour des méthodes d'explication plus complexes [Sun et al., 2022, DEWI et al., 2023] telles que les techniques de rétropropagation de la pertinence et autres, nous estimons qu'il est essentiel d'examiner si ces méthodes chronophages apportent réellement une valeur ajoutée par rapport à des alternatives plus simples [Al-Shouha and Szűcs, 2023, Sahay et al., 2021].

S'appuyant sur les avantages que nous avons observés dans l'espace latent, notre deuxième contribution étend son application à la fois à la génération et à l'évaluation d'explications. Cette approche holistique nous permet d'extraire des informations plus concrètes et tangibles pour élucider les processus de décision en sous-titrage d'images. Nous avons ainsi conçu des variantes de deux méthodes d'explication populaires, LIME et LRP, en les adaptant au paradigme de l'espace latent. Le choix de ces deux méthodes est loin d'être arbitraire. Il visait à étudier les relations causales potentielles entre la qualité des explications, la portée des méthodes d'explication (locale ou globale) et leurs subtilités. En particulier, la méthode BU-LRP se distingue comme une nouvelle contribution dans le domaine, car elle exploite les caractéristiques de type bas-en-haut, en explorant leur potentiel éventuel pour fournir des explications améliorées dans l'espace latent. Grâce à une étude comparative, nous avons évalué l'efficacité de ces méthodes à fournir des explications précises. Les deux approches ont donné des résultats comparables en termes de qualité des explications.

En outre, nous avons introduit le nouveau concept d'ablation latente (masquage latent) pour évaluer la qualité des explications. Contrairement à l'ablation classique, l'ablation latente opère dans les couches latentes de l'architecture. Cela a permis une manipulation plus fine des objets de l'image, offrant plusieurs niveaux d'altération et permettant une évaluation plus précise. Nous avons également examiné l'impact de la complétude du concept, c'est-à-dire de l'objet entier par rapport aux caractéristiques individuelles, sur

la performance et l'évaluation de l'explicabilité. Cela nous a conduits à concevoir deux versions de LIME et d'ablation latente, une se concentrant sur le traitement de l'objet complet ou des caractéristiques individuelles.

Nos résultats indiquent que si le traitement de l'objet n'améliore pas nécessairement la robustesse du modèle d'explication, il reste un élément crucial dans l'évaluation de la qualité de l'explication, en particulier dans le cas de l'ablation latente. La portée d'une méthode explicative s'est avéré moins décisif dans la quête d'une qualité d'explication supérieure, mais joue plutôt un rôle dans la détermination de la granularité des explications produites. Il convient de noter que le choix entre ces deux approches n'est pas uniquement dicté par leur portée, mais s'étend au degré de subtilité des résultats escomptés. Dans ce contexte et en s'appuyant sur l'explication subtile fournie par BU-LRP, le chapitre suivant montrera comment les explications obtenues au niveau des éléments individuels du vecteur de caractéristiques peuvent élargir les horizons de l'explicabilité.

Discernement des concepts latents

Construire des dépendances complexes entre les données pour améliorer les performances globales est devenu relativement simple pour de nombreux modèles d'apprentissage automatique. Cependant, démêler les structures complexes que ces modèles ont tendance à construire au sein de leurs composants internes s'avère être une tâche beaucoup plus difficile. Comme l'ont souligné de nombreuses recherches dans la littérature, l'exploration de l'espace latent constitue un moyen précieux d'étudier la manière dont les modèles d'apprentissage automatique structurent l'information, ainsi que leurs processus d'encodage et de décodage. Étonnamment, seules quelques études se sont aventurées dans cette voie, probablement en raison de sa difficulté inhérente.

Dans cette contribution, nous nous concentrons sur l'étude des concepts sémantiques tels que la position, la taille et la quantité des objets. Cette approche permet de combler le décalage entre l'encodage de bas niveau et la compréhension sémantique de haut niveau des modèles de sous-titrage. À cette fin, nous avons introduit une approche d'exploration qui utilise l'édition d'entrée pour découvrir les facettes cachées du comportement du modèle, en mettant particulièrement l'accent sur le concept quantitatif de l'information visuelle. Cette approche nous a permis d'obtenir des informations précieuses sur le rôle de ce concept et son influence sur les décisions du modèle, tant au niveau de l'encodage de l'entrée que du décodage de la sortie. Nos résultats ont notamment révélé une relation

dynamique entre le nombre d'insertions d'un objet et sa saillance dans les processus de détection visuelle et de prédiction langagière. Nous avons également constaté que pour certaines catégories d'objets, la saillance peut ne pas se manifester même après de multiples insertions, potentiellement en raison d'une incohérence contextuelle ou d'une omission sélective par le décodeur de langage.

