HAL
open science

Robust and comprehensive joint image-text representations

Thi Quynh Nhi Tran

To cite this version:

Thi Quynh Nhi Tran. Robust and comprehensive joint image-text representations. Image Processing [eess.IV]. Conservatoire national des arts et metiers - CNAM, 2017. English. NNT : 2017CNAM1096 . tel-01591614

HAL Id: tel-01591614

https://theses.hal.science/tel-01591614v1

Submitted on 21 Sep 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

le cnam

École doctorale Informatique, Télécommunications et Électronique (Paris)

Centre d'Études et de Recherche en Informatique et Communications

THÈSE DE DOCTORAT

présentée par : **Thi Quynh Nhi TRAN**

soutenue le : **3 mai 2017**

pour obtenir le grade de : **Docteur du Conservatoire National des Arts et Métiers**

Spécialité : **Informatique**

Robust and comprehensive joint image-text representations

THÈSE dirigée par

M. CRUCIANU Michel	*Professeur des Universités, CEDRIC-CNAM, Paris*

RAPPORTEURS

M. GRAVIER Guillaume	*Directeur de Recherche, CNRS-IRISA & INRIA, Rennes*
M. QUÉNOT Geogre	*Directeur de Recherche, CNRS-LIG, Grenoble*

PRÉSIDENT

Mme. HUDELOT Céline	*Professeur des Universités, MICS-Centrale Supélec, Paris*

EXAMINATEURS

M. GREFENSTETTE Gregory	*Docteur, Florida Institute for Human & Machine Cognition*
M. BERRANI Sid-Ahmed	*Docteur, Algérie Télécom, Alger*
M. LE BORGNE Hervé	*Docteur, CEA-LIST, Saclay*

Abstract

This thesis investigates the joint modeling of visual and textual content of multimedia documents to address cross-modal problems. Such tasks require the ability to match information across modalities. A common representation space, obtained by *e.g.* Kernel Canonical Correlation Analysis, on which images and text can be both represented and directly compared is a generally adopted solution. Nevertheless, such a joint space still suffers from several deficiencies that may hinder the performance of cross-modal tasks. An important contribution of this thesis is therefore to identify two major limitations of such a space. The first limitation concerns information that is poorly represented on the common space yet very significant for a retrieval task. The second limitation consists in a separation between modalities on the common space, which leads to coarse cross-modal matching.

To deal with the first limitation concerning poorly-represented data, we put forward a model which first identifies such information and then finds ways to combine it with data that is relatively well-represented on the joint space. Evaluations on *text illustration* tasks show that by appropriately identifying and taking such information into account, the results of cross-modal retrieval can be strongly improved. The major work in this thesis aims to cope with the separation between modalities on the joint space to enhance the performance of cross-modal tasks. We propose two representation methods for bi-modal or uni-modal documents that aggregate information from both the visual and textual modalities projected on the joint space. Specifically, for uni-modal documents we suggest a completion process relying on an auxiliary dataset to find the corresponding information in the absent modality and then use such information to build a final bi-modal representation for a uni-modal document. Evaluations show that our approaches achieve state-of-the-art results on several standard and challenging datasets for cross-modal retrieval or bi-modal and cross-modal classification.

Keywords : common representation, cross-modal retrieval, cross-modal classification, (kernel) canonical correlation analysis, multi-modal representation, image and text.

Résumé

La présente thèse étudie la modélisation conjointe des contenus visuels et textuels extraits à partir des documents multimédia pour résoudre les problèmes intermodaux. Ces tâches exigent la capacité de « traduire » l'information d'une modalité vers une autre. Un espace de représentation commun, par exemple obtenu par l'Analyse Canonique des Corrélations ou son extension à noyaux est la solution généralement adoptée. Sur cet espace, images et textes peuvent être représentés par des vecteurs de même type sur lesquels la comparaison intermodale peut se faire directement. Néanmoins, un tel espace commun souffre de plusieurs insuffisances qui peuvent diminuer la performance des ces tâches. La première concerne les informations très importantes dans le contexte de la recherche intermodale mais qui sont mal représentées sur cet espace appris. La deuxième insuffisance porte sur la séparation entre les projections des différentes modalités sur l'espace commun, ce qui conduit à une qualité de traduction peu satisfaisante entre modalités.

Pour faire face au problème concernant les données mal représentées, nous avons proposé un modèle qui identifie tout d'abord ces informations et les combine ensuite avec des données relativement bien représentées sur l'espace commun. Les évaluations sur une tâche d'*illustration de texte* montrent que la prise en compte de ces informations améliore fortement les résultats de la recherche intermodale. La contribution majeure de la thèse se concentre sur le problème de la séparation entre les modalités sur l'espace commun, afin d'améliorer les performances dans les tâches intermodales. Nous mettons en avant deux méthodes de représentation pour les documents bi-modaux ou uni-modaux qui regroupent à la fois des informations visuelles et textuelles projetées sur l'espace commun. Pour les documents uni-modaux, nous proposons un processus de *complétion* basé sur un ensemble de données auxiliaires pour trouver les informations correspondantes dans la modalité absente. Ces informations complémentaires sont ensuite employées pour construire une représentation bi-modale d'un document uni-modal. Nos approches permettent d'améliorer l'état de l'art pour la recherche intermodale ou la classification bi-modale et intermodale.

Mots clés: espace commun de représentation, analyse canonique des corrélations, recherche intermodale, classification intermodale, représentation multimodale, image et texte.

Au cours des années passées, l'explosion des données multimédia est devenue plus importante que jamais avec l'apparition des sites de médias sociaux comme Facebook, Twitter et des sites de partage de contenu comme YouTube, FlickR, Wikipedia, etc. Différentes collections de grande quantité de données multimédias ont été générées à partir de ces sites. Cela exige des méthodes pour stocker, organiser et traiter efficacement de ces grands volumes de données multimédias.

Dans ce contexte, la recherche et la classification des données multimédia ont reçu la plus grande attention de la communauté multimédia en raison de leurs intérêts pratiques. Ces deux technologies sont au cœur de diverses applications multimédias telles que l'annotation d'image, l'illustration automatique de texte, la détection d'événements ou la catégorisation de documents, etc.

Cette thèse se concentre sur deux modalités de données: modalité visuelle représentée par des images et modalité textuelle représentée par des mots-clés, des étiquettes ou des phrases du langage naturel, etc. Normalement, ces contenus visuels et textuels sont transformés en représentations vectorielles, qui sont après utilisées pour la recherche ou la classification. Initialement, la recherche d'image ou la recherche de document textuel se base sur l'exploration des caractéristiques distinctes de chaque modalité individuelle, c'est à dire soit la modalité visuelle, soit la modalité textuelle. Néanmoins, le contenu visuelle et le contenu textuelle apparaissent souvent ensemble dans des collections multimédia et munissent des informations complémentaires l'une à l'autre. Par exemple, des photos sur FlickR sont souvent caractérisée par des descriptions et/ou des étiquettes fournies par des utilisateurs; ou des articles Wikipedia sont généralement illustré par une ou plusieurs images. Par conséquent, la modélisation conjointe de contenu visuel et son contenu textuel associé peut potentiellement améliorer les performances des systèmes de recherche ou de classification des données multimédia.

Le travail principal de cette thèse considère la modélisation conjointe de l'image et du texte pour résoudre des problèmes intermodales, un paradigme plus récent de la recherche d'informations. Les systèmes intermodaux supportent l'interactivité entre les modalités. Par exemple, une recherche intermodale retrouve des images en réponse à une requête textuelle ou des documents textuels en réponse à une requête visuelle. Ces tâches

sont essentielles à de nombreuses applications d'intérêt pratique, telles que l'illustration automatique de texte, le sous-titrage automatique d'image, etc. Cependant, la modélisation intermodale efficace et efficiente reste encore un défi. La raison principale est le "fossé sémantique", qui est connu comme la différence, du point de vue de la compréhension humaine, entre la représentation visuelle d'image et celle textuelle. Autrement dit, le "fossé sémantique" est considéré comme le manque de coïncidence entre les informations que l'on peut extraire à partir des données visuelles et l'interprétation que les mêmes données ont pour un utilisateur dans une situation donnée. Le fait de réduire le "fossé sémantique" entre les représentations visuelles et les représentations textuelles reste un défi majeur dans la modélisation des systèmes multimédia. En outre, ces systèmes souffrent également de l'hétérogénéité entre différentes modalités. Cela se réfère au fait que les modalités visuelles et textuelles sont habituellement décrites selon des schémas totalement différents et ils résident habituellement dans différents espaces de représentation.

Dans cette thèse, nous adressons le problème susmentionné de la modélisation conjointe de contenu visuel et contenu textuel par le développement d'un espace commun de représentation pour ces deux modalité. Un tel espace latent est principalement considéré pour les tâches intermodales car les images et le texte peuvent être représentés dans cet espace sans aucune distinction. Par exemple la recherche intermodale peut se faire sur des représentations communes en les comparant directement. Aussi, cet espace commun peut être utilisé pour traiter diverses tâches bimodales centrées sur la sémantique.

L'espace commun de représentation pour image et texte peut être servi dans plusieurs tâches intermodales. La performance de ces tâches dépendent fortement de la qualité de cet espace. Ce dernier se reflète dans la qualité de mise en correspondance de données entre deux modalités. Pour cela, le modèle doit rapprocher les images et les textes associées sur son espace commun. Jusqu'à présent, la plupart des travaux existants basés sur des espaces communs de représentation se concentrent sur l'étude de représentations complètes des données dans chaque modalité individuelle et ensuite les utilisent pour construire un modèle robuste. Malgré leurs succès relatifs, cela semble insuffisante pour réduire les fossés entre les modalités visuelles et textuelles en raison de plusieurs limitations de cet espace conjoint. En outre, l'intérêt pratique de tel approche relie au montant de ressources utilisés

pour apprendre les représentations communes. Il est évidemment une grande importance dans la pratique, car l'utilisation d'une très grande quantité de données peut conduire à un calcul insoluble, ou au moins un coût trop élevé pour être intéressant.

La motivation de cette thèse est triple.

- **Représentation robuste sur l'espace conjointe**. Un tel espace de représentation commun souffre de plusieurs limitations qu'on identifie dans la thèse. Une utilisation directe de ces projections se traduit par une qualité limitée de mise en correspondance des données entre les deux modalités. Par conséquent, cela diminue la performance des tâches intermodales ou bimodales. Notre motivation consiste à proposer des représentations plus robustes pour image et texte et puis les utiliser pour adresser ces tâches.

- **Espace commun comme ressource universelle**. Nous espérons développer un espace commun de représentation en tant que ressource universelle, à partir d'un grand ensemble de données bi-modales génériques. Ce dernier peut être servi à traiter des problèmes spécifiques telles que la recherche ou la classification d'information venant du même domaine ou d'une autre domaine de cette ressource. L'intérêt de la ressource universelle est d'éviter de réapprendre un espace commun pour chaque problème à partir d'un ensemble de données spécifiques liées aux problèmes.

- **Modèle conceptuel de plus haut niveau sémantique**. Au-delà de la recherche intermodale de type "un à un" des documents multimédia (par exemple la comparaison d'une image à une autre image ou d'une image à un texte, etc) sur l'espace commun, nous désirons faire aussi la comparaison de type "un à plusieurs" sur cet espace. C'est à dire nous voulons estimer combien une image (ou un texte) est similaire à un groupe de textes (ou des images). Pour cela, un modèle multimédia devrait être en mesure de faire correspondre un document donné aux concepts plus généraux détectés dans un ensemble d'autres documents. En pratique, la première étape consiste simplement à apprendre un concept général à partir de données venant d'une seule modalité, ensuite le tester sur une autre modalité.

Notre but principale est d'étudier des représentations jointes des images et textes qui

sont plus robustes et complètes afin d'adresser des tâches intermodales. Pour cela, notre approche repose sur le développement d'un espace de représentation commun pour la modalité visuelle et textuelle. Généralement, cet espace est obtenu par apprentissage, conjointement à partir des représentations visuelles et ses représentations textuelles associées. Différent principes ont été proposés dans la littérature pour apprendre un tel espace. Nous avons choisi de construire un espace commun de représentation résultant d'une maximisation de la relation entre la modalité visuelle et textuelle. Dans cette direction, notre approche est basé sur le technique de l'analyse canonique des corrélations (ACC) et de son extension en utilisant un noyau. En principe, l'ACC recherche un sous-espace commun des deux espaces qui maximise la corrélation des points projetés à partir des données d'origine. Néanmoins, nous avons identifié deux limitations de cet espace commun.

- La première limitation concerne les données mal représentées sur l'espace commun. Les données sont projetées sur l'espace commun de représentation avec des qualités de représentation différentes. Certaines données sont représentées moins bien que les autres. Malheureusement, ces informations mal représentées peuvent être très importantes dans un contexte de la tâche de recherche d'information. Le fait de ne pas prendre en compte ces informations peut réduire considérablement l'efficacité de l'espace commun de représentation.

- La deuxième insuffisance de cet espace commun porte sur la séparation entre les projections des différentes modalités sur l'espace commun. C'est à dire sur l'espace commun, la projection visuelle et textuelle d'un document donné sont éloignées. Ces projections ont tendance à être regroupées par modalité plutôt que par leur sémantique sur cet espace (voir Figure 1). Cela conduit à une qualité de traduction peu satisfaisante entre deux modalités. La performance des tâches intermodales est donc limitée.

Dans cette thèse, nous proposons trois méthodes de représentation jointe des images et textes dans le but de réduire les imperfections susmentionnées sur l'espace commun de représentation.

Première contribution. Une première contribution traite de la limitation concernant

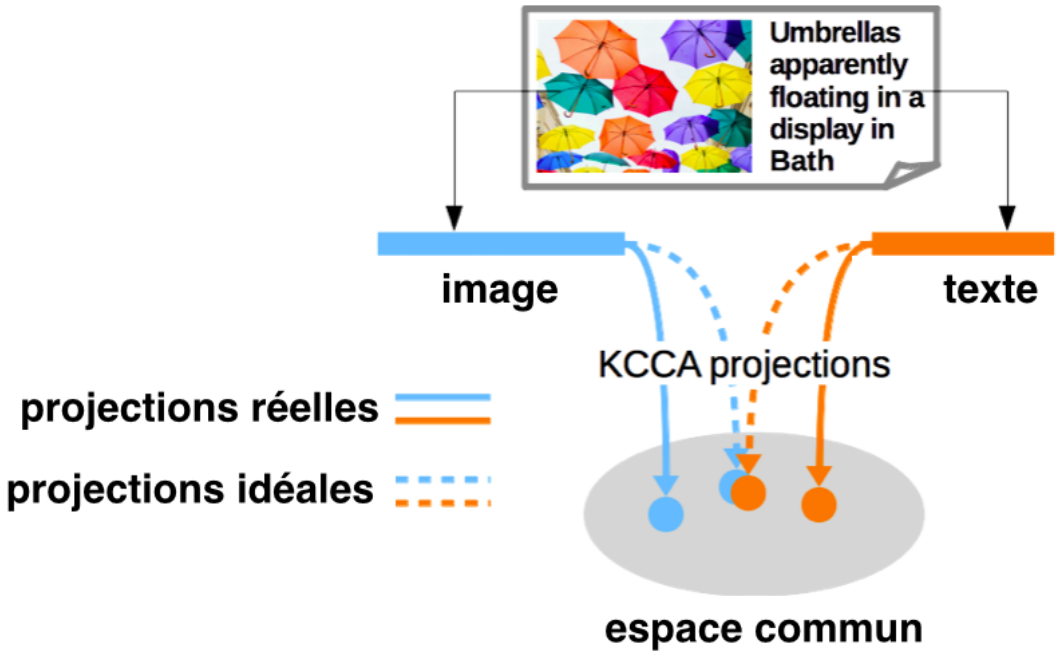

Figure 1: La séparation entre la modalité visuelle et textuelle sur l'espace commun de représentation.

les données pertinentes mais mal représentées sur l'espace commun des images et des textes. Ceci est dû au fait que le développement d'un tel espace commun de représentation basé sur l'ACC repose sur l'extraction de régularités statistiques à partir d'une grande quantité de données d'apprentissage. Les données ayant peu d'occurrences ou des relations très faibles avec d'autres données dans l'ensemble d'apprentissage sont donc ignorées dans cet espace commun. Nous étudions cette limitation de l'espace commun particulièrement dans le contexte de la tâche de recherche d'information. Pour cette tâche, nous disposons une base de références dans laquelle nous cherchons des résultats pertinents pour chaque requête. Ces données de références et des requêtes sont tous projetées sur un espace commun de représentation pour faciliter la recherche. Cet espace est souvent appris à partir d'une base d'apprentissage. Dans des cas pratiques, cette base d'apprentissage est souvent différente de la base de références destinée à la recherche. Notre contribution porte sur la différence

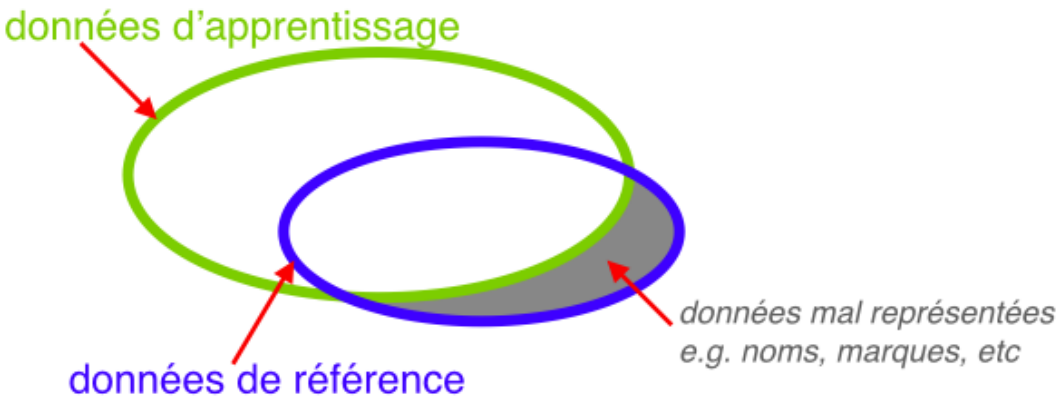

Figure 2: Une première contribution de la thèse concentre sur la différence entre la base d'apprentissage et la base de référence dans le contexte de la tâche de recherche d'information. Cette portion contient potentiellement des données mal représentées sur l'espace commun de représentation appris sur les données d'apprentissage.

entre ces deux bases de données (voir Figure 2). Précisément, nous constatons que des données rares dans la base d'apprentissage ou même nouvelles dans la base de références, telles que *noms* ou *marques*, peuvent être très importantes pour sélectionner des résultats pertinents dans une tâche de recherche d'information. Néanmoins, ces informations sont souvent ignorées sur l'espace jointe de représentation. Si nous comptons seulement sur cet espace, nous ne pouvons pas exploiter ces information. Pour faire face à cette limitation, nous proposons dans une première contribution de cette thèse un modèle qui identifie tout d'abord des informations mal représentées sur l'espace commun et les combine ensuite avec des données relativement bien représentées sur l'espace commun. Concrètement, au lieu de faire la recherche basée seulement sur l'espace commun de représentation, nous la combine avec la recherche sur l'espace initial de représentation. Dans ce travail, nous ne traitons que des informations mal représenté venant de la modalité textuelle.

Nous examinons dans cette contribution comment le modèle proposé est appliqué afin d'adresser la tâche d'illustration automatique de texte. L'illustration de texte consiste à trouver une image appropriée pour illustrer le contenu d'un document textuel donné. Cette tâche a donc la nature de la recherche intermodale entre la modalité visuelle et la

modalité textuelle. Nous montrons que, en identifiant adéquatement des information mal représentées sur l'espace commun de représentation et les prenant en compte, les résultats de la recherche intermodale peuvent être fortement améliorés. Les expériences sont menées sur les trois bases de données différentes: BBC News, Wikipedia2010 et ImageCLEF13. Nous avons montré l'intérêt de notre modèle pour la tâche d'illustration de texte dans différent contextes où la base d'apprentissage utilisée pour apprendre l'espace commun et la base de test pour la recherche d'information sont extraites à partir d'une même collection ou dans un cas plus difficile mais plus réaliste, à partir des collections totalement différentes.

Dans le tableau 1, il est intéressant de noter que plus le nombre de mots textuels mal représentés que le modèle prend en compte, plus l'amélioration du modèle proposé par rapport à l'approche basique basé seulement sur l'espace commun de représentation (appris par l'ACC) est élevée. Par exemple, dans l'évaluation sur des données de Wikipedia2010 où 2,868 mots potentiellement mal représentés sont identifiés, l'amélioration de notre modèle est de +39,1 par rapport à la recherche basée seulement sur l'espace commun appris par l'ACC.

données de test	espace commun appris sur	nombre de mots mal représentés	amélioration p.r.à l'ACC
BBC News	BBC News	41	+4.1
BBC News	ImageCLEF13	208	+18.0
Wikipedia2010	Wikipedia2010	2868	+39.1

Table 1: Comparaison de performance de l'illustration automatique de texte entre notre méthode proposé et l'approche basique basé seulement sur l'espace commun de représentation appris par l'ACC. La base d'apprentissage et la base de références sont extraites à partir d'une même collection ou à partir des collections différentes. Plus la tâche est difficile (en termes de nombre de données potentiellement mal représentées), plus notre méthode est utile.

Deuxième contribution. Une deuxième contribution de la thèse aborde particulièrement la classification intermodale. Cette tâche consiste à apprendre des modèles de classification à partir des données venant d'une modalité et les tester sur des données appartenant d'une autre modalité (voir Figure 3). De cette manière, la tâche de classification intermodale exige les données d'apprentissage uni-modales labellisées dans la première modalité et les données de test uni-modales dans l'autre modalité. Dans le cadre

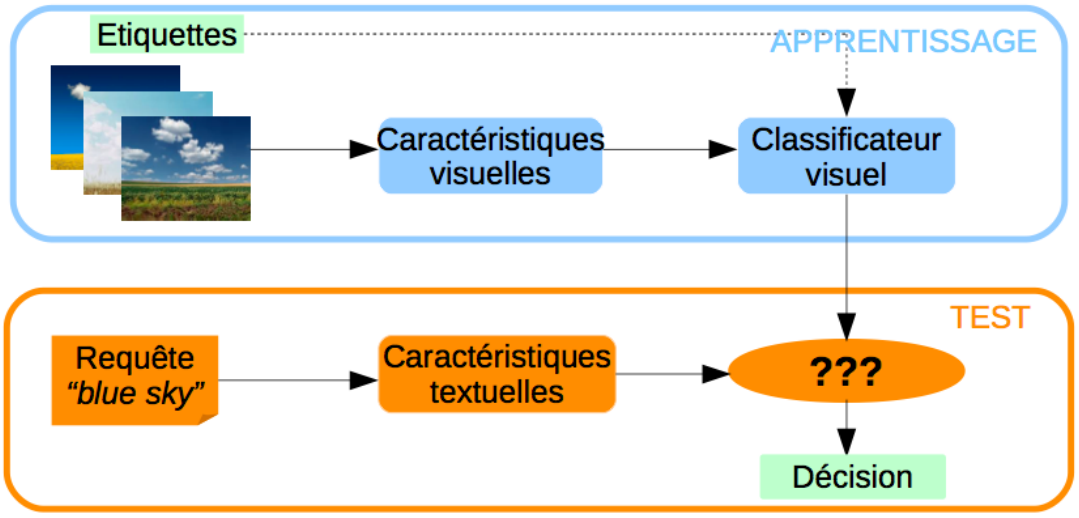

Figure 3: Un exemple de la tâche de classification intermodale. Des classificateurs visuels sont appris à partir des images avec ses étiquettes et sont appliqués pour prédire une requête textuelle.

de cette thèse, nous étudions le problème de classification intermodale pour les contenus visuels et textuels. Cette tâche n'a pas été largement étudiée dans la littérature de la communauté multimédia. Cela est dû au fait que les textes et les images ne sont pas décrits avec les mêmes caractéristiques, et ils ne résident même pas dans le même espace vectoriel, rendant donc la classification intermodale assez incongrue. Cependant, au-delà d'un intérêt académique, nous croyons que cette tâche a également un intérêt pratique. Supposons, par exemple, que les classificateurs pour de nombreux concepts ont pu être appris à partir des données textuelles en raison de la disponibilité massive de telles données labellisées. Nous pourrons souhaiter détecter ces concepts sur un contenu correspondant à une autre modalité, par exemple images, même si les étiquettes de classe (labels) ne sont pas encore disponibles pour ce contenu. Une telle situation peut devenir plus courante avec l'évolution actuelle du micro-blogging, qui passe du contenu purement textuel (Twitter historique) au contenu multimodal (Twitter actuel) ou au contenu purement visuel (Instagram). L'étude de cette tâche de classification intermodale nous permet d'examiner un modèle conceptuel de plus haut niveau sémantique, ce qui est une des motivations de cette thèse.

Afin d'adresser le problème de classification intermodale, un espace commun de représen-

tation de la modalité visuelle et textuelle est une solution appropriée pour surmonter le problème d'incompatibilité entre ces deux modalités. Dans cet espace commun, les informations visuelles et textuelles ont des représentations similaires et deviennent directement comparables. Par conséquent, il est parfaitement concevable d'apprendre un classificateur, par exemple textuel, utilisant les descripteurs qui sont des projections de caractéristiques textuelles et de prédire une image qui est représentée par sa projection sur cet espace commun. Un tel espace commun a été largement examiné dans la littérature récente de la recherche intermodale d'information textuelle et visuelle. Ce sujet est aussi étudié avec beaucoup d'attention dans les contributions présentées dans cette thèse. Cependant, au mieux de notre connaissance, aucune tentative n'a été faite pour employer des classificateurs intermodaux comme ce que nous avons suggéré.

Dans cette contribution, nous cherchons à faire la classification intermodale à partir des projections des données textuelles et visuelles sur l'espace commun. Pour cela, une approche basique est de projeter des données visuelles et textuelles sur cet espace, puis d'apprendre et tester les classificateurs intermodales en utilisant ces projections. La performance de la classification intermodale obtenue avec une telle utilisation directe des projections n'est pas trop faible. Néanmoins, comme identifié précédemment, un tel espace commun de représentation (par exemple appris par l'ACC) présente encore certains limites. Par conséquent, la qualité des projections intermodales n'est pas suffisante pour obtenir une "traduction" robuste entre les deux modalités. Nous croyons donc que la performance de la classification intermodale peut être améliorée une fois que les limites de l'espace commun sont gérées.

Notre but est de proposer des représentations plus robustes et complètes à partir des projections directes sur l'espace commun pour adresser plus efficacement à la tâche de classification intermodale. Dans ce but, nous proposons une méthode de représentation permettant à gérer l'insuffisance de l'espace commun concernant la séparation entre les deux modalités mentionnée précédemment dans la Figure 1. Notre approche consiste à apprendre une représentation bi-modale dans l'espace commun pour toute donnée initialement uni-modale, en faisant la complétion d'une projection uni-modale par un point virtuel dans l'autre modalité. L'idée générale de la complétion est de compter sur un ensemble additif

de données bi-modales, appelé ensemble de données auxiliaires, pour transformer un point de projection d'un document uni-modal en une représentation plus robuste qui prend en compte les deux modalités. Un tel ensemble de données auxiliaires contient des images avec ses textes associés et ses données n'ont pas besoin d'être labellisées. Cet ensemble auxiliaire sert d'ensemble de connexions entre images et textes permettant de relier les modalités visuelle et textuelle dans l'espace commun.

Le problème central de notre approche est de déterminer une représentation complémentaire pertinente de la modalité manquante d'un document uni-modal étudié. Ce dernier est représenté par un point virtuel obtenu à travers l'ensemble des projections des données auxiliaires sur l'espace commun. La représentation bi-modale finale, appelée "Weighted Completion with Averaging" (WCA), est obtenue en faisant la moyenne de ces deux points (la projection uni-modale réelle et le point virtuel associé). L'identification d'un tel point virtuel pertinent est le cœur de notre méthode. Pour cela, nous mentionnons une approche "naïve" et ensuite proposons un schéma légèrement plus sophistiqué pour identifier des informations complémentaires, permettant d'obtenir des résultats nettement meilleurs (voir Figure 4) Dans une approche directe mais "naïve", le virtuel point est obtenu à partir des voisins les plus proches de la projection uni-modale à compléter parmi les projections de données auxiliaires dans la modalité manquante. Par exemple, pour compléter une projection du document initialement textuel, nous cherchons les plus proches voisins de ce point dans l'ensemble de projections visuelles des données auxiliaire. Pour aller plus loin à une telle approche, nous devons considérer les propriétés de l'espace commun. Bien que ce dernier résulte d'une maximisation globale de la relation entre les deux modalités, les projections du contenu textuel et visuel d'un même document sur cet espace ne sont pas nécessairement très proches. Ce fait est également montrée par une des insuffisances concernant la séparation entre les deux modalités sur l'espace commun. Donc, pour une représentation uni-modale donnée, ses voisins directs les plus proches de l'autre modalité ne sont pas le meilleure choix pour compléter la modalité manquante de cette représentation uni-modale. Cependant, nous pensons que les documents ayant un contenu similaire dans une modalité soient susceptibles d'avoir un contenu assez similaire dans l'autre modalité. Nous proposons donc de rechercher tout d'abord les voisins les plus proches de la projection

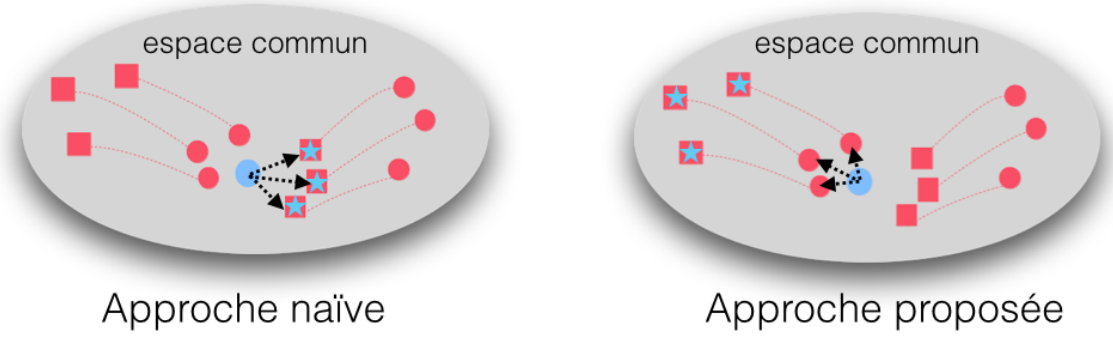

Figure 4: Deux approches pour déterminer des informations complémentaires dans la modalité manquante pour la complétion.
Les carrés et les cercles sont respectivement des projections textuelles et visuelles. Les carrés rouges et les cercles rouges sont les projections textuelles et visuelles des documents de l'ensemble auxiliaire sur l'espace commun. Nous cherchons à compléter une projection visuel représentée par le cercle bleu par des informations dans la modalité textuelle. Des informations identifiées par chaque approche sont marquées par des étoiles dans la figure. Les flèches noires indiques les voisins les plus proches d'un point examiné.

uni-modale parmi les projections des données auxiliaires dans la modalité disponible et ensuite d'utiliser les projections correspondant à ces voisins dans l'autre modalité pour la complétion. Dans ces deux approches, le virtuel point est défini par le barycentre de tous les points complémentaires précédemment retrouvés. Enfin, notre méthode WCA comprend une étape qui agrège en faisant la moyenne du vecteur original issu de la projection d'un document uni-modal examiné avec l'information complémentaire représenté par le point virtuel identifié pour synthétiser un vecteur de représentation unique du document. Cette nouvelle représentation WCA englobe à la fois les informations visuelles et textuelles. Nous proposons ensuite à faire la classification intermodale en utilisant une telle représentation WCA. Pour cela, nous utilisons un ensemble de données bi-modal auxiliaires pour compléter systématiquement des données uni-modales, tant dans l'ensemble d'apprentissage que dans l'ensemble de test, ce qui entraîne des représentations WCA bi-modales plus complètes. Les classificateurs sont appris et testés en utilisant ces représentations.

Pour l'évaluation, nous menons plusieurs expériences sur des ensembles de données publiquement disponibles selon des protocoles expérimentaux standard afin d'adresser

la tâche de classification intermodale. La performance de la tâche est mesurée par une moyenne de deux mAPs (mean Average Precision) correspondant respectivement à la classification Image-Texte (modèles appris sur des images et testés sur des textes) et à la classification Texte-Image (modèles appris sur des textes et testés sur des images). Les résultats (voire Figure 5) montrent que notre méthode WCA améliore l'utilisation directe des projections brutes sur l'espace commun appris par l'ACC. Aussi, l'approche de complétion que nous proposons est plus pertinent que l'approche naïve.

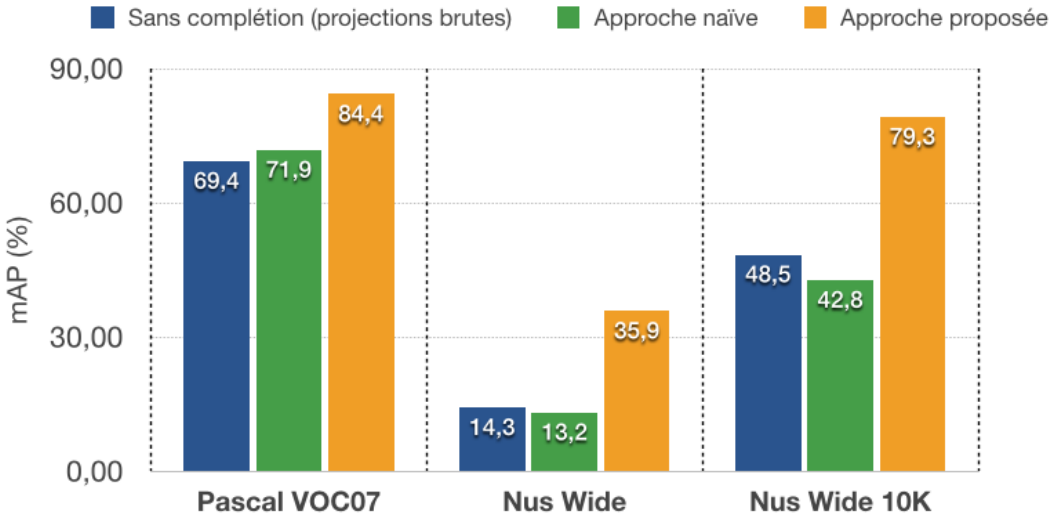

Figure 5: Résultats de la classification intermodale sur trois corpus de données Pascal VOC07, Nus Wide et Nus Wide 10K. Nous comparons la performance de l'approche basique (basée seulement sur l'espace commun en utilisant ses projections brutes) avec deux approches de complétions mentionnées. Pour chaque problème, l'espace commun est appris sur les données d'apprentissage du corpus correspondant et les classificateurs sont appris et testé respectivement sur des données d'apprentissage et de test du problème.

Par ailleurs, nous étudions également l'influence des principaux composants de notre méthode WCA sur la performance de la classification intermodale, à savoir le processus de complétion et la méthode d'agrégation sur la performance de WCA. Particulièrement, nous rappelons qu'une des motivations de cette thèse est de développer un espace commun de représentation comme une "ressource" générique, à partir d'un volume suffisamment grand de données bi-modales, puis d'aborder différents problèmes de classification intermodale utilisant cette ressource. Cela permet d'éviter de réapprendre un espace commun pour

chaque problème spécifique en utilisant ses propres données. Les projections sur une telle ressource bénéficient des relations texte-image génériques obtenues à partir de données utilisées pour apprendre l'espace commun. En outre, un ensemble de données bi-modales différent, potentiellement plus relié au problème cible peut alors être utilisé pour la complétion de la représentation uni-modale, en prenant ainsi en compte les liens spécifiques entre le contenu visuelle et textuelle du problème cible dans la représentation agrégée. Pour

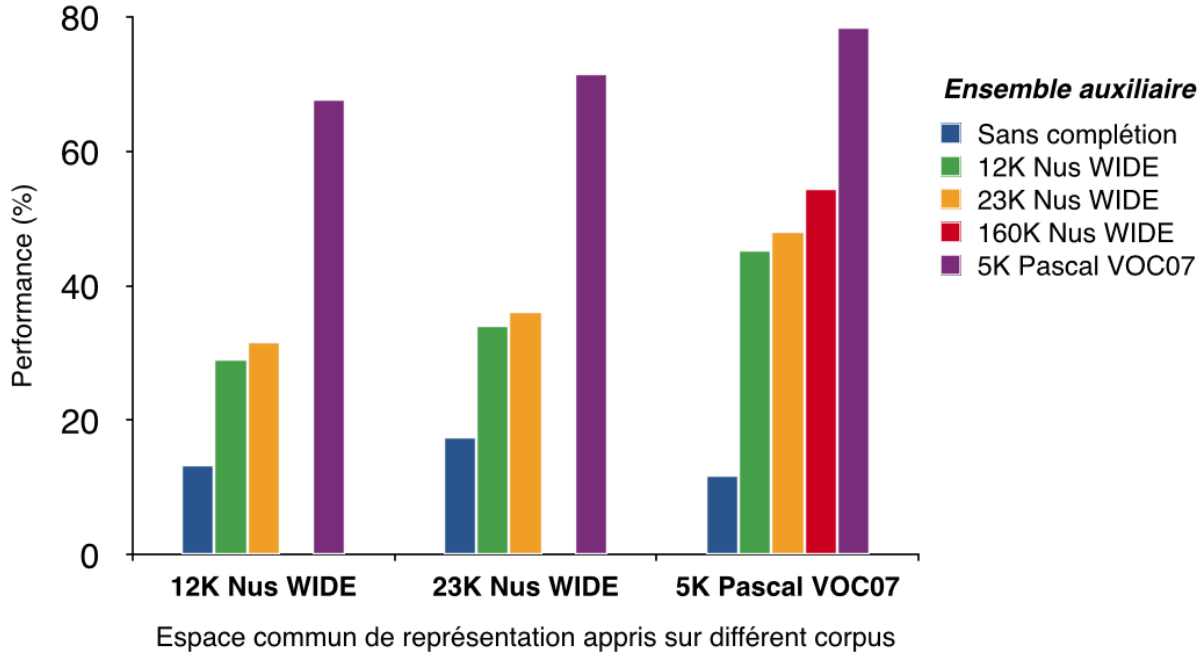

Figure 6: Résultats de la classification intermodale des données Pascal VOC07. Différent espaces communs de représentation sont appris sur des données de différent corpus comme 12,000 données de Nus WIDE (12K Nus WIDE), 23,000 données de Nus Wide (23K Nus WIDE) ou 5,000 données de Pascal VOC07. Différent ensemble auxiliaire (Nus WIDE ou Pascal VOC07) sont utilisés pour la complétion.

explorer cette idée, nous examinons donc la relation entre les deux ensembles de données utilisés pour la fabrication de la représentation WCA: l'ensemble de données auxiliaires et l'ensemble de données utilisées pour apprendre l'espace commun de représentation. A cette fin, nous étudions l'impact de l'utilisation de différentes collections de données pour obtenir l'espace commun et sur chaque espace, différents ensembles de données auxiliaires pour compléter les représentations uni-modales. Les résultats montrent que la performance de

WCA dépend à la fois de la taille de l'ensemble de données auxiliaire et de l'accord entre des données de cet ensemble auxiliaire et le problème spécifique considéré. Quelques résultats de la classification intermodale des données du corpus cible Pascal VOC07 sont décrits dans la Figure 6. La performance est améliorée si nous augmentons la taille de l'ensemble auxiliaire (résultats avec différent sous-ensembles de données Nus WIDE). Pourtant, l'utilisation des données du problème cible (seulement 5,000 données de Pascal VOC07) comme des données auxiliaires donne des meilleures performance.

Classification	Pascal VOC07	Nus-WIDE	Nus-WIDE 10K
Uni-modale (Image)	86.10	50.38	78.53
Uni-modale (Texte)	82.50	46.57	70.20
Bi-modale (Image+Texte)	86.16	50.87	82.89
Intermodale (Texte-Image)	85.49	37.80	79.53
Intermodale (Image-Texte)	83.38	34.02	79.15

Table 2: Comparaison en termes de mAP (%) entre différent scénarios de classification telle que la classification uni-modale, bi-modale et la classification intermodale proposée.

Enfin, nous également comparons la classification intermodale avec les différent tâches de l'état de l'art telles que la classification uni-modale et bi-modale, qui sont des tâches plus classiques et moins difficiles que la tâche intermodale proposé dans cette thèse. Les résultats obtenus pour la classification intermodale sont relativement proches de ceux obtenus pour la classification uni-modale ou bi-modale (voire Tableau 2). Un tel niveau de performance rend notre approche de la classification intermodale un choix convaincant pour les applications réelles, telles que l'apprentissage des classificateurs à partir d'une grande quantité de données textuelles annotées disponibles et leur application au contenu visuel, pour annoter des images par exemple.

Troisième contribution. Cette contribution présente une autre méthode de représentation conjointe des modalité texte et image, appelée «agrégation de composantes multimédia corrélées» (MACC), pour les documents bi-modaux ou uni-modaux. Comme WCA, cette méthode cherche à réduire la séparation de la modalité visuelle et textuelle sur leur espace commun de représentation. De la même façon que WCA, la représentation MACC regroupe à la fois des informations visuelles et textuelles projetées sur l'espace commun pour fabriquer une représentation bi-modale unique de donnée. La différence est sur la

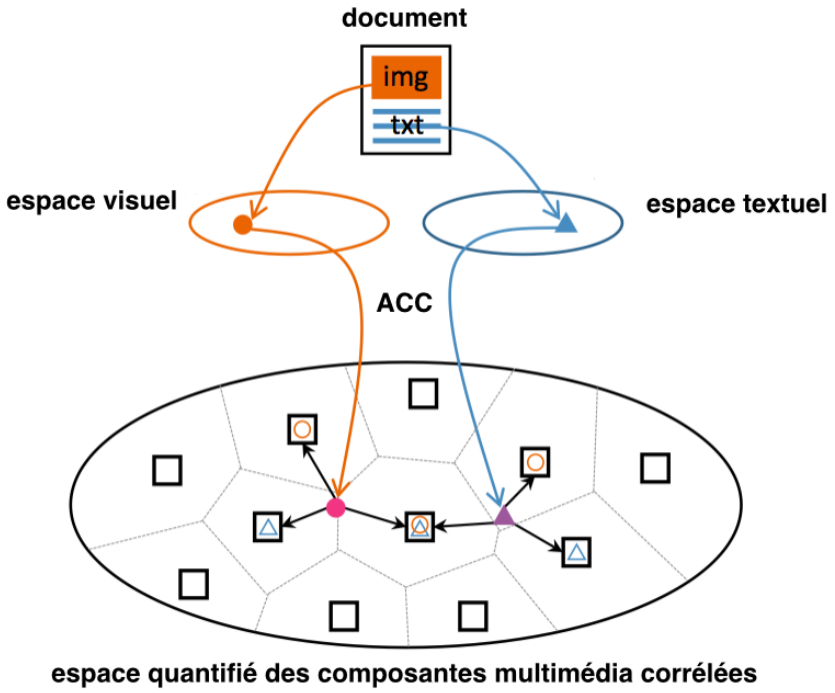

Figure 7: L'illustration de la méthode MACC.
Les contenus visuels et textuels d'un document sont projetés sur un espace commun précédemment quantifié. Les deux projections, correspondant au même document, sont encodées selon un vocabulaire commun avant leur agrégation.

quantification de l'espace commun en plusieurs contextes multimédia et la représentation des données selon ces contextes (Figure 7). Concrètement, pour un ensemble de documents multimédia, nous construisons d'abord l'espace commun de représentation (par exemple par l'ACC) et ensuite apprenons un vocabulaire (obtenu par exemple par $k - means$ clustering) à partir de toutes les projections visuelles et textuelles de ces documents sur leur espace commun. Pour chaque document bi-modal, ses caractéristiques visuelles et textuelles sont projetées sur cet espace commun et puis chaque projection (visuelle et textuelle) est codée par un vecteur des différences entre cette projection et les centres des clusters. Enfin, ces deux vecteurs de différences sont agrégées en un seul vecteur MACC qui est la représentation multimédia du document. Particulièrement, dans le cas d'un document uni-modal où une seule modalité est disponible, nous suggérons également de compléter la modalité absente en utilisant les données d'un ensemble de données auxiliaires

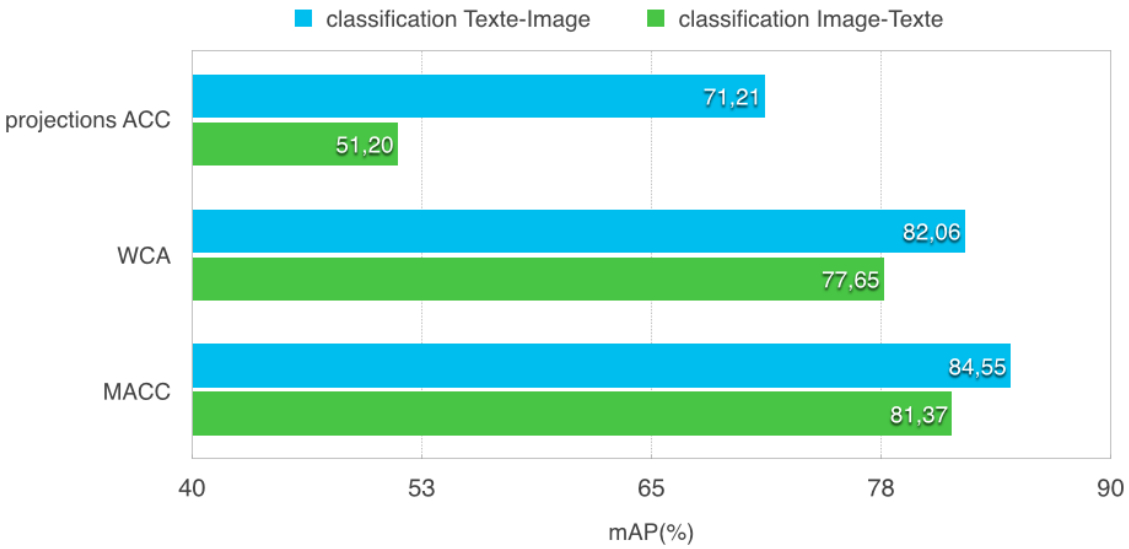

Figure 8: Les performances de la classification intermodale sur les données de Pascal VOC07.

suite à la procédure de complétion décrite dans la méthode précédente WCA. Par la suite, nous combinons les descripteurs de deux modalités pour construire la représentation MACC du document initialement uni-modal.

La méthode est évaluée sur les collections VOC 20017, FlickR8K et FlickR30K, en classification bi-modale, en classification intermodale et en recherche d'image. Dans les trois cas, la performance observée est au niveau de celle de l'état de l'art ou plus élevée. Une expérience importante porte sur la comparaison entre différent types de représentation basé sur l'espace commun de représentation telles que les projections (ACC) brutes, la WCA et la MACC (Figure 8). Pour cela, nous évaluons la performance de la tâche de la classification intermodale sur la collection Pascal VOC07. Le résultat montre que la MACC et la WCA améliorent l'utilisation directe des projections sur l'espace commun. De plus, les meilleures performances sont obtenues avec la représentation MACC. Pour la tâche de recherche d'image sur les données de FlickR 8K, la représentation MACC améliore légèrement la performance des résultats actuels de l'état de l'art (Figure 9). D'ailleurs, de nombreuse expériences sont également menées afin d'étudier l'impact des paramètres de la MACC telles que la méthode de codage, la taille du vocabulaire et l'ensemble auxiliaire

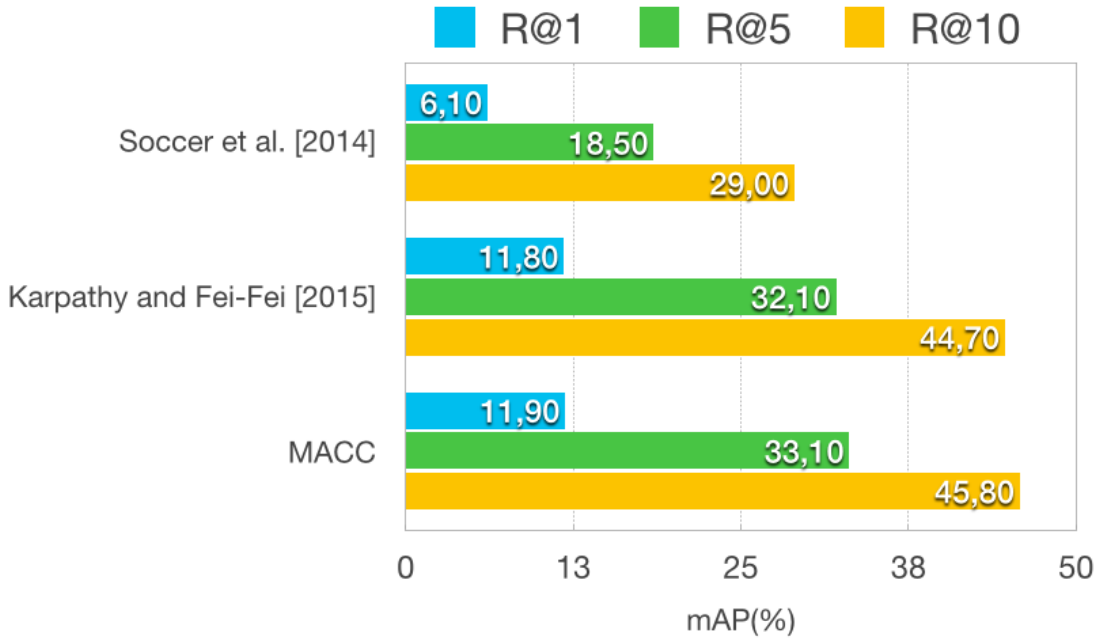

Figure 9: Comparaison avec l'état de l'art de la recherche d'image sur les données de FlickR 8K.

utilisé pour la complétion sur la performance des tâches intermodales.

Enfin, pour terminer la présente thèse, nous discutons éventuellement des points qui peuvent être inspirés par les problèmes de recherche présentés. En particulier, nous considérons l'extension de nos contributions au cas des espaces communs fondé sur d'autres principes que la maximisation de la corrélation.

Acknowledgements

Firstly, I would like to express my sincere gratitude to my advisors Prof. Michel Crucianu and Dr. Hervé Le Borgne for giving me an opportunity to work with them as their PhD student. I am very grateful to them for their continuous support, patience, motivation and knowledge in all the time of my research and writing of this thesis. It would not have been possible to accomplish this research without their precious support and encouragement.

Besides, I would like to thank DR Guillaume Gravier, DR Georges Quénot, Dr. Gregory Grefenstette, Dr. Céline Hudelot and Dr. Sid-Admed Berrani for agreeing to review my thesis work and to serve on my thesis committee. I am grateful for their insightful comments and questions.

My sincere thanks to all my labmates, in particular to Alexandru Ginsca who was always available to give me good advice. My gratitude extends to my wonderful Vietnamese friends that I was fortunate to meet at Nano Innov. They all made my PhD life a marvelous experience.

Last but not the least, I would like to thank my family. To my beloved parents, my mother in law and my sister, who have been so caring and supportive to me all the time and especially during this last year of my thesis. I owe thanks to my husband Binh-An and my little daughter Camille for their love, support and understanding during my pursuit of PhD degree that made the completion of this thesis possible.

Contents

1 Introduction **39**

1.1 Motivations 42

1.2 Goals 44

1.3 Contributions 44

1.4 Organization of the thesis 48

2 State-of-the-art **51**

2.1 Introduction 52

2.2 Single-media representation 53

2.2.1 Visual Features 53

2.2.2 Textual Features 60

2.3 Multimedia information retrieval and classification 63

2.3.1 Uni-modal problems 63

2.3.2 Multi-modal problems 65

2.3.3 Cross-modal problems 69

2.4 Common representation for text and image 72

2.4.1 Correlation learning approaches 73

2.4.2 Topic modeling approaches 79

2.4.3 Rank-based approaches 82

2.5 Multi-modal datasets 86

2.6 Conclusion 91

3 Common representation space and its limitations 93

3.1 Introduction 94

3.2 KCCA: common representation space for image and text 95

3.2.1 Common representation space learning with (K)CCA 96

3.2.2 Projections onto (K)CCA subspace 97

3.2.3 Cross-modal matching on (K)CCA subspaces 98

3.3 Limitations of common representation subspaces 99

3.3.1 Poorly-represented data on the common space 99

3.3.2 Separation between modalities on the common subspace 104

3.4 Conclusion 108

4 Combining generic and specific information for cross-modal retrieval 109

4.1 Introduction 110

4.2 Text illustration problem 111

4.3 Combining generic and specific information 114

4.3.1 Specific information identification 114

4.3.2 Generic and specific information combination for cross-modal retrieval 117

4.3.2.1 Text illustration problem formalization 117

4.3.2.2 Generic and specific information combination 118

4.4 Experimental evaluation 122

4.4.1 Experimental Setting 123

4.4.2 Results on BBC News 124

4.4.3 Results in a domain transfer context 126

4.4.4 Results on Wikipedia 2010 127

4.5 Conclusion 129

5 Uni-modal data completion with the missing modality 131

5.1 Introduction 132

5.2 Weighted Completion with Averaging (WCA) 136

5.2.1 Relevant completion information identification 137

5.2.2 Completion of the missing modality 139

5.2.3 Aggregated representation construction 141

5.3 Experimental evaluation 142

5.3.1 Experimental settings 142

5.3.2 Proposed completion *vs.* naive completion *vs.* no completion 144

5.3.3 Influence of the completion and aggregation methods 145

5.3.4 Impact of the auxiliary data and the common space 147

5.3.4.1 Auxiliary dataset $\mathcal{A}$ 147

5.3.4.2 Common space training dataset $\mathcal{T}$ 148

5.3.5 Comparison to the state-of-the-art 150

5.4 Conclusion 153

6 Aggregating image & text quantized correlated components 155

6.1 Introduction 156

6.2 Proposed approach 157

6.2.1 Multimedia Aggregated Correlated Components 157

6.2.1.1 Codebook learning 158

6.2.1.2 MACC representation 159

6.2.2 MACC completion with the missing modality 160

6.3 Experimental evaluation 162

6.3.1 Image classification on Pascal VOC07 163

6.3.1.1 Classification of bi-modal documents. 163
6.3.1.2 Classification in a cross-modal context. 165
6.3.2 Image retrieval on FlickR 8K and FlickR 30K 167
6.3.2.1 FlickR 8K image retrieval. 168
6.3.2.2 FlickR 30K: benefit of auxiliary dataset. 171
6.4 Conclusion 173

7 Conclusion and Perspectives **175**
7.1 Conclusion 176
7.2 Perspectives for future research 180
7.2.1 Short-term perspectives 180
7.2.2 Longer-term perspectives 181

Bibliography **183**

Appendices **199**

A Tags cleaning **199**

B List of publications **201**

List of Tables

1 Comparaison de performance de l'illustration automatique de texte entre notre méthode proposé et l'approche basique basé seulement sur l'espace commun de représentation appris par l'ACC. La base d'apprentissage et la base de références sont extraites à partir d'une même collection ou à partir des collections différentes. Plus la tâche est difficile (en termes de nombre de données potentiellement mal représentées), plus notre méthode est utile. 12

2 Comparaison en termes de mAP (%) entre différent scénarios de classification telle que la classification uni-modale, bi-modale et la classification intermodale proposée. 19

2.1 Summarization of multi-modal datasets considered in the thesis 86

3.1 Average distances between projections on KCCA space. 106

3.2 Distribution of textual and visual KCCA-projected points into clusters. . . 106

4.1 Text illustration results on BBC News dataset 126

4.2 Results on BBC News with domain transfer 127

4.3 Results on Wikipedia 2010 127

4.4 Comparative performance of our method to the baseline CCA for different experimental settings. The more difficult the task (in term of number of specific words) the more our method is useful. 129

4.5 Results with top-10 evaluation 130

5.1 Cross-modal classification results (mAP%) on Pascal VOC07. 144

5.2 Cross-modal classification results (mAP%) on Nus-WIDE and Nus-WIDE 10K. 145

5.3 Results on Pascal VOC07 with the common space obtained from the Pascal VOC07 training set. Different auxiliary datasets $\mathcal{A}$ are used for WCA (with $d = 4000, \mu = 5$). 147

5.4 Results on Pascal VOC07 with different datasets $\mathcal{T}$ to learn the common space and different auxiliary datasets $\mathcal{A}$ (with $d = 100$, $\mu = 5$) to connect the modalities in the common space. 149

5.5 Comparison in terms of mAP(%) with uni-modal and bi-modal classification results. 151

5.6 mAP@50 for cross-modal *retrieval* and for cross-modal *classification* on NUS-WIDE 10K. We implemented our method (WCA) and report the results of the original paper for Ngiam et al. [2011]; Feng et al. [2014]. Experimental protocols are coherent with these last (see text for details). 152

6.1 Pascal VOC07: comparison with published results. 164

6.2 Pascal VOC07: comparison with baselines. 164

6.3 Pascal VOC07: mAP (%) for different values of k and n. 165

6.4 Pascal VOC07: classification in a cross-modal context using the completion mechanism for MACC representations. MACC and WCA representations rely on the 150-dimensional KCCA common representation space. 167

6.5 Image retrieval results on FlickR 8K. 168

6.6 Image retrieval results on FlickR 30K. 172

7.1 Average Euclidean distances between image and text representations on Pascal VOC07 data. All representations are calculated on the 150-dimensional KCCA spaces. $k = 16, n = 5, \mu = 15$ for MACC and $\mu = 15$ for WCA. . . . 178

7.2 Average Euclidean distances between image and text representations on 1000 FlickR 8K testing data. All representations are calculated on the 50-dimensional KCCA spaces. $k = n = 32, \mu = 64$ for MACC. 179

List of Figures

1 La séparation entre la modalité visuelle et textuelle sur l'espace commun de représentation. 10

2 Une première contribution de la thèse concentre sur la différence entre la base d'apprentissage et la base de référence dans le contexte de la tâche de recherche d'information. Cette portion contient potentiellement des données mal représentées sur l'espace commun de représentation appris sur les données d'apprentissage. 11

3 Un exemple de la tâche de classification intermodale. Des classificateurs visuels sont appris à partir des images avec ses étiquettes et sont appliqués pour prédire une requête textuelle. 13

4 Deux approches pour déterminer des informations complémentaires dans la modalité manquante pour la complétion. 16

5 Résultats de la classification intermodale sur trois corpus de données Pascal VOC07, Nus Wide et Nus Wide 10K. Nous comparons la performance de l'approche basique (basée seulement sur l'espace commun en utilisant ses projections brutes) avec deux approches de complétions mentionnées. Pour chaque problème, l'espace commun est appris sur les données d'apprentissage du corpus correspondant et les classificateurs sont appris et testé respectivement sur des données d'apprentissage et de test du problème. 17

6 Résultats de la classification intermodale des données Pascal VOC07. 18

7 L'illustration de la méthode MACC. 20

8 Les performances de la classification intermodale sur les données de Pascal VOC07. 21

9 Comparaison avec l'état de l'art de la recherche d'image sur les données de FlickR 8K. 22

2.1 The flowchart of the Bag-of-Visual-Words (BoVW) generation scheme including three steps of codebook learning, local features coding and pooling. Figure extracted from [Znaidia 2014]. 54

2.2 The CNN-based features brought a significative breakthrough in image classification task on ImageNet dataset (from [Russakovsky et al. 2015]) . . 57

2.3 CNN-based feature extracted from pre-trained CNN model 59

2.4 Illustration of a content-based image retrieval (CBIR) system 63

2.5 Semantic gap between low-level visual features offered by CBIR systems and semantic concepts identified by users 65

2.6 Multi-modal approach based on early fusion. 67

2.7 Multi-modal approach based on late fusion. 68

2.8 The framework of the approach proposed by Wang et al. [2009]. 69

2.9 Cross-modal retrieval problem. 70

2.10 Different categories of common representation learning approaches for text and image. 72

2.11 Semantic Correlation Matching model [Costa Pereira et al. 2014] 76

2.12 Difference between two-stage methods and Corr-AE: Corr-AE incorporates representation learning and correlation learning into a single process while two-stage methods separate the two processes [Feng et al. 2014] 77

2.13 Difference between Corr-AE [Feng et al. 2014] and DCCA [Andrew et al. 2013]. Corr-AE correlates hidden representations of a pair of auto-encoder while DCCA correlates output layers of two deep networks describing visual and textual modalities. 78

2.14 DeViSe model (center) initialized with parameters pre-trained at the lower layers of a visual object categorization network with a softmax output layer (left) and a skip-gram language model (right) [Frome et al. 2013] 83

2.15 Illustration of the Deep Fragment Embeddings for Bidirectional Image Sentence Mapping model [Karpathy et al. 2014] 84

2.16 A sample from our BBC News database. Each entry contains an image, a caption for the image, and the accompanying document with its title. . . . 87

2.17 A sample from our Wikipedia database. Image in Wikipedia article is given with corresponding metadata. 88

2.18 Example Pascal VOC 07 images with their associated tags collected using AMT and their corresponding ground-truth labels. 88

2.19 Example Nus-WIDE images with their associated tags and their corresponding ground-truth labels. Several tags are very noisy. For instance, "colorartaward" and "platinumheartaward" are respectively for "color art award" and "platinum heart award". 89

2.20 Example FlickR 8K images with their five associated descriptions. 90

3.1 Canonical correlation analysis between image and text modalities 97

3.2 Quality of representation on the common subspace. 100

3.3 Quality of representation of words collected from BBC News. 103

3.4 Separation between modalities on the KCCA space. 104

3.5 Separation between modalities on the KCCA space. 107

4.1 Automatic text illustration problem. 111

4.2 Ill-represented data on the common representation space involved in the difference between training and reference databases 116

4.3 CCA_{cap} results (on the BBC News test set) and dimension selection (cross-validation on the training set) 125

4.4 Top three images with their captions which are proposed by our model to illustrate the BBC News documents about "Extent of school failure disputed" and "Thames Water heads pollution list" 128

4.5 Illustration by author may not be the unique and the best choice to describe the actual content of BBC News article 130

5.1 Illustration of *Image-Text* cross-modal classification problem. 132

5.2 The proposed WCA approach for cross-modal classification 136

5.3 Naive completion (a) *vs.* proposed completion (b). 139

5.4 Results of different completion and aggregation methods on Pascal VOC07, showing the mAP(%) with respect to the number of neighbor points μ used in the auxiliary dataset $\mathcal{A}$. For each method, the curves are the average of the Text-Image and Image-Text tasks. Parameter d is set to 4000. 146

6.1 Visual and textual contents of a document are projected onto a common space that has been previously quantized. Both projections, corresponding to the same document, are encoded according to a common vocabulary before their aggregation. 158

6.2 Coding methods comparison for MACC representation. 170

6.3 FlickR 8K image retrieval: stability of MACC representations over a large range of parameters. Evaluation follows the protocol of Chen and Zitnick [2015]. 171

A.1 Python program to clean the tags 200

Chapter 1

Introduction

During the past decade, the explosion of multimedia data has become more important than ever with the rise of social media sites like Facebook, Twitter and content-sharing sites like YouTube, FlickR, Wikipedia, etc. Issued from this, a lot of huge yet increasing public collections of multimedia content have been generated. This has led to a surge of research activity in how to store, organize, access and search these immense multimedia resources.

In this context, multimedia *retrieval* and *classification* have been receiving most attention of the multimedia community due to their practical interests. These latter are the core of various real-world multimedia applications such as image annotation, automatic text illustration, hot topic detection, event detection, document categorization, etc. The efforts of the multimedia community have become increasingly widespread, due in part to the introduction of large-scale research and evaluation benchmarks such as TRECVID [Smeaton et al. 2006] and ImageCLEF [Müller et al. 2010] involving datasets that span multiple modalities.

This thesis focuses on two popular modalities of data, namely *visual modality* represented by *e.g.* images and *textual modality* represented by *e.g.* tags, keywords, natural language phrases, etc. Earlier research such as image retrieval or document retrieval focused on exploring the distinct characteristics of each individual modality *e.g.* image or text. Nevertheless, image and text usually appear together in multimedia collections and provide complementary information to each other. For example, the visual content of a FlickR photo is characterized by user-provided descriptions and tags or the content of a Wikipedia article is usually illustrated by one or several images. Therefore, *jointly modeling* visual and the associated textual content can potentially improve the performance of many multimedia *retrieval* and *classification* systems.

The research on multimedia retrieval and classification has been very active in the past decades. However, the latter is still a very challenging problem yet to be solved due to the "semantic gap" [Smeulders et al. 2000] between features and semantics. Images are made of pixels and represented by low-level features *e.g.* color, shape, texture, etc. Textual content consists of keywords, concepts or natural language phrases representing higher-level semantics of images. Bridging the "semantic gap" between the image low-

level features and high-level semantics remains a major challenge in modeling multimedia systems. Furthermore, these systems also suffer from the "heterogeneity gap" across different modalities. This refers to the fact that visual and textual modalities are usually described according to totally different schemes, and usually reside in different feature spaces.

The main work of this thesis considers the *joint modeling* of image and text to address *cross-modal* problems, a more recent paradigm of information retrieval. Cross-modal systems support interactivity *across* modalities. For instance, a cross-modal retrieval task finds images in response to a query text, or text documents in response to a query image. These tasks are central to many applications of practical interest, such as finding on the web the picture that best illustrates a given text *e.g.* in automatic text illustration, finding the texts that best describe a given picture *e.g.* in image captioning task, etc. Furthermore, the availability of rich information pertaining to various modalities makes users expect to have a free choice of which content modality they submit as query and which one they want to receive. In this context, a cross-modal system is considered as a natural way of searching multimedia content and becomes increasingly important. Meanwhile, effective and efficient cross-modal modeling remains a challenge, due to the *heterogeneity gap* across different modalities and the *semantic gap* between low-level features and semantics.

Recently, jointly modeling image and text for cross-modal systems has continuously attracted much attention in the multimedia community. Since images and text reside in different feature spaces, one core issue of cross-modal models is how to reduce the differences between these heterogeneous features. This can be accomplished through the development of a *common representation space* resulting from a maximization of the relatedness between the different modalities. Such a joint space typically relies on Canonical Correlation Analysis (CCA) or its kernel extension [Hardoon et al. 2004; Costa Pereira et al. 2014; Hwang and Grauman 2012a; Gong et al. 2014] that seek a common subspace of both feature spaces that maximises the correlation of the projected points from the original datasets. Several alternatives based on deep learning also exist [Ngiam et al. 2011; Srivastava and Salakhutdinov 2012; Frome et al. 2013; Feng et al. 2014; Karpathy and Fei-Fei 2015; Wang et al. 2015], that usually retain a different criterion to create the

common space. In this space, both modalities are described similarly, with regards to new latent variables whose definitions depend on the approach employed. Such a latent space is mainly considered for cross-modal tasks as both images and text can be represented in this space without any distinction. Meanwhile, it can be used to address various bi-modal tasks that focus on semantics.

Since both modalities are homogeneous when projected onto the joint space, various cross-modal tasks can be addressed, such as image or text retrieval [Hardoon et al. 2004; Hwang and Grauman 2012a; Gong et al. 2014], social event detection [Ahsan and Essa 2014] or image classification [Costa Pereira et al. 2014]. The performance of such tasks thus highly depends on the quality of the common representation space. The choice of the method to build a common space is driven by two main considerations:

- The first is the core problem in designing cross-modal models and deals with the quality of "matching" across modalities. The models aim to bring text and images close together for a high quality of "matching" of information across modalities. Also, common representation spaces need to favor inter-related information that usually highlights semantics and discounts modality-specific information. Until now, most of the existing work based on common representation spaces focuses on investigating complete representations of data in each individual modality for building a robust common space. Despite their relative successes, the resulting common representation seems to be insufficient to narrow the gap between visual and textual modalities due to several limitations of this joint space.

- The second consideration relates to the amount of resources used to learn a robust common representation. It has of course a great importance in practice since the use of a very large amount of data can lead to intractable computation, or at least a cost that is too high to be interesting.

1.1 Motivations

The motivation of this thesis is three-fold.

First, we aim to obtain a *robust representation* of multimedia documents on the

common representation space. A robust representation of data is a key to success in several multimedia tasks such as classification, cross-modal retrieval, etc. As mentioned above, multimedia documents are often described by image and text content which might belong to the same semantics. Information extracted from these different modalities are often semantically related and complementary. This multi-modal aspect is thus an opportunity to learn a better representation of data by simultaneously analyzing these modalities. The representation of data on the common latent space accounts for the relatedness between images and text and further reduces semantic and heterogeneity gaps betweens these two modalities. However, such a common representation space suffers from several limitations. A direct use of these projections results in a limited quality of matching between modalities and consequently hampers performance in cross-modal or bi-modal tasks. Our motivation is to propose a robust common representation method that relates visual and textual modalities more closely on their joint space. Another aspect of this motivation is that a common representation should not only support cross-modal problems, but remain appropriate for uni-modal (image or text only) and bi-modal (both image and text) problems. In other words, we desire a representation that is robust for cross-modal problems and at least "agnostic" (*i.e.* lead to equal or better results) for uni-modal and bi-modal problems.

The second motivation relates to the development of a *universal resource* for different specific problems. We expect to develop a common representation space as a generic resource, from a large and general bi-modal dataset, then address specific retrieval or classification problems using this resource. The interest of the "universal resource" is to avoid re-learning a common space for each problem from a specific problem-related dataset.

The third motivation concerns the design of a conceptual multimedia model that allows to attain a higher semantic level. Going beyond the one-to-one matching of multimedia documents on the common space, a multimedia model is expected to be able to match a given document to "more general concepts" that result from a set of other documents. In practice, the first step would simply consists in learning a "general concept" from one modality only, while being able to test it on another modality. This motivation is supported by an increasing practical interest for this type of tasks. We assume that many concepts

have the massive availability of corresponding labeled data in one modality such as text. However, this is not always the case in other modalities such as image. In this case, it is not possible to test a visual content against these concepts because of the missing labeled visual information for learning. An ideal multimedia model allows one to detect these concepts on visual content even if class labels are not (yet) available for this content (or only available for a very limited amount of this content). In such a context, our motivation is to take advantage of the high-level semantics of the textual resource in order to apply it on the visual content. In particular, such a situation may become more common with the current trends in micro-blogging, that evolves from purely textual content (*e.g.* historical Twitter) to multi-modal content (*e.g.* current Twitter) or purely visual content (*e.g.* Instagram).

1.2 Goals

In this thesis, our objective is to propose methods that address bi-modal and cross-modal problems by effectively and efficiently combining, simultaneously, visual and textual information. More precisely, we address the following issues:

- Building a latent common representation space for images and text that supports the "matching" of information from one modality to another to address cross-modal problems.

- Identifying the limitations of the method used to build the common space in terms of quality of "matching" between visual and textual modalities. Each "matching deficiency" on the latent common representation space might hamper the performance in multimedia tasks.

- Proposing methods to exploit heterogeneous visual and textual content in order to enrich the multimedia document representation and thus enhance the performance of retrieval and classification tasks.

- Designing efficient multimedia systems that use as few resources as possible.

1.3 Contributions

This thesis investigates the problem of learning robust representations of multimedia document to address the challenging cross-modal tasks. Our research relies on a common representation space for visual and textual modalities. While the main work focuses on cross-modal tasks, several parts of this thesis consider bi-modal tasks.

In practice, all the methods proposed in this thesis have been evaluated on the joint spaces built using Canonical Correlation Analysis (CCA) or its kernelised version Kernel Canonical Correlation Analysis (KCCA). Indeed, three reasons motivated this choice of a common representation learning method. First, the CCA has been introduced quite a long time ago by Hotelling [1936] and its theoretical foundations are well understood. Besides, when I began my thesis, several significant works in the multimedia community were revisiting and broadly investigating with success this method for cross-modal tasks [Hardoon et al. 2004; Hwang and Grauman 2012a; Costa Pereira et al. 2014; Gong et al. 2014]. The last motivation concerns the objective of our research. Actually, the thesis is not particularly interested in the method of building the common representation space for image and text. Instead, our aim is to make use of such a joint space to develop robust representations allowing to enhance the performance in cross-modal problems. In such a context, basing our work on a method like (K)CCA that maximizes the correlation between original modalities, is a reasonable solution. Furthermore, this is a sufficiently "minimal hypothesis" to expect that the generic proposed methods would still be relevant with more complex common representation space learning approaches. These motivations of our choice are discussed in greater detail in Section 3.2. Also, a discussion of the choice of the joint space learning approach is provided as one of our perspectives presented in Section 7.2.

Deciding to approach cross-modal problems by developing a common representation space for image and text, we identify the two major limitations of such a space as follows.

- *Poorly-represented data on the common space*

  The first limitation relates to poorly-represented data on the common space. The development of such a latent common space relies on extracting statistical regularities

from a large amount of training data. Any fragment of textual data, *e.g.* words, having very few occurrences or weak relations to other data is thus ignored in the joint model. However, the poorly represented information can be very significant in a retrieval context. Disregarding such information may strongly reduce the effectiveness of the joint representation space.

- *Separation between textual and visual modalites on the common space*
  For any given multi-modal document, the projections of its visual and respectively textual features fall far apart. These projections tend to be grouped by modality rather than according to their semantic on the common representation space. A direct use of these projections results in a limited quality of "translation" between modalities.

Our following contributions consider these two limitations in order to improve the performance on cross-modal and/or bi-modal tasks. Our first contribution introduced in Chapter 4 addresses the problem of ill-represented data on the joint space. Then, two other contributions detailed in Chapter 5 and Chapter 6 deal with the separation between image and text modalities. A summary of these contributions is given below, each of them corresponding to an article published during the thesis.

Combining generic and specific information for cross-modal retrieval. In the first contribution, originally published in Tran et al. [2015], we propose a joint model that is able to include *"non regular but likely to be relevant"* information, for the textual modality in particular. The proposed model first identifies such information (mainly words that are rare in the dataset used to learn the common space) and distinguishes it from noise. Then, it finds ways to combine it with the evidence provided by the joint representation model. We examine how the proposed model is applied to address *text-illustration*, a typical problem of cross-modal *retrieval*. This task consists in finding an appropriate image to illustrate the content of a given textual document. By appropriately identifying and taking such information into account, the results of cross-modal *retrieval* can be strongly improved. The proposed approach is compared to others on a previously published benchmark [Feng and Lapata 2010] and shown to produce better results.

Uni-modal data completion with the missing modality. In the second contribution, originally published in Tran et al. [2016b], we consider *cross-modal classification*, a task that was not widely investigated in the multimedia community. It consists in training models on data from one modality and applying them to predict data from another modality. To address this specific problem, we proposed a method relying on a text-image common representation space, called Weighted Completion with Averaging (WCA). A key aspect of this contribution is the use of a bi-modal dataset, called auxiliary dataset, that acts as a set of connections between the modalities within the joint space. We suggest to rely on this auxiliary dataset to find the complementary information in the missing modality of a uni-modal document. Once this complement has been identified, a more complete bi-modal representation of any uni-modal data can be built from information in both modalities. Experiments have been conducted on well-known datasets including the Pascal VOC07 image collection with tags collected by the work of Hwang and Grauman [2012a] and the NUS Wide dataset [Chua et al. 2009]. The evaluation shows that the WCA representation method significantly improves the results compared to the use of a latent space alone. Also, the level of performance achieved with respect to classical tasks *e.g.* bi-modal classification and cross-modal retrieval, makes cross-modal classification a convincing choice for real applications.

Aggregating Image and Text Quantized Correlated Components. In the third contribution, originally published in Tran et al. [2016a], we put forward a robust representation method, called Multimedia Aggregated Correlated Components (MACC) that aggregates information provided by the projections of both visual and textual modalities on their joint subspaces. MACC representations aim to reduce the separation between the projections of visual and textual features by embedding them in a local context reflecting the data distribution in the common space. More precisely, a "unified vocabulary" (codebook) is obtained by quantizing the projection space. Both visual and textual projections of a document are then encoded with respect to one or several codewords and sum pooled to get the final MACC representation. This representation can be employed for bi-modal and uni-modal documents. In the case of uni-modal documents, the uni-modal completion process relying on an auxiliary dataset (introduced in the second contribution) is performed

to provide the corresponding complementary information in the missing modality of the document. Extensive experimental evaluations have been conducted on three challenging datasets including Pascal VOC07 with tags collected in the work of Hwang and Grauman [2012b] for bi-modal and cross-modal classification, FlickR 8K [Rashtchian et al. 2010] and FlickR 30K [Young et al. 2014] for cross-modal image retrieval. Obtained results show that our proposed MACC representation allows to reach state-of-the-art performance in various multimedia tasks such as bi-modal and cross-modal classification and image retrieval.

1.4 Organization of the thesis

The rest of this dissertation is divided into six chapters.

In **Chapter** 2, we review some of the most indicative work on the joint modeling of the image and text modalities for multimedia retrieval and classification, especially in a cross-modal context. A main part of this chapter presents a thorough survey on common representation learning for image and text. We end up the chapter by giving an overview of the datasets of the state-of-the-art employed to evaluate models proposed in this thesis.

In **Chapter** 3, we study the characteristics of Kernel Canonical Correlation Analysis method on which we rely to build common representation space. Thereafter, in this first contribution, we identify two limitations of such a joint space, that are highlighted through several experiments. The results of these experiments allow to better understand the organization of the modalities within the common space, and consequently the reason why cross-modal and bi-modal tasks are not "fully solved" by such an approach. Hence, this chapter constitutes a kind of initial experimental motivation for proposing solutions to the identified limitations, that are further developed in the three next chapters.

Chapter 4 describes our second contribution in which we address the limitation concerning ill-represented data on the common representation space. For this purpose, we put forward a method to combine poorly-represented information with other relatively well-represented information in order to enhance the performance of cross-modal retrieval. Our proposal is evaluated in the context of a challenging text illustration task.

Chapter 5 describes our third contribution in which we introduce the cross-modal classification problem. We then propose the WCA representation to address this task. Relying on a bi-modal auxiliary dataset, WCA *completes* a uni-modal document and then builds a more comprehensive bi-modal representation for this document. Experimental evaluations show that the bi-modal WCA representation significantly improves the "translation" between modalities and consequently the performance in cross-modal classification compared to the direct use of the original common representations.

Chapter 6 describes our fourth contribution. With the aim to reduce the separation between text and image on the joint space, we introduce in this chapter the bi-modal MACC representation. We propose a method to build this representation for bi-modal documents and uni-modal documents. Finally, extensive experiments are conducted for image retrieval and various (bi-modal and cross-modal) classification tasks, showing the effectiveness of our proposed MACC representation.

Chapter 7 concludes this dissertation. We first summarize the motivations and the contributions of this thesis. We eventually discuss the perspectives that can be inspired by the presented research problems. In particular, we consider the extension of our contribution to the case of common spaces built on other principles than correlation maximization.

Chapter 2

State-of-the-art

2.1 Introduction

This chapter reviews some of the most indicative work in the literature of multimedia retrieval and classification in the context of social media. We consider two modalities of data: the visual modality represented by images and the textual modality represented by tags or descriptions associated to each image.

We start by reviewing various techniques of single-media content representation, respectively for visual and textual modality in Section 2.2. Afterwards, we briefly overview different multimedia information retrieval and classification problems covering uni-modal 2.3.1, multi-modal 2.3.2 and cross-modal 2.3.3 paradigms. For each of these problems, we introduce several relevant related work that were proposed over the last few years. We center around those involved in classification, retrieval and annotation tasks for visual and textual data.

Our main research focuses on the problem of cross-modal between image and text modalities. The recent literature in computer vision and multimedia has shown that learning a common representation to these two modalities is a relevant solution to address such a problem. We thus continue this chapter by proposing a brief state-of-the-art of joint embedding approaches for image and text modalities in Section 2.4. In general, these approaches aim to learn a mapping from the original visual or textual space to a common representation that preserves the relatedness of data between the different modalities. According to the fundamental principle of the resulting common space, we further classify these approaches into three categories: correlation learning (Section 2.4.1), topic modeling (Section 2.4.2) and rank-based approaches (Section 2.4.3). Our considerable attention has been paid to those relying on correlation learning scheme.

Lastly, we conclude the chapter by a summary on various multi-modal datasets including BBC News, Wikipedia articles, Pascal VOC07, Nus WIDE and FlickR 8K/30K databases in Section 2.5. They are treated in our contribution for different multimedia problems such as text illustration, cross-modal image retrieval or cross-modal classification.

2.2 Single-media representation

Feature extraction and representation is a crucial stage in multimedia processing. We briefly review several visual and textual representations which have been successfully used in multimedia retrieval over the last few years.

2.2.1 Visual Features

Bag-of-Visual-Words. A standard approach to describe an image is to extract a set of local patch descriptors, encode them into a high dimensional vector and then pool them into an image-level signature. The most common patch encoding strategy consists in quantifying the local descriptors into a finite set of prototypical elements (called codebook). This leads to the popular Bag-of-Visual-Words(BoVW) representation [Sivic and Zisserman 2003; Csurka et al. 2004a]. Before the rise of convolutional neural networks, BoVW has been the dominant feature trend for image representation in many computer applications *e.g.* TRECVID video retrieval [Over et al. 2014] and has been seen as one of the state-of-the-art representation for visual content. BoVW was inspired by the traditional Bag-of-Words (BoW) method proposed by Salton and McGill [1986]. While BoW represents a textual document by a vector of the occurrences of each word in the document, BoVW aggregates local descriptors extracted from interest points (image patches) into a fixed-size vector that describes the global properties of the image. An example of these local features is the well-known dense SIFT descriptors [Lowe 2004].

The pipeline of BoVW is described as follows (see Figure 2.1).

Codebook learning. From every image in a training dataset, local features are extracted. BoVW learns a codebook for example by performing a k-means algorithm on these local features. Each cluster is treated as a discrete visual word.