Conclusion

Tout au long de ce travail qui s'étend sur plusieurs chapitres, nous avons relevé le défi d'améliorer l'explicabilité dans les systèmes de sous-titrage d'images. Chaque chapitre aborde une facette distincte de cette tâche, et nos recherches nous ont amenés à utiliser différentes méthodologies. Nous avons principalement étudié les architectures de sous-titrage d'images basées sur le cadre Encoder-Attention-Decoder, qui incorpore des composants d'apprentissage profond tels que les CNN et les LSTM. Il est essentiel de souligner une facette distinctive notable de l'ensemble du présent travail : notre concentration sur l'exploitation de l'espace latent pour la recherche et la dérivation d'explications ce qui représente une démarche ambitieuse dans le domaine de l'explicabilité en sous-titrage d'images (XIC).

Les divers résultats de cette thèse ont révélé des pistes prometteuses pour la poursuite de la recherche et de l'exploration pouvant être classées en plusieurs catégories. Une piste intéressante à envisager consiste à mener une étude approfondie des sources potentielles de biais et d'hallucinations dans les architectures de sous-titrage. Une attention particulière devrait également être accordée à la compréhension de la manière dont les objets saillants sont identifiés et distingués de l'arrière-plan des scènes visuelles. Ces facteurs peuvent jouer un rôle déterminant dans les décisions prises par les systèmes de sous-titrage et méritent d'être étudiés en profondeur. Bien qu'une partie de ce travail se soit concentrée sur les méthodes d'explication LRP et LIME en raison de leur adoption généralisée dans le domaine, l'approche de l'espace latent développée dans cette thèse peut être étendue pour englober d'autres techniques d'explication. Par exemple, des études futures pourraient explorer l'applicabilité des explications de Shapley et évaluer leur compatibilité avec le cadre proposé. Notre exploration de l'aspect de la quantification et de sa corrélation potentielle avec l'aspect de la saillance a posé des bases solides pour des recherches ultérieures. Des études approfondies peuvent porter sur d'autres attributs visuels, tels que la position et la taille des objets, et évaluer leurs relations potentielles avec l'aspect de la saillance. Les résultats de cette thèse, en conjonction avec les directions de recherche prospectives mentionnées

ci-dessus, promettent une applicabilité élargie à des architectures plus complexes telles que celles basées sur les transformers ou les graphes. Une telle expansion comprendrait une évaluation approfondie et une exploration explicative comparative à travers un spectre diversifié d'architectures de sous-titrage d'images existantes, offrant ainsi un aperçu plus approfondi de leurs mécanismes opérationnels uniques et de leurs caractéristiques distinctives.

Dans cette thèse, nous avons poursuivi l'objectif de rendre les architectures de sous-titrage plus interprétables, tant au niveau fondamental qu'au niveau de l'application, ce qui a permis d'ouvrir la voie à d'autres recherches permettant d'aborder les problèmes croissants de l'explicabilité des systèmes de sous-titrage en particulier et multimodaux en général.

Bibliography

[Adadi and Berrada, 2018] Adadi, A. and Berrada, M. (2018). Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, 6(c):52138–52160.

[Ahmadi and Agrawal, 2023] Ahmadi, S. and Agrawal, A. (2023). An examination of the robustness of reference-free image captioning evaluation metrics. *arXiv preprint arXiv:2305.14998*.

[Al-Shouha and Szűcs, 2023] Al-Shouha, M. and Szűcs, G. (2023). Pic-xai: Post-hoc image captioning explanation using segmentation. In *2023 IEEE 17th International Symposium on Applied Computational Intelligence and Informatics (SACI)*, pages 000033–000038.

[Anderson et al., 2016] Anderson, P., Fernando, B., Johnson, M., and Gould, S. (2016). Spice: Semantic propositional image caption evaluation. In Leibe, B., Matas, J., Sebe, N., and Welling, M., editors, *Computer Vision – ECCV 2016*, pages 382–398, Cham. Springer International Publishing.

[Anderson et al., 2018] Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., and Zhang, L. (2018). Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[Bach et al., 2015] Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., and Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLOS ONE*, 10(7):1–46.