Coding. For each image, local features are mapped to visual words into compact descriptors during the *coding* step. Different coding methods have been investigated in the literature. In the original BoVW model, *hard coding* [Csurka et al. 2004b] maps the local feature to its nearest visual word. However, this coding often introduces large quantization errors. van Gemert et al. [2010] proposed *soft coding* method that assigns a local feature

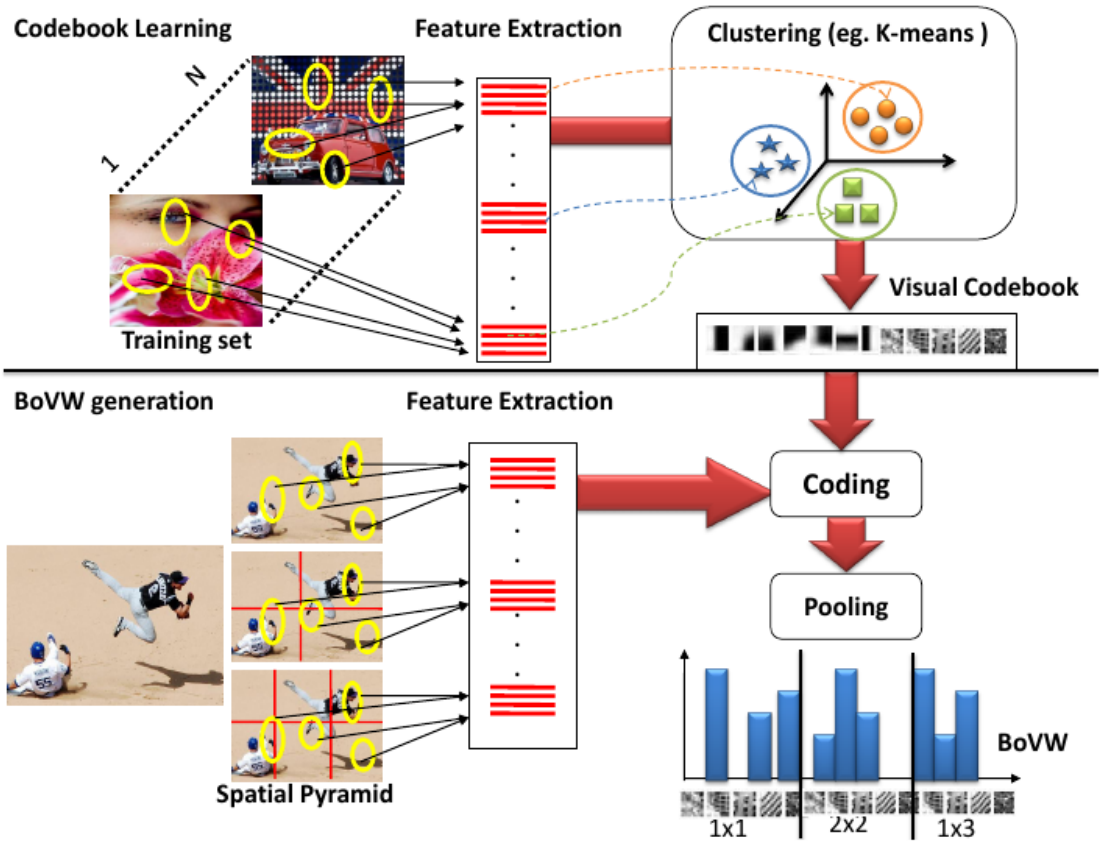

Figure 2.1: The flowchart of the Bag-of-Visual-Words (BoVW) generation scheme including three steps of codebook learning, local features coding and pooling. Figure extracted from [Znaidia 2014].

to all codewords, according to the similarity between this local feature and each codeword. While the quantization errors has been reduced, no proof showed that the use of the entire codebook is optimal. Wang et al. [2010] proposed an efficient *locality constrained linear coding (LLC)* by mapping the local feature to only the L-nearest codewords. This promising image representation has shown its computational advantage and yielded good performance for image classification [Wang et al. 2010; Liu et al. 2011].

Pooling. In the pooling step, these local descriptors are aggregated into a unique image-level representation using a pooling function. The latter can be the average, the sum [Lazebnik et al. 2006] or the maximum function [Yang et al. 2009] of all local descriptors (component by component) of an image. The sum-pooling is the sum of the coding coefficients obtained on local features while the average-pooling is its normalized form.

Several work [Boureau et al. 2010; Wang et al. 2010; Yang et al. 2009; Liu et al. 2011] indicate that max-pooling that chooses the largest coefficient for a visual word can improve the performance of classification and recognition task. Recently, an extension of pooling scheme, called BossaNova, was proposed by Avila et al. [2013] in order to enhance image representation. The technique keeps a histogram of distances between the local descriptors of the image and those in the codebook and hence preserves important information about the distribution of the local descriptors around each codeword. BossaNova produces a distance distribution instead of compacting all information concerning a codeword into a single scalar.

Spatial Pyramid Matching. Since the classic BOVW is an orderless signature that disregards the location of the visual words in the image, the spatial pyramid matching (SPM) [Lazebnik et al. 2006] is an interesting way to incorporate some global spatial contextual information into the signature. SPM consists in partitioning an image into sub-regions in different manners. For each partition, BoVW is computed for each sub-region. A pooling operator is then conducted on these region-relative BoVWs to build a unique representation of image relying on this division. The final representation of image is a concatenation of all representations resulting from different image partition. SPM shows significantly improved performance on several multimedia tasks such as the challenging scene categorization [Lazebnik et al. 2006] or image classification [Bosch et al. 2007].

Fisher Vector. The Fisher Vector (FV) extends the BoVW by going beyond counting, *i.e.* 0-order statistics, to encode second order statistics. The FV representation is computed by characterizing local descriptors by their corresponding deviations from an "universal" generative Gaussian Mixture Model (GMM) of the loglikelihood of the problem. The GMM model can thus be seen as a "probabilistic visual vocabulary". The deviation is measured by computing the gradient of the sample log-likelihood with respect to the model parameters. Originally designed for classification, Perronnin et al. [2010b] further greatly improved the retrieval performance of FV by applying a set of normalization strategies *e.g.* L2 or power normalization to FV and combining this representation with the spatial pyramids. The FV representation shows many advantages in comparison to the BoVW such as its lower

computational cost by using much smaller vocabularies or its higher performance even with simple linear classifiers. However, while the BoVW is quite sparse, the FV is almost dense.

Vector of Locally Aggregated Descriptors. Jégou et al. [2010] proposed a simple efficient way of aggregating local image descriptors into a vector of limited dimension for large-scale applications, called Vector of Locally Aggregated Descriptors (VLAD). VLAD can be seen as a simplified version of the Fisher vector representation, that uses a (quite small) codebook in place to the GMM of the log-likelihood to represent the "universal vocabulary", and replace the gradient by a simple point-wise difference of the local vector and the codewords.

Assuming $c_1, c_2, .., c_k$ a codebook learned using k-means. Each local descriptor x_t ($d-$dimensional) is associated to its nearest visual words $NN(x_t)$ in the codebook. For each codeword c_i, the differences $x_t - c_i$ of the vectors x_t assigned to c_i are accumulated

$$v_i = \sum_{x_t \quad such\ that \quad NN(x_t)=c_i} (x_t - c_i) \tag{2.1}$$

The VLAD representation v is the concatenation of the $d-$dimensional vectors v_i and has $D = d \times k$ dimensions.

By employing an asymmetric product quantization scheme for the vector compression part, Jegou et al. [2012] jointly optimized the dimensionality reduction and compression of VLAD representation. This image representation can be reduced to a few dozen bytes while preserving high accuracy. Searching a 100 million image dataset takes about 250 ms on one processor core. However, while improving scalability, this aggressive compression significantly decreases accuracy compared to the use of full FV.

CNN-based deep visual features. Before the rise of the convolutional neural networks (CNNs), Fisher Vector and Vector of Locally Aggregated Descriptors have been powerful shallow representations for image retrieval and classification, particularly on the PASCAL VOC 2007, Caltech101 [Chatfield et al. 2011, 2014] and on the renowned ImageNet dataset [Russakovsky et al. 2015]. Since 2011, CNNs significantly improved the state of the art in many computer vision tasks. CNN networks which are trained for classification

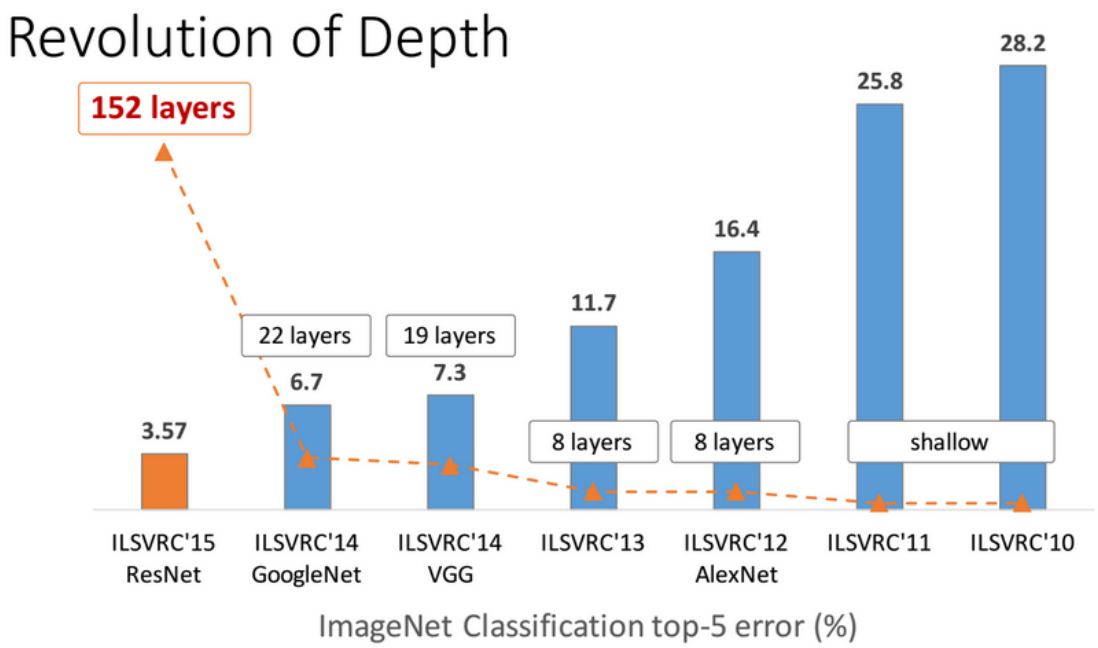

Figure 2.2: The CNN-based features brought a significative breakthrough in image classification task on ImageNet dataset (from [Russakovsky et al. 2015])

on ILSVRC have been used as feature extractors by removing the output layer (thus using the output of the last fully connected layer as image representation), then used in a transfer learning scheme on other classification benchmarks. Several CNN-based deep visual features extracted from these pre-trained networks have been proposed to the computer vision community [Jia et al. 2014]. These deep features have shown to perform excellently over standard classification and detection benchmarks [He et al. 2015a; Simonyan and Zisserman 2014; Szegedy et al. 2015; Ioffe and Szegedy 2015; He et al. 2016].

Concretely, both powerful performance and revolution of the CNNs features has been clearly shown through image classification task on the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [Russakovsky et al. 2015] (see Figure 2.2). In ILSVRC 2011 challenge, Fisher Vector reached the state-of-the-art with a top-5 classification error of 25.8%. Afterward, variants of CNN models have achieved increasingly better performance by bringing down the classification error to 16.4% with AlexNet [Krizhevsky et al. 2012], 11.7% with Clarifai [Zeiler and Fergus 2013], 7.3% with VGG [Simonyan and Zisserman 2014], 6.7% with GoogLeNet [Szegedy et al. 2015], 4.9% with PReLU-net [He et al. 2015a]

and recently 3.6% with ResNet [He et al. 2016]. The humain expert for this task is 5.1%.

Also, the classification performance evolution on PASCAL VOC datasets has been exploited in the work of Chatfield et al. [2014]. On the Pascal VOC 07 benchmark, using shallow visual features, the mean Average Precision (mAP) was 54.48% for the BoVW in 2008 and 61.7% for the improved Fisher Vector in 2010 [Perronnin et al. 2010b]. Recently, the use of CNN-based features has further increased the mAP score to 82.4 [Chatfield et al. 2014], 85.2 [Wei et al. 2014], 86.1 [Simonyan and Zisserman 2014], 88.2 [Tammazousti et al. 2016].

Above image classification, Pepik et al. [2015] has indicated that a direct usage of the CNN-trained features can yield top performing results on tasks such as object detection [Girshick et al. 2013], pose estimation [Chen and Yuille 2014], face recognition [Schroff et al. 2015], object tracking [Li et al. 2014], keypoint matching [Fischer et al. 2014], stereo matching [Zbontar and LeCun 2015], optical flow [Dosovitskiy et al. 2015], boundary estimation [Xie and Tu 2015], and semantic labeling [Long et al. 2015].

Beside the performance obtained, another advantage of CNN-based features is that their extractors were publicly released. In other words, we can extract and use CNN-based features without requiring the knowledge or computing infrastructure for training a convolutional neural network from scratch. OverFeat [Sermanet et al. 2013], Caffe [Jia et al. 2014] and VGG [Simonyan and Zisserman 2014] were the first CNN off-the-shelf features. These CNNs are both trained using ImageNet data associated to 1,000 concepts (classes) of the ILSVRC challenge.

The typical architecture of a CNN is composed of three cascaded stages: convolution, non-linearity and pooling [Krizhevsky et al. 2012]. The convolutional layers output feature maps, each element of which is obtained by computing a dot product between the local region it is connected to in the input feature maps and a set of weights (filters). In general, an elementwise non-linear activation function is applied to these feature maps. One of the most used is the rectified linear unit (ReLU) that implement $f(x) = \max(0, x)$. The pooling layers perform a downsampling operation along the spatial dimensions of feature maps via computing the maximum on a local region. The fully-connected (FC) layers finally follow several stacked convolutional and pooling layers, and the last fully-connected

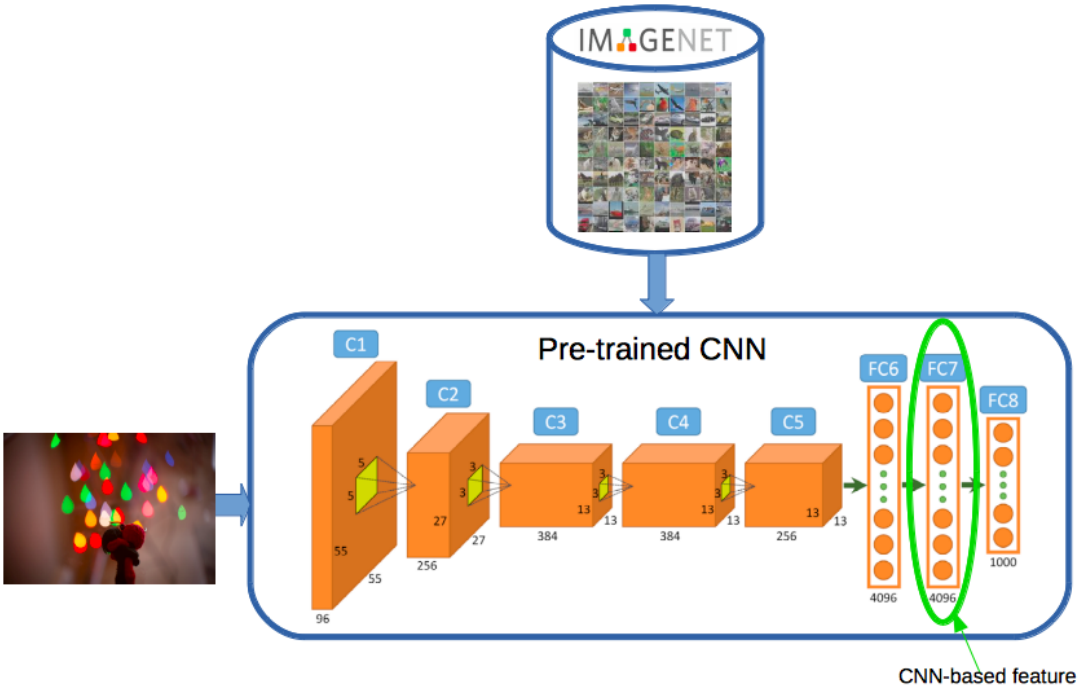

Figure 2.3: CNN-based feature extracted from pre-trained CNN model

layer is a Softmax layer that computes the scores for each defined class. CNNs transform the input image from original pixel values to the final class scores through the network in a feedforward manner. The parameters of CNNs (i.e., the weights in convolutional and FC layers) are trained with classic stochastic gradient descent[1] and uses the backpropagation algorithm [Williams and Hinton 1986] to efficiently compute the gradient. Most of CNNs such as AlexNet [Krizhevsky et al. 2012], CaffeNet [Jia et al. 2014], VGG-Net [Simonyan and Zisserman 2014] and [Zhou et al. 2014] have this architecture.

In this dissertation, we do not get into the technical details on how CNNs models are trained in each stage. We are more interested in the direct use of the CNN features for computer vision tasks. The internal layers of a CNN can act as a generic extractor of image representations, regardless on the architecture of networks (*e.g* the number of convolutional of fully-connected layers, the size of the sliding window used for the convolution operation, the image transformations or the regularization used), see Figure 2.3. Most of current

[1]Actually, instead of using only one sample, the gradient is computing by averaging several samples, grouped into a so-called minibatch

CNN-based visual feature extractors are trained on 1,2 million images from 1,000 concepts of the ImageNet dataset. The network presented in the Figure 2.3 is a simplified view of AlexNet proposed by Krizhevsky et al. [2012]. The size of the parameter vectors (the number of neurons) is respectively 290400 for the first layer C1, 186624 for C2, 64896 for C3 and C4, 43264 for C5, 4096 for the full-connected layers fc6 and fc7. The last fully full connected layer fc8 (size of 1000) offers predictions for the 1000 ImageNet classes on which the CNN was trained. In our work, we use the deep visual features extracted from OverFeat and VGG network. While these CNNs are designed in different manner, the feature extraction procedure is the same as described above.

2.2.2 Textual Features

Bag-of-Words. Bag-of-Words (BoW) model [Salton and McGill 1986] is the most basic but nevertheless widely employed technique for text representation. This representation has been applied in various textual tasks such as retrieval or classification. The Bag-of-Words first obtains a textual vocabulary from all of the available documents. Several preliminary steps are usually required for vocabulary learning such as reducing inflectional forms and sometimes derivationally related forms of a word to a common base form *e.g.* stemming and lemmatization, removing words that are known to be usefulness to discriminate, such as articles and determinants, some adverbs, etc. A document is modeled by a vector where each component is a function of the frequency of appearance of a word of the dictionary in this document, followed by a L2-normalization. For instance, this function can be term frequency (TF), term frequency–inverse document frequency (TF-IDF) or BM25.

As in the basic BoW model, this function is term frequency (TF) referring to how many times a dictionary's word is present in the document. Another way than to judge the topic of an document by the words it contains consists on Term frequency–inverse document frequency weighting (TF-IDF). This measure reflects how important each word is to a document in a collection or corpus. The importance increases proportionally to the number of times a words appears in the document but is offset by the frequency of the word in the corpus. TF-IDF weights are designed to give more importance to terms frequent in the document while penalizing words appearing in too many documents. Other weighting

method *e.g.* BM25 [Robertson et al. 2009] can yield better performance than TF-IDF.

These representations have been successfully used in many applications such as document retrieval, classification, topic modeling or image annotation, etc. Specially, Hwang and Grauman [2012a] integrated the order of image tags provided by user in the BoW representation to leverage the "relative importance" of objects in the scene.

Applying BoW model for text representation, two documents (or a query and a document) are considered similar if they are exactly composed of the same words by computing a cosine similarity between two representations. However, the vector space representation suffers from the classics problems of natural languages: synonymy and polysemy. Synonymy refers to a case where certain topic can be expressed with different words, *e.g. car and automobile*. Polysemy on the other hand refers to the case where a given word can be used in totally different context.

Probabilistic topic models. These models attempt to take into account the links between words in order to address the shortcomings of the classical vector representation. Topic-based representation methods have been successfully applied to many text mining tasks such as retrieval, summarization, categorization and topic models.

Among topic models, Latent Semantic Analysis (LSA) [Deerwester et al. 1990] is a well-known method for text representation. LSA represents the documents with respect to latent topics identified from the correlations between the occurrences of terms in the vocabulary. LSA maps the standard vector space representation of a document to a lower dimension latent space where each dimension can be seen as a topic. For this purpose, one first determines the document-term matrix which describes the frequency of terms occurring in a collection of documents. In such matrix, rows correspond to documents in the collection and columns correspond to terms. There are various schemes *e.g.* TF-IDF for determining the value that each entry in the matrix should take. LSA consists in applying a Singular Value Decomposition (SVD) to obtain a lower-dimensional approximation of this documents-terms matrix.

The idea of identifying a number of latent topics from the set of terms in a vocabulary learnt from a corpus or collection of documents has been subsequently developed by using

other techniques such as Probabilistic Latent Semantic Analysis (pLSA) [Hofmann 2001] or Latent Dirichlet Allocation (LDA) [Blei et al. 2003]

Explicit Semantic Analysis. The mentioned textual representation methods are based on purely statistical techniques that did not make use of *a priori* world knowledge. To improve text representation with massive amounts of world knowledge, Gabrilovich and Markovitch [2007] proposed a novel textual representation method, called Explicit Semantic Analysis (ESA). ESA represents a text (document) as a weighted mixture of a predetermined set of natural concepts derived from Wikipedia, the largest encyclopedia in existence. In this way, the meaning of a text fragment is thus interpreted in terms of its affinity with a host of Wikipedia concepts. This semantic analysis technique is explicit in the sense that it manipulates explicit concepts grounded in human cognition, rather than "'latent concepts"' used by LSA, pLSA or LDA.

Word2Vec: a neural probabilistic language model. The previous representation methods consist in describing a document (text) using the set of words (terms) that it contains. It is worth that information relying on the context of words is ignored in such a representations. However, the context of a word in a sentence allows characterizing quite well the word on both syntactic and semantic sides. Therefore, such information is useful to improve the textual representation.

The idea of word's context modeling is employed in Word2Vec [Mikolov et al. 2013], a recent word embedding trained from two-layer neural network. Word2Vec takes as input a large corpus of text and produces a high-dimensional vector space, typically of several hundred dimensions. Using Word2Vec, each unique word in the corpus is assigned to a corresponding vector in the space. These words vectors are positioned in the vector space such that words that share common contexts in the corpus are located in close proximity to one another in the space. With the proposed Skip-gram model, the goal of Word2Vec consists in finding representations that allow to predict the best possible context of input word.

The Word2Vec representation can be developed not only for individual words but

also for a short sentence containing several words. In such cases (*e.g.* phrase, list of tags), a possible representation can be computed as the center of gravity of Word2Vec representations of words in the document.

A Word2Vec model pre-trained on part of Google News dataset of about 100 billions words has been publicly released [Mikolov et al. 2013]. Using this model, each word is modeled by a 300-dimensional vector. Unlike the basic representations such as TF-IDF or ESA, the Word2Vec representation is fairly low-dimensional (*e.g.* 300) and dense.

Recently, an effective textual representation based on the original Word2Vec representations have been proposed in the work of Klein et al. [2015]; Wang et al. [2016b]. The novel technique aims to use the Fisher Vector [Perronnin et al. 2010b] to represent sentences by pooling the Word2Vec embedding of each word in the sentence.

2.3 Multimedia information retrieval and classification

In this section, we review work on multimedia information retrieval and classification for image and text, concerning uni-modal, multi-modal and cross-modal problems.

2.3.1 Uni-modal problems

Many extensive researches have been conducted on the problems of image and text retrieval in the fields of information retrieval, computer vision and multimedia. In the early years of these domains, the emphasis has been placed on uni-modal approaches.

Uni-modal retrieval is a classical problem in multimedia retrieval [Salton and McGill 1986; Smeulders et al. 2000; Shen et al. 2000; Srihari et al. 2000; Vasconcelos 2004; Datta et al. 2008] where query and retrieved documents in the reference base are represented according to the same modality. Text-based and content-based retrieval are the popular uni-modal problems which were introduced in the early years of image retrieval. For example, in [Shen et al. 2000] a query text and in [Vasconcelos 2004] a query image is used to retrieve respectively similar text documents and images. An example of uni-modal problem is illustrated in Figure 2.4.

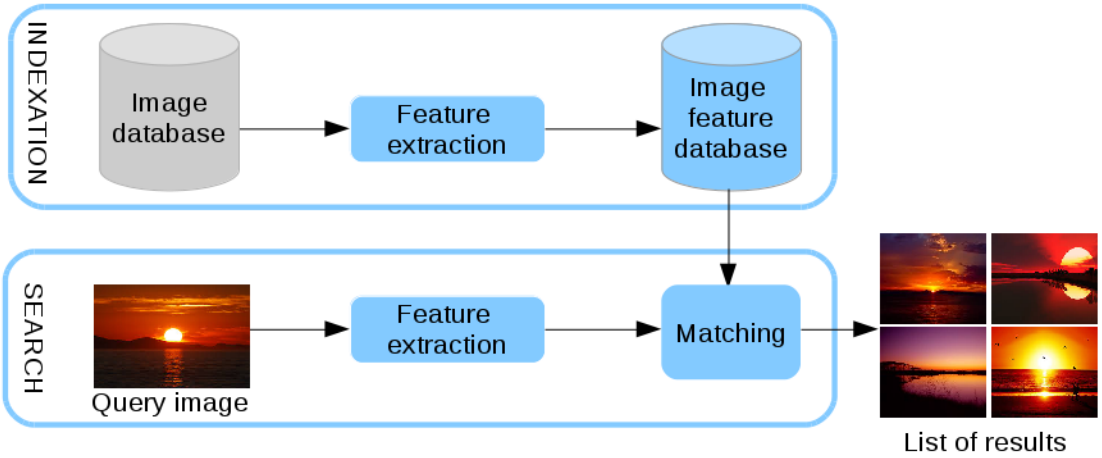

Figure 2.4: Illustration of a content-based image retrieval (CBIR) system

The first contributions in image retrieval centered around the text-based (keyword-based) retrieval systems which used keywords as descriptors to index an image [Salton and

McGill 1986; Shen et al. 2000; Srihari et al. 2000]. The keywords are extracted from the texts surrounding image such as image title, image caption or content of article containing image. For instance, Shen et al. [2000] explored the context of web pages as potential annotations for the images in the same pages. Srihari et al. [2000] proposed to extract named entities from the surrounding text to index images. Afterward, the semantic of an image is represented by a textual representation, for instant using TF-IDF [Salton and McGill 1986], Word2Vec [Mikolov et al. 2013], from these relevant keywords. During the retrieval stage, the system evaluates the similarity between textual representations of query text and retrieved images. The major constraint of text-based image retrieval methods is that it requires high-quality text information of image, *e.g.* annotation quality and completeness. In many situation, this requirement may not be satisfied. Having humans manually annotated images by entering keywords or metadata in a large database can be time consuming and may not capture the keywords desired to describe the actual content of the image.

A few years later, content-based image retrieval (CBIR) has been employed as an alternative to text-based image retrieval. CBIR concentrates on the contents of image rather than the metadata *e.g.* keywords, tags, descriptions associated with the image which are used in text-based approaches. More particularly, it regards information extracted from image itself such as colors, shapes, textures, etc. Hence, such paradigm allows the ability to query by example, which means that users express their queries by providing examples, *e.g.* images, of what they are looking for, and items, *e.g.* images, in reference database are retrieved by similarity to these user-provided examples. Various types of visual features including low-level and high-level features have been investigated in CBIR. At early years, low-level features such as color, texture, shape, spatial relations or combination of above features were used. A comprehensive survey on these approaches is given by Liu et al. [2007]. More recently, advanced features such as Fisher Vector [Perronnin and Dance 2007] or those extracted from a neural network [Jia et al. 2014; Simonyan and Zisserman 2014] have efficiently improved the performance of CBIR systems.

Despite these advantages, it is important to recognize the shortcomings which make CBIR not effective for all multimedia search problems. One problem with most CBIR

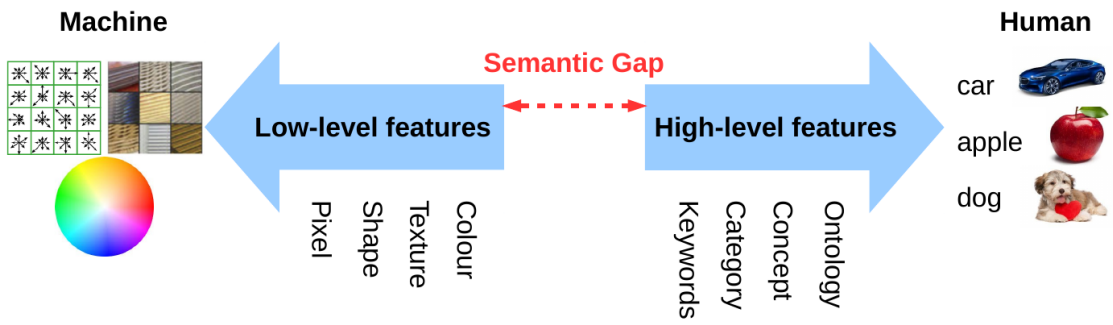

Figure 2.5: Semantic gap between low-level visual features offered by CBIR systems and semantic concepts identified by users

approaches is the reliance of visual similarity to judge semantic similarity. This fact may be problematic due to the well-known "semantic gap" between low-level content and higher-level concepts. In a review of the early years of CBIR, Smeulders et al. [2000] defined the latter as "the lack of coincidence between the information that one can extract from the visual data and the interpretation that the same data have for a user in a given situation". The latter remains a major problem in CBIR and severely hampers the performance of uni-modal image retrieval systems.

The semantic gap manifests typically between the image semantic that a user requires and the low-level features that CBIR systems offer (Figure 2.5). While humans interpret images at different levels, both in low-level features (color, texture, shape, spatial layout) and high-level features (keywords, text descriptors), a machine is only able to interpret images based on low-level image features. Extensive experiments on CBIR show that low-level image features often fail to describe the high-level semantic concepts in user's mind [Zhou and Huang 2000]. On the other hand, while CBIR systems index images using low-level features, humans prefer to express their information needs as queries at the high-level semantic of human natural language instead of the level of preliminary image features. For instance, users' queries may be "find an image of sunset" rather than "find an image contains red and yellow colors". Recently, CNN-based feature is a hope for bridging the semantic gap in CBIR by an embedding from low-level feature extracted from images to concepts of higher level of semantics.

2.3.2 Multi-modal problems

Over the last decades, with the advance of computer network and multimedia technologies, large amounts of different types of media data, including images, texts and videos have been rapidly generated, shared and accessed on social networks such as FlickR, Twitter, YouTube and Wikipedia. It is common that different types of data are used to describe the same events or topics. For example, a web page usually contains not only textual document but also images or videos used to illustrate the article's content. Such different types of data are referred as multi-modal data, which exhibit heterogeneous properties.

In multi-modal problems, queries and entries in the reference base combines multiple content modalities. For example, text and image of a news article are both used to retrieve entries with the same combination of modalities (*e.g.* text and image) of documents in the reference base. Many applications for multi-modal data have been introduced and exploited in the multimedia community such as hot topic detection, personalized recommendation, video retrieval or event detection. These efforts of multi-modal approaches have become increasingly widespread, due in part to large-scale research and evaluation benchmarks such as TRECVID [Smeaton et al. 2006] and ImageCLEF [Müller et al. 2010], involving datasets that span multiple data modalities.

Classical uni-modal approaches are not able to deal with multi-modal problem because they only perform similarity search of the same media type, such as text-based retrieval or content-based retrieval, etc. Multi-modal approaches aim to integrate the use of data from multiple modalities so that they can support the similarity search for multi-modal data.

Since there exists *correlative* and *complementary* relations between different modalities of data such as image and text, a fusion of data from these modalities can improve the performance of a uni-modal model. Various attempts have already been proposed in the literature to fuse multiple modalities for multi-modal tasks [Li et al. 2009; Barnard et al. 2003; Wang et al. 2009; Liu et al. 2013]. The approaches are categorized into two distinct multi-modal fusion schemes: early fusion and late fusion [Snoek et al. 2005]. In early fusion methods, fusion is performed by combining the low-level features of multimedia object and then using the fused features for further processing *e.g.* classification or retrieval.

On the other hand, in late fusion methods, individual uni-modal learning methods are first used in each modality separately and their high-level results and decisions are then fused.

Early fusion

Early fusion is performed at the feature level (Figure 2.6). The approach first extracts uni-modal features for image and text and then combines these features into a single multimedia representation. Classification or retrieval task can be performed on these multimedia representations.

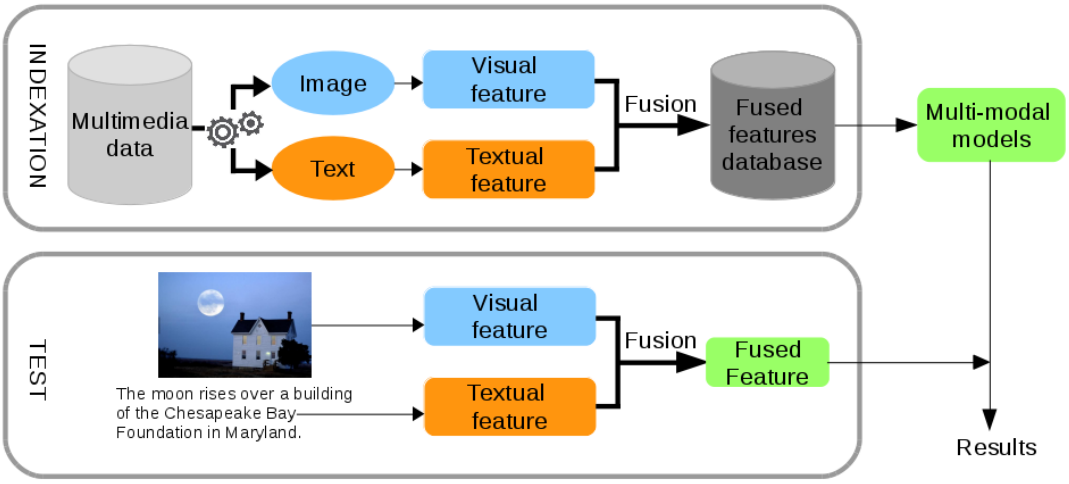

Figure 2.6: Multi-modal approach based on early fusion.

A simple and widely employed method to combine the single-media representations in the early fusion is to concatenate these representations. Li et al. [2009] learned support vector machine (SVM) models for landmark classification using the concatenation of visual and tag features. The classification performance showed the early fusion of text and image leaded to significant improvements compared to uni-modal classification on textual or visual features. However, such a simple concatenation only allows to exploit the complementary relation between text and image and ignores the correlation that can exist between these two modalities.

Other early fusion approaches attempted to take into account the correlation between image and text. Barnard et al. [2003] introduced multi-modal models which learn the

joint distribution of image regions and words. The models were used to predict words associated with whole images (annotation) and corresponding to particular image regions (region naming).

Despite the simplicity of the early fusion approaches, one of their disadvantages is the large size of the representations issued from a concatenation operator of single-media features. The induced curse of dimensionality becomes a bottleneck for the learning task. Another disadvantage of such approaches is the difficulty in combining features of different natures (*e.g.* image and text) into a common homogeneous representation since each dimension of the resulting vector does not correspond to the same underlying type of information. For instance, one dimension can correspond to a quantity of "green color" while another can be a normalized weight of the word "leaves". Minkowsky normalization *e.g.* dividing by the Euclidean norm of each vector before concatenation can at least make the range of the dimension relatively similar.

Late fusion

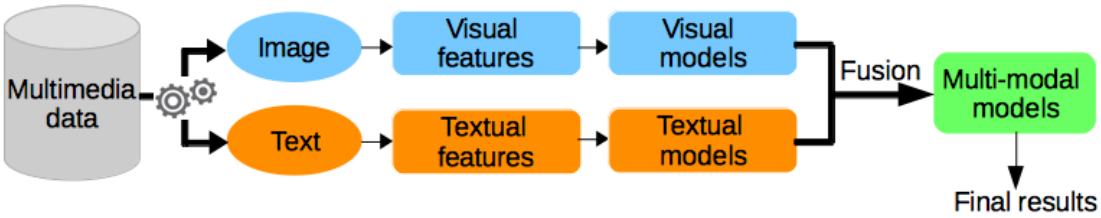

Figure 2.7: Multi-modal approach based on late fusion.

Late fusion approaches are competitive alternatives to overcome the drawbacks of early fusion. While early fusion is performed at the feature level, late fusion is performed at decision level (Figure 2.7). The late fusion scheme consists in integrating the scores predicted by individual classifiers of different modalities through a fusion operator. The fusion can be made from a statistical rule (such as sum, average, max, min, majority voting) or a classification-based approach. The late fusion approaches not only provide a trade-off between preservation of information and computational efficiency but also perform

favorably compared to early fusion methods in several comparative studies, for instant on visual concept detection in video sequences [Snoek et al. 2005; Liu et al. 2013] or on video retrieval [Amir et al. 2004].

Wang et al. [2009] proposed to build a textual feature for an untagged image and then merge both textual and visual content for object image classification task (see Figure 2.8). The textual feature is built using an auxiliary dataset of images annotated with tags *e.g.* downloaded from FlickR. For each image, the model extracts visual features and finds its nearest neighbor images from the auxiliary set. Text associated with these near neighbor internet images is used to build the textual features. For image classification task, the model train two classifiers corresponding to concepts separately, one for image and the other one for text. In the fusion step, a third classifier is trained to combine the classification scores of the two initial classifiers into a final prediction. This classifier uses logistic regression and is learned on a validation set.

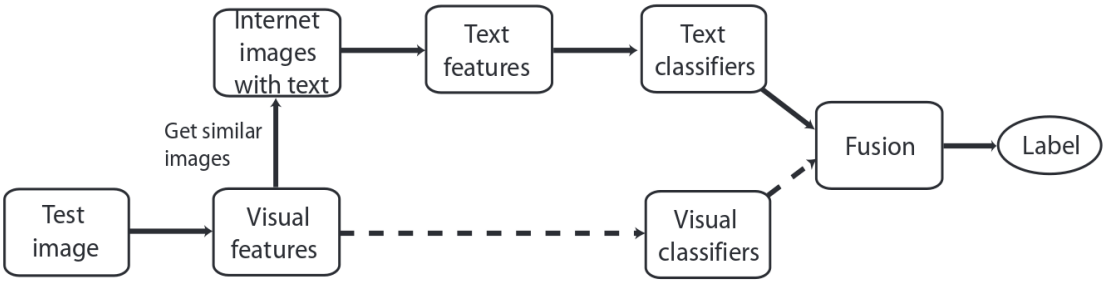

Figure 2.8: The framework of the approach proposed by Wang et al. [2009].

For each image, the model extracts its visual features and finds the most similar images from auxiliary dataset. The text associated with these near neighbor auxiliary images is summarized to build the textual features. Textual and visual classifiers are separately trained and their scores are fused to perform the final classification.

Liu et al. [2013] proposed a fusion scheme for visual concept recognition, called Selective Weighted Late Fusion (SWLF). SWLF automatically selects and weights the scores from the best features among a set of textual and visual features according to the concept to be classified. Concretely, this approach includes a training stage and a testing stage. The training stage trains models using SVM on training data for each pair of concept and type of features. The SWLF is learned by optimizing the overall mean average precision of these

models on validation data in order to selectively choose the best classifiers applied for an testing data. In the testing stage, the approach extracts various types of features of the input image and then applies the corresponding fusion scheme proposed by SWLF for each concept to deliver a recognition decision.

2.3.3 Cross-modal problems

Recently, the research attention in multimedia community has largely focused on cross-modal problems. In cross-modal model, query from one type (modality) of data can be matched to entries from other types of data in the reference base. That is, in the context of image and text, given a text document, it can find the most related images ; or given an image, it finds the words (sentence) that best describe the image. These tasks are central to many applications of practical interest, such as finding the picture that best illustrates a given text (e.g., to automatically illustrate a page of a story book), finding the texts that best match a given picture (e.g., to automatically generate a caption or a description of a picture), or furthermore searching using a combination of text and images. The general scheme of cross-modal retrieval problem is illustrated in Figure 2.9.

Cross-modal model is more flexible than uni-modal and multi-modal models, since by submitting either one or multiple media objects in cross-modal model, we can obtain all of the related media objects of different media types. Furthermore, since different types of media provide complementary information, a cross-modal model can help to search in a more natural way. For instance, given a query image of the Eiffel tower, besides retrieving the images of Eiffel tower, cross-modal model can also suggest the related media contents of different media types such as text description, *e.g.* the travel guide about this site.