[Banerjee and Lavie, 2005] Banerjee, S. and Lavie, A. (2005). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine*

Translation and/or Summarization, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

[Barredo Arrieta et al., 2020] Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., and Herrera, F. (2020). Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 58:82–115.

[Bastani et al., 2017] Bastani, O., Kim, C., and Bastani, H. (2017). Interpretability via model extraction. *CoRR*, abs/1706.09773.

[Beddiar and Oussalah, 2023] Beddiar, R. and Oussalah, M. (2023). Chapter 12 - explainability in medical image captioning. In Benois-Pineau, J., Bourqui, R., Petkovic, D., and Quénot, G., editors, *Explainable Deep Learning AI*, pages 239–261. Academic Press.

[Bengio et al., 2013] Bengio, Y., Courville, A., and Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828.

[Borch and Min, 2022] Borch, C. and Min, B. H. (2022). Toward a sociology of machine learning explainability: Human–machine interaction in deep neural network-based automated trading. *Big Data & Society*, 9(2):20539517221111361.

[Borkar and Karam, 2019] Borkar, T. S. and Karam, L. J. (2019). Deepcorrect: Correcting dnn models against image distortions. *IEEE Transactions on Image Processing*, 28(12):6022–6034.

[Buchanan and Smith, 1988] Buchanan, B. G. and Smith, R. G. (1988). Fundamentals of expert systems. *Annual review of computer science*, 3(1):23–58.

[Burkart and Huber, 2021] Burkart, N. and Huber, M. F. (2021). A survey on the explainability of supervised machine learning. *Journal of Artificial Intelligence Research*, 70:245–317.

[Carvalho et al., 2019] Carvalho, D. V., Pereira, E. M., and Cardoso, J. S. (2019). Machine learning interpretability: A survey on methods and metrics. *Electronics (Switzerland)*, 8(8):1–34.

[Casalicchio et al., 2019] Casalicchio, G., Molnar, C., and Bischl, B. (2019). Visualizing

the feature importance for black box models. In Berlingerio, M., Bonchi, F., Gärtner, T., Hurley, N., and Ifrim, G., editors, *Machine Learning and Knowledge Discovery in Databases*, pages 655–670, Cham. Springer International Publishing.

[Cau et al., 2020] Cau, F. M., Spano, L. D., Tintarev, N., et al. (2020). Considerations for applying logical reasoning to explain neural network outputs. In *CEUR WORKSHOP PROCEEDINGS*, volume 2742, pages 96–103. CEUR-WS.

[Chen et al., 2020] Chen, Z., Bei, Y., and Rudin, C. (2020). Concept whitening for interpretable image recognition. *Nature Machine Intelligence*, 2(12):772–782.

[Clark et al., 2019] Clark, K., Khandelwal, U., Levy, O., and Manning, C. D. (2019). What does bert look at? an analysis of bert's attention. *arXiv preprint arXiv:1906.04341*.

[Clinciu and Hastie, 2019] Clinciu, M. and Hastie, H. (2019). A survey of explainable ai terminology. In *Proceedings of the 1st workshop on interactive natural language technology for explainable artificial intelligence (NL4XAI 2019)*, pages 8–13.

[Cui et al., 2018] Cui, Y., Yang, G., Veit, A., Huang, X., and Belongie, S. (2018). Learning to evaluate image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[Daisuke, 2018] Daisuke, W. (2018). Self-driving uber car kills pedestrian in arizona, where robots roam. *The New York Times*. `https://www.nytimes.com/2018/03/19/technology/uber-driverless-fatality.html`. Accessed: 2023-19-09.

[Deutsch et al., 2022] Deutsch, D., Dror, R., and Roth, D. (2022). On the limitations of reference-free evaluations of generated text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10960–10977, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

[Devlin et al., 2019] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

[DEWI et al., 2023] DEWI, C., CHEN, R.-C., YU, H., and JIANG, X. (2023). Xai for

image captioning using shap. *Journal of Information Science & Engineering*, 39(4).

[Doshi-Velez and Kim, 2017] Doshi-Velez, F. and Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

[Dosovitskiy et al., 2021] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

[Došilović et al., 2018] Došilović, F. K., Brčić, M., and Hlupić, N. (2018). Explainable artificial intelligence: A survey. In *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, pages 0210–0215.

[Elguendouze et al., 2022] Elguendouze, S., de Souto, M. C. P., Hafiane, A., and Halftermeyer, A. (2022). Towards explainable deep learning for image captioning through representation space perturbation. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.

[Elguendouze et al., 2023] Elguendouze, S., Hafiane, A., de Souto, M. C., and Halftermeyer, A. (2023). Explainability in image captioning based on the latent space. *Neurocomputing*, 546:126319.

[Fong and Vedaldi, 2017] Fong, R. C. and Vedaldi, A. (2017). Interpretable explanations of black boxes by meaningful perturbation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 3429–3437.

[Gilpin et al., 2018] Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., and Kagal, L. (2018). Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 80–89.

[Guidotti et al., 2018] Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., and Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM Comput. Surv.*, 51(5).

[Guo and Farrell, 2021] Guo, P. and Farrell, R. (2021). Semantic network interpretation.

[Gurumoorthy et al., 2019] Gurumoorthy, K. S., Dhurandhar, A., Cecchi, G., and Aggar-

wal, C. (2019). Efficient data representation by selecting prototypes with importance weights.

[Han and Choi, 2018] Han, S.-H. and Choi, H.-J. (2018). Explainable image caption generator using attention and bayesian inference. In *2018 International Conference on Computational Science and Computational Intelligence (CSCI)*, pages 478–481.

[He et al., 2017] He, K., Gkioxari, G., Dollar, P., and Girshick, R. (2017). Mask r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

[He et al., 2016] He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[Hendricks et al., 2016] Hendricks, L. A., Akata, Z., Rohrbach, M., Donahue, J., Schiele, B., and Darrell, T. (2016). Generating visual explanations. In Leibe, B., Matas, J., Sebe, N., and Welling, M., editors, *Computer Vision – ECCV 2016*, pages 3–19, Cham. Springer International Publishing.

[Herdade et al., 2019] Herdade, S., Kappeler, A., Boakye, K., and Soares, J. (2019). Image captioning: Transforming objects into words. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 32, pages 11137–11147. Curran Associates, Inc.

[Hessel et al., 2021] Hessel, J., Holtzman, A., Forbes, M., Le Bras, R., and Choi, Y. (2021). CLIPScore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7514–7528, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

[Honegger, 2018] Honegger, M. (2018). Shedding light on black box machine learning algorithms: Development of an axiomatic framework to assess the quality of methods that explain individual predictions. *arXiv preprint arXiv:1808.05054*.

[Huang et al., 2019] Huang, L., Wang, W., Chen, J., and Wei, X.-Y. (2019). Attention on attention for image captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4634–4643.

[Huang and Marques-Silva, 2023] Huang, X. and Marques-Silva, J. (2023). The inade-

quacy of shapley values for explainability. *arXiv preprint arXiv:2302.08160*.

[Huk Park et al., 2017] Huk Park, D., Hendricks, L. A., Akata, Z., Rohrbach, A., Schiele, B., Darrell, T., and Rohrbach, M. (2017). Attentive Explanations: Justifying Decisions and Pointing to the Evidence (Extended Abstract). *arXiv e-prints*, page arXiv:1711.07373.

[Ilinykh and Dobnik, 2021] Ilinykh, N. and Dobnik, S. (2021). What does a language-and-vision transformer see: The impact of semantic information on visual representations. *Frontiers in Artificial Intelligence*, 4.

[Jeanneret et al., 2023] Jeanneret, G., Simon, L., and Jurie, F. (2023). Adversarial counterfactual visual explanations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16425–16435.

[Jiang et al., 2019] Jiang, M., Huang, Q., Zhang, L., Wang, X., Zhang, P., Gan, Z., Diesner, J., and Gao, J. (2019). TIGEr: Text-to-image grounding for image caption evaluation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2141–2152, Hong Kong, China. Association for Computational Linguistics.

[Karpathy and Fei-Fei, 2015] Karpathy, A. and Fei-Fei, L. (2015). Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[Kiros et al., 2014] Kiros, R., Salakhutdinov, R., and Zemel, R. S. (2014). Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539*.

[Klerings et al., 2019] Klerings, A., Tang, S., and Chen, Z. (2019). Structuralizing disaster-scene data through auto-captioning. In *Proceedings of the 2nd ACM SIGSPATIAL International Workshop on Advances on Resilient and Intelligent Cities*, ARIC'19, page 29–32, New York, NY, USA. Association for Computing Machinery.

[Kusner et al., 2015] Kusner, M., Sun, Y., Kolkin, N., and Weinberger, K. (2015). From word embeddings to document distances. In Bach, F. and Blei, D., editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 957–966, Lille, France. PMLR.