Besides, cross-modal is still a difficult problem in defining how to measure the content similarity between different types (modalities) of data. In terms of images and texts, cross-modal model directly addresses the well-known problem of *"semantic gap"* [Smeulders et al. 2000] which is described in Section 2.3.1, and refers to the interpretation inconsistency between the high-level semantic description of visual content and the extracted low-level image descriptors. Cross-modal model requires cross-media relation modeling to bridge the *"semantic gap"* so that users can retrieve what they want by submitting what they

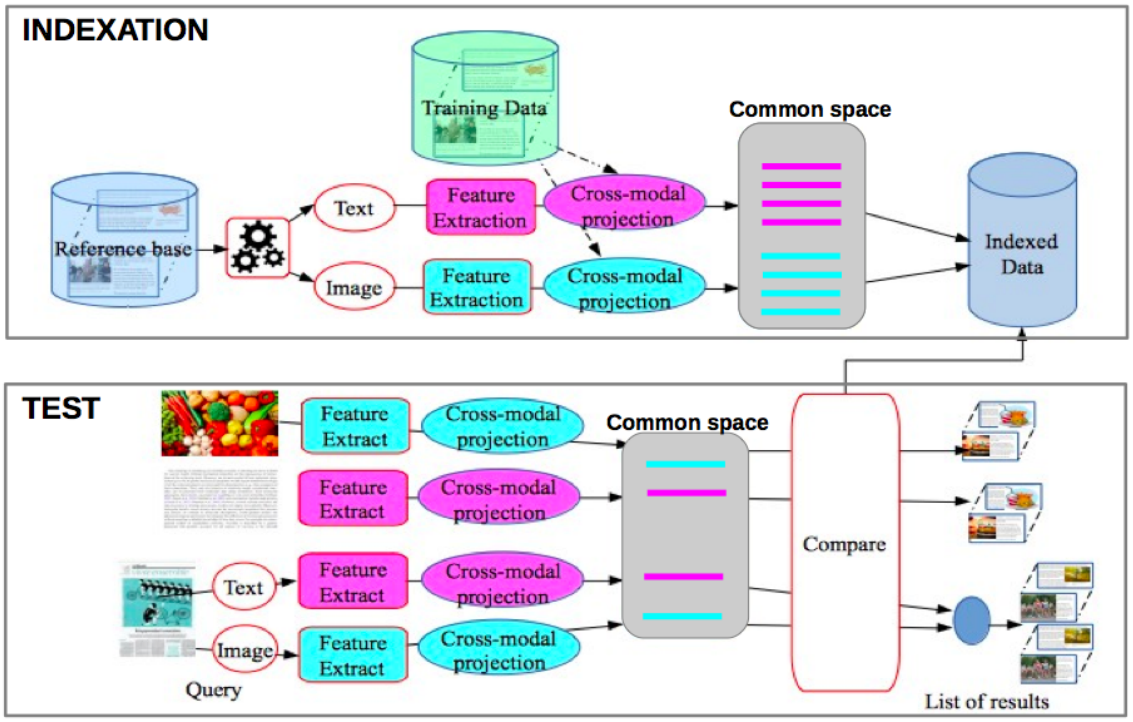

Figure 2.9: Cross-modal retrieval problem.
Queries and retrieved items in reference base belong to different modalities.

have [Wang et al. 2016a].

Another important requirement for cross-modal model is to reduce the *heterogeneity gap* across modalities. This gap refers to the fact that modalities are usually described according totally different scheme and that they usually do not even lie in the same vectorial space. In fact, images live in a continuous feature space whereas text (*e.g.* tags, sentences, keywords) are discrete.

To reduces these gaps existing across different modalities, recent studies have concentrated on learning a common space jointly for image and text on which data from these modalities have unique representations. One popular solution is to learn view-specific projection directions using paired samples from different views (modalities) to project samples from different views into a common latent space. In this procedure, feature extraction for multi-modal data is considered as the first step to represent various modalities of data. Based on these representations of multi-modal data, cross-modal correlation modeling is performed to learn common representations for various modalities of data. On the joint

representation space, the learned features can be directly measured between modalities and preserve the correlation across modalities. Therefore, the common representations enable the cross-modal problems such as cross-modal retrieval, cross-modal classification, etc. The details of such approaches are presented in the following Section 2.4

2.4 Common representation for text and image

Recently, various *cross-modal* applications such as text illustration, image captioning, or visual question answering have attracted considerable research attention. A core problem of these applications is how to measure the semantic similarity between visual data, *e.g.* image or regions, and text data, *e.g.* a sentence or tags. The most popular solution aims to learn a joint embedding space where text and image modalities can be both represented and directly compared. More precisely, these approaches learn view-specific projections using paired samples from different modalities *e.g.* text and image, to project samples from these views into a common latent space. Feature extraction for multi-modal data is considered as the first step to represent various modalities of data. Based on these representations of multi-modal data, cross-modal relation modeling is performed to learn common representations for various modalities. The embedding space is usually of low dimension and is very suitable for cross-modal tasks.

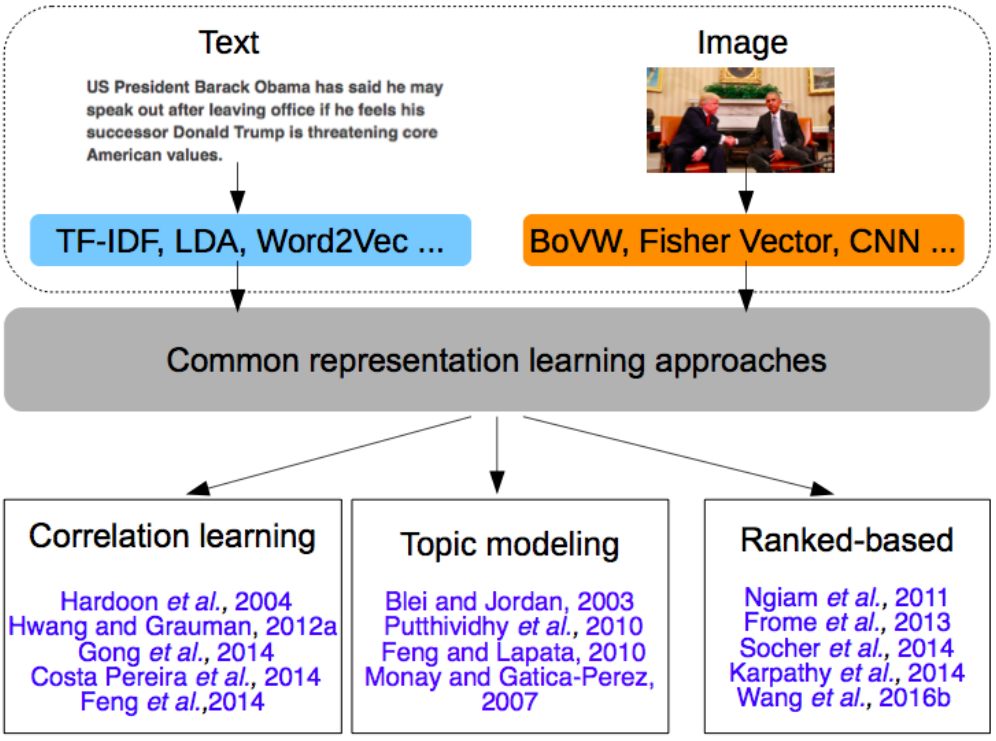

Figure 2.10: Different categories of common representation learning approaches for text and image.

According to the fundamental principle of the resulting common space, we categorize common representation learning approaches into three groups: *correlation learning* approaches, *topic modeling* approaches, and *rank-based* approaches. In what follows, we describe significant work for each category. Our particular interest lies on the correlation learning approaches for visual and textual content.

2.4.1 Correlation learning approaches

Subspace learning consists of the most popular methods for cross-modal problem. Such approaches aim to learn a common subspace shared by different modalities of data on which the similarity between these modalities can be directly measured. Subspace learning methods enforce pair-wise closeness between different modalities in the common subspace.

Canonical correlation analysis (CCA) is one of the most popular unsupervised subspace learning method for establishing inter-modal relationships between data from various modalities. CCA learns common representation subspace between two sets of data (*e.g.* data from two modalities) where the correlation between these two sets is maximized. CCA and its extensions, *e.g.* Kernel CCA (KCCA), have been widely used for cross-modal retrieval between image and text [Hardoon et al. 2004; Hwang and Grauman 2012a; Gong et al. 2014; Costa Pereira et al. 2014], cross-lingual retrieval [Udupa and Khapra 2010] and other vision problems such as automatic face recognition [Li et al. 2011].

A review on the principle of CCA and its kernelized version KCCA for common representation space learning is further described in 3.2. In this section, we review significant work that employ (K)CCA to build a joint space to address the cross-modal problem between image and text. These approaches can be either unsupervised [Hardoon et al. 2004; Hwang and Grauman 2012a; Hodosh et al. 2013; Yao et al. 2015] or supervised using the label information of data [Gong et al. 2014; Costa Pereira et al. 2014; Rasiwasia et al. 2014; Sharma et al. 2012].

CCA and its kernelised version KCCA were first introduced in cross modal retrieval problem in the seminal work of [Hardoon et al. 2004]. The principle is to compute a common latent space from both visual and textual features such that the correlation between the projections of both modalities for a given bi-modal dataset is maximized. All

the documents of a reference database are then projected onto the common latent subspace. When a query is processed, it is also projected onto this space and its nearest neighbors can be found directly, independently of their original modality (visual or textual), according to their similarity in the latent space.

A refinement was proposed by [Hwang and Grauman 2012a] to take into account the objects present in a scene with their relative significance within that scene. This is modeled by the rank of the tags used by an "ordinary user" to describe the scene. Then, KCCA is employed with an average kernel over three features to describe the visual aspect and three other features for the textual aspect, including the relative and absolute tag rank.

Hodosh et al. [2013] evaluated the performance of the (K)CCA on the much more stringent task of image description. The task consists in associating images with sentences, that describe what is depicted in them, from a large predefined pool of image descriptions. The authors proposed to map both images and sentences into the same space learned by (K)CCA and then frame the image description as the task of ranking the given pool of captions on the common space. The work compares a number of text kernels that capture different linguistic features as the input for learning the common space between image and text. For example, their final models extend beyond the standard basic bag-of-words representation of the captions by utilizing subsequence kernels and kernels that capture semantic similarity to increase the quality of the induced space. Experimental results demonstrate the importance of robust textual representations that consider the semantic similarity of words, and hence take the linguistic diversity of the different captions associated with each image into account. They also test the model with a number of relatively simple image description systems.

Recently, Yao et al. [2015] proposed a ranking canonical correlation analysis (RCCA) for learning query (text or image) and image similarities. The goal of RCCA is to improve the performance of a real image search engine by taking into account click-through data which is served as a reliable and implicit feedback for understanding both the query and the user's intent for image search. RCCA initially finds a CCA common subspace between queries and corresponding images clicked by user from a real image search engine. Furthermore, RCCA simultaneously learns a bi-linear query-image similarity function and adjusts the

subspace to preserve the preference relations implicit in the click-through data. Once the subspace is finalized, query-image similarity can be computed by the bi-linear similarity function on their mapping in this subspace. RCCA has been shown to be powerful for image search with superior performance over several state-of-the-art methods on both text-to-image and image-to-image retrieval tasks.

While (K)CCA has been popular for its simplicity and efficiency, it has several drawbacks. First and foremost is the inability of the classic (K)CCA to account for additional high-level semantic information such as the class label (concept) of the data. Recently, several work have successfully addressed this shortcoming by proposing alternatives and extensions of CCA to account for label information [Gong et al. 2014; Rasiwasia et al. 2014; Sharma et al. 2012; Ranjan et al. 2015]. These approaches aim to learn a more discriminative subspace which is better suited for cross-modal problem. These attempts have been made to enforce different-class samples to be mapped far apart while the same-class samples lie as close as possible on the learning subspace.

Sharma et al. [2012] showed that the classical CCA method only cares about pair-wise closeness in the common subspace so they are not well suited for classification or retrieval. Especially, when within-class variance is large, these methods are bound to perform poorly for classification/retrieval because classification and retrieval both require that within-class samples are united. The authors proposed instead a supervised extension of CCA, called Generalized Multiview Analysis (GMA) to address cross-modal problems. GMA formulates the problem of finding correlated subspaces as that of jointly optimizing covariance between sets and separating the classes in the respective feature spaces. The proposed approach is general and has the potential to replace CCA whenever cross-modal problem is the purpose and label information is available. In particular, this work investigated cross-modal classification by using a k-NN classification scheme for pose-invariant face recognition. It consists in classifying a sample by a majority vote of its neighbors, with the case being assigned to the class most common among its k nearest neighbors measured by a distance function. In this work, the parameter k is set to 1 (1-NN) which means simply to assign the sample to the class of its nearest neighbor on the latent space using the normalized correlation score as a metric.

In Semantic Correlation Matching (SCM) [Costa Pereira et al. 2014], the image and text features projected onto a KCCA space are used to build semantic features (Figure 2.11). It means that a document is represented as a set of supervised classifiers learned from projections on the common latent space. It is worth to note that these classifiers are only employed to represent a uni-modal document and the contribution addresses cross-modal retrieval alone. The authors demonstrate that the SCM model satisfies two hypotheses of a joint representation space: correlation and abstraction. The correlation hypothesis is that explicit modeling of low-level correlations between the different modalities is important for the success of the joint models. The abstraction hypothesis is that model benefits from semantic abstraction, *i.e.* the representation of images and text in terms of semantic (rather than low level) descriptors.

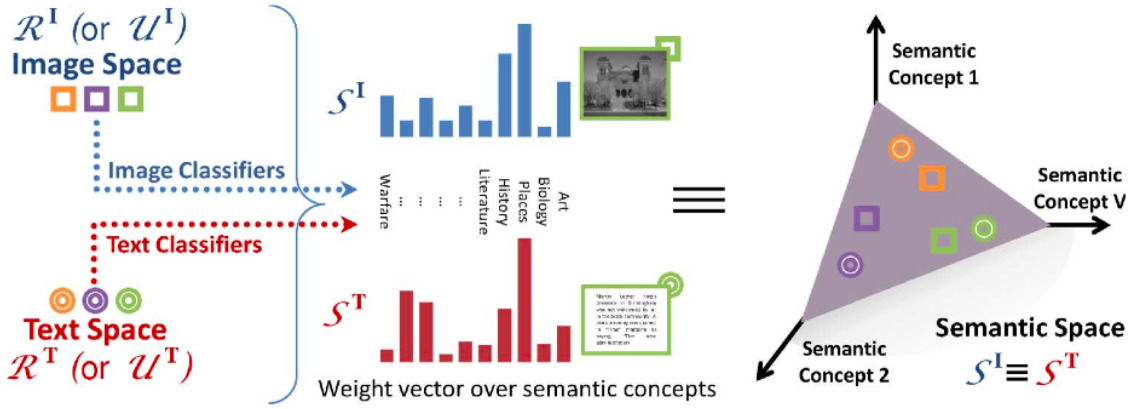

Figure 2.11: Semantic Correlation Matching model [Costa Pereira et al. 2014]

Another extension of CCA taking account label information of data is cluster-CCA proposed by Rasiwasia et al. [2014]. The proposed cluster-CCA learns discriminant low-dimensional subspaces that maximize the correlation between two modalities while segrageting the different classes on the learnt subspaces. Cluster-CCA introduces correspondence between each sample from any class in the first modality to all the same class sample in the second modality. Once this correspondence is established, the standard CCA is used to learn the projections onto the common subspaces.

An important extension to the standard CCA method of [Hardoon et al. 2004] was put forward in [Gong et al. 2014], where a third view (modality) was added to the traditional

two-view algorithm. Above the visual and textual view, semantic classes are also considered. In a *supervised* scenario, they are derived from the ground-truth annotations (labels), the search keywords used to download the images. However, in most realistic situations, the ground-truth annotations and keywords for the third view are very noisy or are absent completely. In such an *unsupervised* scenario, the authors demonstrate that the third view can be derived by a set of clusters obtained from the tags. KCCA is reformulated as a linear CCA applied to the kernel space, following the idea of approximate kernel maps [Perronnin et al. 2010a].

However, the strategies proposed in [Gong et al. 2014; Rasiwasia et al. 2014; Sharma et al. 2012] assume that the data is annotated with a single label. Ranjan et al. [2015] introduced an approach, called multi-label canonical correlation analysis (ml-CCA), that accounts for multi-label images. Unlike the standard CCA or the multi-view CCA [Gong et al. 2014] that require correspondence information across the modalities, ml-CCA does not rely on explicit pairings between modalities. Instead, it uses the multi-label information to establish correspondences. They also present fast-ml-CCA, a computationally efficient version of ml-CCA which is able to handle large scale dataset.

Feng et al. [2014] addresses cross-modal retrieval by training a "correspondence" auto-encoder, called Corr-AE, between visual and textual features. Most of cross-modal strategies such as [Costa Pereira et al. 2014; Rasiwasia and Vasconcelos 2009; Gong et al. 2014] involves a two-stage framework: feature extraction (or feature learning) and common representation learning. While such approaches separate correlation learning from representation learning, the proposed Corr-AE is more effective by incorporating these two stages into a single process (see Figure 2.12). For this, the authors introduced a loss function including the reconstruction losses of different auto-encoders for all modalities and the correlation loss between different modalities. Corr-AE is furthermore extended to two correspondence models: Corr-Cross-AE by replacing the basic auto-encoder by cross-modal auto-encoder and Corr-Full-AE using a combination of a basic auto-encoder and cross-modal auto-encoder.

Inspired by representation learning using deep networks, Andrew et al. [2013] presented Deep Canonical Correlation Analysis (DCCA), a nonlinear extension of the linear CCA. It

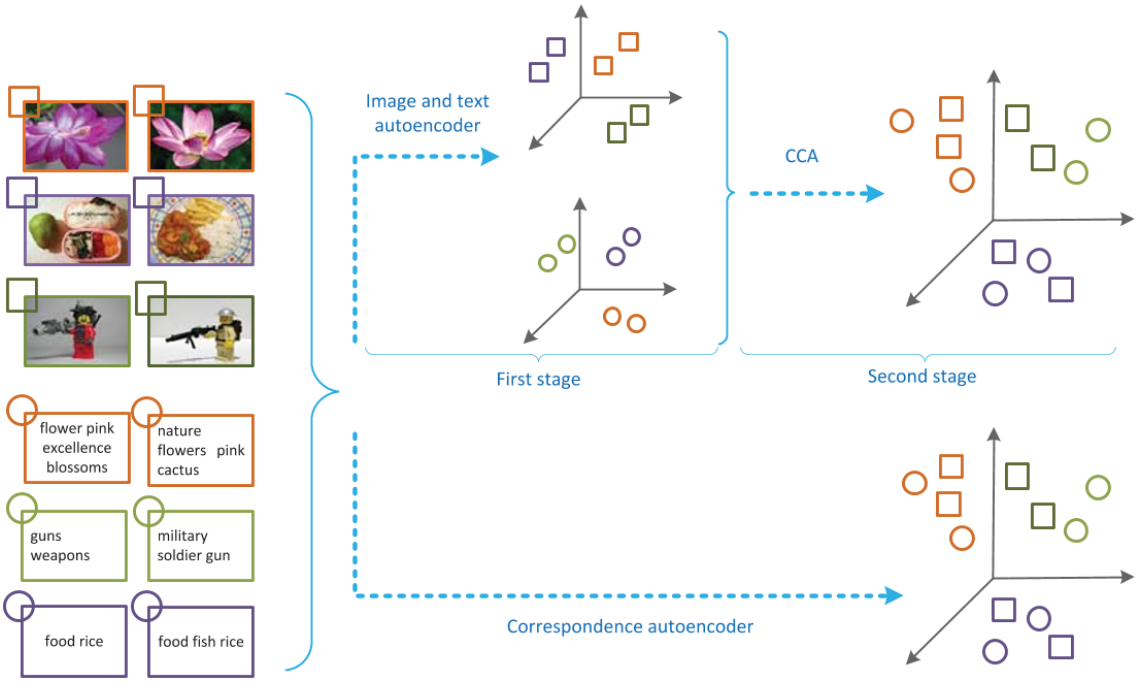

Figure 2.12: Difference between two-stage methods and Corr-AE: Corr-AE incorporates representation learning and correlation learning into a single process while two-stage methods separate the two processes [Feng et al. 2014]

is an alternative to the non-parametric KCCA for learning correlated non-linear transformations. The approach is closely related to [Ngiam et al. 2011; Srivastava and Salakhutdinov 2012]. The key difference is that DCCA learns two separate deep networks, with the objective that the output layers (topmost layer of each network) are maximally correlated. Experimental results show that DCCA learns representations with significantly higher correlation than those learned by CCA and KCCA. However, the high dimensionality of the input features introduces a great challenge in terms of memory and speed complexity of DCCA framework. To address this issue, Yan and Mikolajczyk [2015] presented an alternative end-to-end learning scheme to make DCCA applicable to high dimensional image and text representations and large datasets by resolving non-trivial complexity and overfitting issues. They proposed a GPU implementation with CUDA libraries, which the efficiency is several orders of magnitude higher than CPU implementations. The proposed approach is successfully employed for image-caption matching task.

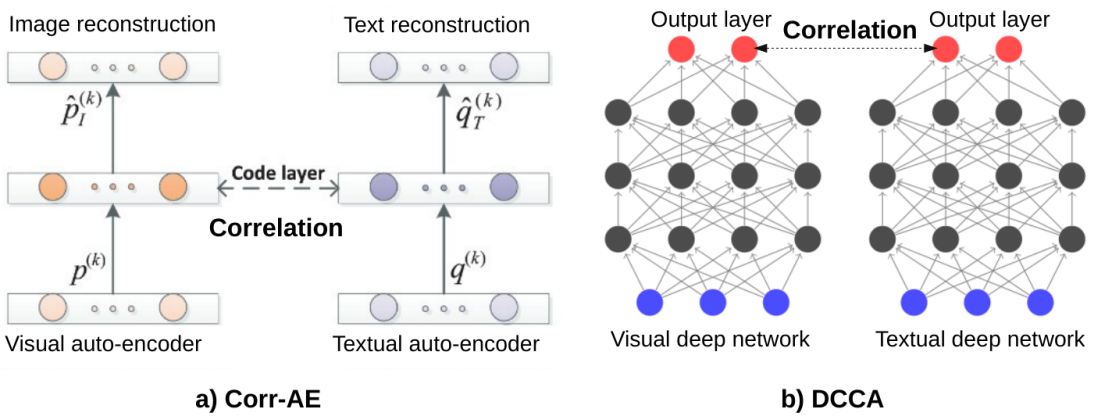

Figure 2.13: Difference between Corr-AE [Feng et al. 2014] and DCCA [Andrew et al. 2013]. Corr-AE correlates hidden representations of a pair of auto-encoder while DCCA correlates output layers of two deep networks describing visual and textual modalities.

2.4.2 Topic modeling approaches

This category consists of probabilistic approaches in which topic modeling is a well-known method. The topic modeling approaches, also called topic-based approaches, has been widely applied to solve different cross-modal problems such as image annotation or text illustration [Blei and Jordan 2003; Putthividhy et al. 2010; Feng and Lapata 2010].

Topic modeling was originally investigated and widely used in the literature of natural language processing [Blei et al. 2003; Hofmann 1999]. Topic modeling approaches assume that each data point result from multiple hidden "topics". The key idea of such approaches is to map high-dimensional representation space, *e.g.* issued from term frequency vectors arising in the vector space representation of text documents [Salton and McGill 1986], to a lower-dimensional semantic representation space defined by the hidden topics. In comparison with classical representations like bag of words, representations on this latent semantic space are more robust to the problems of *polysemy* that a single word may represent different content and *synonym* that different words may represent the same content. Moreover, the resulting lower-dimensional latent space speeds up learning computation (*e.g.* SVM) on these topic-based representations.

Latent Dirichlet Allocation (LDA) [Blei et al. 2003] and probabilistic Latent Semantic Analysis (pLSA) [Hofmann 1999] are the popular techniques in this direction. LDA and

pLSA both consider a document as a mixture of various "hidden" topics which are detected from data. The difference is that LDA assumes topic distribution having a Dirichlet prior.

Recently, LDA and pLSA techniques have been extended to learn the joint distribution of multi-modal data to capture the correlation between images and text [Blei and Jordan 2003; Putthividhy et al. 2010; Feng and Lapata 2010; Monay and Gatica-Perez 2007].

Corr-LDA [Blei and Jordan 2003] uses a set of shared latent variables to represent the underlying causes of cross-correlations in multi-modal data. In this model, the visual modality drives the definition of the latent space to which the textual modality is linked. Concretely, the model first generates a set of hidden variables (topics) that generate the regions of an image, decomposing an image into a mixture of latent variables. A subset of these latent topics is then selected to generate the text caption, which intuitively corresponds to the natural process of image annotation. Consequently, Corr-LDA can model the joint distribution of an image and its caption, the conditional distribution of words given an image or an particular region of an image. It has been successfully employed in different multimedia tasks such as automatic image annotation, automatic image region annotation and text-based image retrieval.

While Corr-LDA shares a set of latent topics between the textual and visual modalities, Tr-mm-LDA [Putthividhy et al. 2010] learns two separate sets of hidden topics, respectively for image and text. The number of topics in the two modalities can be different. A regression module is then introduced to correlate these two sets. As a result, one set of topics can be linearly predicted from the other. Tr-mm-LDA has shown its power in the task of image and video annotation. This model outperformed the Corr-LDA approach.

Monay and Gatica-Perez [2007] proposed a new dependence between words and image regions based on pLSA method [Hofmann 1999]. This approach learns a pLSA model from a set of concatenated representations of the textual and visual modalities of annotated images. Using such concatenated representations, this approach attempts to simultaneously model visual and textual modalities. Furthermore, in this model, visual and textual features can have equal importance or one of the two modalities can dominate in defining the latent space. Contrary to the Corr-LDA model in which textual features are linked in the latent space learned from visual features, the authors demonstrated that the textual modality is

more appropriate to learn a semantically meaningful latent space, which translates into improved image retrieval and annotation performance.

In the same direction, Feng and Lapata [2010] proposed a probabilistic image annotation model, called Mix-LDA that learns to automatically label images despite the noisy nature of data. Mix-LDA model exploits the redundancy inherent in multi-modal documents (*e.g.* news articles, Wikipedia articles) by assuming that images and their surrounding text are generated by a shared set of latent topics. Concretely, Feng and Lapata [2010] described documents and images by a *common* multi-modal vocabulary consisting of textual and visual words. Then, using LDA, the model represents visual and textual meaning *jointly* as a probability distribution over a set of topics. The Mix-LDA model brought improvements over competitive models such as the previously presented Corr-LDA, pLSA in image annotation and text illustration tasks. The Mix-LDA has been seen as the first multimodal distributional semantic model. The idea of this approach was recaptured in a more general manner in a recent work proposed by Bruni et al. [2014].

Furthermore, Jia et al. [2011] proposed a new probabilistic representation for image and text, called Multi-modal Document Random Field (MDRF). While Corr-LDA requires a full correspondence between modalities and Tr-mm-LDA assumes that an image is associated with a text description, MDRF tackles more realistic scenarios where a narrative text is only loosely related to an image and where only a few image-text pairs are available. MDRF learns a set of shared topics across the modalities. The model defines a Markov random field on the document level which allows modeling more flexible document similarities. The effectiveness of MDRF was evaluated in image retrieval from a loosely related text.

Rasiwasia and Vasconcelos [2008] also introduced an intermediate space based on a low dimensional semantic "topics" image representation. The overall proposed representation is similar to a topic model, but where topics are explicitly defined instead of being learnt in an unsupervised manner (*e.g.* using LDA or pLSA) from the features representations. An image is annotated with a subset of the topics that it actually contains. The number of semantic topics used defines the dimensionality of the intermediate space, henceforth referred to as "semantic space". Each topic induces a probability density on the space of low-level features, and the image is represented as the vector of posterior topics probabilities.

The scene classification results show that the proposed low-dimensional representation correlates well with human scene understanding, outperforms the unsupervised latent-space approaches and achieves performance close to the state-of-the-art method which uses a much higher dimensional image representation.

However, this representation on the semantic space suffers from a certain amount of contextual noise, due to the inherent ambiguity of classifying image patches. Hence, Rasiwasia and Vasconcelos [2009] introduced a second level of representation that operates in the semantic space. The proposed model enables robust inference in the presence of this noise by modeling the distribution of each concept in the semantic space. This distribution is referred to as the contextual model for the concept. Image are the represented by their posterior probabilities with respect to a set of contextual models. Evaluations on scene classification and image annotation showed that besides quite simple to compute, the proposed context models attained superior performance than state-of-the-art systems in both tasks.

Recently, Wang et al. [2014] proposed a multi-modal mutual topic reinforce modeling (M^3R) approach, which seeks to learn correlated but discriminative latent representations for multi-modal data by introducing of topic interaction and label information. M^3R learns separately modality-specific topics from multi-modal documents, *e.g.* textual-specific topics and image-visual topics and then detects sharing cross-modal topics by multi-modal reinforcement modeling. The cross-modal topics means the topics are simultaneously remarked by images and texts within the same multi-modal documents. The proposed M^3R encourages these mutually consistent cross-modal topics with a relatively high priority, while discourages but still preserves the remaining modality-specific topics. M^3R gains interpretable latent representations for multi-modal retrieval and is effective for cross-modal retrieval.

2.4.3 Rank-based approaches

As an alternative of correlation learning method, this category approaches the cross-modal problems by learning a joint representation space with a *ranking loss*. Depending on the objective defined in the ranking function, such approaches can be regrouped in

two categorized: *single-directional* and *bi-directional*. The *single-directional* models aim to ensure that correct texts for each training image get ranked above incorrect ones. In addition to this criteria, the *bi-directional* approaches also ensures that for each text, the image described by that text gets ranked above ones described by other texts. Most of the ranked-based approaches involve deep learning architectures. These models have been successfully used in a wide range of practical applications such as image annotation, image captioning, visual question answering, etc.

Single-directional approaches. The first approaches of ranked-based aim to learn linear transformations of visual and textual features to a joint representation space using *single-directional* ranking loss [Weston et al. 2011; Frome et al. 2013]. These approaches apply a margin-based penalty to incorrect annotations that get ranked higher than correct ones for each training image.

Weston et al. [2011] introduced a model, called W_{SABIE}, for image annotation. The approach attempts to represent images and annotations jointly in a low-dimensional embedding space and to optimize precision at the top of the ranked list of annotations for a given image. The model is trained with Weighted Approximate-Rank Pairwise (WARP) loss function. This was the first data analysis of image annotation that reported results on a larger scale than ever previously reported (10 million training examples and 100 thousand annotations).

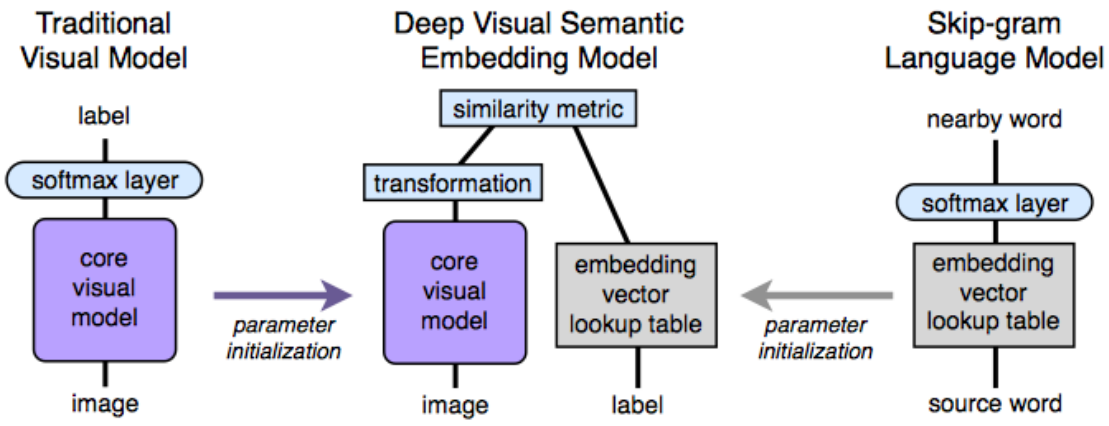

Figure 2.14: DeViSe model (center) initialized with parameters pre-trained at the lower layers of a visual object categorization network with a softmax output layer (left) and a skip-gram language model (right) [Frome et al. 2013]

While W_{SABIE} is a shallow embedding method, DeViSe model [Frome et al. 2013] is inspired by the progress of deep learning to learn embedding space of images and words. The objective of DeViSe is to leverage semantic knowledge learned in the text domain, and transfer it to a model trained for visual object recognition. The authors pre-train a simple neural language model well-suited for learning semantically-meaningful, dense vector representations of words [Mikolov et al. 2013]. In parallel, they pre-train a state-of-the-art deep neural network for visual object recognition [Krizhevsky et al. 2012], complete with a traditional softmax output layer. The DeViSe model is constructed by taking the lower layers of the pre-trained visual object recognition network and re-training them to predict the vector representation of the image label as learnt by the language model (Figure 2.14). The model is applied to visual object classifiers and especially it leverages visual and semantic similarity to correctly predict object category labels for unseen categories, *i.e.* "zero-shot" classification.

Bi-directional approaches. As a more powerful objective function, other recent work have proposed a *bi-directional* ranking loss [Karpathy et al. 2014; Karpathy and Fei-Fei 2015; Socher et al. 2014; Chen and Zitnick 2015; Wang et al. 2016b]. Karpathy et al. [2014] proposed a deep neural bi-directional network that embeds fragments of images (objects) and fragments of sentences (typed dependency tree relations) into a common space and explicitly reasons about their latent, inter-modal correspondences (Figure 2.15). The authors then formulated a structured max-margin objective allowing to explicitly associate these fragments across modalities.

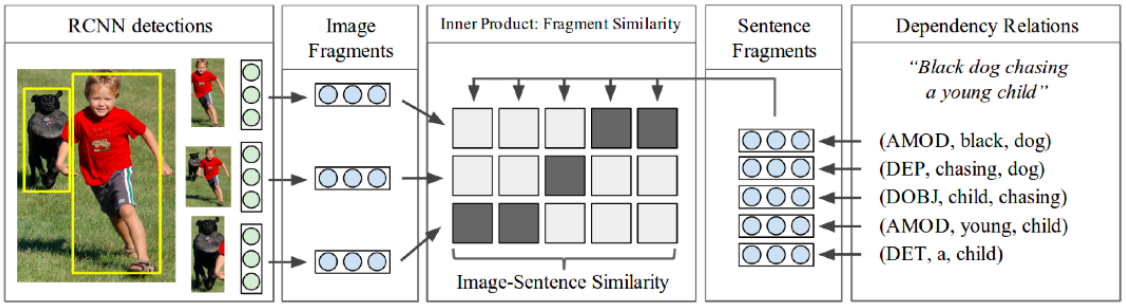

Figure 2.15: Illustration of the Deep Fragment Embeddings for Bidirectional Image Sentence Mapping model [Karpathy et al. 2014]

Adopting a similar approach, Socher et al. [2014] described a Dependency Tree Recursive Neural Network (DT-RNN) which maps sentences and images into a common embedding space in order to be able to retrieve one from the other. The DT-RNN learns vector representations for text (sentences, phrases) based on dependency trees. These vectors capture more of the meaning of sentences, where they define meaning in terms of similarity to a "visual representation" of the textual description.

More recently, Karpathy and Fei-Fei [2015] presented a model to generate natural language descriptions of images and their regions. The alignment model is based on a combination of Convolutional Neural Networks over image regions, bidirectional Recurrent Neural Networks over sentences, and a structured objective that aligns the two modalities through a multi-modal embedding. Region-word pairwise similarities are computed with inner products and used for defining the corresponding image-sentence score.

Chen and Zitnick [2015] uses a bi-directional mapping from visual features to words and words to visual features in a recurrent neural network model to generate descriptions based on visual feature, and to reconstruct visual descriptions given a description. The global objective is to maximize the likelihood of a word and the observed visual features given the previous words and their visual interpretation. The approach is evaluated on sentence generation, sentence retrieval and image retrieval.

Recently, Wang et al. [2016b] proposed to learn an image-text embedding using a two-view neural network with two layers of nonlinearities on top of any representations of the image and text views. These representations can be given by the outputs of two pre-trained networks, off-the-shelf feature extractors, or trained jointly end-to-end with the embedding. The authors use a bi-directional ranking loss function similar to [Karpathy et al. 2014; Karpathy and Fei-Fei 2015] together with within-view neighborhood structure preservation constraints for learning the model. Specifically, in the learned latent space, images (resp. sentences) with similar meaning are expected to be close to each other. This work demonstrates that these constraints provide a useful regularization term for the cross-modal matching task.

2.5 Multi-modal datasets

In this section, we describe the multi-modal datasets that we used to evaluate the effectiveness and the robustness of our proposed approaches for different multimedia tasks such as cross-modal retrieval, bi-modal classification or cross-modal classification. Since the core research of this thesis focuses on the joint embedding of visual and textual content to improve the results of such multimedia tasks, we mainly consider the multi-modal datasets in which images and text are both available.

The datasets we investigate in this thesis contain real-world social images with their associated text. The textual content can be human annotated tags, image captions, and/or a textual document related to the content of image. A global statistic on datasets that we considered in this dissertation is given in Table 2.1.

Dataset	Sample	#training/testing	#labels
BBC News	image-caption-document	3,121 / 240	*n/a*
Wikipedia	image-caption-document	106,582 / 240	*n/a*
Pascal VOC07	image-description	5,011 / 4,952	20
NUS WIDE	image-description	161,789 / 107,859	80
FlickR 8K	image-5 descriptions	6,000 / 1000	*n/a*
FlickR 30K	image-5 descriptions	29,783 / 1000	*n/a*

Table 2.1: Summarization of multi-modal datasets considered in the thesis

In the following, we briefly describe each of these multi-modal datasets.

BBC News dataset. This dataset was introduced in Feng and Lapata [2010] for image annotation and text illustration. It consists of 3121 articles for training and 240 for testing, downloaded from the BBC News website[2]. The articles cover a wide range of topics including national and international politics, advanced technology, sports, education, etc.

Each BBC News article is accompanied with an image and associated caption. The dataset thus consists of *image-caption-document* tuples. An example of an entry in our database is illustrated in Figure 2.16. The average caption length is 5.35 words and the average document length is 133.85 words.

[2]http://news.bbc.co.uk/

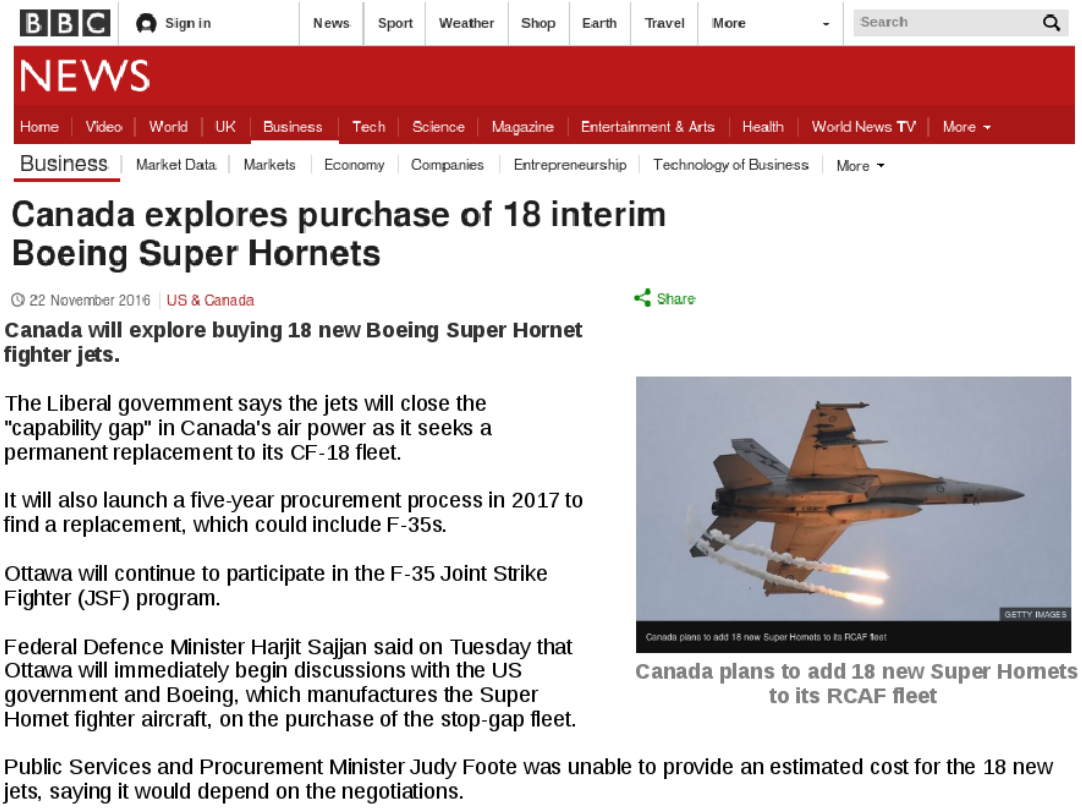

Figure 2.16: A sample from our BBC News database. Each entry contains an image, a caption for the image, and the accompanying document with its title.

This dataset is especially challenging for text illustration because of its small size and the quite indirect relation between textual and visual content in most news articles. For many articles, other images in the collection can be objectively considered more relevant to the article's content than the image selected by the author.