[Laugros et al., 2019] Laugros, A., Caplier, A., and Ospici, M. (2019). Are adversarial robustness and common perturbation robustness independant attributes? In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0.

[Lee et al., 2021] Lee, H., Yoon, S., Dernoncourt, F., Bui, T., and Jung, K. (2021). UMIC: An unreferenced metric for image captioning via contrastive learning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 220–226, Online. Association for Computational Linguistics.

[Letham et al., 2015] Letham, B., Rudin, C., McCormick, T. H., and Madigan, D. (2015). Interpretable classifiers using rules and Bayesian analysis: Building a better stroke prediction model. *The Annals of Applied Statistics*, 9(3):1350 – 1371.

[Li et al., 2019] Li, G., Zhu, L., Liu, P., and Yang, Y. (2019). Entangled transformer for image captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

[Li et al., 2018] Li, O., Liu, H., Chen, C., and Rudin, C. (2018). Deep learning for case-based reasoning through prototypes: A neural network that explains its predictions. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).

[Lin, 2004] Lin, C.-Y. (2004). ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

[Lin et al., 2014] Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In Fleet, D., Pajdla, T., Schiele, B., and Tuytelaars, T., editors, *Computer Vision – ECCV 2014*, pages 740–755, Cham. Springer International Publishing.

[Liu et al., 2023] Liu, J., Jin, H., Xu, G., Lin, M., Wu, T., Nour, M., Alenezi, F., Alhudhaif, A., and Polat, K. (2023). Aliasing black box adversarial attack with joint self-attention distribution and confidence probability. *Expert Systems with Applications*, 214:119110.

[Liu et al., 2022] Liu, M., Hu, H., Li, L., Yu, Y., and Guan, W. (2022). Chinese image caption generation via visual attention and topic modeling. *IEEE Transactions on Cybernetics*, 52(2):1247–1257.

[Liu et al., 2020] Liu, M., Li, L., Hu, H., Guan, W., and Tian, J. (2020). Image caption generation with dual attention mechanism. *Information Processing & Management*, 57(2):102178.

[Liu et al., 2021] Liu, W., Chen, S., Guo, L., Zhu, X., and Liu, J. (2021). Cptr: Full transformer network for image captioning.

[Lu et al., 2017] Lu, J., Xiong, C., Parikh, D., and Socher, R. (2017). Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 375–383.

[Lundberg and Lee, 2017] Lundberg, S. and Lee, S. (2017). A unified approach to interpreting model predictions. *CoRR*, abs/1705.07874.

[Madhyastha et al., 2019] Madhyastha, P., Wang, J., and Specia, L. (2019). VIFIDEL: Evaluating the visual fidelity of image descriptions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6539–6550, Florence, Italy. Association for Computational Linguistics.

[Miller, 1995] Miller, G. A. (1995). Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41.

[Miller, 2019] Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38.

[Mokady et al., 2021] Mokady, R., Hertz, A., and Bermano, A. H. (2021). Clipcap: Clip prefix for image captioning. *ArXiv*, abs/2111.09734.

[Moore and Swartout, 1988] Moore, J. D. and Swartout, W. R. (1988). Explanation in expert systemss: A survey. Technical report, University of Southern California Marina del Rey Information Sciences Inst.

[Mori et al., 2021] Mori, Y., Hirakawa, T., Yamashita, T., and Fujiyoshi, H. (2021). Image captioning for near-future events from vehicle camera images and motion information. In *2021 IEEE Intelligent Vehicles Symposium (IV)*, pages 1378–1384. IEEE.

[Nguyen et al., 2016] Nguyen, A. M., Dosovitskiy, A., Yosinski, J., Brox, T., and Clune, J. (2016). Synthesizing the preferred inputs for neurons in neural networks via deep generator networks. *CoRR*, abs/1605.09304.

[Otani et al., 2023] Otani, M., Togashi, R., Sawai, Y., Ishigami, R., Nakashima, Y., Rahtu, E., Heikkilä, J., and Satoh, S. (2023). Toward verifiable and reproducible human evaluation for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14277–14286.

[Pan et al., 2020] Pan, Y., Yao, T., Li, Y., and Mei, T. (2020). X-linear attention networks for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10971–10980.

[Papineni et al., 2002] Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

[Pavlopoulos et al., 2019] Pavlopoulos, J., Kougia, V., and Androutsopoulos, I. (2019). A survey on biomedical image captioning. In *Proceedings of the Second Workshop on Shortcomings in Vision and Language*, pages 26–36, Minneapolis, Minnesota. Association for Computational Linguistics.

[Pelka et al., 2021] Pelka, O., Ben Abacha, A., G. Seco de Herrera, A., Jacutprakart, J., Friedrich, C. M., and Müller, H. (2021). Overview of the imageclefmed 2021 concept & caption prediction task. *Proceedings of the CLEF 2021 Conference and Labs of the Evaluation Forum - working notes*, (CONFERENCE):Pp. 1101–1112.

[Radford et al., 2021] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. (2021). Learning transferable visual models from natural language supervision. In Meila, M. and Zhang, T., editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, PMLR.

[Ras et al., 2022] Ras, G., Xie, N., van Gerven, M., and Doran, D. (2022). Explainable deep learning: A field guide for the uninitiated. *Journal of Artificial Intelligence Research*, 73:329–397.

[Ren et al., 2015] Ren, S., He, K., Girshick, R., and Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28:91–99.

[Rennie et al., 2017] Rennie, S. J., Marcheret, E., Mroueh, Y., Ross, J., and Goel, V. (2017). Self-critical sequence training for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[Ribeiro et al., 2016] Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). "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*, KDD '16, page 1135–1144, New York, NY, USA. Association for Computing Machinery.

[Ribeiro et al., 2018] Ribeiro, M. T., Singh, S., and Guestrin, C. (2018). Anchors: High-precision model-agnostic explanations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).

[Robnik-Šikonja and Bohanec, 2018] Robnik-Šikonja, M. and Bohanec, M. (2018). Perturbation-based explanations of prediction models. *Human and Machine Learning: Visible, Explainable, Trustworthy and Transparent*, pages 159–175.

[Rudin et al., 2022] Rudin, C., Chen, C., Chen, Z., Huang, H., Semenova, L., and Zhong, C. (2022). Interpretable machine learning: Fundamental principles and 10 grand challenges. *Statistics Surveys*, 16(none):1 – 85.

[Sahay et al., 2021] Sahay, S., Omare, N., and Shukla, K. K. (2021). An approach to identify captioning keywords in an image using lime. In *2021 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS)*, pages 648–651.

[Sai et al., 2022] Sai, A. B., Mohankumar, A. K., and Khapra, M. M. (2022). A survey of evaluation metrics used for nlg systems. *ACM Comput. Surv.*, 55(2).

[Schallner et al., 2020] Schallner, L., Rabold, J., Scholz, O., and Schmid, U. (2020). Effect of superpixel aggregation on explanations in lime–a case study with biological data. In *Machine Learning and Knowledge Discovery in Databases: International Workshops of ECML PKDD 2019, Würzburg, Germany, September 16–20, 2019, Proceedings, Part I*, pages 147–158. Springer.

[Segaran, 2007] Segaran, T. (2007). Collective intelligence-building smart web 2.0 applications. *Newton: O'Reilly*.

[Selvaraju et al., 2017] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017). Grad-cam: Visual explanations from deep networks via

gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 618–626.

[Simonyan and Zisserman, 2015] Simonyan, K. and Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In Bengio, Y. and LeCun, Y., editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

[Sun et al., 2022] Sun, J., Lapuschkin, S., Samek, W., and Binder, A. (2022). Explain and improve: Lrp-inference fine-tuning for image captioning models. *Information Fusion*, 77:233–246.

[Sundararajan et al., 2017] Sundararajan, M., Taly, A., and Yan, Q. (2017). Axiomatic attribution for deep networks. In Precup, D. and Teh, Y. W., editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3319–3328. PMLR.

[Tan et al., 2017] Tan, S., Caruana, R., Hooker, G., and Lou, Y. (2017). Detecting bias in black-box models using transparent model distillation. *arXiv preprint arXiv:1710.06169*.

[Tan et al., 2020] Tan, S., Soloviev, M., Hooker, G., and Wells, M. T. (2020). Tree space prototypes: Another look at making tree ensembles interpretable. In *Proceedings of the 2020 ACM-IMS on Foundations of Data Science Conference*, FODS '20, page 23–34, New York, NY, USA. Association for Computing Machinery.

[Tavanaei, 2020] Tavanaei, A. (2020). Embedded encoder-decoder in convolutional networks towards explainable ai. *arXiv preprint arXiv:2007.06712*.

[Tenney et al., 2019] Tenney, I., Xia, P., Chen, B., Wang, A., Poliak, A., McCoy, R. T., Kim, N., Van Durme, B., Bowman, S. R., Das, D., et al. (2019). What do you learn from context? probing for sentence structure in contextualized word representations. *arXiv preprint arXiv:1905.06316*.