Wikipedia dataset. This dataset consists of 106,822 articles including 106,582 samples for training and 240 for testing. Theses samples are english articles that we collected from the ImageCLEF 2010 Wikipedia collection[3]. Each article is accompanied by an image and associated caption. Each image is provided with its metadata in a XML files. The metadata is often an image description or image caption in different languages such as English, French, German. We are only interested in images which have corresponding

[3]http://imageclef.org/2010/wiki

```
<?xml version="1.0" encoding="UTF-8" ?>
<image id="997" file="images/1/997.png">
 <name>Coat of Arms of Asha (Chelyabinsk oblast).png</name>
 <text xml:lang="en">
  <description>Asha (Chelyabinsk oblast), coat of arms</description>
  <comment>Asha (Chelyabinsk oblast), coat of arms</comment>
  <caption article="text/en/4/339180">Coat of arms of Asha</caption>
 </text>
 <text xml:lang="de">
  <description/>
  <comment />
  <caption />
 </text>
 <text xml:lang="fr">
  <description />
  <comment />
  <caption article="text/fr/4/536916"/>
 </text>
 <comment>({{en|Asha (Chelyabinsk oblast), coat of arms}}  {{ru|Аша (Челябинская
 область), герб}}  [http://vector-images.com/image.php?epsid=4151 Source]  {{Vector-
 Images.com}}  [[Category:Coats of arms of cities and villages of Chelyabinsk
 )</comment>
 <license>Public Domain</license>
</image>
```

Figure 2.17: A sample from our Wikipedia database. Image in Wikipedia article is given with corresponding metadata.

Wikipedia articles and image captions in English. An example that illustrates images in the Wikipedia collection and their metadata is provided in Figure 2.17

Pascal VOC 07 dataset. This dataset was introduced in the Pascal VOC 07 challenge [Everingham et al. 2010]. It includes $5,011$ training and $4,952$ testing images collected from Flickr without their original user tags. Each image has between 1 and 6 labels from a set of 20 labels. These labels describe vehicles (car, bus, bicycle...), animals

Images				
Tags	buildings, windows, cars, tree, streetlight, road	window, chair, table, image	cloth, cat	bus, banner, road, car
Labels	car	chair, diningtable, pottedplant	cat	car, bus, person

Figure 2.18: Example Pascal VOC 07 images with their associated tags collected using AMT and their corresponding ground-truth labels.

Images				
Tags	monochrome, car, vintage, mercedes, benz, cilest, kurt, vehicle	dog, husky, wolf, perro, lobo, roja, caperucita, siberiano, vorfas	blue, summer, sky, umbrella, holidays, ysplix, flickelite, colorartaward, platinumheartaward	ocean, sunset, sea, water, animals, whale, whales, orca, whalewatching, whaletail, orcawhale, whalephotos
Labels	vehicle	dog	sky	ocean, sunset, water, whales

Figure 2.19: Example Nus-WIDE images with their associated tags and their corresponding ground-truth labels. Several tags are very noisy. For instance, "colorartaward" and "platinumheartaward" are respectively for "color art award" and "platinum heart award".

(cat, dog, horse...), household (sofa, tv/monitor, chair ...) and persons.

In the PASCAL VOC challenge, the original dataset is not multi-modal and only images are available. Using Amazon Mechanical Turk, several tags were also made available for each image by the work of Hwang and Grauman [2012b]. These tags are provided on the authors' web page[4]. Each image is associated to 1 to 75 tags for training (6.9 on average) and between 1 and 18 tags for testing (3.7 on average). Several examples of the Pascal VOC 07 are given in Figure 2.18

Nus-WIDE dataset. This dataset was introduced in the work of Chua et al. [2009]. The NUS-WIDE dataset includes $161,789$ training and $107,859$ testing Flickr images with both user tags and "ground truth" labels according to 81 concepts. The 81 concepts are divided into six categories including people, objects, scene or location, event or activities, program and graphics. Example images with their associated user tags and labels are shown in 2.19

Moreover, the authors proposed several rules to filter the original tags and create smaller and specialized sub-datasets (Objects, Scenes,...). They deleted tags with too

[4]http://vision.cs.utexas.edu/sungju/pascal_twkim.zip

Images		
Descriptions	A black dog is running after a white dog in the snow Black dog chasing brown dog through snow Two dogs chase each other across the snowy ground Two dogs play together in the snow Two dogs running through a low lying body of water	A boy smiles in front of a stony wall in a city . A little boy is standing on the street while a man in overalls is working on a stone wall . A young boy runs across the street . A young child is walking on a stone paved street with a metal pole and a man behind him . Smiling boy in white shirt and blue jeans in front of rock wall with man in overalls behind him .

Figure 2.20: Example FlickR 8K images with their five associated descriptions.

low frequency (number of occurrence in the dataset is less than a threshold). The low frequency threshold is set to 100. They also removed tags that do not exist in WordNet which is a large lexical database of English. The final tag list has 5,018 unique tags. The original tags are preserved and also available for users. Meanwhile, we observe that some tags remain nevertheless noisy after the tags filter process proposed by the authors. For instance, some tags are concatenated and result into a unique (non existing) word (*e.g. sunsetoverthesea*). This fact infers a shortcoming in textual feature extraction. For example, the term "*sunsetoverthesea*" is naturally absent from the Word2Vec vocabulary; hence the Word2Vec model will not able to correctly represent it. To improve the quality of textual features, we automatically separate the words (producing *e.g. sunset over the sea*) before employing techniques of textual features extraction. For this, each tag is matched to the previously mentioned tag dictionary of 5,018 terms and we retain only the valid largest sub-strings. The exact proposed process is described in Appendix A.1 (Python code).

FlickR 8K and FlickR 30K datasets. The FlickR 8K [Rashtchian et al. 2010] and FlickR 30K [Young et al. 2014] datasets contain 8000 and 31783 images respectively. Each image was annotated by 5 sentences using Amazon Mechanical Turk. These datasets have the same 1000 images for validation and 1000 images for testing. While the training set of FlickR 8K contains 6000 images, the one of FlickR 30K is much larger containing 29783 images.

2.6 Conclusion

In this chapter, we presented a survey of the state-of-the-art approaches involved in retrieval and classification problems in the context of social media. During the chapter, we investigated two modalties of data including visual modality represented by images or their regions and the textual modality represented by tags, keywords, caption, or descriptions associated to images.

This section aims to provide the reader an overview of various multimedia information retrieval or classification problems such as uni-modal, multi-modal and particularly cross-modal one. An important stage in these paradigms consists in learning to represent content of data, covering single-media with the availability of only one modality (*i.e.* text or image) and multi-modal media with both two modalities of data (*i.e.* text and image). We thus provided a brief review on important techniques of media representation.

We are particularly interested in cross-modal problem to which learning a common representation space (embedding space) for visual and textual data is a relevant solution. We introduced in this chapter different categories of image and text embedding approaches and furthermore described how they have been employed in different cross-modal tasks such as image annotation, text illustration, image captioning and visual question answering.

Finally, we presented several multi-modal datasets that are commonly used in the state of the art and that we employed for testing our proposed models in the following chapters of this thesis.

Chapter 3

Common representation space and its limitations

3.1 Introduction

In this chapter, we first propose to go through in details of *(Kernel) Canonical Correlation Analysis*, the most popular method relying on correlation learning to find common representation for image and text. The common representation subspace enables the "matching" of information from one modality to another. In Chapter 2, the effectiveness of the common representation subspace has been clearly shown in several cross-modal and multi-modal problems between image and text. Nevertheless, in this chapter, we identify two limitations of such common space, especially one learned from the KCCA. We also demonstrate that these identified imperfections lead to an important decrease in performance of multimedia retrieval and classification problems.

The first limitation relates to ill-represented data on the common space. The development of the KCCA space relies on extracting statistical regularities from a large amount of training data. Any piece of data having very few occurrences or weak relation to other data is thus ignored in the joint model. However, the poorly represented information can be very significant in a retrieval context. Disregarding such information highly obstructs the effectiveness of the joint representation space.

The second limitation concerns the separation of data projections between different modalities on the common space. By observing the distribution of projections on the common representation space obtained with KCCA, we found that this space only provides a very coarse association between modalities. For any given multi-modal document, the projections of its visual and respectively textual features fall far apart. A direct use of these projections results in limited quality of "translation" between modalities.

The Chapter is organized as follows. Section 3.2 aims to review the principle of the (K)CCA method. We next demonstrate each of the identified limitations of the common representation space learnt by KCCA in Section 3.3 followed by their investigations on real cases of BBC News, Pascal VOC07 and FlickR 8K datasets.

3.2 KCCA: common representation space for image and text

While different categories of approaches have been introduced in Chapter 2 to approach the problem of image and text common representation learning, our main research particularly focuses on correlation learning method. The latter embeds different modalities (*e.g.* image and text) of data from their original feature spaces into a lower-dimensional common representation subspace by learning inter-modal relationships between data from these modalities.

In this thesis, we decided to rely on Canonical Correlation Analysis (CCA) for common representation learning. Indeed, three reasons motivated this choice of such a method.

- First, Canonical correlation analysis (CCA) is the most popular method situated among correlation learning approaches. The method has been introduced quite a long time ago by Hotelling [1936] and its theoretical foundations are relatively well formulated.

- Secondly, at the beginning of the thesis in 2013, several significant works in the multimedia community were revisiting and widely investigating with success the CCA and its kernelized extension KCCA to address various multi-modal and cross-modal problems of visual and textual modalities [Hardoon et al. 2004; Hwang and Grauman 2012a; Costa Pereira et al. 2014; Gong et al. 2014]. Particularly, the experiments concerning (K)CCA are relatively easy to reproduce.

- Finally, a common representation space allows to reduce *semantic* and *heterogeneity* gaps across visual and textual modalities and such common representation enables multi-modal and cross-modal tasks. However, the thesis is not particularly interested in how to build a common representation space for image and text. Our considerable attention aims to rely on a text-image joint space to develop robust representation for multi-modal and cross-modal tasks. In this context, basing our work on a method like (K)CCA, which maximizes the correlation between modalities, appeared to be a reasonable and sufficient solution. A discussion on the choice of the joint space

learning approach is further provided as an interesting perspective (Section 7.2).

This rest of this section aims to review the principle of CCA and its kernelized version method KCCA for common representation subspace learning in 3.2.1, the projection of data onto the KCCA subspace in 3.2.2 and finally the matching of data from different modalities onto the KCCA subspace in 3.2.3.

3.2.1 Common representation space learning with (K)CCA

CCA was first introduced by Hotelling [1936] and then applied to solve cross-modal problem in the seminar work of [Hardoon et al. 2004]. CCA is similar to the well-known Principal Components Analysis (PCA) in the sense that it attempts to map data from original feature spaces into a *lower-dimensional* (sub)spaces spanned by a set of canonical components. These components are issued from linear combinations of features on the original spaces. The difference between the two methods is that PCA detects the internal relationships among one set of variables and CCA detects the relationship between two different sets of variables.

Let X^T and X^I be two random variables, taking values in $\mathbb{R}^{d_T}$ and respectively $\mathbb{R}^{d_I}$. Consider N samples $\{(x_i^T, x_i^I)\}_{i=1}^N \subset \mathbb{R}^{d_T} \times \mathbb{R}^{d_I}$. For each pair of data (x_i^T, x_i^I), $x_i^T \in \mathbb{R}^{d_T}$ has a link with $x_i^I \in \mathbb{R}^{d_I}$ and we expect to conserve this relation onto the common space.

CCA learns the d-dimensional subspaces $\mathcal{U}^T \sqsubseteq \mathbb{R}^{d_T}$ for text and $\mathcal{U}^I \sqsubseteq \mathbb{R}^{d_I}$ for image where the correlation between two modalities is maximal (Figure 3.1). Concretely, CCA simultaneously seeks directions $w_T \in \mathbb{R}^{d_T}$ and $w_I \in \mathbb{R}^{d_I}$ that maximize the correlation between the projections of each x^T onto w_T with its corresponding x^I onto w_I,

$$w_T^*, w_I^* = \underset{w_T, w_I}{\arg\max} \frac{{w_T}' \, C_{TI} \, w_I}{\sqrt{{w_T}' \, C_{TT} \, w_T \, {w_I}' \, C_{II} \, w_I}} \tag{3.1}$$

where C_{TT}, C_{II} denote the auto-covariance matrices of X^T and X^I respectively, while C_{TI} is the cross-covariance matrix.

The solutions w_T^* and w_I^* are eigenvectors of $C_{TT}^{-1} C_{TI} C_{II}^{-1} C_{IT}$ and respectively $C_{II}^{-1} C_{IT} C_{TT}^{-1} C_{TI}$. The set of d eigenvectors associated to the d largest eigenvalues $\{w_{T,k}\}_{k=1}^d$ and $\{w_{I,k}\}_{k=1}^d$ define a basis of the maximally correlated d-dimensional subspaces $\mathcal{U}^T$ and

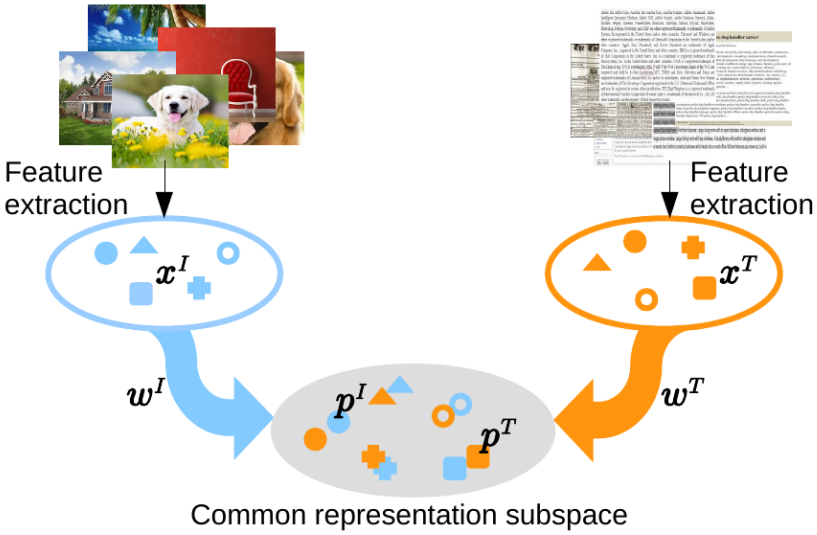

Figure 3.1: Canonical correlation analysis between image and text modalities

$\mathcal{U}^I$. Even though these are linear subspaces of two different spaces, they are often referred to as "common" representation space.

Kernel CCA (KCCA, see *e.g.* [Hardoon et al. 2004]) aims to remove the linearity constraint by using the "kernel trick" to first map the data from each initial space to the reproducing kernel Hilbert space (RKHS) associated to a selected kernel and then looking for correlated subspaces in these RKHS.

KCCA then seeks vectors of coefficients $\alpha_T, \alpha_I \in \mathbb{R}^N$ that allow to define these maximally correlated subspaces. α_T, α_I are solutions of

$$\alpha_T^*, \alpha_I^* = \underset{\alpha_T, \alpha_I}{\arg\max} \frac{\alpha_T^{'} K_T K_I \alpha_I}{V(\alpha_T, K_T)\, V(\alpha_I, K_I)} \tag{3.2}$$

where $V(\alpha, K) = \sqrt{\alpha^t (K^2 + \kappa K) \alpha}$, $\kappa \in [0, 1]$ is a regularization parameter and K_T, K_I denote the $N \times N$ kernel matrices obtained from $\{x_i^T\}_{i=1}^N$ and $\{x_i^I\}_{i=1}^N$. Finding the solutions amounts to solving a generalized eigenvalue problem and keeping the d highest eigenvalues together with their associated eigenvectors, $\{\alpha_{T,k}\}_{k=1}^d$ and $\{\alpha_{I,k}\}_{k=1}^d$.

3.2.2 Projections onto (K)CCA subspace

Image $x^I \in \mathbb{R}^{d_I}$ and text $x^T \in \mathbb{R}^{d_T}$ are represented by their projections p^I and p^T onto the subspaces $\mathcal{U}^I$ and $\mathcal{U}^T$ respectively.

In the case of CCA, the projection p^I (respectively p^T) is linear thus obtained by computing the *dot products* between the vector representing the image $x^I \in \mathbb{R}^{d_I}$ (or text $x^T \in \mathbb{R}^{d_T}$) and the image (or text) basis vectors, *e.g.* $\{w_{I,k}\}_{k=1}^d$ or $\{w_{T,k}\}_{k=1}^d$.

In the case of KCCA, the projections p^T of x^T and p^I of x^I onto their subspaces are obtained as:

$$p_k^T = [\mathcal{K}_T(x^T, x_1^T) \dots \mathcal{K}_{\mathcal{T}}(x^T, x_N^T)]\alpha_{T,k} \quad k \in \{1, .., d\} \tag{3.3}$$

and respectively:

$$p_k^I = [\mathcal{K}_I(x^I, x_1^I) \dots \mathcal{K}_{\mathcal{I}}(x^I, x_N^I)]\alpha_{I,k} \quad k \in \{1, .., d\} \tag{3.4}$$

3.2.3 Cross-modal matching on (K)CCA subspaces

The similarity between two data points $x^I \in \mathbb{R}^{d_I}$, $x^T \in \mathbb{R}^{d_T}$ from two different feature spaces can not be directly computed because of the difference in dimensionality, nature of data content, etc. The common representation subspaces allow establishing the similarity between these two points x^I, x^T through the proximity between their projections p^I, p^T.

A natural invertible mapping between the projections onto $\mathcal{U}^I$ and $\mathcal{U}^T$ follows from the correspondence between the d-dimensional bases of the subspaces. This results in a compact, efficient representation of both modalities, where vectors p^I and p^T are coordinates in two isomorphic *d-dimensional* subspaces $\mathcal{U}^I$ and $\mathcal{U}^T$.

Given an image query x^I with projection p^I, the text x^T that most closely matches it is that for which p^T minimizes the distance between p^I and p^T on the d-dimensional common representation space. In this thesis, we mainly employ the Euclidean distance. However, different types of distance *e.g.* Kullback-Leibler divergence, normalized correlation, centered normalized correlation measure [Costa Pereira et al. 2014] can be considered.

Similarly, given a query text x^T with projection p^T, the closest image match x^I is that for which p^I minimizes the distance on the common space between p^I and p^T.

3.3 Limitations of common representation subspaces

The common representation subspaces enable the "matching" of information from one modality to another. In Chapter 2, the effectiveness of the common representation subspaces such as those issued from (K)CCA has been widely shown in cross-modal and multi-modal problems between visual and textual modalities of data. Nevertheless, we identified two limitations of such common subspace. We also perceive that these limitations lead to a significant loss in performance of multimedia retrieval and classification problems. The latter involves the quality of representation when method like CCA attempts to map data from original spaces onto common subspaces with *fewer* dimensions. As (K)CCA method uses lower-dimensional representations to summarize initially complete representations, several data and relations contained in original spaces may not be preserved on their common subspaces. The first limitation consists of relevant data that are ill-represented in the common space (Section 3.3.1) and the second is about the separation of projections between different modalities in the common subspace (Section 3.3.2). In the following, we clarify each of these imperfections followed by evaluations on real data such as BBC News, FlickR 8K and Pascal VOC07.

3.3.1 Poorly-represented data on the common space

This section aims to demonstrate the limitation concerning ill-represented data on the common representation space. Particularly, we investigate this problematic in a context of cross-modal retrieval where such limitation has a strong influence on a limited retrieval performance.

The limitation mentioned in this section relies on the quality of representation into the common representation space. This is a common problem for dimensionality reduction methods such as Principle Component Analysis (PCA) and Canonical Correlation Analysis (CCA). Using projections issued from PCA or CCA methods, each individual in the original space is summarized by a projection on the lower-dimensional subspace. One of the most crucial points is the quality of representation which means the reliability of the representation of each individual on this common subspace. The latter aims to estimate

whether the projection point is a good approximation of the original data.

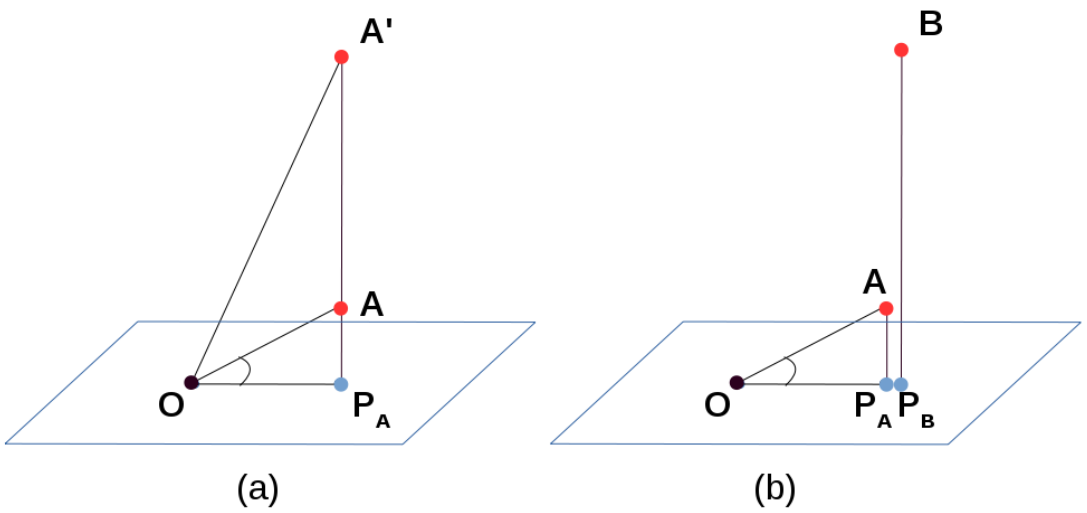

Figure 3.2: Quality of representation on the common subspace.
(a): two different data points A, A' have the same projection P_A. This projection is a good approximation of the close (hence well-represented) point A but not the very far (poorly-represented) point A'. (b): two data points A, B are far from each other on the original space but have very close projections P_A,P_B because of the poor quality of representation of the point B.

For a better visualization, an illustration is given in Figure 3.2. For a projection P_A with a fixed distance OP_A on the subspace learnt from CCA or PCA, its original data point may have any distance to this subspace. The latter can be close (*e.g.* A) or arbitrarily far (*e.g.* $A^{'}$) from the subspace (Figure 3.2.a). Accordingly, the angle between the vector issued from the individual and its projection on the subspace can be small *e.g.* $\widehat{AOP_A}$ or large *e.g.* $\widehat{A'OP_A}$. Only points that are close to the subspace *e.g.* A (the corresponding angle *e.g.* $\widehat{AOP_A}$ is small) are reasonably faithfully represented on the common space. This is explained by the fact that the projection vector (*e.g.* OP_A) of such point (*e.g.* A) express relatively well the variance of the data (OA). In this case, the distance between the origin to projection of a point gives a suitable approximation of the distance between the origin to this point on the common space. On the contrary when the data point *e.g.* $A^{'}$ is very far, the distance OP_A between the origin and its projection is not a good approximation of the true distance OA' of the original point A' to the origin. Furthermore, the proximity

among projections is only correctly judged provided that the corresponding original data points are well-represented. In Figure 3.2.b), two data points A, B are far from each other on the original space but they have very close projections P_A,P_B. The proximity between these two points is not adequately estimated on the subspace due to the poor quality of representation of the point B.

The most widely used measure to judge the quality of representation of each individual thus relies on the angle between vector issued from the individual and its projection on the joint space *e.g.* $\widehat{AOP_A}$ or $\widehat{A'OP_A}$ in Figure 3.2.a). In general, the cosine of this angle is estimated to evaluate the quality of representation of each individual. If this cosine is large (close to one), this individual is close to the subspace and therefore will be well represented. In this condition, we can then examine the position of its projection onto the subspace with respect to other projection points.

In what follows, we investigate the limitation involving the quality of representation of data on the common subspace learned with CCA on the BBC News data. These data contain both visual and textual information, however we only examine in this work the quality of CCA representation of *textual* data. In particularly, we investigate the quality of representation of words collected from this dataset onto their common representation space computing from text and images.

Assume $\mathcal{W} = \{w_1, w_2, .., w_{|\mathcal{W}|}\}$ the set of $|\mathcal{W}| = 23,617$ unique words collected from *training* documents in the BBC News corpus. A vector x_i $(i = 1, .., |\mathcal{W}|)$ is the Vector Space Model (VSM) representation with TF-IDF weighting [Salton and McGill 1986] of a word w_i with respect to the vocabulary $\mathcal{W}$. In such case, the $tfidf_i$ value of each word w_i simply becomes idf_i since $tf_i = 1$. The VSM representation x_i of the word w_i can be rewritten as $x_i = (0, .., 0, idf_i, 0, .., 0)$. The ith component is defined by idf_i which reflects the popularity (common or rare) of the word w_i across all documents in the BBC News corpus. The more common a word is, the lower its *idf* value.

In this experiment, the CCA common space is learnt from BBC News training documents and has 2,500 dimensions. We note that the collection $\mathcal{W}$ of 23,617 words is a (textual) part of what we use to learn the CCA space. Each word w_i is then projected onto the common representation space into p_i.

To study the quality of representation of each words $w_i \in \mathcal{W}$, we estimate the cosine value of the angle between vector issued from the individual and its projection. This measure is the ratio between ℓ^2-norms of vectors issued from the projection p_i of a word w_i onto the common representation space and the TF-IDF representation x_i of this word. It is determined as follows

$$\text{Quality of representation}(w_i) = \frac{\|p_i\|}{\|x_i\|} \tag{3.5}$$

where $\|a\|$ is the ℓ^2-norm of the $k-$dimensional vector a and $\|a\| = \sqrt{a_1^2 + a_2^2 + .. + a_k^2}$.

We report in Figure 3.3 the relation between the quality of representation computed by Eq. 3.5 and the *idf* value of each word in the BBC News collection. An important observation is there exists a relation between the quality of representation of a word on the subspace and its popularity within the corpus. Common words that we can easily find in news articles such as "*govern, home, nation, verdict, wild, economy, divorce*" are relatively well-represented on the common subspace. It is well worth noting that words with low quality of representation have higher *idf* values with respect to other words. In other words, the words which are poorly represented on the common subspace are relatively rare in the corpus. As we can see in the figure, several instances of these poorly-represented words "*Bingham, Christianne, Britney, Fett, Gustard, Wikimedia*" which are likely names or proper nouns. In this way, they are rare because their meanings are more particular and/or individual than other common words. An explanation for the poor quality of representation of these relatively rare words is that, because the development of a latent joint representation relies on extracting statistical regularities from a preferably large amount of data, any piece of data having very few occurrences or very weak relations with other data is ignored in the resulting joint model.

While some of specific words are quite ill-represented on the common representation subspace, they are nevertheless very important indicators to select relevant results in *retrieval* context. The fact that these words is ill-represented on the embedding space may severely hamper the performance of cross-modal retrieval problem. For example, a corpus of news articles covers a large number of topics such as business, world news, technology, health, entertainment, etc. The vocabulary of words used in such corpus is huge. Many

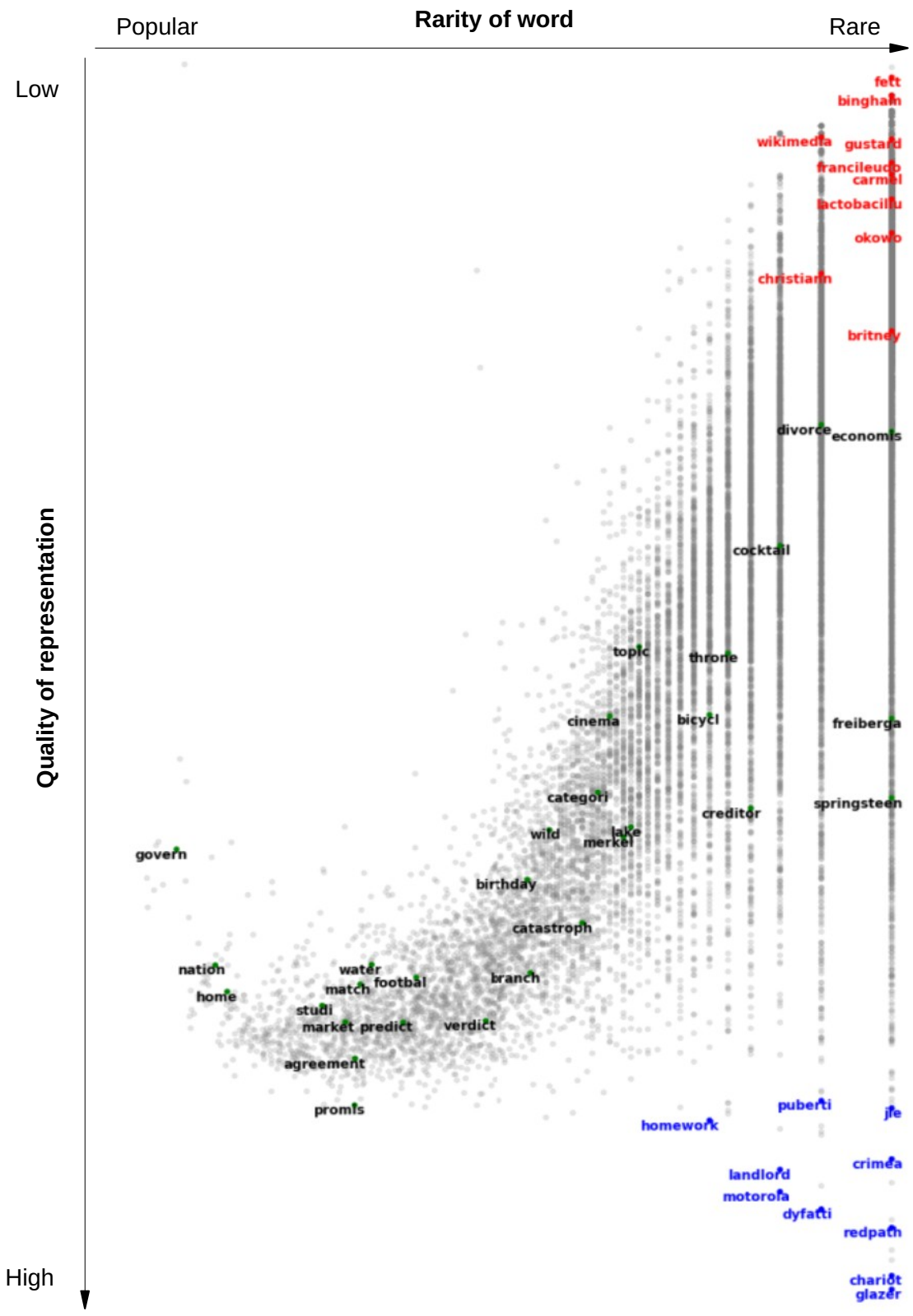

Figure 3.3: Quality of representation of words collected from BBC News.

words in this vocabulary (*e.g. students, nation, homes, cars, economy, migration, flight, tourist*) appear million times by their popular utilization. Meanwhile, we can also see the word "Maldives" that is present only in a couple of articles. This word is ignored in the common representation space for image and text relying on CCA method due to its infrequency with respect to the others in the vocabulary. Meanwhile, for a query text "Maldives island", the ignored word "Maldives" can help cross-modal retrieval model refine the search results to give a best-matched image answer.

3.3.2 Separation between modalities on the common subspace

In this section, we focus on the second limitation of the common representation space with regard to the separation of projected data between modalities. Furthermore, we conduct several preliminary experiments to study this shortcoming of the joint space on various real-world data such as in Pascal VOC07 and FlickR 8K.

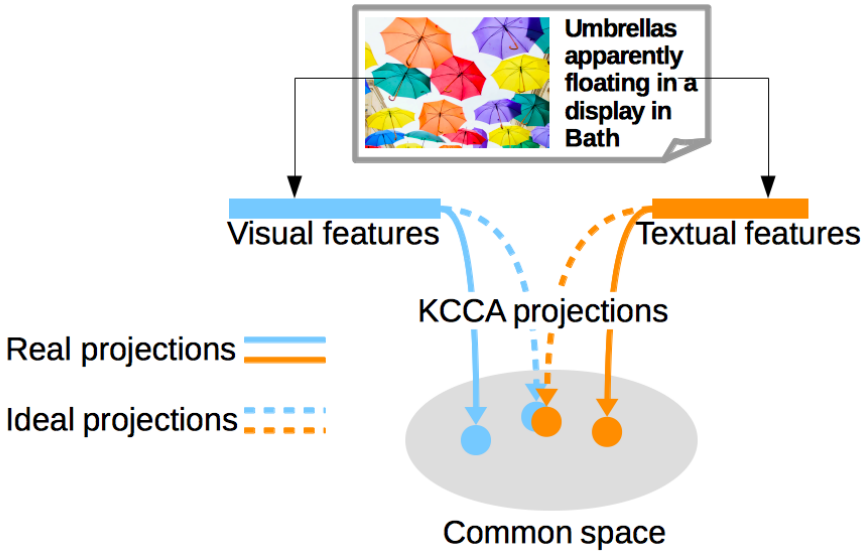

Figure 3.4: Separation between modalities on the KCCA space.

An important observation is that the textual and visual projections tend to be grouped by modality rather than according to their semantic on the common representation space obtained with KCCA. This observation is illustrated in Figure 3.4. Given a bi-modal document containing both image and text. Visual features and textual features of this

document are first extracted and then projected onto the common representation space in visual and textual projections respectively. Ideally, these two projections should be as close as possible. This is because they reflect the same semantics as they are textual document and illustration image of a unique multi-modal content. However, the real projections that we actually obtain on the common space usually fall far apart.

Furthermore, by observing the distribution of data projected on the common representation space, we found that the projections consequently establish themselves two separate projection clouds. One of the clouds contains almost only textual projections and the rest contains almost only visual projections. In this way, the common representation space only provides a coarse association between modalities. As a result, a direct use of these projections leads to a limited quality of "translation" between image and text. The performance of cross-modal or multi-modal tasks on such space is consequently restricted.

In what follows, we study the imperfection of the KCCA projection on Pascal VOC07 and FlickR 8K datasets. We show several data analysis results that highlight the mentioned problem.

First, we investigate the distribution of data projections on the common space by measuring distances between these projections. For this purpose, we respectively compute the *intra-modality* distances between projections belonging to a modality and the *inter-modality* distance between projections concerning different modalities.

Table 3.1 reports the average distances between KCCA projections. The results are reported on the sets of both textual and visual projections of *training* data, respectively 10,022 points in Pascal VOC07 and 12,000 points in FlickR 8K. We denote by $d_{\text{intramodality}}(I)$ and $d_{\text{intramodality}}(T)$ the average within-modality distances between image and respectively text projected points. Next, we distinguish two types of *inter-modality* distances: the distance between a visual projection and its associated textual projection on the KCCA space of a training sample and the distance between visual and textual projections over all training data (this last is the "classical" inter-class distance). The average of these distances are respectively denoted as $d_{\text{intermodality}}(\text{sample})$ and $d_{\text{intermodality}}(\text{overall})$. The values obtained in Table 3.1 show that projected points are closer to their within-modality neighbors than to their corresponding points in the other modality. The latter confirms

Average Distance	Pascal VOC07	FlickR 8K
$d_{\text{intramodality}}(I)$	1.18 ± 0.16	1.17 ± 0.13
$d_{\text{intramodality}}(T)$	1.11 ± 0.19	0.75 ± 0.13
$d_{\text{intermodality}}(\text{sample})$	1.39 ± 0.07	1.02 ± 0.12
$d_{\text{intermodality}}(\text{overall})$	1.42 ± 0.06	1.28 ± 0.10

Table 3.1: Average distances between projections on KCCA space.

Dataset	# projected points	k	# visual clusters	# textual clusters
Pascal VOC07	10022	16	12	4
FlickR 8K	12000	8	6	2

Table 3.2: Distribution of textual and visual KCCA-projected points into clusters.

our observation about the imperfection of the common space on that data are regrouped by modality rather than by their semantic. The fact that retrieval "can work" in the common space directly results from $d_{\text{intermodality}}(\text{sample})$ being lower $d_{\text{intermodality}}(\text{overall})$. On average, in the common space, the two corresponding projected points from a given document are closer than the average distance between the modalities. However, the difference between these two average distances is much larger than the average intra-modalities distances. There is thus margin for improvement.

For a better visualization about such separation on the common space, we report the distribution of the textual and visual projections of training data. For that, we computed the centers of gravity of the visual and respectively textual points, then projected all the points onto the line that joins these two centers. In Figure 3.5, we report the distribution (histograms) of these projected points. The separation in the KCCA space between data points from the two modalities appears very clearly, for both Pascal VOC07 and FlickR 8K datasets.

The last analysis we mention here consists of the statistic on clusters obtained with k-means from the collection of both textual and visual projections of training data on the common subspace. Table 3.2 shows the number of clusters associated to each modality in Pascal VOC07 and FlickR 8K. Given the separation between modalities on the common space, the resulting clusters contain mostly data from a single modality, *i.e.* image or

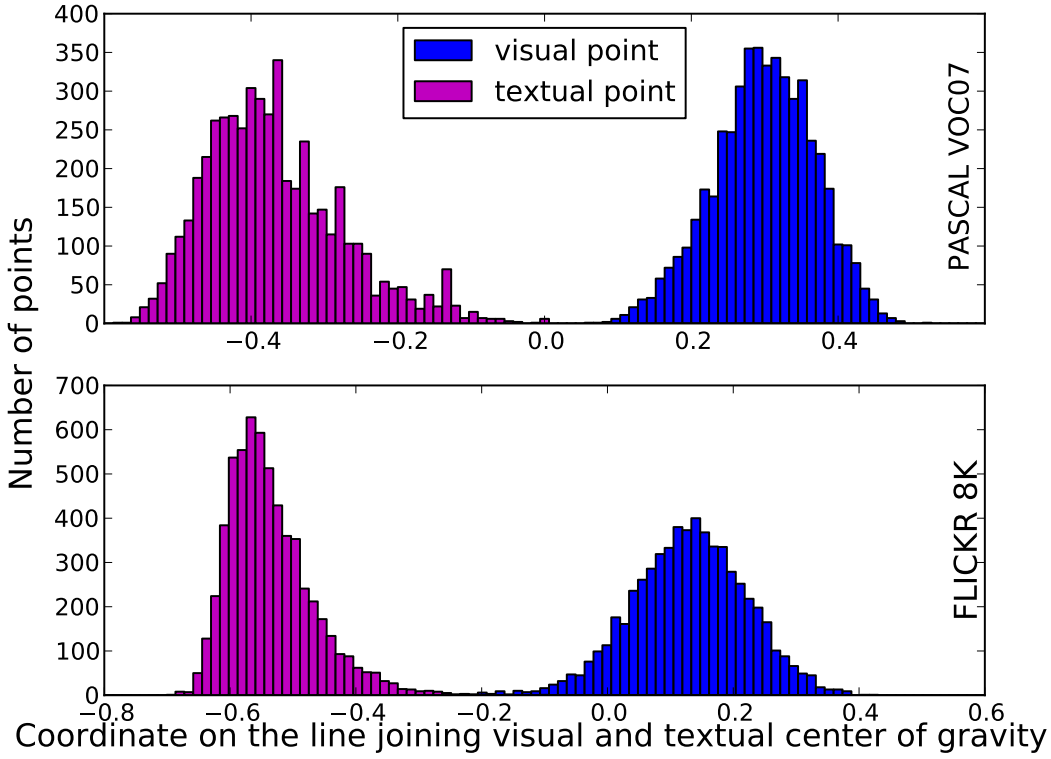

Figure 3.5: Separation between modalities on the KCCA space.

text. They are qualified as "visual" or "textual" according to the majority of points they contain. For more details, fifteen among the sixteen (15/16) clusters obtained from Pascal VOC07 projections contain data from single modality including eleven visual and four textual clusters. Only one cluster has simultaneously visual (99.02%) and textual (0.98%) projections which is hence classified into visual cluster category. Similarly, FlickR 8K contains two pure visual and two pure textual clusters. Fours cluster has both visual and textual projections. The major modality occupies at least 95% the numbers of projections in each of these bi-modal clusters.

3.4 Conclusion

In this chapter, we proposed to review briefly the theory of Canonical Correlation Analysis (CCA) and its kernelized extension Kernel CCA (KCCA) for learning common representation subspaces from visual and textual modalities of data. Three principle steps are mentioned including how to learn this common subspace using (K)CCA, how to project data from the original features spaces onto the KCCA subspace and finally how to perform a cross-modal matching on this subspace.

Despite the success of KCCA in the recent literature of image and text retrieval and classification, we have identified two major limitations of the common representation subspaces issued from this method. These limitations are related to the quality of data representation on the common subspaces. The first limitation is about several relevant data that are yet poorly represented by the joint model. An explication is that, because the development of KCCA subspace relies on extracting statistical regularities from a large amount of data, any piece of data having very few occurrences or weak relation to other data is ill-represented and thus ignored in the joint subspace. Unfortunately, ignored data is nevertheless very significant to select pertinent results in a retrieval context. Disregarding them therefore strongly obstructs the effectiveness of the system. The second limitation is the separation of data projections between visual and textual modalities on the common subspace. This coarse association between modalities make the direct use of projections resulting in limited quality of "translation". These identified imperfections of common representations lead to an important decrease in performance of multimedia retrieval and classification problems. Besides identifying the imperfections of such common subspaces, we furthermore show how they manifest on real-world data such as BBC News, Pascal VOC07 and FlickR 8K.