[Tjoa and Guan, 2021] Tjoa, E. and Guan, C. (2021). A survey on explainable artificial intelligence (xai): Toward medical xai. *IEEE Transactions on Neural Networks and Learning Systems*, 32(11):4793–4813.

[van Aken et al., 2019] van Aken, B., Winter, B., Löser, A., and Gers, F. A. (2019). How does bert answer questions? a layer-wise analysis of transformer representations. In

Proceedings of the 28th ACM International Conference on Information and Knowledge Management, CIKM '19, page 1823–1832, New York, NY, USA. Association for Computing Machinery.

[Van Lent et al., 2004] Van Lent, M., Fisher, W., and Mancuso, M. (2004). An explainable artificial intelligence system for small-unit tactical behavior. In *Proceedings of the national conference on artificial intelligence*, pages 900–907. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999.

[Vaswani et al., 2017] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. (2017). Attention is all you need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

[Vedantam et al., 2015] Vedantam, R., Zitnick, C. L., and Parikh, D. (2015). Cider: Consensus-based image description evaluation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4566–4575.

[Wang et al., 2021] Wang, S., Yao, Z., Wang, R., Wu, Z., and Chen, X. (2021). Faier: Fidelity and adequacy ensured image caption evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14050–14059.

[Weissman, 1976] Weissman, M. S. (1976). Decisiveness and psychological adjustment. *Journal of Personality Assessment*, 40(4):403–412. PMID: 957087.

[Wexler et al., 2020] Wexler, J., Pushkarna, M., Bolukbasi, T., Wattenberg, M., Viégas, F., and Wilson, J. (2020). The what-if tool: Interactive probing of machine learning models. *IEEE Transactions on Visualization and Computer Graphics*, 26(1):56–65.

[Wu and Song, 2019] Wu, T. and Song, X. (2019). Towards interpretable object detection by unfolding latent structures. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6032–6042.

[Wu et al., 2022a] Wu, T., Wang, X., Qiao, S., Xian, X., Liu, Y., and Zhang, L. (2022a). Small perturbations are enough: Adversarial attacks on time series prediction. *Information Sciences*, 587:794–812.

[Wu et al., 2022b] Wu, T., Yang, N., Chen, L., Xiao, X., Xian, X., Liu, J., Qiao, S., and

Cui, C. (2022b). Ergcn: Data enhancement-based robust graph convolutional network against adversarial attacks. *Information Sciences*, 617:234–253.

[Wu et al., 2023] Wu, T.-W., Huang, J.-H., Lin, J., and Worring, M. (2023). Expert-defined keywords improve interpretability of retinal image captioning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1859–1868.

[Wu et al., 2019] Wu, Y., Kirillov, A., Massa, F., Lo, W.-Y., and Girshick, R. (2019). Detectron2. 2019. *URL https://github. com/facebookresearch/detectron2*, 2(3).

[Xian et al., 2021] Xian, X., Wu, T., Qiao, S., Wang, W., Wang, C., Liu, Y., and Xu, G. (2021). Deepec: Adversarial attacks against graph structure prediction models. *Neurocomputing*, 437:168–185.

[Xu et al., 2015] Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., and Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. In Bach, F. and Blei, D., editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2048–2057, Lille, France. PMLR.

[Xu et al., 2018] Xu, X., Chen, X., Liu, C., Rohrbach, A., Darrell, T., and Song, D. (2018). Fooling vision and language models despite localization and attention mechanism. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4951–4961.

[Yang et al., 2021] Yang, Q., Zhu, X., Fwu, J.-K., Ye, Y., You, G., and Zhu, Y. (2021). Mfpp: Morphological fragmental perturbation pyramid for black-box model explanations. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 1376–1383.

[Yao et al., 2017] Yao, T., Pan, Y., Li, Y., Qiu, Z., and Mei, T. (2017). Boosting image captioning with attributes. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

[You et al., 2016] You, Q., Jin, H., Wang, Z., Fang, C., and Luo, J. (2016). Image captioning with semantic attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 4651–4659.

[Young et al., 2014] Young, P., Lai, A., Hodosh, M., and Hockenmaier, J. (2014). From

image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.

[Yu et al., 2022] Yu, B., Zhong, Z., Qin, X., Yao, J., Wang, Y., and He, P. (2022). Automated testing of image captioning systems. In *Proceedings of the 31st ACM SIGSOFT International Symposium on Software Testing and Analysis*, ISSTA 2022, page 467–479, New York, NY, USA. Association for Computing Machinery.