Chapter 4

Combining generic and specific information for cross-modal retrieval

4.1 Introduction

This chapter addresses the limitation concerning relevant but poorly-represented data on the joint space for image and text. This is due to the fact that the development of such a CCA-based latent joint representation relies on extracting statistical regularities from a preferably large amount of training data; any piece of data having very few occurrences or very weak relations with other data in the training collection is consequently ignored in the resulting joint model. Unfortunately, in a *retrieval* context, pieces of data that are rare in the training set or even new in the test set, such as *names* or *trademarks*, can be very significant in selecting the relevant results. Disregarding such information may strongly obstruct the effectiveness of the joint representation space.

The work aims to manage the mentioned deficiency of the joint space to enhance the cross-modal retrieval performance. For this purpose, it is necessary to extend the retrieval framework beyond the joint model that one may be able to include *"non regular but likely to be relevant"* information. We put forward a model which first identifies such information (words) by particularly distinguishing it from noise and then finds ways to combine it with the evidence provided by the joint representation model. This chapter furthermore examines how the proposed model is applied to address *text-illustration*. This task consists in finding an appropriate image to illustrate the content of a given textual document. We show that, by appropriately identifying and taking such information into account, the results of cross-modal *retrieval* can be strongly improved.

The rest of this chapter is organized as follows. Section 4.2 gives a brief overview of text illustration and significant work addressing this problem. Section 4.3 outlines the proposed model, including how to identify *specific* information that is poorly-represented on the joint model in 4.3.1 and then how to combine it with *generic* information that is relatively well-represented on the joint model in 4.3.2. Section 4.4 presents the experimental results. BBC News and Wikipedia datasets are used to evaluate our models for text illustration. One of the experiments considers domain transfer, emphasizing the need to make use of information that may be absent from the training data. Our conclusions are drawn in the final section 4.5.

4.2 Text illustration problem

Text and images usually appear together in multimedia content. For example, a children's book or a news article is usually attached with pictures which aim to describe the content of the text. Often the pictures themselves become more important than the surrounding text as they can punctuate the effect of stories. In such a scenario, many automatic text illustration (also called story picturing systems) have been investigated in order to choose one or more representative images from an available large image collection to accompany a text document (Figure 4.1). For instance, one may assists news writers in complementing their text without manually searching pictures from a large corpus of images to illustrate the content on their articles. The selected images highlight the content mentioned by the text and furthermore allow readers to quickly catch the main messages or topics that authors aim to deliver.

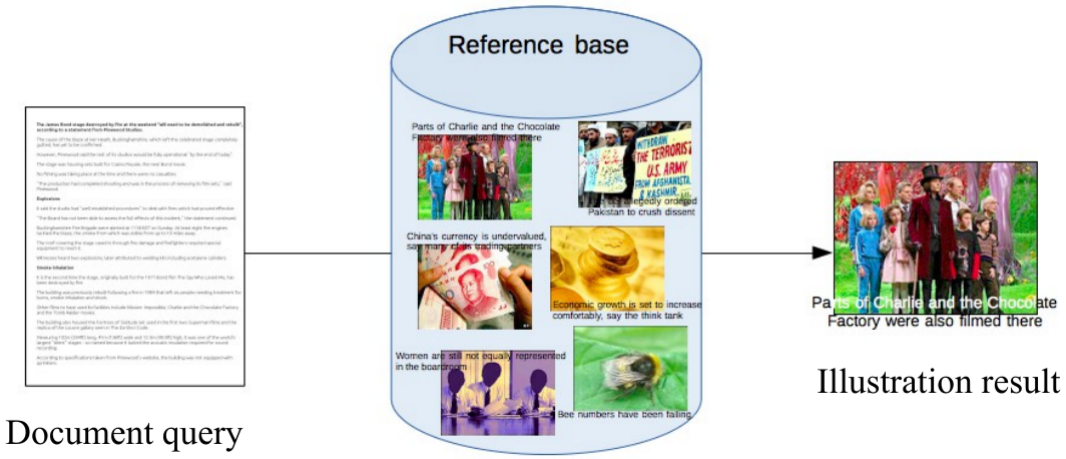

Figure 4.1: Automatic text illustration problem.

Considering textual data as the query to retrieve relevant visual data, text illustration can be seen as a particular application of *cross-modal retrieval* problem between image and text modalities. A variety of methods have been investigated to address this task. Most of them have been developed relying on techniques from image processing, information retrieval and more importantly natural language processing. A very classical approach to this task consists on uni-modal retrieval paradigm adopting only *textual* information. In

this direction, Delgado et al. [2010] proposed an application to help people reading news by illustrating the news story. The latter consists in finding the suitable images for each scene and next selecting the best set of illustrations to improve the story sequence. Image tags and story text are represented as the textual vectors, *i.e.* Vector Space Model (VSM), modeled with the term frequency-inverse document frequency TF-IDF weights and then compared by computing their similarities. The linguistic ontology WordNet [Miller 1995] is also used to refine the text-image relationship by adopting a semantic expansion on story text and image tags before the text-image comparison phase.

Meanwhile, the uni-modal solution is not usually effective for multi-modal data because of the existence of the well-known semantic gap between current image representations and those adopted by humans. To get out of this, text illustration model needs to take both visual and textual modalities of data into account for its higher retrieval accuracy. One popular approach following this direction relies on re-ranking techniques. One performs first an uni-modal retrieval using data from one modality (*e.g.* text) and next uses other modality (*e.g.* image) to re-rank the results of the previous uni-modal search. [Joshi et al. 2004, 2006] presented an unsupervised approach to automated story picturing. Semantics keywords are first extracted from the story and then used to retrieve image in an annotated image database. Thereafter, the importance of each candidate image is determined by an image ranking processing which takes both lexical annotations and visual content into account. The image ranking and selection is based on mutual reinforcement and discrete Markov chain model. The highest ranked images are selected to illustrate principle ideas conveyed by the story. Another cross-media re-ranking based illustration approach was proposed by Coelho and Ribeiro [2011] in order to assist writers with content enrichment. The model first performs a textual search followed by a scoring model to refine the results in the pool of image candidates. The images issued from the textual search are then re-ranked using visual information through a clustering scheme. This stage allows to eliminate very distinct photos may belong to unrelated events that were not filtered in the first step.

An inconvenience of the above mentioned approaches is the requirement of an *annotated* image collection in which textual search can be performed to match document query and annotated image(s). In most of cases, this condition is not met. As alternatives, several

works have attempted to directly match visual content of image(s) with textual content of document. Ones have approached the problem by describing documents and images by a common multi-modal vocabulary consisting of both textual and visual information. These representations can be learnt with Bag of Multimedia Words [Znaidia et al. 2012] or probabilistic multi-modal model *mixLDA* [Feng and Lapata 2010]. Based on Latent Dirichlet Allocation (LDA) [Barnard et al. 2003]-a probabilistic model of text generation, Feng and Lapata [2010] proposed *mixLDA* model representing visual and textual meaning jointly as a probability distribution over a set of topics. *mixLDA* model uses concatenated representations of words and images features assuming that the two modalities have equal importance in defining the latent space. The latter is built for the purpose of automatic text illustration and image annotation. *mixLDA* models the probability of each visual term in the vocabulary to a given text query through hidden topics and then delivers a ranked list of visual terms according to the query. Images having highest overlap with the top visual terms in the list are considered as the text illustration results.

4.3 Combining generic and specific information

Differing from the existing work, we approach the text illustration problem by learning a common correlated representation space between visual and textual modalities, *e.g.* using CCA method and then performing a retrieval process in this space. The CCA subspace enables matching either between document query and image itself or between document query and text associated to image such as caption, tags, descriptions. As shown in 3.3.1, such a space suffers from its imperfection while neglecting several pieces of data, called specific information, which is relatively rare yet relevant for the retrieval process. To achieve a good illustration performance, we put forward a retrieval method which combines both generic and specific information. In Section 4.3.1, we aim to identify such specific information that is ill-represented onto the common subspace and then show how these pieces of data can bias the CCA-based approach in a cross-modal retrieval context like text illustration. In Section 4.3.2, we subsequently introduce a model that fixes this drawback by combining the identified pieces of data and those which are well-represented on the common representation space.

4.3.1 Specific information identification

In Section 3.3.1, the imperfection of common representation subspace has been thoroughly investigated and demonstrated the existence of *relevant* yet *poorly-represented* pieces of data on such space. By examining the relation between the frequency of appearance of data in the collection and its corresponding quality of representation on the common space, we indicated that ill-represented information probably involves infrequent data. More precisely, such pieces of data are relatively rare and have weak relation to other data in the collection. Sometimes, they carry important and discriminant information and then the fact that they are ignored on the common representation space potentially has negative impact on *cross-modal retrieval* performance. It is thus important to identify such pieces of data while distinguishing them from noise. According to the rare and discriminant properties of such data, we called them *"specific"* information what is different from *"generic"* information being well-represented on the common space.

For a better illustration, we refer the reader back to Figure 2.9 in Chapter 2 where a complete *cross-modal retrieval* problem was described. Such a system is composed of two phases: one for indexing and the other for retrieval (or test). The first phase consists in learning a common representation space, *e.g.* with KCCA, for both visual and textual modalities and then map data into this latter for retrieval. The test phase aims to search in the indexed database a relevant (the most similar) entry for a given query. The imperfection about ill-represented information is involved in the indexation phase. Concretely, two databases are available during the learning phase: *training* base and *reference* base. The *training* base contains multi-modal documents that are used to learn cross-modal projections, *e.g.* KCCA projections, that map data from original visual and textual feature spaces onto a common representation space. Training data samples require the presence of both visual and textual modalities. Another dataset is the *reference* base contains documents used for information retrieval. The reference dataset can be uni-modal (including only textual documents or only images) or multi-modal (with both texts and images available). Assuming both of these databases contain multi-modal documents. In general, entries from these two sets can be either identical or completely different. Meanwhile, in the particular context of cross-modal retrieval, the reference base is potentially different from the training one. The major imperfection concerning ill-represented information on the common representation space relies on the *difference* between data in the two datasets. In the context of a real application, the training database is used to learn a common space, that is a fixed resource. It can then be used with several use cases, each of them having their own reference database. The fixed resource being included *e.g.* in a commercial product, one can not always adapt it to the reference database (*i.e.* recomputing the projection). Thus, it makes sense to study the effect of their difference.

The development of a latent joint representation obtained with KCCA for two or more modalities of data relies on extracting statistical regularities from a preferably large amount of training data. In this way, any piece of data having very few occurrences in the training set or very weak relations with other training data is ignored in the resulting joint representation model. Consequently, we define as "generic" information what is included in the training base used for KCCA learning. This "generic" portion is thus

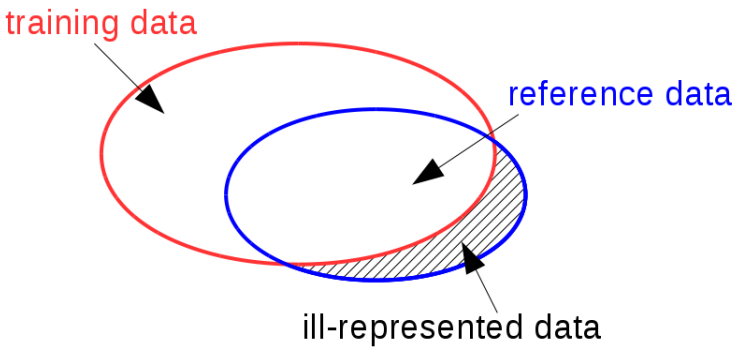

Figure 4.2: Ill-represented data on the common representation space involved in the difference between training and reference databases

relatively well-represented (regarding to the quality of representation) onto the common representation subspace. On the contrary, ill-represented "specific" information includes what is relatively *rare* with respect to other training data or even *absent* from the training base. Specific information is defined as what is present in the reference base but not in the training base, see Figure 4.2. In the scope of the thesis, we investigate the problem of ill-represented information only in the *textual* modality.

As shown in Section 3.3.1, the identified "specific" information consists of *names, proper nouns, trademarks or scientific terms*. In a *retrieval* context, the latter can be seen as important indication to help the cross-modal model find the relevant retrieval result. Furthermore, query and reference entries are likely to have a correspondence if they share one or a set of these "specific" information. For example, a corpus of news articles covers a large number of topics such as business, world news, technology, health, entertainment, etc. The vocabulary of words used in such corpus is immense. Many words in this vocabulary (*e.g. students, nation, homes, cars, economy, migration, flight, tourist*) appear million times by their popular utilization. Meanwhile, we can also see the word "Maldives" that presents in several articles. This word is ignored in the common representation space for image and text relying on CCA method due to its infrequency with respect to the others in the vocabulary. For a query text "Maldives island", the ignored word "Maldives" can help cross-modal retrieval model refine the search results to give a best-matched image answer.

4.3.2 Generic and specific information combination for cross-modal retrieval

The fact that the "specific" portion containing "non regular but likely to be relevant" information is poorly represented on the embedding space severely hampers the cross-modal retrieval performance. Our contribution aims to put forward a model that combines "generic" and "specific" information to achieve effectiveness of the examining cross-modal problem, in particular for the text illustration task.

4.3.2.1 Text illustration problem formalization

Consider a reference base $\mathcal{L}$ of *multi-modal* documents comprising image and text components

$$\mathcal{L} = \{D_1, D_2, .., D_{|\mathcal{L}|}\} \tag{4.1}$$

where D_j is the jth document in $\mathcal{L}$ and $j = 1, .., |\mathcal{L}|$.

Each reference document contains an image and its associated text (*e.g.* description, caption)

$$D_j = (x_j^I, x_j^T) \tag{4.2}$$

where $x_j^I \in R^I$ and $x_j^T \in R^T$ are thus respectively vectors of representation of the jth document in visual and textual feature spaces (here, we directly assimilate the image and text to their features).

In the scope of this chapter, we investigate *text illustration* - a particular case of cross-modal retrieval problem where query only belongs to textual modality. For a query text $q \in R^T$, the model aims to find a set of illustration image(s) $\mathcal{I}(p)$ retrieved from the reference collection $\mathcal{L}$ to illustrate the content of q. The images in $\mathcal{I}(p)$ must have the most similar content to the query q. The illustration system hence needs to estimate the similarity between the query and images or its associated textual descriptions. According to different strategies in estimating this measure, we distinguish two approaches for selecting relevant images. Uni-modal similarity, *e.g. Text-Text*, consists in computing the similarity between a query text and a textual description of image. Cross-modal similarity, *e.g. Text-Image*, estimates instead the similarity between a query text and an image itself.

Let note $Sim(q, x_j^I)$ (respectively $Sim(q, x_j^T)$) the similarity function between the query text q and the representation of an image x_j^I (respectively of a textual description x_j^T) in the reference base. The illustration $\mathcal{I}(p)$ for the query p is defined as follows.

- In the case of *Text-Image*

  The similarity between the query text $q \in R^T$ and a document $D_j = (x_j^I, x_j^T)$ is evaluated through the similarity between the query q and the image representation x_j^I of the document D_j. In this case, we directly use the most similar image x^I to illustrate the content of the query text q.

$$\mathcal{I}(q) = \{x_{j^*}^I\} \quad \text{such that} \quad j^* = \underset{j \in \{1,..,|\mathcal{L}|\}}{\arg\max} \left(Sim(q, x_j^I)\right) \tag{4.3}$$

- In the case of *Text-Text*

  The similarity between the query $q \in R^T$ and the textual description x_j^T is instead estimated to evaluate the similarity between the query text q and the document $D_j = (x_j^I, x_j^T)$. The image of the most similar document to the query text q is chosen as illustration of q.

$$\mathcal{I}(q) = \{x_{j^*}^I\} \quad \text{such that} \quad \begin{cases} j^* = \underset{j \in \{1,..,|\mathcal{L}|\}}{\arg\max} \left(Sim(q, x_j^T)\right) \\ D_j = (x_j^I, x_j^T) \end{cases} \tag{4.4}$$

4.3.2.2 Generic and specific information combination

The design of the text illustration system thus reduces to the design of effective similarity functions, *i.e.* $Sim(q, x_j^I)$ in Eq. 4.3 and $Sim(q, x_j^T)$ in Eq. 4.4, which allows us to define the relevant image in $\mathcal{I}(p)$ to illustrate the query text q. To facilitate the matching between a query text $q \in R^T$ and a reference image $x_j^I \in R^I$ or a reference text $x_j^T \in R^T$, we map two feature spaces R^I and R^T into a common representation subspace, *e.g.* using the CCA method. A natural way to perform the text illustration task within the CCA space relies on estimating *cosine similarity* between vectors of projection of the query q and the reference data x_j^T or x_j^I onto this space. However, we found that the direct use of these projections

is not pertinent. As specific data is poorly represented on the common space, evaluating the similarity between the query and a reference data relying on their projections onto this space neglects important information provided by specific data. Our goal consists in modeling effective similarity functions for the text illustration system by taking into account both generic and specific information.

In what follows, we always consider a reference base $\mathcal{L}$ of multi-modal documents that is *a priori* different from the learning multi-modal base $\mathcal{T}$ used to compute the common space by CCA model.

Let us now investigate the consequence of poorly-represented data in such a scheme. In what follows, we consider the visual modality I and the textual modality T of data. The modality $T = (G, S)$ is a textual feature vector that is composed of subspace G of the *generic* vocabulary well represented by the training data and a subspace S of *specific* vocabulary that is *infrequent* in training data, thus poorly represented by the CCA model. The dimensionality of G and S is respectively d_G and d_S.

Since data in S is infrequent, we assume the cross-covariance matrix between G and S is null, *i.e.* $C_{GS} \approx 0$ and the auto-covariance of S is the identity, *i.e.* $C_{SS} = Id_S$. The cross-covariance between T and I is

$$C_{TI} = \begin{bmatrix} C_{GI} \\ C_{SI} \end{bmatrix}$$

auto-covariance of T is written as

$$C_{TT} = \begin{bmatrix} C_{GG} & 0 \\ 0 & Id_S \end{bmatrix}$$

In what follows, we consider two CCA spaces. The first one, named CCA(I, T) is the full CCA space learnt from the data on taken from R^I (image) and R^T (text with both generic and specific information). The second one, named CCA(I, G), refers to CCA space learnt from R^I (image) and R^G (text accounting only generic information). Assuming W_T the projection matrix from the feature space T onto the CCA(I, T) space and W_G the projection matrix from the feature space G onto the CCA(I, G) space. Each column of W_T (respectively W_G) contains respectively a direction w_T (respectively w_G) obtained by resolving the corresponding eigen-problem relying on the matrix M_T (respectively M_G).

M_T and M_G are defined as follows

$$M_G = C_{GG}^{-1} C_{GI} C_{II}^{-1} C_{IG} \tag{4.5}$$

and

$$\begin{aligned} M_T &= C_{TT}^{-1} C_{TI} C_{II}^{-1} C_{IT} \\ &= \begin{bmatrix} C_{GG}^{-1} & 0 \\ 0 & Id_S \end{bmatrix} \begin{bmatrix} C_{GI} \\ C_{SI} \end{bmatrix} C_{II}^{-1} \begin{bmatrix} C_{IG} & C_{IS} \end{bmatrix} \\ &= \begin{bmatrix} C_{GG}^{-1} C_{GI} C_{II}^{-1} C_{IG} & C_{GG}^{-1} C_{GI} C_{II}^{-1} C_{IS} \\ C_{SI} C_{II}^{-1} C_{IG} & C_{SI} C_{II}^{-1} C_{IS} \end{bmatrix} \end{aligned} \tag{4.6}$$

With such approximation proposed in the hypothesis, the projection matrix W_T becomes

$$W_T = \begin{bmatrix} W_G & 0 \\ 0 & R \end{bmatrix} \tag{4.7}$$

where W_G is the projection matrix from the feature space G onto the common space obtained by CCA learnt from (I, G) and R a random matrix of size $(d_S \times d_S)$.

Accordingly, the cosine similarity in the case of *Text-Text* retrieval (Eq. 4.4) between the query text $q = \binom{g_q}{s_q}$ and a document D_j where its textual content is described by $x_j^T = \binom{g_j}{s_j}$ is defined as

$$\underset{CCA(I,T)}{Sim}(q, x_j^T) = \frac{< W_T^{'} q, W_T^{'} x_j^T >}{\|W_T^{'} q\| \left\|W_T^{'} x_j^T\right\|} = \frac{q^{'} W_T W_T^{'} x_j^T}{\|W_T^{'} q\| \left\|W_T^{'} x_j^T\right\|} \tag{4.8}$$

where A' denotes the transpose of a matrix A and $\|v\|$ the ℓ^2-norm of a vector v.

Combining (4.7) and (4.8) results into:

$$\underset{CCA(I,T)}{Sim}(q, x_j^T) = \frac{g_q^{'} W_G W_G^{'} g_j + s_q^{'} R R^{'} s_j}{\left\|\binom{W_G^{'} g_q}{R^{'} s_q}\right\| \left\|\binom{W_G^{'} g_j}{R^{'} s_j}\right\|} \tag{4.9}$$

Since $\left\|W_G^{'} g_q\right\| \ll \left\|R^{'} s_q\right\|$, one can keep the two first terms of the Taylor series of the denominator of Eq 4.9, then the similarity between the query q and a text x_j^T can be

rewritten as follows:

$$\begin{aligned} \underset{CCA(I,T)}{Sim}(q, x_j^T) &= \frac{g_q^{'} W_G W_G^{'} g_j + s_q^{'} R R^{'} s_j}{\|W_G^{'} g_q\| \, \|W_G^{'} g_j\|} (1 - \frac{1}{2} \frac{\left\|R^{'} s_q\right\|^2}{\|W_G^{'} g_q\|^2})(1 - \frac{1}{2} \frac{\left\|R^{'} s_j\right\|^2}{\|W_G^{'} g_j\|^2}) \\ &= (\underset{CCA(I,G)}{Sim}(g_q, g_j) + \frac{s_q^{'} R R^{'} s_j}{\|W_G^{'} g_q\| \, \|W_G^{'} g_j\|})(1 - \frac{1}{2} \frac{\left\|R^{'} s_q\right\|^2}{\|W_G^{'} g_q\|^2})(1 - \frac{1}{2} \frac{\left\|R^{'} s_j\right\|^2}{\|W_G^{'} g_j\|^2}) \end{aligned} \tag{4.10}$$

Since the number of specific words is much smaller than generic one (*i.e.* $d_S \ll d_G$) and the representation on the specific space (*e.g.* s_q, s_j) is relatively sparse with respect to one on the generic space (*e.g.* g_q, g_j), the ratio between the ℓ^2-norms of $R^{'} s_q$ and $W_G^{'} g_q$ (similarly between $R^{'} s_j$ and $W_G^{'} g_j$) is very small, closing to 0. In this case, the similarity between q and x_j^T in Eq. 4.10 becomes

$$\underset{CCA(I,T)}{Sim}(q, x_j^T) = \underset{CCA(I,G)}{Sim}(g_q, g_j) + \frac{s_q^{'} R R^{'} s_j}{\|W_G^{'} g_q\| \, \|W_G^{'} g_j\|} \tag{4.11}$$

On the right hand side of Eq. 4.11 , the first term is the similarity according to the CCA model computed on well-represented data, *i.e.* $CCA(I, G)$ from image modality and generic textual information. In the second term, the impact of specific data in S is biased by the CCA-based denominator. To fix this, we propose to remove the CCA-based weighting from the second term. In other words, we keep only the quantity of $s_q^{'} R R^{'} s_j$. In the simplest case, we consider the random matrix R an identity matrix of size $(d_S \times d_S)$. We furthermore propose to use a boolean model for this specific information, *i.e.* s_q and s_j binary vectors. The second term hence becomes $s_q^{'} \cdot s_j$ and simply reflects the number of common specific dimensions (corresponding to infrequent words) that are shared by the query q and the reference textual content x_j^T.

However, with such a model, when two documents D_1, D_2 sharing the same number of specific dimensions ($s_q^{'} \cdot s_1 = s_q^{'} \cdot s_2$) their relative similarity to the query only depends on the first term (CCA similarity). This may be inaccurate in this case since the documents have specific dimensions. Hence, we propose to weight the second term by a better adapted measure of similarity, given by the well-known TF-IDF model.

Finally, the similarity function of our proposed model, denoted CCA* taking both

specific and generic information into account can be written as

$$\underset{CCA^*}{Sim}(q, x_j^T) = \underset{CCA(I,G)}{Sim}(g_q, g_j) + s_q^{'} \cdot s_j \cdot \underset{TF-IDF}{Sim}(g_q, g_j) \tag{4.12}$$

When $x_j^T = \binom{g_j}{s_j}$ does not contain any specific dimension, that means every dimension of s_j (then $s_q^{'} \cdot s_j$) is equal to zero, our model is equivalent to the classic CCA-based retrieval model. On the contrary, when s_j is different from the zero vector, the second term may become dominant in the similarity estimation.

It is worth noting that our model supports cross-modal retrieval since the similarity in the CCA space can be estimated from the projection of any feature image x_j^I or text x_j^T. Similarly, for the case of *Text-Image* retrieval (as in 4.3), we define the cosine similarity between the query text $q = \binom{g_q}{s_q}$ and a document D_j with its visual content x_j^I as

$$\underset{CCA^*}{Sim}(q, x_j^I) = \underset{CCA(I,G)}{Sim}(g_q, x_j^I) + s_q^{'} \cdot s_j \cdot \underset{TF-IDF}{Sim}(g_q, g_j) \tag{4.13}$$

The difference between the uni-modal similarity 4.12 and the cross-modal similarity 4.13 only involves the similarity in the CCA space, which is in the first terms on the right hand side of these equations. More precisely, the proposed models computes the similarity between the query text q and the textual representation x_j^T of the reference document D_j in *Text-Text*, respectively the visual representation x_j^I in *Text-Image*. While specific information is ignored in the computation of similarity in the CCA space, the latter is handled, if available, in the second term of the equations. In both uni-modal (*Text-Text*) and cross-modal (*Text-Image*) retrieval cases, our proposed model manages either generic information, which is well represented on the CCA common space and specific information, which is almost ignored by this space.

4.4 Experimental evaluation

The proposed method for automatic text illustration is evaluated on the BBC News dataset and the Wikipedia dataset that are presented in Section 2.5. In the following, they are respectively noted "*bbc*" and "*wp*". We start by comparing our method with several baselines on the BBC News illustration task. Then, we specifically study the impact of

information that is absent from the training data by considering a *domain transfer* context, that is closer from a real situation: the common space is learned from a dataset that is different from the reference dataset used in the evaluation experiment. We end up by a comparison on the Wikipedia 2010 collection, that contains a large amount of specific information in comparison to the first two experiments.

4.4.1 Experimental Setting

Evaluation method. We adopt the evaluation methodology proposed in the work of Feng and Lapata [2010], based on top-1 accuracy. For a query article, the system is expected to rank first the image that was selected by the original author of the article. The reported accuracy is thus the percentage of successfully matched image-article pairs in the test set (240 documents). From this last, we obtain a reference database made of images aligned with their caption, while the associated article is used as query. The BBC News training dataset ($3,121$ documents) is used to learn the common space only. In that last case, article and caption are concatenated to get a unique textual content aligned with the image.

Content representation. To represent visual content we use OverFeat [Sermanet et al. 2013] which has been widely known to provide powerful features for several classification tasks. More precisely, we employ 3072-dimensional vectors which are the layer-18 outputs at the stage 6 of the fast OverFeat network and further ℓ^2-normalize them. For text features, we learn the dictionary by removing stop words, stemming the remaining words and filtering the stems by their frequency. Accordingly, vocabulary on $\mathcal{V}_{bbc}$ has 23,617 words that are either stems appearing at least twice in the training set or proper nouns from this set. The vocabulary on $\mathcal{V}_{wp}$ has 19,653 words, each appearing at least 5 and at most 1500 times. The filtering thresholds have been chosen such that the size of the dictionary is about the same order of magnitude for both datasets ($\sim 20k$). Texts have a TF-IDF representation, followed by ℓ^2-normalization.

Baselines. The proposed method is compared with *three* text illustration methods presented in Feng and Lapata [2010]. The baselines *Overlap* and *Vector Space Model*

disregard visual content.

- *Overlap* selects the image whose caption has the largest number of words in common with the test document.
- *Vector Space Model* (VSM) first represents articles and image captions using TF-IDF vectors. The cosine similarity measure is then performed to find the image whose caption is most similar to the test article. We report the result of VSM baseline $\text{VSM}_{\mathcal{V}}$ introduced in Feng and Lapata [2010]. This work used a vocabulary $\mathcal{V}$ of 6,300 words for text representation. In $\text{VSM}_{\mathcal{V}_{bbc}}$ baseline, we reproduce the VSM baseline using our textual vocabulary $\mathcal{V}_{bbc}$ of 23,617 words.
- *mixLDA* [Feng and Lapata 2010] considers both visual and textual content in defining a latent space. The method consists in computing the probability of each visual term in the visual vocabulary to a given text query through hidden topics and delivering a ranked list of relevant visual terms for the query. The image having the highest overlap with the top 30 visual terms in the list is considered as the best image to illustrate the text.

Notations. We denote by CCA_{cap} (resp. CCA_{img}) the basic CCA illustration model in which document-to-caption (resp. document-to-image) nearest neighbor search with cosine similarity measure is applied on document and caption (image) projections. The models in Eq.(4.12) and Eq.(4.13) are denoted by CCA^*_{cap} and CCA^*_{img} respectively. In our experiments, CCA spaces are constructed from images and text that concatenate documents (articles) and captions.

4.4.2 Results on BBC News

We first compare the basic CCA model to the three baselines. CCA dimension selection is performed via 10-fold cross-validation on the 3121 articles *training* set of BBC News. We retain 2500 as the projection dimension since it corresponds to the highest score, see Fig. 4.3. As shown in Table 4.1, the corresponding accuracy of CCA_{cap} on the test set is 76.7%, higher than the accuracy of all the baselines including those presented in Feng and

Lapata [2010] as well as the Vector Space $\text{VSM}_{\mathcal{V}_{bbc}}$ model with the larger $\mathcal{V}_{bbc}$ vocabulary.

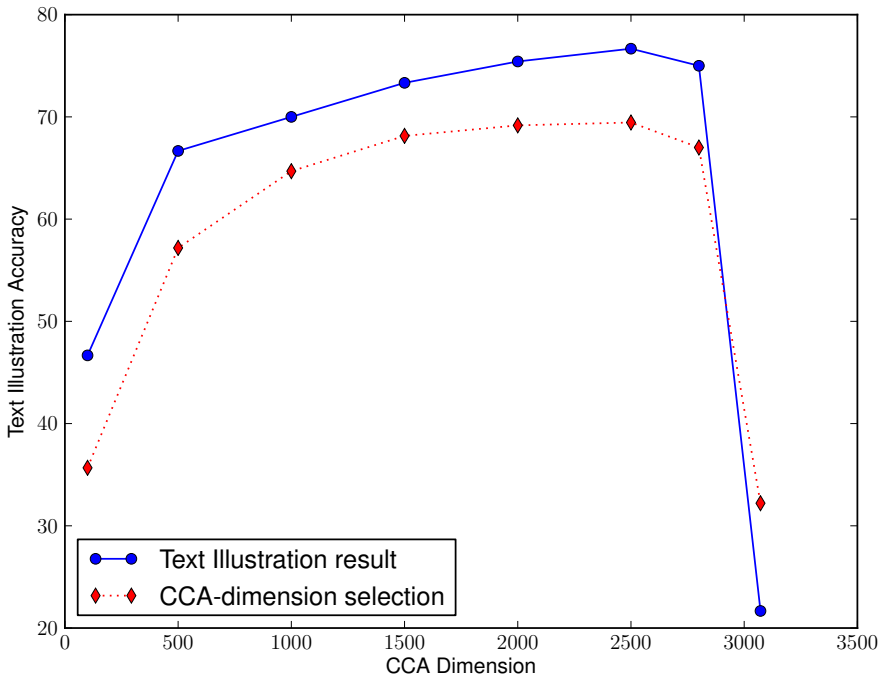

Figure 4.3: CCA_{cap} results (on the BBC News test set) and dimension selection (cross-validation on the training set)

The large improvement of $\text{VSM}_{\mathcal{V}_{bbc}}$ (73.8%) over $\text{VSM}_{\mathcal{V}}$[Feng and Lapata 2010] (38.7%) highlights the importance of an appropriate text representation for this task. The fact that the textual baseline $\text{VSM}_{\mathcal{V}_{bbc}}$ yields a high illustration performance, together with the weak score of CCA_{img} (5.4%), are indications that in the BBC News dataset the visual and textual contents are rather poorly related, so the latent representation learned by CCA is not so reliable. Meanwhile, the VSM model can easily take advantage of the connection between documents and captions, that appears to be strong.

To apply the CCA^* model we proposed in Section 4.3, we identify the vocabulary $\mathcal{S}_{bbc}$ of 41 specific words that are present in 240 testing captions but not in the initial vocabulary $\mathcal{V}_{bbc}$. As shown in Table 4.1, our CCA^*_{cap} model achieves the best score with 80.8% accuracy.

The third column in Table 4.1 reports the results of proposed models using a 10-fold cross-validation on 3,361 BBC News data. For each fold, 3,121/3,361 documents are employed for learning the CCA common space and the remaining 240/3,361 are used for test. The number of specific words varies from 26 to 42 over these folds while its average value is 35. The results show that the CCA^*_{cap} model outperforms the others.

Model	Accuracy(%)	Accuracy(%) 10-fold CV
Overlap [Feng and Lapata 2010]	31.3	*n/a*
$\text{VSM}_{\mathcal{V}}$ [Feng and Lapata 2010]	38.7	*n/a*
mixLDA [Feng and Lapata 2010]	57.3	*n/a*
$\text{VSM}_{\mathcal{V}_{bbc}}$	73.8	72.8 ± 2.1
CCA_{img}	5.4	8.9 ± 1.7
CCA_{cap}	76.7	74.7 ± 2.5
CCA^*_{img}	15.8	19.0 ± 2.0
CCA^*_{cap}	80.8	78.9 ± 2.4

Table 4.1: Text illustration results on BBC News dataset

4.4.3 Results in a domain transfer context

In the second experiment, we aim to study the impact of having a larger difference between the vocabularies of the target set and of the training set (used to learn the common space), which is an important issue in practical applications. We thus used an independent dataset to learn the latent space, collected from the large ImageCLEF 2013 Photo Annotation and Retrieval dataset[1] and later noted "*ic*".

The ImageCLEF 2013 collection includes 250,000 images downloaded from the Internet. Each image has at most 100 tags (words) extracted near the position of the image in the webpage it appears [Villegas et al. 2013]. These tags are considered as textual content associated to images and thus used for learning the textual vocabulary $\mathcal{V}_{ic}$ for the ImageCLEF 2013 training dataset. We follow the processing of vocabulary learning proposed in Section 4.4.1 including removing stop words, stemming the remaining words and filtering the stems by their frequency steps. The final vocabulary $\mathcal{V}_{ic}$ contains 18,003 unique words occurring each between 5 and 1500 times.

[1] http://www.imageclef.org/2013/photo/annotation

Model	Accuracy(%)
$\text{VSM}_{\mathcal{V}_{ic}}$	53.8
CCA_{cap}	43.3
CCA^*_{cap}	61.3

Table 4.2: Results on BBC News with domain transfer

The CCA space is learned from images and tags using OverFeat features of size 3072 and TF-IDF textual representations of size $18,003$. When used for retrieval, this textual representation induces a Vector Space Model noted $\text{VSM}_{\mathcal{V}_{ic}}$. In this experiment, the dimension of the latent space is set to 3072.

In this domain transfer context, there are 208 specific words $\mathcal{S}_{ic}$ present in the BBC News test set captions but not in $\mathcal{V}_{ic}$. Table 4.2 reports the performance of three models in this context: $\text{VSM}_{\mathcal{V}_{ic}}$, CCA_{cap} and CCA^*_{cap}. By taking the specific information into account, our model significantly improves the result of basic CCA, showing that it can be quite effective in a domain transfer context.

One can note that the score obtained by CCA_{cap} in this domain transfer context (43.3%) is much lower than that obtained when the latent space is learned on the BBC News training set (80.8%). This reveals that the two datasets, ImageCLEF 2013 and BBC News, present rather different relations between images and words. However, when one uses our contribution CCA^*_{cap}, it can fix about half of the performance loss, reaching 61.3%

4.4.4 Results on Wikipedia 2010

Model	Accuracy(%)
$\text{VSM}_{\mathcal{V}_{wp}}$	20.8
CCA_{img}	9.2
CCA_{cap}	16.3
CCA^*_{img}	58.3
CCA^*_{cap}	55.4

Table 4.3: Results on Wikipedia 2010

For the third experiment we consider the Wikipedia 2010 collection that can be directly employed for text illustration while being larger than the BBC News dataset. As for the first experiment, the CCA space is obtained from images and texts that accumulate

Figure 4.4: Top three images with their captions which are proposed by our model to illustrate the BBC News documents about "Extent of school failure disputed" and "Thames Water heads pollution list"

documents and captions, using the features described in Section 4.4.1. From the testing set captions, we determined 2868 specific words out of the training vocabulary $\mathcal{V}_{wp}$.

The results of the Vector Space model baseline, of the basic CCA model and of the proposed CCA* model are shown in Table 4.3. Our approach CCA^*_{cap} improves the text illustration accuracy over basic CCA from 16.25% to 55.4% and significantly outperforms the Vector Space model (20.8%). The cross-modal model CCA^*_{img} of Eq. (4.13) with visual features obtains even better results (58.3%).

The score are globally lower than in the experiment on the BBC News dataset, showing that this larger benchmark is more challenging. It is nevertheless remarkable that our proposal still conduct to a larger improvement when the task is more difficult.

4.5 Conclusion

We proposed a new approach for CCA-based cross-modal retrieval that takes advantage of specific information that is poorly represented in the training data but likely to be relevant for the task. Our contribution consists in first identifying specific information and then leveraging both specific information and generic information which is well-represented on the CCA common representation space for cross-modal retrieval.

Benchmark	Common space learned on	Nb specific words	Improv. over VSM	Improv. over CCA
BBC News	BBC News	41	+7.0	+4.1
BBC News	ImageCLEF13	208	+7.5	+18.0
Wikipedia2010	Wikipedia2010	2868	+37.5	+39.1

Table 4.4: Comparative performance of our method to the baseline CCA for different experimental settings. The more difficult the task (in term of number of specific words) the more our method is useful.

We showed the interest of our model in the context of the challenging text illustration task formulated as top-1 cross-modal retrieval. The proposed approach was compared to others on a previously published benchmark and shown to produce better results. We also proposed two new benchmarks that are more realistic in the sense that they contain more data that is new in the test set with respect to the training set. The results show that the proposed method improves even more effectively over the performance of CCA in these cases. Table 4.4 shows that the more number of specific information the model takes into account, the higher the improvement of the model over the basic CCA or the VSM model is. For instance, in the case of Wikipedia benchmark where 2,868 specific words are accounted, the improvement of our proposed model is +37.5 over the VSM model and +39.1 over the basic CCA model.

For discussion, it is worth noting that evaluation based on the top-1 result alone as proposed in Feng and Lapata [2010] is quite strict for such a text illustration task. The system must return as first relevant result the very image chosen by the author, but it may not be the best one to illustrate the document.

Actually, in many cases, for both BBC News and Wikipedia 2010, other images in

Figure 4.5: Illustration by author may not be the unique and the best choice to describe the actual content of BBC News article

Model	BBC News	Wikipedia
VSM	94.2	27.1
CCA^*_{cap}	95.0	70.8

Table 4.5: Results with top-10 evaluation

the collection are at least as relevant for the document. Several examples are shown in Figure 4.5. In the cases of a) and b), illustrations proposed by authors are placed in red frames while our system suggested other relevant images that are also related to the content of article. However, these images are ignored because they are not the original illustration used by authors. It is thus important to also evaluate the methods based on the accuracy of top-k results, with $k > 1$. As shown in Table 4.5 for $k = 10$, the proposed method compares well with the Vector Space model on both datasets.

Finally, we found that BBC News dataset is especially challenging for text illustration because of its small size and the quite indirect relation between textual and visual content in most news articles. For many articles, as in example c) of Figure 4.5, the original illustration does not relate to the actual content of article. However, other images in the collection can be objectively considered more relevant to the article's content than the image selected by the author.

Chapter 5

Uni-modal data completion with the missing modality

5.1 Introduction

In this chapter, we are interested into *cross-modal classification*, a task which has not been widely investigated in the multimedia community. It consists in training models on data from one modality and applying them to predict data from another modality. In the scope of this thesis, we investigate the cross-modal classification problem for image and text modalities, thus we consider the cases where training is performed on labelled textual-only data and testing on visual data or, symmetrically, training on labelled visual data and testing on textual data. Cross-modal classification is illustrated in Figure 5.1.

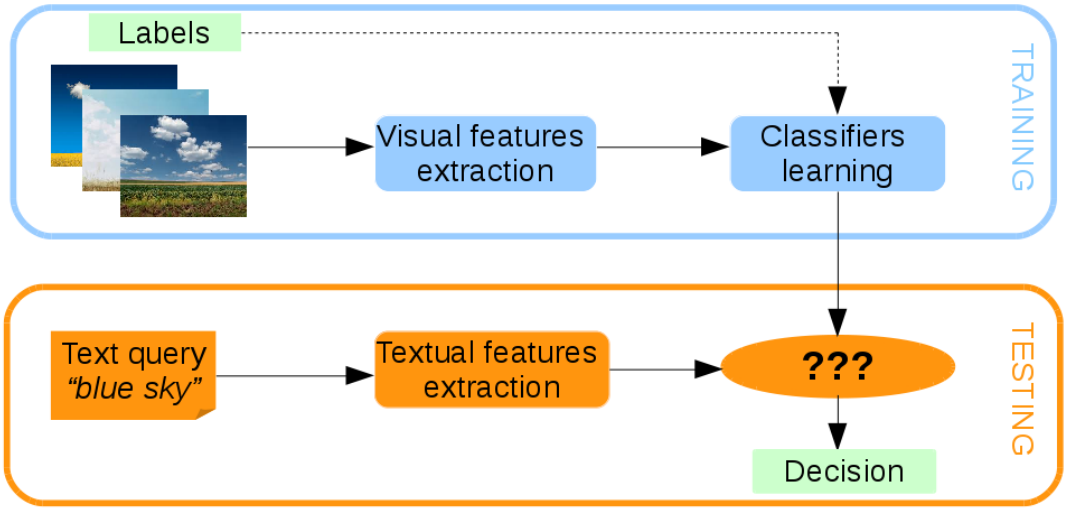

Figure 5.1: Illustration of *Image-Text* cross-modal classification problem. Models are trained on images and applied to predict a text.

This task has not been extensively investigated in the literature, first and foremost because text and images are usually not described with the same features, and usually not even in the same vector space, making the task quite incongruous. However, beyond an academic interest, we believe this task also has an increasing practical interest. Suppose, for example, that classifiers for many concepts could been learned from textual data because of the massive availability of such labeled data. One could wish to detect these concepts on content corresponding to another modality, *e.g.* images, even if class labels are not (yet) available for this content. Such a situation may become more common with the current

evolution of micro-blogging, that changes from purely textual content (historical Twitter) to multi-modal content (current Twitter) or purely visual content (Instagram). Furthermore, the study of the cross-modal classification task allows to explore in a more clear setting methods that aim to make the best use of the many datasets that *mix* uni-modal and bi-modal data.

Cross-modal classification is different but has connections with classical multimedia problems between image and text such as bi-modal image classification or cross-modal retrieval. The latter can be seen as a step beyond bi-modal image classification that usually considers images associated to keywords or sentences (*e.g.* captions) as input data and uses both visual and textual content to solve the task. This task is further related to cross-modal retrieval tasks such as text illustration or image annotation that require matching the information from one modality to the other.

The bi-modal classification and cross-modal retrieval tasks both need to relate text and image modalities. Various approaches have been extensively exploited in the literature to this purpose. To address bi-modal classification, several fusion approaches were proposed to combine the two modalities, *e.g.* Wang et al. [2009]. However, for cross-modal retrieval it was necessary to devise methods that are able to relate the two modalities more closely. The development of a common latent representation space, resulting from a maximization of the "relatedness" between the different modalities, is a generally adopted solution [Ngiam et al. 2011; Sharma et al. 2012; Srivastava and Salakhutdinov 2012; Hwang and Grauman 2012a; Srivastava and Salakhutdinov 2012; Gong et al. 2014; Feng et al. 2014; Costa Pereira et al. 2014].

In the context of cross-modal classification, such a common latent representation space is a suitable solution to overcome the incompatibility problem between different feature spaces. In this space, visual and textual information have similar representations and become directly comparable. Hence, it is perfectly conceivable to train a classifier on vectors of the common space that are projections of textual features and predict an output for a vector that is a projection of a visual feature. A common space was widely employed for cross-modal retrieval, *i.e.* information retrieval with both uni-modal and multi-modal queries. Several significant work in the recent literature on this topic have been reviewed

in Section 2.4. In particular, our contributions presented in Chapter 4 and Chapter 6 focus on investigating a robust common representation space to address cross-modal retrieval task. However, to the best of our knowledge, no attempt was made to employ cross-modal classifiers as we suggested. This is precisely the question we investigate in this chapter.

The problem investigated in this work is related to the work of Sharma et al. [2012] and Costa Pereira et al. [2014]. These models learn classifiers onto joint space of image and text (detailed in Section 2.4.1). To our knowledge, the cross-view classification was first mentioned by Sharma et al. [2012] to address the pose-invariant face recognition task. This work approaches cross-view classification using a k-NN classification scheme on their Generalized Multiview Analysis (GMA) joint space. It consists in classifying a sample by a majority vote of its neighbors, with the case being assigned to the class most common among its k nearest neighbors (k-NN) measured by a distance function. In this work, the parameter k is set to 1 (1-NN) which means simply to assign the sample to the class of its nearest neighbor on the latent space using the normalized correlation score as a metric. This approach is different from our research problem consisting on learning classifiers using features from one modality *e.g.* image and directly applying for another modality *e.g.* text on the common space. Furthermore, our work is close to the Semantic Correlation Matching (SCM) [Costa Pereira et al. 2014] in which classifiers are trained using projections on the joint representation space learnt by KCCA. However, this work used the output scores of these classifiers to build a semantic representation of uni-modal data to address the cross-modal retrieval alone.

More importantly, as it has been shown in the Chapter 3, the quality of cross-modal projections such as those obtained with KCCA is not good enough to achieve a robust "translation" between the two modalities. As shown in Section 5.3, the performance of cross-modal classification obtained with a direct use of KCCA projections is not too low. However, we believe that this latter can be improved once the limitations of the joint space described in Section 3 are managed.

Our contribution mainly consists in "completing" the projection of a uni-modal feature on the common space with information coming from the other modality. For this, we propose to rely on an auxiliary multi-modal dataset that acts as a set of connections

between the modalities within the common latent space. Such a dataset is always available, as it is required to obtain the common space. However, we also consider the case where the auxiliary dataset is totally different from that employed to learn the common space. While we mention a "naive" approach based on the auxiliary dataset, we propose a slightly more sophisticated scheme to identify the complementary information, leading to significantly better results. Last, our method includes a step that aggregates the original vector coming from the projection of a uni-modal feature with the identified complementary information to synthesize a unique representative vector of the document. This new representation thus embeds both modalities. In what follows, we call **"Weighted Completion with Averaging" (WCA)** the resulting proposed representation method. Consequently, learning a classifier with such WCA representation and applying it to uni-modal documents naturally leads to much better results than the "direct" approaches.

The rest of this chapter is organized as follows. In Section 5.2, we present the proposed method WCA including the "completion" of the missing modality (Section 5.2.1, 5.2.2) and the construction of an aggregated mutli-modal representation (Section 5.2.3) of a uni-modal projection onto the common representation space. Section 5.3 reports the evaluation results on Pascal VOC07 and NUS-WIDE. Comparisons are performed with two baselines, showing that the proposed method leads to significant performance improvements (Section 5.3.2). The impact of different parameters employed in our approach are investigated in Section 5.3.3, 5.3.4. We eventually compare the cross-modal classification results we obtained to state-of-the-art results concerning cross-modal retrieval, as well as uni-modal and bi-modal classification, showing that the performance level attained in cross-modal classification makes it a convincing choice for real applications (Section 5.3.5).

5.2 Weighted Completion with Averaging (WCA)

Cross-modal classification consists in training models on data from one modality (*e.g.* text) and applying them to data from another modality (*e.g.* image). This requires uni-modal labeled training data from the first modality and uni-modal testing data from the other modality. However, to *relate* the two modalities, one can rely on an additive *bi*-modal dataset. There is no need for this dataset to have class labels. As in many cross-modal retrieval methods, this additive dataset can be employed to learn a "common" latent space for the two modalities. The projection of uni-modal data on this common space makes data representations for the two modalities directly comparable. Nevertheless, as shown in Chapter 3, this common representation space suffers from several limitations that may hinder the performance of task *e.g.* cross-modal classification relying on such representation.

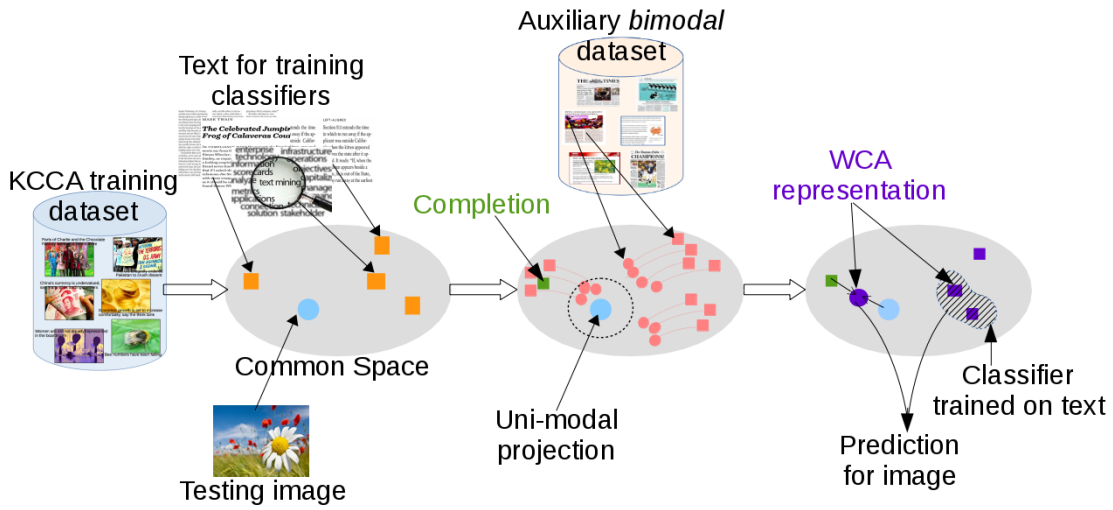

Figure 5.2: The proposed WCA approach for cross-modal classification

What we suggest here is to use another *auxiliary* bi-modal dataset to reflect the common space projection distortions. It is thus possible to rely on this dataset to "fix" the distortion. In practice, we propose to *build a bi-modal representation* in the common space for any data originally uni-modal, by completing a uni-modal projection with a virtual point from the other modality. The virtual point is obtained through the auxiliary dataset. The resulting average of these two points (real and virtual) is named "Weighted Completion with

Averaging" (WCA) in the following. The idea of using auxiliary dataset for uni-modal data completion has been previously introduced in the work of Wang et al. [2009]. However, this work addressed only the multi-modal classification and not the cross-modal classification as we suggest here.

To train the classifiers, such a bi-modal representation is first obtained for each uni-modal labeled training example and then learning is performed with these synthetic bi-modal WCA representations. For each uni-modal testing example (in the other modality than the one used for training), the bi-modal representation is built with the help of the auxiliary dataset and then the available classifiers are applied to this representation. Figure 5.2 illustrates this approach to cross-modal classification.

We consider here two cross-modal classification tasks: *Text-Image* (T-I) and *Image-Text* (I-T). In the *Text-Image* task, the classifiers are trained with documents that have only textual content and then evaluated on documents in which only the visual modality is available. Symmetrically, in the *Image-Text* task, classifiers are trained with visual-only documents and tested on textual-only documents.

5.2.1 Relevant completion information identification

Let us consider an *auxiliary* dataset of m documents, each having both visual and textual contents. Let $\mathcal{A}$ be the set of pairs of KCCA projections of the visual and textual features of these documents on the common space, with $\mathcal{A} = \{(q^T, q^I)\}$, $q^T \in \mathcal{A}^T$, $q^I \in \mathcal{A}^I$, $|\mathcal{A}| = m$. Dataset $\mathcal{A}$ can be seen as a sample of pairs of "linked" points, each concerning one modality. If the points are considered in the original spaces of visual and textual features, these links may be loose because part of the visual content of a document is unrelated to its textual content and conversely. The links should be stronger between the projections of the visual and textual features of the documents on the common space such as KCCA on which the relatedness between modalities are maximized. The sample $\mathcal{A}$ provides information regarding the relations between the two modalities. The more representative this sample is, the more reliable is the information.

The general idea of completion is to rely on this auxiliary dataset to transform a projection point from a uni-modal document into a more robust representation that

takes into account both modalities. For this purpose, one needs to determine a relevant complementary representation from the missing modality of the examining uni-modal document. The latter is represented by a virtual point obtained through the set of projections of auxiliary data $\mathcal{A}$ on the common space. The process of identifying such relevant complementary representation is explained in what follows.

Let us consider a document D with textual content only, described by a feature vector x^T that is projected as p^T on the KCCA space. The method described here (and in Sections 5.2.2 and 5.2.3) for a textual-only document can be symmetrically applied to a document having only visual content.

A direct but "naive" choice would be to complete p^T with a vector obtained from its μ nearest neighbors among the points of the auxiliary dataset projected from the *other* modality (visual one here), $NN^{\mu}_{\mathcal{A}^I}(p^T)$, because this is the missing modality for D. This naive choice, considered in Section 5.3 as a second baseline, can be expressed as

$$\mathcal{M}_c(p^T) = \{q^I_j\} \quad \text{such that} \quad q^I_j \in NN^{\mu}_{\mathcal{A}^I}(p^T) \tag{5.1}$$

To go further to such an approach, we need to consider the properties of the common space. While it results from an overall maximization of the relatedness between the two modalities, the projections of the textual and of the visual content of a same document on this space are not necessarily very close. So, given the uni-modal representation of a document D, its direct nearest neighbors within the other modality are not the best source for "filling in" the missing modality of D. However, we expect that documents having similar content according to one modality are likely to have quite similar content according to the other modality.

So, we propose to find the auxiliary documents having similar projected content with D in the *available* modality for D (textual modality in this case) and to use the projections of the visual content of these documents to complete p^T, see Figure 5.3. Formally, we define the set of contributors to the "modality complement" of p^T as

$$\mathcal{M}_c(p^T) = \{q^I_j\} \quad \text{such that} \quad \begin{cases} q^T_j \in NN^{\mu}_{\mathcal{A}^T}(p^T) \\ (q^T_j, q^I_j) \in \mathcal{A} \end{cases} \tag{5.2}$$

where the condition $(q^T_j, q^I_j) \in \mathcal{A}$ means that q^T_j and q^I_j are the projections of two feature

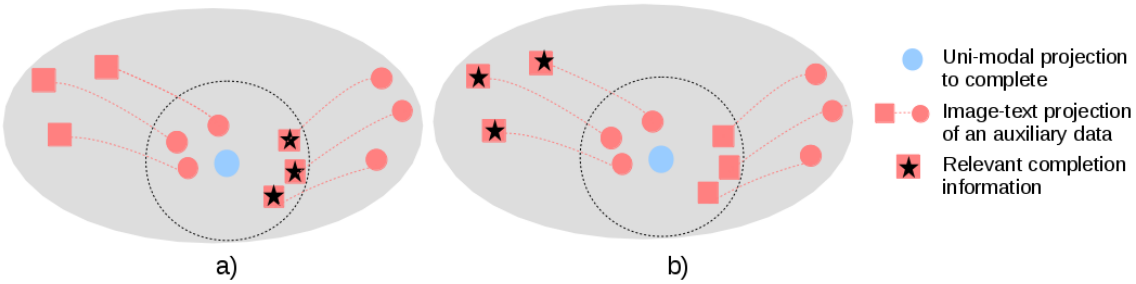

Figure 5.3: Naive completion (a) *vs.* proposed completion (b).
Squares and circles are text and resp. image projections. Connected red squares and circles represent bi-modal documents in $\mathcal{A}$. The blue circle is the projection of an image-only document. The naive approach seeks neighbors in the missing modality directly, while our proposal looks for them in the *available* modality.

vectors extracted from the *same* bimodal document. Note that $\left|\mathcal{M}_c(p^T)\right| = \mu$.

In practice, the auxiliary dataset $\mathcal{A}$ can be the training data employed to obtain the KCCA space and denoted by $\mathcal{T}$ in Section 5.3. However, we also consider and evaluate the use of *different* datasets for building the common space and for determining the relevant completion information within the missing modality of a uni-modal document. This can have a practical interest when, for example, the common space is an open resource but the dataset employed to build it is private or no longer available. Alternately, the dataset used to obtain the common space may be too large and generic, thus a smaller but "better focused" auxiliary dataset would be preferable to better model the characteristics of a narrow target domain.

5.2.2 Completion of the missing modality

Once the relevant complementary information regarding the missing modality of a document D has been collected on the common space as $\mathcal{M}_c(p^T)$, we employ it for building a representation for the missing modality of D.

Let $\hat{p}^I$ be the representation of this missing modality (the visual modality here) on the common space. A solution is to obtain $\hat{p}^I$ as the centroid of the q_j^I in $\mathcal{M}_c(p^T) = \{q_j^I\}$, *i.e.*

$$\hat{p}^I = \frac{1}{\mu} \sum_{q_j^I \in \mathcal{M}_c(p^T)} q_j^I \tag{5.3}$$

With several neighbors ($\mu > 1$), the neighborhood of p^T is better sampled, making the representation more robust. This is confirmed by experiments in Section 5.3.

The use of the centroid gives equal importance to all the μ neighbors. However, the similarity between p^T and each point $q_j^T \in NN_{\mathcal{A}^T}^{\mu}(p^T)$ should have an impact on the construction of the representation $\hat{p}^I$ of the missing modality. If, within the available modality (textual modality here), p^T is closer to a textual point $q_{j_1}^T$ than to $q_{j_2}^T$ (with $q_{j_1}^T, q_{j_2}^T \in NN_{\mathcal{A}^T}^{\mu}(p^T)$), then within the missing modality (visual modality here) the representation $\hat{p}^I$ should be closer to the corresponding visual point $q_{j_1}^I$ than to $q_{j_2}^I$ (with $q_{j_1}^I, q_{j_2}^I \in \mathcal{M}_c(p^T)$). Consequently, we prefer to define the representation $\hat{p}^I$ of the missing modality for p^T as a *weighted* centroid:

$$\hat{p}^I = \sum_{q_j^I \in \mathcal{M}_c(p^T)} \omega_j q_j^I \tag{5.4}$$

where ω_j is the weight of q_j^T. Among the μ nearest neighbors of p^T considered, some may be very close to p^T and others comparatively far away. The weighting method should allow to take into account the close neighbors and ignore the others, so the weight should quickly drop when the distance increases. We consequently define the weights as:

$$\omega_j = \frac{\sigma(p^T, q_j^T)}{\sum\limits_{q_j^I \in \mathcal{M}_c(p^T)} \sigma(p^T, q_j^T)} \tag{5.5}$$

with $\sigma(p^T, q_j^T) = 1/(\epsilon + \|p^T - q_j^T\|)$. Here, $\|p^T - q_j^T\|$ is the Euclidean distance between p^T and each $q_j^T \in \mathcal{M}_c(p^T)$. Also, ϵ is set to 10^{-16} to avoid marginal singularities for the points that may actually belong to the auxiliary dataset. In particular, this is important when the auxiliary dataset has data in common with the training set of the target task. The representative point into the missing modality is thus built from the complementary points of the neighbors found in the available modality and weighted according to the similarity computed in the available modality as well.

To complete the missing modality of a document D with the help of the auxiliary dataset $\mathcal{A}$ according to Eq. 5.4, it is necessary to retrieve $NN_{\mathcal{A}^T}^{\mu}(p^T)$, the μ nearest neighbors of p^T among the points in $\mathcal{A}^T$. If the auxiliary dataset is relatively small ($\|\mathcal{A}\| \leq 10^6$), exact exhaustive search is fast enough. For larger $\mathcal{A}$ a sublinear approximate retrieval method can be employed, *e.g.* Joly and Buisson [2011]; Novak et al. [2015].

5.2.3 Aggregated representation construction

For any unimodal document D originally described by p^T alone, after building the representation $\hat{p}^I$ of the missing modality, we aggregate p^T and $\hat{p}^I$ to obtain a unique descriptor p of D. Various aggregation methods can be used and several are compared in Section 5.3.

A widely employed method is the concatenation of the components, in this case of p^T and $\hat{p}^I$, resulting in a vector of size $2d$. This "unfolded" representation allows the classifier to process the textual and visual components separately but doubles the dimension of the description space.

Max-pooling consists in building a descriptor where the ith element is the maximum between p_i^T and $\hat{p}_i^I$. This method has already been used with good results for bag-of-visual-words (BoVW) representations, see *e.g.* Boureau et al. [2010]. We also evaluate max-pooling here, even though quantization is not employed for p^T and $\hat{p}^I$.

Averaging is also considered in Section 5.3. It obtains the aggregated description as the element-wise average of the two components p^T and $\hat{p}^I$:

$$p = (p^T + \hat{p}^I)/2 \tag{5.6}$$

The approach presented in sections 5.2.1, 5.2.2 and 5.2.3 for a textual-only document can be symmetrically employed for a visual-only document.

5.3 Experimental evaluation

We conduct several experiments on publicly available datasets according to standard experimental protocols. Beyond the raw performance of the proposed WCA method and its comparison to baselines, we study the influence of the main components of WCA, namely the completion process, the aggregation method and the relation between the auxiliary data $\mathcal{A}$ and the $\mathcal{T}$ dataset used for KCCA. We then compare the cross-modal classification results of WCA to state-of-the-art results concerning cross-modal retrieval. To better situate the performances attained by WCA on cross-modal classification, we compare them to uni-modal and bi-modal classification results of the state of the art.

5.3.1 Experimental settings

Dataset descriptions. We evaluate the proposed WCA approach for cross-modal classification task on Pascal VOC07, Nus-WIDE and Nus-WIDE 10K data. We refer the readers back on Chapter 2 for the detail description of Pascal VOC07 and Nus-WIDE dataset. Nus-WIDE 10K is a subset of the original Nus-WIDE dataset that we collected following the protocol proposed in Feng et al. [2014]. Only the following ten concepts are chosen: *animal*, *clouds*, *flowers*, *food*, *grass*, *person*, *sky*, *toy*, *water* and *window*. For each of these concepts we select 1000 image-text pairs (800 for training, 100 for validation and 100 for testing) that only belong to this single concept. This dataset is considered for the sake of comparison to the state of the art.

In what follows, we denote the full Nus-WIDE training set of 161,789 samples by NW160K. We also selected two smaller subsets of NW160K, NW12K of nearly 12,000 images and NW23K of nearly 23,000 images, both containing training images of the 81 concepts. NW23K and NW12K contain maximum 300 images and respectively 150 images for each of 81 concept.

Content representation. To represent visual content we use the 4096-dimensional features of the Oxford VGG-Net [Simonyan and Zisserman 2014], L2-normalized. They are extracted from a fully-connected layer (fc7, 16th layer) of a CNN architecture trained

on the ILSVRC 2012 dataset [Russakovsky et al. 2015] that contains 1.2 million images annotated according to 1,000 classes. These VGG features were shown to provide very good results in several classification and retrieval tasks [Razavian et al. 2014].

For texts (sets of tags or sentences) we employ Word2Vec [Mikolov et al. 2013], an efficient method for learning vector representations of words from large amounts of unstructured text. In our experiments, textual features are 300-dimensional, L2-normalized vectors. Following Mikolov et al. [2013], a single vector is obtained from several tags or a sentence associated to a given image, by summing the vectors of the individual words.

Baselines. We compare WCA to two cross-modal classification baselines. The first, denoted by KCCA_0, is simply the direct use of the projections on the KCCA space. The common space is learned from the dataset $\mathcal{T}$ and the two cross-modal tasks are performed *without* any completion, both for training and testing. More explicitly, classifiers are trained with the projections of one modality on the common KCCA space and tested with projections of the other modality on this space.

The second baseline, denoted by KCCA_{nc} (nc stands for "naive completion"), employs the "naive" completion method following Eq.(5.1). For either training or testing, the available modality is projected on the KCCA space and this projection is then completed, according to Eq. (5.1), with a vector obtained by the centroid method (Eq. 5.3) from its μ nearest neighbors among the points in $\mathcal{A}$ projected from the other modality. The averaging aggregation method of Eq. (5.6) is employed.

Common space and classifier settings. In all the experiments we use the KCCA implementation in Hardoon et al. [2004] to build the common space, with a regularization parameter $\kappa = 0.1$ and a Gaussian kernel with standard deviation set to $\sigma = 0.2$. These are the default values, also employed in other references Hodosh et al. [2013].

For each category, an SVM classifier with a linear kernel is trained, following a one-vs-all strategy. In practice, we use the implementation proposed by Bottou [2010]. It provides fast computations and is very efficient in terms of memory footprint since it is based on averaged stochastic gradient, an approach that is asymptotically efficient after a single pass

on the training set.

5.3.2 Proposed completion *vs.* naive completion *vs.* no completion

We first study the effectiveness of the completion mechanism for cross-modal classification on all the datasets. In the *Text-Image* task, the classifiers are trained with documents (of the training set) from which the visual content was removed and then evaluated on testing documents from which the textual content was removed. Symmetrically, in the *Image-Text* task, the classifiers are trained with image-only documents and then evaluated on text-only documents. Table 5.1 and 5.2 report the results obtained on these tasks by WCA and compares them to the KCCA_0 and KCCA_{nc} baselines on Pascal VOC07 and Nus-Wide datasets respectively.

Method	Pascal VOC07		
	T-I	I-T	Average
KCCA_0	78.98	59.88	69.43
KCCA_{nc}	75.07	68.77	71.92
WCA	**85.49**	**83.38**	**84.44**

Table 5.1: Cross-modal classification results (mAP%) on Pascal VOC07. Training set of Pascal VOC07 was employed for KCCA learning and for auxiliary dataset. Parameters $d = 4000$ and $\mu = 15$.

On Pascal VOC07, we employ the training examples (5011 image-text pairs) both as training data $\mathcal{T}$ for learning the KCCA space and as auxiliary data $\mathcal{A}$ for the modality completion stage. The best performances of the KCCA_0 baseline (78.98% for Text-Image and 59.88% for Image-Text) are obtained with $d = 4000$ dimensions. For the sake of comparison, the results of the KCCA_{nc} baseline and of WCA are reported in Table 5.1 for this 4000-dimensional common space. With $\mu = 15$, WCA yields a better performance than the two cross-modal classification baselines (+15% and +12.5% on average compared to KCCA_0 and KCCA_{nc} respectively).

On Nus-WIDE and Nus-WIDE 10K, the common space is learned from the data in NW23K. Subsequently, the 161,789 training and 107,859 testing examples of Nus-WIDE (respectively the 8,000 training and 1,000 testing data of Nus-WIDE 10K) are projected onto the common space to perform cross-modal classification tasks. We use NW23K as

Method	NUS-WIDE			NUS-WIDE 10K		
	T-I	I-T	Average	T-I	I-T	Average
$KCCA_0$	16.97	11.69	14.33	53.34	43.69	48.51
$KCCA_{nc}$	14.87	11.61	13.24	46.41	39.28	42.85
WCA	**18.81**	**17.90**	**18.36**	**58.62**	**52.77**	**55.69**

Table 5.2: Cross-modal classification results (mAP%) on Nus-WIDE and Nus-WIDE 10K. The common representation spaces were learned using NW23K collection. For auxiliary data, NW23K was employed for Nus-WIDE benchmark while 9,000 (training and validation) data in Nus-WIDE 10K was employed for Nus-WIDE 10K. Parameters $d = 10$ and $\mu = 10$.

auxiliary data $\mathcal{A}$ to complete uni-modal data in the Nus-WIDE benchmark. The 8,000 training plus 1,000 validation data in NUS-WIDE 10K are employed together as auxiliary data $\mathcal{A}$ for the NUS-WIDE 10K benchmark. In this experiment, the number of neighbors μ used for completion is set to 10 both for the $KCCA_{nc}$ baseline and for WCA. The best performances of $KCCA_0$ and $KCCA_{nc}$ are obtained with $d = 10$ for the two datasets. In this 10-dimensional common space, WCA (with $\mu = 10$) outperforms these two baselines by reaching a mAP of 18.81% for the Text-Image task and 17.9% for the Image-Text task on the NUS-WIDE dataset, and respectively 58.62% and 52.77% on NUS-WIDE 10K.

5.3.3 Influence of the completion and aggregation methods

We study the influence of the different completion and aggregation methods described in Section 5.2 on the performance obtained on Pascal VOC07, with the same parameters as in Section 5.3.2. The training examples in Pascal VOC07 were employed here both for KCCA learning ($d = 4000$) and as auxiliary dataset $\mathcal{A}$. WCA uses the weighted centroid for completion, and aggregation by averaging. "Weighted+Concatenation" combines the weighted centroid for completion with aggregation by concatenation. "Weighted+Max" also employs the weighted centroid for completion but max-pooling for aggregation. The "Centroid+Average" method uses the unweighted centroid (Eq. 5.3) for completion and average-pooling for aggregation. For each method, we report in Figure 5.4 the average of the mAP values obtained for the Text-Image and Image-Text tasks with $\mu \in \{1, 5, 10, 15\}$.

Both WCA and "Centroid+Average" perform significantly better than $KCCA_{nc}$, showing

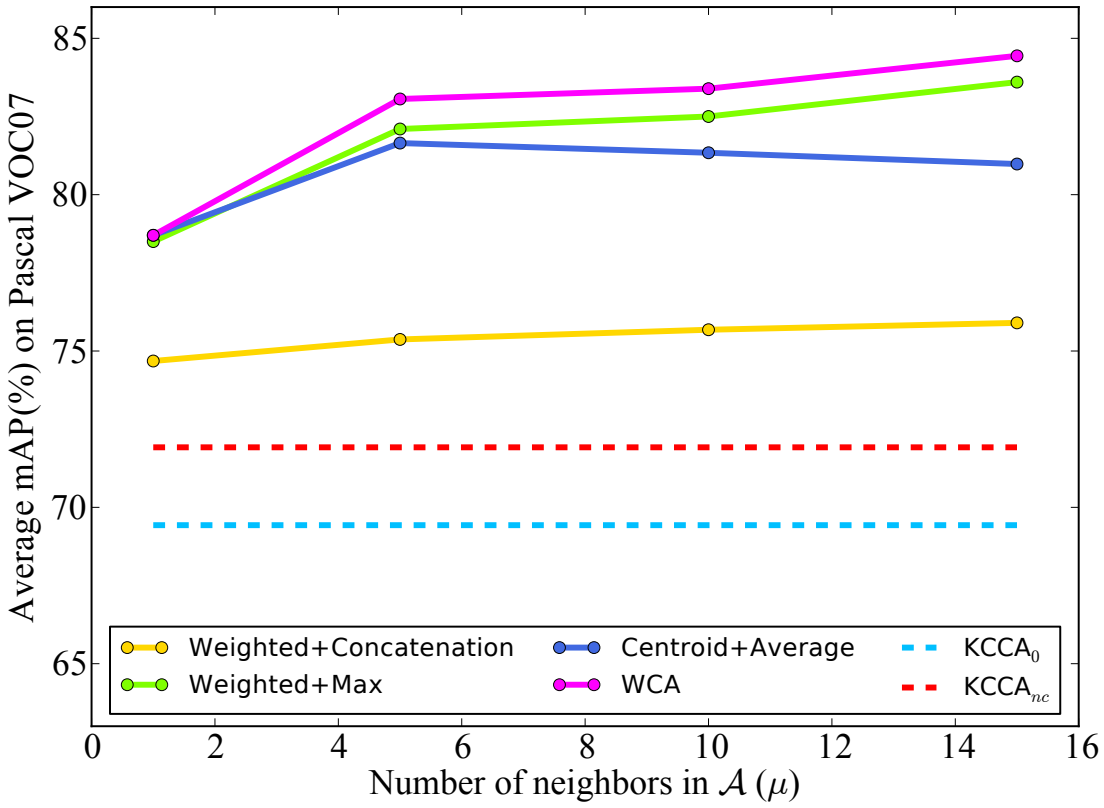

Figure 5.4: Results of different completion and aggregation methods on Pascal VOC07, showing the mAP(%) with respect to the number of neighbor points μ used in the auxiliary dataset $\mathcal{A}$. For each method, the curves are the average of the Text-Image and Image-Text tasks. Parameter d is set to 4000.

the interest of the proposed completion method of Eq. (6.4) in comparison to the naive completion of Eq. (5.1). Averaging is consistently better than max-pooling but the difference is small. Both averaging and max-pooling are significantly better than concatenation.

Performance increases with the number of neighbors μ for the three aggregation methods if the weighted centroid is employed, while with the unweighted centroid mAP slightly diminishes beyond $\mu = 5$. Indeed, for higher values of μ some neighbors that contribute to the completion of the missing modality are not enough near to be representative, thus can be considered as noise. In Eq. (5.5), the close neighbors are taken into account while those that are not "near enough" are ignored thanks to the weighting. With such a scheme, the number of neighbor μ can thus be increased without any risk of performance loss.

5.3.4 Impact of the auxiliary data and the common space

One of our motivations is to develop a common representation space as a generic "resource", from a large and general bimodal dataset $\mathcal{T}$, then address different cross-modal classification problems with this resource. This allows to avoid re-learning a common space for each problem, using a specific problem-related dataset. Projections onto this space benefit from the generic text-image relations learned from $\mathcal{T}$. A different, potentially more problem-related dataset $\mathcal{A}$ can then be employed for representation completion, taking thus into account problem-specific text-image links in the aggregated data representation.

To explore this idea, we study in this section the impact of using different datasets for obtaining the common KCCA space (dataset $\mathcal{T}$) and for completing the unimodal representations (dataset $\mathcal{A}$). All the experiments in this section concern Text-Image and Image-Text cross-modal classification on Pascal VOC07.

5.3.4.1 Auxiliary dataset $\mathcal{A}$

We fix the dataset $\mathcal{T}$ employed for learning the KCCA space as the bimodal training set of Pascal VOC07. The dimension of the common space is set to $d = 4000$ because the baseline KCCA_0 reaches its best performance for this value. While in the previous experiments the auxiliary dataset $\mathcal{A}$ was the same as $\mathcal{T}$, here we successively evaluate as auxiliary dataset NW12K, NW23K, NW160K and eventually $\mathcal{T}$. The number μ of nearest neighbors in $\mathcal{A}$ used for data completion is set to 5. The cross-modal classification results on the Pascal VOC07 test set are reported in Table 5.3.

Method	$\mathcal{A}$	T-I	I-T	Average
KCCA_0	-	78.98	59.88	69.43
KCCA_{nc}	VOC07	75.07	68.77	71.92
WCA	NW12K	65.06	57.30	61.18
	NW23K	69.34	60.49	64.92
	NW160K	73.77	64.47	69.12
	VOC07	83.79	81.33	82.56

Table 5.3: Results on Pascal VOC07 with the common space obtained from the Pascal VOC07 training set. Different auxiliary datasets $\mathcal{A}$ are used for WCA (with $d = 4000, \mu = 5$).

The results show that the performance of WCA depends both on the size of the

auxiliary dataset $\mathcal{A}$ and on the "agreement" between $\mathcal{A}$ and the specific classification problem considered. As expected, with NW12K, NW23K and NW160K as auxiliary datasets, the larger $\mathcal{A}$, the better the performance. Nevertheless, the mAP value obtained when $\mathcal{A}$ is the comparatively small (5011 bimodal documents) Pascal VOC07 training dataset is 82.56%, significantly higher than the one obtained when $\mathcal{A}$ is the much larger NW160K dataset (only 69.1%). A first potential explanation is that NUS-WIDE does not sample well the domain in the common space covered by Pascal VOC07. Consequently, given a projection of a unimodal document in Pascal VOC07, its μ nearest neighbors in NW12K, NW23K or NW160K are not as close as the ones in the Pascal VOC07 training set, so completion is less reliable with NUS-WIDE data.

A second potential explanation is that NUS-WIDE is not so well represented by the projections on the common space obtained with KCCA performed on the small Pascal VOC07 training set, because text-image relations may differ between the two datasets. Yet another explanation is that the NUS-WIDE data remains noisy even after separating the concatenated tags. This is shown by the fact that the cross-modal classification results obtained on NUS-WIDE are significantly lower than those attained on Pascal VOC07.

5.3.4.2 Common space training dataset $\mathcal{T}$

To get a better understanding of the relations between the dataset $\mathcal{T}$ employed for learning the common space, the dataset $\mathcal{A}$ used for data completion and the specific classification problem, a second experiment is performed. In this experiment, we still consider cross-modal classification tasks on Pascal VOC07, but we vary both $\mathcal{T}$ and $\mathcal{A}$. When $\mathcal{T}$ is NW12K or NW23K, the baseline KCCA_0 reaches its best performance for $d = 100$. To support comparisons, we consider $d = 100$ and $\mu = 5$ for all experiments.

Table 5.4 reports the cross-modal classification results on the Pascal VOC07 dataset for each common space learned from $\mathcal{T} \in \{$NW12K, NW23K, VOC07$\}$ and $\mathcal{A} \in \{$NW12K, NW23K, VOC07$\}$. The results for KCCA_{nc} were omitted from Table 5.4 because they are very close to those of KCCA_0.

As seen from Table 5.4, performance improves when the data in $\mathcal{T}$ is problem-related rather than some other dataset (Pascal VOC07 training set instead of NW23K). This is true

Method	$\mathcal{T}$	$\mathcal{A}$	T-I	I-T	Avg.
$KCCA_0$	NW12K	-	11.03	15.62	13.33
WCA	NW12K	NW12K	21.10	37.12	29.11
		NW23K	23.05	40.33	31.69
		VOC07	59.92	75.48	67.70
$KCCA_0$	NW23K	-	14.93	19.99	17.46
WCA	NW23K	NW12K	26.37	41.84	34.11
		NW23K	29.32	43.11	36.22
		VOC07	67.82	75.24	71.53
$KCCA_0$	VOC07	-	11.68	11.82	11.75
WCA	VOC07	NW12K	45.67	44.63	45.15
		NW23K	50.31	45.76	48.04
		NW160K	56.50	52.49	54.50
		VOC07	80.70	76.23	78.47

Table 5.4: Results on Pascal VOC07 with different datasets $\mathcal{T}$ to learn the common space and different auxiliary datasets $\mathcal{A}$ (with $d = 100$, $\mu = 5$) to connect the modalities in the common space.

even though NW23K is more than four times larger than the training set of Pascal VOC07. Also, with a same $\mathcal{A}$, cross-modal classification results improve for larger $\mathcal{T}$ sampled from the NUS-WIDE data (NW23K instead of NW12K). Using more data for obtaining the common space does improve performance, even if this data (NW12K, NW23K) is not related to the specific problem to be solved (in this case, cross-modal classification on Pascal VOC07).

An interesting observation is that the results are significantly better when $\mathcal{T}$ is NW23K (respectively NW12K) and $\mathcal{A}$ is the training set of Pascal VOC07 than when $\mathcal{T}$ is the training set of Pascal VOC07 and $\mathcal{A}$ is NW23K (respectively NW12K). Using problem-related data as auxiliary dataset $\mathcal{A}$, *i.e.* for completing the unimodal representations, has a much larger positive impact than using problem-related data for obtaining the common space. Together with the fact that the increase in performance from $\mathcal{T}$ = NW12K to $\mathcal{T}$ = NW23K is relatively high, this makes us optimistic about the possibility that, with a much larger but generic $\mathcal{T}$, results can improve beyond the level attained when $\mathcal{T}$ is problem-related.

Another observation is that regardless of the dataset $\mathcal{T}$ used for learning the common space, the highest performance is always obtained with Pascal VOC07 training data as

auxiliary dataset $\mathcal{A}$. The result obtained in Table 5.3 with problem-related $\mathcal{T}$ is thus extended to the use of a $\mathcal{T}$ that is not related to the problem. A smaller but "better focused" auxiliary dataset supports more reliable completion of unimodal representations, with a significant positive impact on cross-modal classification performance. This is also important from a complexity perspective. Indeed, our completion mechanism requires nearest-neighbor retrieval from the projections of the points in $\mathcal{A}$, according to the available modality. If good results can be obtained with a relatively small $\mathcal{A}$ then retrieval can be very fast and sublinear solutions may not be needed.

5.3.5 Comparison to the state-of-the-art

To our knowledge, cross-modal classification for text and image data was not previously investigated. It is *not* directly comparable, in principle, to the more classical uni-modal and bi-modal classification scenarios where classifiers are trained and tested with information of a same nature (same modality for the uni-modal case, both modalities together for the bi-modal case). Since it is nevertheless useful to have an idea of the relative levels of performance attained in these different scenarios, we compare in Table 5.5 the performance of WCA on cross-modal tasks to state-of-the-art results obtained on uni-modal and bi-modal classification.