[Zellers et al., 2018] Zellers, R., Yatskar, M., Thomson, S., and Choi, Y. (2018). Neural motifs: Scene graph parsing with global context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5831–5840.

[Zhang et al., 2020a] Zhang, S., Wang, Z., Xu, X., Guan, X., and Yang, Y. (2020a). Fooled by imagination: Adversarial attack to image captioning via perturbation in complex domain. In *2020 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6.

[Zhang et al., 2020b] Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y. (2020b). Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

[Zhong et al., 2020] Zhong, Y., Wang, L., Chen, J., Yu, D., and Li, Y. (2020). Comprehensive Image Captioning via Scene Graph Decomposition. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 12359 LNCS, pages 211–229. Springer Science and Business Media Deutschland GmbH.

[Zilke et al., 2016] Zilke, J. R., Loza Mencía, E., and Janssen, F. (2016). Deepred–rule extraction from deep neural networks. In *Discovery Science: 19th International Conference, DS 2016, Bari, Italy, October 19–21, 2016, Proceedings 19*, pages 457–473. Springer.

Sofiane ELGUENDOUZE

Approches d'Intelligence Artificielle eXplicables pour le Sous-titrage d'Images

Résumé : L'évolution rapide des modèles de sous-titrage d'images, impulsée par l'intégration de techniques d'apprentissage profond combinant les modalités image et texte, a conduit à des systèmes de plus en plus complexes. Cependant, ces modèles fonctionnent souvent comme des boîtes noires, incapables de fournir des explications transparentes de leurs décisions. Cette thèse aborde l'explicabilité des systèmes de sous-titrage d'images basés sur des architectures Encodeur-Attention-Décodeur, et ce à travers quatre aspects. Premièrement, elle explore le concept d'*espace latent*, s'éloignant ainsi des approches traditionnelles basées sur l'espace de représentation originel. Deuxièmement, elle présente la notion de *caractère décisif*, conduisant à la formulation d'une nouvelle définition pour le concept d'influence/décisivité des composants dans le contexte de sous-titrage d'images explicable, ainsi qu'une approche par perturbation pour la capture du caractère décisif. Le troisième aspect vise à élucider les facteurs influençant la qualité des explications, en mettant l'accent sur la portée des méthodes d'explication. En conséquence, des variantes basées sur l'espace latent de méthodes d'explication bien établies telles que LRP et LIME ont été développées, ainsi que la proposition d'une approche d'évaluation centrée sur l'espace latent, connue sous le nom d'Ablation Latente. Le quatrième aspect de ce travail consiste à examiner ce que nous appelons la *saillance* et la représentation de certains *concepts visuels*, tels que la quantité d'objets, à différents niveaux de l'architecture de sous-titrage.

Mots clés : Intelligence Artificielle Explicable, Sous-titrage d'Image, Sous-titrage d'Image Explicable, Espace de représentation / Espace latent, Caractère décisif, Saillance.

Explainable Artificial Intelligence approaches for Image Captioning

Abstract: The rapid advancement of image captioning models, driven by the integration of deep learning techniques that combine image and text modalities, has resulted in increasingly complex systems. However, these models often operate as black boxes, lacking the ability to provide transparent explanations for their decisions. This thesis addresses the explainability of image captioning systems based on Encoder-Attention-Decoder architectures, through four aspects. First, it explores the concept of the *latent space*, marking a departure from traditional approaches relying on the original representation space. Second, it introduces the notion of *decisiveness*, leading to the formulation of a new definition for the concept of component influence/decisiveness in the context of explainable image captioning, as well as a perturbation-based approach to capturing decisiveness. The third aspect aims to elucidate the factors influencing explanation quality, in particular the scope of explanation methods. Accordingly, latent-based variants of well-established explanation methods such as LRP and LIME have been developed, along with the introduction of a latent-centered evaluation approach called Latent Ablation. The fourth aspect of this work involves investigating what we call *saliency* and the representation of certain *visual concepts*, such as object quantity, at different levels of the captioning architecture.

Keywords: Explainable Artificial Intelligence (XAI), Image Captioning (IC), Explainable Image Captioning (XIC), Representation space / Latent space, Decisiveness, Saliency.

- Laboratoire d'Informatique Fondamentale d'Orléans (LIFO) -
6 Rue Léonard de Vinci, 45067 Orléans, France