As introduced earlier, we employed VGG [Simonyan and Zisserman 2014] to represent the visual content and Word2Vec (W2V) [Mikolov et al. 2013] for text since these features led to state-of-the-art results in the literature, on several tasks. In the following, we situate the performance of our proposed cross-modal classification with respect to the more classical problems such as uni-modal and bi-modal classification based on these presented features. In uni-modal classification, for the visual-only (denoted by VGG) and respectively textual-only (W2V) case, classifiers are trained and tested on VGG (resp. W2V) features alone. For bi-modal classification, in the VGG+W2V case of Table 5.5, representations for both training and testing data are produced by concatenating VGG and W2V features. The good results obtained in uni-modal classification, also very close to those of bi-modal classification with VGG+W2V, show the high effectiveness of the features employed.

On Pascal VOC07, the WCA results is obtained on 4000-dimensional KCCA space

learned from 5,011 Pascal VOC07 training data. This collection is also employed as auxiliary data and the parameter μ is set to 15. The results of both cross-modal classification tasks are lower but quite close to those of uni-modal classification with VGG or bi-modal classification with VGG+W2V. On Nus-WIDE and Nus-WIDE 10K, the WCA performances are obtained on 1000-dimensional KCCA space learned from NW23K. The parameter μ is set to 10. In Nus-WIDE benchmark, NW23K is also employed as auxiliary data while in Nus-WIDE 10K, we use its corresponding 9,000 training and validation. On Nus-WIDE, the difference is larger between cross-modal classification and uni-modal or bi-modal task. We suspect that this may be due to a comparatively weaker link between the two modalities on this dataset. On NUS-WIDE 10K, WCA provides slightly better results than uni-modal classification and weaker performance than bi-modal classification. We believe that the protocol put forward in Feng et al. [2014] selects data where the visual and textual modalities are better related. The mechanism we proposed for completing the uni-modal features with complementary information in the missing modality appears to have a very significant contribution in bringing the performance of cross-modal classification closer to the state-of-the-art in uni-modal and bi-modal classification. We also compare WCA to Chen et al. [2010] that reported previous state-of-the-art results for bimodal classification on NUS-WIDE. WCA significantly outperforms this method.

Classification type	Method	Pascal VOC07	Nus-WIDE	Nus-WIDE 10K
Uni-modal	VGG	86.10	50.38	78.53
	W2V	82.50	46.57	70.20
Bi-modal	VGG+W2V	86.16	50.87	82.89
	Chen et al. [2010]	n/a	19.30	n/a
Cross-modal	WCA (T-I)	85.49	37.80	79.53
	WCA (I-T)	83.38	34.02	79.15

Table 5.5: Comparison in terms of mAP(%) with uni-modal and bi-modal classification results.

Cross-modal *retrieval* is another well-known task and it may be interesting to see how the cross-modal *classfication* approach proposed here compares to this task. For cross-modal retrieval, the query is an item described along one modality and the ranked answers belong to the other modality. In [Ngiam et al. 2011; Feng et al. 2014], the cross-modal retrieval results reported on NUS-WIDE 10K employed the available concepts (our class

Cross-modal Task	Method	I-T	T-I	Avg.
Retrieval	Ngiam et al. [2011]	25.0	29.7	27.4
	Feng et al. [2014]	33.1	37.9	35.5
Classification	WCA	89.2	89.7	89.5

Table 5.6: mAP@50 for cross-modal *retrieval* and for cross-modal *classification* on NUS-WIDE 10K. We implemented our method (WCA) and report the results of the original paper for Ngiam et al. [2011]; Feng et al. [2014]. Experimental protocols are coherent with these last (see text for details).

labels) as ground-truth for computing the mAP@50. For our cross-modal classification, the "query" is a decision boundary learned in one modality and the ranked answers are items described along the other modality. Table 5.6 shows both the mAP@50 results of cross-modal retrieval and of cross-modal classification on NUS-WIDE 10K. Note that Ngiam et al. [2011]and Feng et al. [2014] employed "classical" low or medium-dimensional features such as color histograms or bag of SIFT descriptors for images and bag of words for text, while we made use of VGG and W2V. The reader should however keep in mind that these two tasks are different, so Table 5.6 should be interpreted with care.

5.4 Conclusion

In this contribution, we put forward an approach, called "Weighted Completion with Averaging" (WCA) that addresses *cross-modal classification* for visual and textual data. This tasks consists in training classifiers with data from one modality and testing with data from the other modality. In line with recent literature on cross-modal *retrieval*, this approach relies on the development of a common latent representation space where image and text possess same representations. The novelty of our approach lies in the use of an auxiliary bi-modal dataset to systematically *complete* unimodal data, both for training and testing, resulting in more comprehensive bi-modal representations. The completion method we propose goes beyond a more direct completion solution that we also mention.

We provide an in-depth study of several aspects of our approach and compare it to recent work in the literature. It outperforms two cross-modal classification baselines that employ the raw KCCA data projections onto the common space. The proposed approach also provides interesting results compared to recent cross-modal retrieval methods. Furthermore, the performance level we attain on cross-modal classification also compares well to state-of-the-art results of uni-modal and bi-modal classification, which are more classical classification tasks. Such a performance level makes our approach to cross-modal classification a convincing choice for real applications, such as learning classifiers from an existing large amount of annotated textual data and applying them to visual content, to annotate images for example.

Chapter 6

Aggregating image & text quantized correlated components

6.1 Introduction

As previously mentioned, a common representation space can improve the performance of cross-modal and bi-modal tasks. However, Chapter 3 discusses that this space provides a very coarse association between modalities. A direct use of these projections therefore leads to limited quality "matching" between modalities.

To address this problem, our contribution is to put forward a new representation method for the projections on the common space, called **Multimedia Aggregated Correlated Components (MACC)**. MACC representation aims to reduce the gap between the projections of visual and textual features by embedding them in a local context reflecting the data distribution in the common space. Given a database of multimedia documents, we first perform KCCA and build a codebook from all the projections of visual and textual features on the KCCA common space. Subsequently, for each multimedia document, visual and textual features are projected on this common space, then coded using the codebook and eventually aggregated into a single MACC vector that is the multimedia representation of the document.

Specially, when a document is uni-modal, we further suggest to complete the absent modality using data from an auxiliary dataset following the completion procedure described in the Section 5. Subsequently, we combine the descriptors from two modalities to build the MACC representation of the initially uni-modal document.

In our experiments, we show that MACC representations allow to reach state-of-the-art performance in classification tasks on Pascal VOC07 and in image retrieval on FlickR8K and FlickR30K.

The remainder of this chapter is organized as follows. In Section 6.2, we focus on the construction of MACC representations, involving an aggregation of the projections of visual and textual content represented on a common vocabulary. We also introduce the MACC representation when data is missing for one of the modalities. The evaluation in Section 6.3, conducted on three datasets, concerns both image classification and cross-modal retrieval.

6.2 Proposed approach

In Section 6.2.1, we describe a new representation of multimedia documents relying on an aggregation of the projections of visual and textual content defined on a common vocabulary. Since (K)CCA aims to find a projection space where the correlation between modalities is maximized, we named this new representation **"Multimedia Aggregated Correlated Components" (MACC)**. In Section 6.2.2 we propose an extension for completing the MACC representations of documents for which only one modality is available. While MACC addresses problems with the representation of bi-modal documents, this extension focuses on actual cross-modal cases.

6.2.1 Multimedia Aggregated Correlated Components

Let us consider a document with a textual and a visual (image) content. A feature vector x^T is extracted from its textual content and another feature vector x^I from the visual one. In what follows, we assimilate a document to a couple of feature vectors (x^T, x^I). A set of such data is a set of couples $\mathbf{X} = \{(x_i^T, x_i^I), i = 1 \dots N\}$.

By applying KCCA to this data, we obtain $2N$ points (vectors) belonging to a "common" vector space where the two modalities are maximally correlated. In this space, a document (x^T, x^I) is represented by two points, p^T that is the projection of x^T and p^I the projection of x^I. Ideally, since they represent the same document, p^T and p^I should be closer to each other than to any other point in the projection space. However, in practice, this is far from being the case as shown in Chapter 3. It is thus quite problematic for a given document to be represented by two very distinct points for multimedia recognition tasks.

We propose to create a unified representation for each document, by the following process:

1. Define a unifying vocabulary in the projection space,

2. Describe both p^T and p^I according to this vocabulary,

3. Aggregate both descriptions into a unique representative vector of the document.

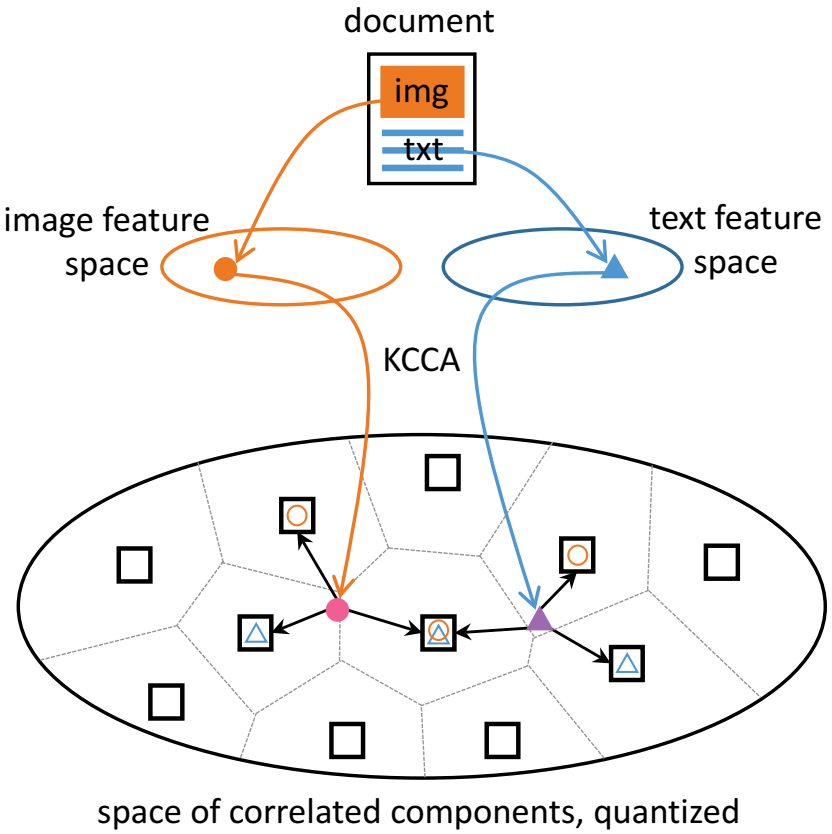

Figure 6.1: Visual and textual contents of a document are projected onto a common space that has been previously quantized. Both projections, corresponding to the same document, are encoded according to a common vocabulary before their aggregation.

Our approach is illustrated in Figure 6.1. Simply said, the "unified vocabulary" is obtained by quantizing the projection space, then p^T and p^I are projected to this codebook and sum pooled to get the final representation. Since it is well known that in computer vision devil is in the details [Chatfield et al. 2011, 2014], this is further explained below.

6.2.1.1 Codebook learning

As for the bag of words (BoW) model, we learn a codebook $\mathcal{C} = \{c_1, .., c_k\}$ of k codewords with k-means directly in the projection space. A crucial point is that *all* the projected points, coming from both textual and visual modalities, are employed as input to the k-means algorithm. Hence, the clustering potentially results into three types of codewords

(that are centers of the clusters). Some are representative of textual data only, others of visual data only, while some clusters contain both textual and visual projection points. The codebook is thus intrinsically cross-modal and can serve as "common vocabulary" for all the points in the projection space, whether they result from the projection of a textual content or of a visual one.

6.2.1.2 MACC representation

A bi-modal document (x^T, x^I) is projected on the KCCA projection space of dimension d into (p^T, p^I). Each of these points is then encoded by its *differences* with respect to its nearest codewords:

$$v_i^T = p^T - c_i; \quad c_i \in NN^n(p^T) \tag{6.1}$$

$$v_i^I = p^I - c_i; \quad c_i \in NN^n(p^I) \tag{6.2}$$

where $i = 1, .., k$ denotes the index of the k codewords of the vocabulary and $NN^n(p)$ denotes the set of the n nearest codewords of p. The modality-specific representations v^T and v^I result from the concatenation of the d-dimensional vectors v_i^T and respectively v_i^I.

The MACC representation v is then obtained by aggregating the visual and textual descriptors v^I, v^T by sum pooling, leading to:

$$\begin{aligned} v &= [v_1, v_2, \ldots, v_i, \ldots, v_k] \quad s.t. \\ v_i &= (p^T - c_i)\mathbb{1}_{NN^n(p^T)}(c_i) + (p^I - c_i)\mathbb{1}_{NN^n(p^I)}(c_i) \end{aligned} \tag{6.3}$$

where $\mathbb{1}_A(.)$ is the indicator function. Vector v is subsequently L2-normalized. The projection space obtained with KCCA has dimension d, so the *modality-specific encoded vectors* v^T and v^I, as well as the MACC vector v, have a size of $D = d \times k$, where k is the size of the codebook $\mathcal{C}$.

The vectors v^T and v^I are component-wise differences of p^T and p^I with some codewords. When $n = 1$, such a gradient can be seen as a simplified non-probabilistic version of a Fisher Vector (FV) representation. The FV representation is itself an extension of the BoW model resulting from a Maximum Likelihood estimation of the gradient with respect

to the parameters of a Gaussian Mixture that models the log-likelihood of data used to learn the codebook [Jegou et al. 2012].

However, in our case we show in the experimental Section 6.3 that choosing $n > 1$ is advantageous. In some cases, the best results are even obtained with $n = k$. With respect to the vocabulary of a BoW model [Chatfield et al. 2011], we could say that Jegou et al. [2012] uses a *hard coding* ($n = 1$) while we prefer *soft coding* ($n = k$) or possibly *local soft coding* ($1 < n < k$). The benefits of soft coding are well known in the BoW context [Huang et al. 2014] but have not been proven in the context of FV-like signatures (*i.e.* when one uses component-wise gradients with respect to the codebook).

There is also another advantage in our context, where some codewords may be representative of "modality-specific" Voronoi cells, *i.e.* clusters that contain projected points of only one modality after k-means (see Chapter 3). Therefore, by encoding p^T and p^I according to several codewords, it is more likely to include information from both modalities. Hence, the "modality vectors" v^T and v^I are not exactly modality-specific since they benefit from a sort of "modality regularization" with the multimodal codebook. Yet another advantage is that if p^T and p^I are close enough then they certainly share one or several nearest codewords. These codewords will then be enforced by Eq. 6.3 in the final vector v.

All this indicates that the MACC representation is a *soft synthesis* of the contributions of both modalities that compensates for the imperfection of the KCCA projection space in the context of bi-modal tasks.

6.2.2 MACC completion with the missing modality

The MACC representation proposed in the previous section is defined when the multimedia document it describes has both a visual and a textual content. However, this condition does not hold for several important multimedia tasks. This reflects particularly in cross-modal problems, where data in the reference base and/or the query usually come from one modality.

In this section, our contribution consists in extending the original MACC representation method so that it can deal with uni-modal documents. The idea here is to firstly complete

the uni-modal data with suitable information that concerns the missing modality. The completion process is performed with the use of an *auxiliary dataset*, as introduced in Chapter 5. Once the complementary elements have been identified, we consider the initially uni-modal document with its complementary part in the missing modality as a whole bi-modal document. The MACC representation of the uni-modal document is therefore easily estimated following its original process.

We formulate the problem as follows. Let us consider a document with textual content only, described by a feature vector x^T which is projected as p^T onto the KCCA common space. Similarly to the completion process described in Chapter 5, we consider a set $\mathcal{A}$ of pairs of KCCA projections of the visual and textual features of the bi-modal documents from the auxiliary dataset.

$$\mathcal{A} = \{(q^T, q^I)\} \quad with \quad q^T \in \mathcal{A}^T, q^I \in \mathcal{A}^I, |\mathcal{A}| = m$$

Our goal is to construct the bi-modal MACC representation of the document, given only its textual content x^T and the corresponding projection p^T.

In the modality completion stage, we look for the complementary information in the missing modality of an uni-modal document. Since our proposal for completion has shown its effectiveness in the previous contribution in Chapter 5, we also employ it here for modality completion. According to Eq. 6.4, we identify a set $\mathcal{M}_c(p^T)$ of contributors to the "modality completion" of the missing modality of p^T. In the case under study, each element of $\mathcal{M}_c(p^T)$ is a visual projection of a document in the auxiliary dataset on the common space.

$$\mathcal{M}_c(p^T) = \{q_j^I\} \quad \text{such that} \quad \begin{cases} q_j^T \in NN^{\mu}_{\mathcal{A}^T}(p^T) \\ (q_j^T, q_j^I) \in \mathcal{A} \end{cases}$$

In the next stage, we estimate the MACC representation of the initially textual-only document from its original representation p^T and the identified set of complementary information $\mathcal{M}_c(p^T)$. As information provided in p^T and $\mathcal{M}_c(p^T)$ is complementary and related to each other, we consider a document containing p^T together with $\mathcal{M}_c(p^T))$ as an extension in bi-modal content of p^T. Hence, the MACC representation method proposed in 6.2.1 can be absolutely applied for such a bi-modal document.

The only difference is that the document we studied in the previous section always has one visual projection and one textual projection, while in the case we consider here a document is described by the original projection (*e.g.* in the text modality) and *several* complementary projections (*e.g.* in the image modality). In particular, we use μ projections regarding the missing modality for completion. In order to take this difference into account for MACC construction, we propose to first encode the μ complementary projections with respect to the codebook $\mathcal{C}$ and then describe the complementary part by the (element by element) average of these descriptors. The aggregation is always performed using sum pooling of visual and textual representations.

Formally, the MACC representation of the initial textual-only document described by p^T is obtained as

$$\begin{aligned} v =& [v_1, v_2, \ldots, v_i, \ldots, v_k] \quad s.t \\ v_i =& (p^T - c_i)\mathbb{1}_{NN^n(p^T)}(c_i) \\ &+ \frac{1}{\mu} \sum_{q_j^I \in \mathcal{M}_c(p^T)} (q_j^I - c_i)\mathbb{1}_{NN^n(q_j^I)}(c_i) \end{aligned} \tag{6.4}$$

We note that the same development could be symmetrically applied to a document having only visual content.

6.3 Experimental evaluation

To evaluate the effectiveness and the robustness of the proposed MACC representation, we conduct experiments for image classification on Pascal VOC07 and image retrieval on FlickR8K and FlickR30K.

We refer the reader back to Section 2.5 for dataset details regarding the number of images, number of classes, number of tags per image, etc. For content representation, we employ the same features as in previous WCA contribution. This means we use the VGG features [Simonyan and Zisserman 2014] to represent images and Word2Vec [Mikolov et al. 2013] to represent text.

Importantly, we remind that the limitation of the KCCA projections on these three datasets have been highlighted and explained in Chapter 3. In the following, we demonstrate

the effectiveness of MACC representations in improving the "matching" quality between text and image. Our contribution is first evaluated for *bi-modal* and also *cross-modal* classification (introduced in Chapter 5) on Pascal VOC07 in Section 6.3.1. In Section 6.3.2 we then show that MACC establishes a new state of the art in cross-modal retrieval, improving former results on FlickR 8K and FlickR 30K.

6.3.1 Image classification on Pascal VOC07

The KCCA is learnt on the 5011 training data, with both visual and textual content. We used the seminal KCCA implementation [Hardoon et al. 2004], with a regularization parameter $\kappa = 0.1$ and a Gaussian kernel with standard deviation $\sigma = 0.2$. The dimension of the "common" projected space is set to $d = 150$. All 5011 training data are then projected on this common space and a codebook $\mathcal{C}$ is learnt with k-means from this set ($2 \times 5011 = 10022$ points) for $k \in \{8, 16, 32\}$.

6.3.1.1 Classification of bi-modal documents.

The first evaluation considers the classification of documents having both a visual and a textual content, such that a MACC representation (of size $d \times k$) of each document is directly obtained from Eq. 6.3, using the previously built codebook. The parameter n in Eq. 6.3 varies in $\{1, 2, 5, 16, 32\}$. For each category, we learn a SVM classifier with linear kernel, following a one-versus-all strategy.

With such settings, the best result we obtain on the testing set is a mAP of 90.37, with $(k = 16, n = 5)$, resulting into a 2400-dimensional MACC representation. However, when a full cross-validation is conducted on the training set, we obtain a mAP of **90.12** with $(n = 5, k = 32)$.

Table 6.1 compares this performance to other results in the literature. We report superior performance with respect to methods that use only the original (visual) data of the Pascal VOC07 challenge, such as BoVW and Fisher Vectors (FV) [Sánchez et al. 2013; Chatfield et al. 2014]. Our approach also outperforms methods employing additional information sources for training, such as text [Znaidia et al. 2012], ground-truth bounding box information [Dong et al. 2013], or based on deep learning [Perronnin and Larlus 2015;

He et al. 2015b; Chatfield et al. 2014; Wei et al. 2014; Simonyan and Zisserman 2014].

Approach	mAP (%)
BoVW	54.5
FV [Sánchez et al. 2013]	63.9
improved FV[Chatfield et al. 2014]	68.0
BoMW [Znaidia et al. 2012]	67.8
AGS [Dong et al. 2013]	71.1
FV+CNN [Perronnin and Larlus 2015]	76.2
[He et al. 2015b]	82.4
[Chatfield et al. 2014]	82.4
HCP-2000C [Wei et al. 2014]	85.2
VGG NetD&NetE [Simonyan and Zisserman 2014]	89.7
Our MACC	**90.12**

Table 6.1: Pascal VOC07: comparison with published results.

We further compare our image classification result to several baselines that employ the same features as MACC in Table 6.2. For the VGG-Net (respectively Word2Vec) baseline, classifiers are trained and tested on VGG-Net (respectively Word2Vec) features only, *i.e.* using the visual (respectively textual) content alone. For the VGG-Net+Word2Vec baseline, representations for both training and testing data are obtained by early fusion, *i.e.* by concatenating VGG-Net features and Word2Vec features.

For the KCCA_{img} (respectively KCCA_{txt}) baseline, the visual (respectively textual) features are first projected on the KCCA common space for both training and testing data and then used for classifiers learning. We consider two different sizes of the KCCA common space, 150 and 2400, so that the results can be compared to our 2400-dimensional MACC representation (built from a 150-dimensional common space, with 16 codewords). The results in Table 6.2 show that the MACC approach outperforms all the mentioned baselines.

We report in Table 6.3 the results obtained with the MACC approach for different values of k and n (for $d = 150$). We note that the results are quite stable and consistently above the performance of the previously mentioned baselines for this entire range of parameters. Furthermore, these results show that (local) soft coding ($n > 1$) is more effective than hard coding ($n = 1$) to build the MACC representations.

Baseline	Size of representation	mAP (%)
VGG-Net	4096	86.10
Word2Vec	300	82.50
VGG-Net+Word2Vec	4396	86.16
KCCA_{img}	150	84.84
KCCA_{img}	2400	85.29
KCCA_{txt}	150	82.01
KCCA_{txt}	2400	82.60
MACC	2400	**90.12**

Table 6.2: Pascal VOC07: comparison with baselines.

	k=8	k=16	k= 32
n=1	88.75	87.73	86.33
n=2	90.1	89.71	89.18
n=5	89.96	**90.37**	90.10
n=16	-	89.68	90.33
n=32	-	-	89.68

Table 6.3: Pascal VOC07: mAP (%) for different values of k and n.

6.3.1.2 Classification in a cross-modal context.

In this experiment, we investigate the performance of cross-modal classification task introduced in Chapter 5 using the proposed representation MACC.

Let us now consider a scenario where a global resource is available, consisting of a projection space obtained by KCCA and a codebook built on this space. One may wish to train classifiers on new classes, using new data for which only *one* modality is available, and then run these classifiers on other data that may also have only one modality available (and maybe not the same as the one used for training).

Thanks to the completion mechanism (Eq. 6.4), the MACC representation addresses not only classical cross-modal tasks *e.g.* cross-modal retrieval but also such a scenario, that is tested in the following.

As in Chapter 5, we consider the *Text-Image* and *Image-Text* cross-modal classification tasks. In the *Text-Image* task, the SVM classifiers are trained with documents from the training set of Pascal VOC07 but the visual content was removed. Each document, originally described by its textual content alone, has its MACC representation completed

with a visual part following the procedure described in Section 6.2.2, with the training set of PascalVOC 07 chosen as auxiliary dataset $\mathcal{A}$. Hence, the visual part of the signature is not computed from the original visual content of that document but results from combining the contributions of the visual parts of its nearest neighbors according to the textual modality (the document itself is *not* considered among its μ nearest neighbors). The resulting classifiers are then evaluated on the testing documents of Pascal VOC07 but where the textual content was removed and the MACC representations completed following the procedure in Section 6.2.2. The *Image-Text* task is symmetric to the *Text-Image* task: classifiers are trained with documents without textual content and tested on documents without visual content, all being completed according to Eq. 6.4.

In what follows we use the same 150-dimensional projection space obtained by KCCA from the bi-modal training data of Pascal VOC07 and the codebook learnt on this space ($k = 16$) for MACC representations. Consequently, all MACC representations in Table 6.4 are 2400-dimensional vectors.

The results obtained on these two novel tasks are shown in Table 6.4 for several values of the parameter μ and compared to two baselines. For the KCCA_0 baseline, classifiers are trained with the direct projections (without any completion) of one modality on the common KCCA space and tested with the projections of the other modality on this space. For the sake of comparison, the representation of KCCA_0 baseline are issued from the 2400-dimensional KCCA joint space. For the MACC_{rand} baseline, the MACC representation is completed with randomly selected data point along the missing modality.

Without completion ($\mu = 0$), the performance of MACC representations is very low. However, as soon as the completion is considered, the performance is significantly above that of the baseline.

We also compare our MACC representation to the WCA representation proposed in Chapter 5. As in MACC representation, WCA representation also relies on 150-dimensional KCCA common representation space. We set the parameter μ to 15 for WCA on which the MACC representation gets best performance among the different values of μ that we reported in Table 6.4. The MACC representation improves the performance of the WCA for this task, respectively +2.49 and +3.72 for Text-Image and Image-Text tasks.

Representation	mAP (%) Text-Image	mAP (%) Image-Text
$KCCA_0$	71.21	51.20
$MACC_{rand}$	7.76	7.33
WCA($\mu = 15$)	82.06	77.65
MACC($\mu = 0$)	12.03	10.04
MACC($\mu = 1$)	79.00	76.88
MACC($\mu = 3$)	81.72	79.18
MACC($\mu = 5$)	82.17	78.82
MACC($\mu = 8$)	82.18	78.65
MACC($\mu = 15$)	84.55	81.37

Table 6.4: Pascal VOC07: classification in a cross-modal context using the completion mechanism for MACC representations. MACC and WCA representations rely on the 150-dimensional KCCA common representation space.

Lastly, it is not surprising that the results of MACC representation obtained in this cross-modal scenario are not as good as those obtained in the bi-modal task (90.12%, see Table 6.1). However, the difference is not so large and the improvement with respect to the baselines is significant.

6.3.2 Image retrieval on FlickR 8K and FlickR 30K

In this section, we investigate image retrieval task on FlickR 8K and FlickR 30K datasets. This task has been considered in many recently published work such as Socher et al. [2014]; Hodosh et al. [2013]; Karpathy and Fei-Fei [2015]; Chen and Zitnick [2015].

Evaluation protocols. Since a FlickR image is associated to five different sentences (texts), different evaluation methods for image retrieval task on these datasets have been proposed and considered in the literature. In our work, we mainly followed the evaluation protocol introduced by Chen and Zitnick [2015]. This latter aims to return the best ranked image among candidate images retrieved from 5 sentences. Five sentences queries are employed to obtain a corresponding list of candidate images and the image with the best rank is selected as relevant image. Nevertheless, in several experiments, we also tested our proposed representation using the more strict evaluation method proposed by Karpathy and Fei-Fei [2015]. This protocol considers each of five sentences of an image as an individual

Approach	R@1	R@5	R@10
Socher et al. [2014]	6.1	18.5	29
Hodosh et al. [2013]	7.6	20.7	30.1
Karpathy and Fei-Fei [2015]	11.8	32.1	44.7
KCCA$_{(VGG^*+W2V)}$	9.8	27.9	38.1
MACC(VGG*)	**11.9**	**33.1**	**45.8**
Chen and Zitnick [2015]	17.3	42.5	57.4
KCCA$_{(VGG+W2V)}$	26.1	53.7	65.6
MACC(VGG)	**27.6**	**55.6**	**69.4**

Table 6.5: Image retrieval results on FlickR 8K.
Two protocols of evaluation are considered. The first block of results employs the protocol proposed by Karpathy and Fei-Fei [2015] and the second block employs the protocol proposed by Chen and Zitnick [2015]. See text for the details of these evaluation protocols.

query, resulting to 5,000 queries text in the test set. The system must retrieve the image associated to a sentence.

Common representation space setting. In the following experiments for both the FlickR 8K and FlickR 30K benchmarks, the common representation space is learned on the 6,000 FlickR 8K training documents with both visual and textual content. To select the parameters, a grid search is performed employing the validation set of 1,000 documents. This leads to use a Gaussian kernel with a standard deviation $\sigma = 2$, a regularization parameter $\kappa = 1$ and only $d = 50$ dimensions for the projected space. The visual and textual features of the training documents are then all projected on this common space and a codebook is learnt from this set of 12,000 ($= 2 \times 6000$) points.

6.3.2.1 FlickR 8K image retrieval.

For the text-to-image retrieval task the training dataset of FlickR 8K is used as auxiliary dataset $\mathcal{A}$. Parameters being cross-validated on the FlickR 8K training data leads to the choice of $k = n = 32$ and $\mu = 64$.

We compare our proposed approach to several significant work for image retrieval on FlickR 8K. Two blocks of results are considered in Table 6.5, corresponding to two different evaluation protocols. The first group follows the "strict" protocol of Karpathy and Fei-Fei

[2015] while the second group follows the protocol proposed by Chen and Zitnick [2015]. As shown in Table 6.5, the proposed MACC has higher R@1, R@5 and R@10 than the other image retrieval methods in the recent literature on the FlickR 8K dataset for both evaluation methods. We obtain **R@1=11.9%** with the protocol of [Karpathy and Fei-Fei 2015] and **R@1=27.6%** using the evaluation protocol proposed by Chen and Zitnick [2015].

In this table, $\text{KCCA}_{(VGG+W2V)}$ and MACC_{VGG} rely on the KCCA common representation learning with the features VGG-Net [Simonyan and Zisserman 2014] while $\text{KCCA}_{(VGG^*+W2V)}$ and MACC_{VGG^*} employ the more recent features proposed by Tamaazousti et al. [2017]. These features are extracted from a network relied on the deep architecture VGG [Simonyan and Zisserman 2014]. The difference is that the network is learned from larger dataset (*i.e.* ImageNet instead of ILSVRC dataset) on a diversified set of 4,000 categories (1,000 categories for VGG). The weights of the network are initialized by those of the pre-trained VGG network.

The work of Hodosh et al. [2013] was also based on cross retrieval in the KCCA space but their visual and textual representations are simply described by several specific kernels on classical features such as color, texture or GIST descriptors for images, and bag of words for texts. In the $\text{KCCA}_{(VGG^*+W2V)}$ baseline, we apply the image retrieval method of [Hodosh et al. 2013] with our KCCA space built from the visual features [Tamaazousti et al. 2017] and Word2Vec, leading to better performance (9.8%) than Hodosh et al. [2013]. Our method also outperforms several recent deep learning approaches [Socher et al. 2014; Karpathy and Fei-Fei 2015; Chen and Zitnick 2015]. Furthermore, the MACC representation achieves better results than the current state-of-the-art results [Karpathy and Fei-Fei 2015; Chen and Zitnick 2015] on FlickR 8K for the two evaluation methods.

We studied the impact of different coding parameters on the effectiveness of MACC representations. We note that in the following experiments, the performance of our approach is evaluated using the protocol of Chen and Zitnick [2015]. Codebook size being fixed to $k = 64$, Figure 6.2 reports the performance with hard coding ($n = 1$), local soft coding ($1 < n < k$) and soft coding ($n = k$). Since soft coding provides a better location of data points in feature space (with respect to all k codewords, not only to one or to a few of

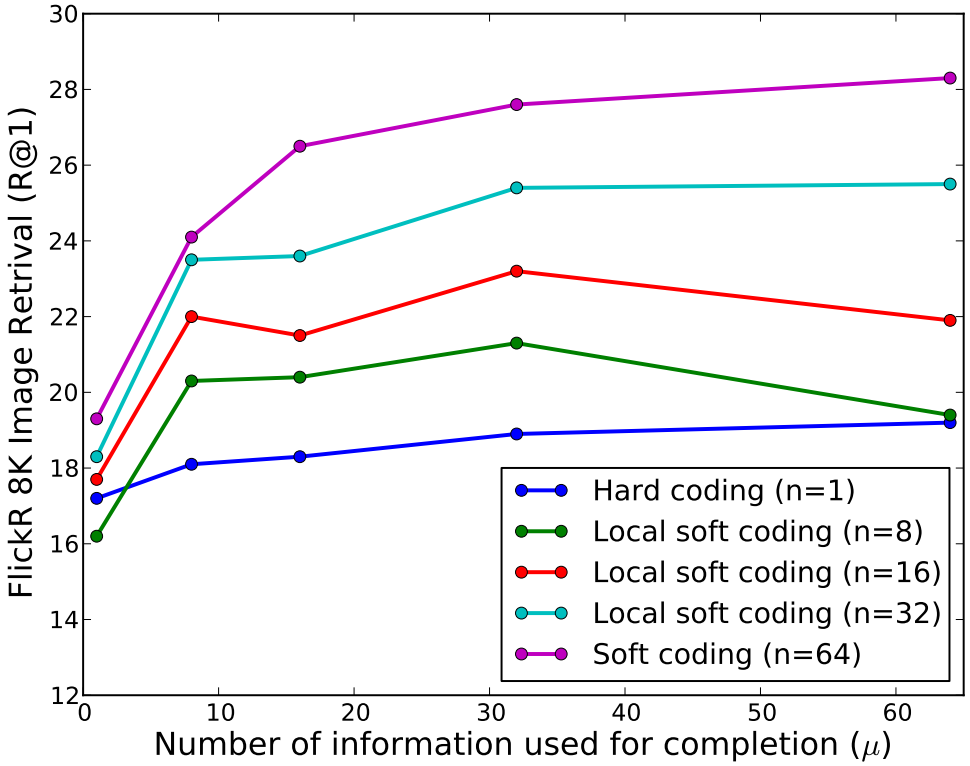

Figure 6.2: Coding methods comparison for MACC representation. Codebook size is fixed at $k = 64$. Evaluation follows the protocol of Chen and Zitnick [2015].

them), it usually performs better for retrieval.

The most important result is nevertheless that our method achieves better performance than the state-of-the-art [Chen and Zitnick 2015] as soon as $\mu > 10$. On the FlickR 8K benchmark, the performances are quite stable with any given coding scheme for $\mu > 20$.

In a third experiment we study the stability of our approach with regard to k, n and μ. Figure 6.3 reports performance on FlickR 8K while varying these parameters. Following the conclusion of the second experiment, soft coding ($n = k$) is employed in this experiment for its effectiveness. The results firstly show that even when the size of codebook and resulting MACC representations is very small, we consistently achieve better performance than the other methods in Table 6.5. For instance, our approach has a first rank recall (R@1) of

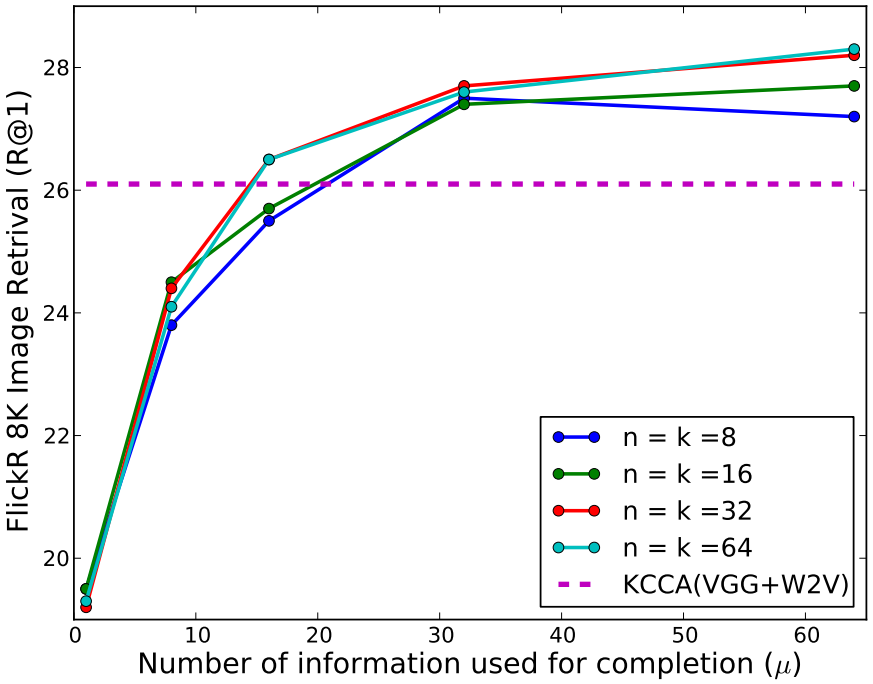

Figure 6.3: FlickR 8K image retrieval: stability of MACC representations over a large range of parameters. Evaluation follows the protocol of Chen and Zitnick [2015].

18.5% with k as low as 8 (the corresponding MACC representation is only 400-dimensional).

Besides, an interesting observation is that with a sufficiently large number μ of contributors to MACC completion, the proposed approach yields superior performance over the robust KCCA$_{(VGG+W2V)}$ baseline regardless of the size of the codebook. These results show the stability of our approach over a large range of parameters.

6.3.2.2 FlickR 30K: benefit of auxiliary dataset.

To study the impact of the auxiliary dataset $\mathcal{A}$ used for MACC completion in cross-modal tasks, we conducted an experiment on FlickR 30K that has the same validation and testing sets as FlickR 8K but a larger training set. The experimental protocol was the same as for FlickR 8K (same KCCA space and codebook) except for the choice of $\mathcal{A}$, where we used the full training set of FlickR 30K ($29,783$ images) instead of the training set of FlickR 8K ($6,000$ images).

Approach	R@1	R@5	R@10
Socher et al. [2014]	8.9	29.8	41.1
Karpathy and Fei-Fei [2015]	15.2	37.7	50.5
Our MACC (F8K)	14.5	35.8	48.8
Chen and Zitnick [2015]	18.5	45.7	58.1
Our MACC (F8K)	**33.9**	**65.6**	**77.5**
Our MACC (F30K)	**35.3**	**66.0**	**78.2**

Table 6.6: Image retrieval results on FlickR 30K.
MACC parameters are cross-validated on FlickR 8K (F8K) or FlickR 30k (F30K).

Following the evaluation method proposed by Chen and Zitnick [2015], the MACC approach yields a significant improvement of 6 points (from 27.6% to 33.9%) on 1,000 testing data, which is thus due to the larger auxiliary dataset. The improvement increases when the parameters are cross-validated on FlickR30k training set (35.3%). While the improvement in the previous state-of-the-art [Chen and Zitnick 2015] is from 17.3% on FlickR 8k to 18.5% on FlickR 30K, in our case it is from 27.6% to 33.9% using the parameters cross-validated on Flickr 8K data. This result shows a better use of the extended training dataset, at a limited cost as KCCA and the codebook are always computed on FlickR 8K.

In Table 6.6, we also compare our MACC approach with other previous methods following the evaluation proposed by Karpathy and Fei-Fei [2015]. Our approach significantly outperforms recent methods such as Socher et al. [2014]. However, the performance obtained with MACC is weaker (14.5% with respect to 15.2% at R@1) than the current state-of-the-art [Karpathy and Fei-Fei 2015]. In Karpathy and Fei-Fei [2015], each image is described by a set of 20 vectors corresponding to the whole image and the top 19 most relevant image regions detected using a CNN pre-trained on ImageNet and finetuned on the 200 classes of the ImageNet Detection Challenge. These visual features are computationally costly. Our approach reports the performance using the description of the whole image which is lighter yet yields competitive results with respect to the state-of-the-art.

6.4 Conclusion

In this chapter, we proposed a new representation, called Multimedia Aggregated Correlated Components (MACC) to describe a multimedia document including both visual and textual content. The MACC representation aggregates information provided by the projections of both modalities, *e.g.* image and text, on the joint space. Especially, projections are encoded relatively to several codewords of a vocabulary learnt on the common space before their aggregation.

Furthermore, we extended the MACC representation method to support the uni-modal document where only visual or textual content is available. For this purpose, we re-employed the idea of data completion onto the common space using auxiliary data proposed in the previous contribution of Chapter 5.

The proposed representation approaches can address different multimedia tasks such as bi-modal or cross-modal classification and cross-modal retrieval. Experimental results show that the MACC brought improvement in term of performance in comparison with a direct use of data projections for these last tasks. This shows that our proposed approaches successfully reduced the gap between two modality onto the common space, which is one of the most important shortcomings of such common representation based approach.

The interest of our approach was demonstrated in bi-modal classification, cross-modal classification on Pascal VOC07 data and in cross-modal retrieval on FlickR 8K/30K data. In these benchmarks, our method provides state-of-the-art performances.

Chapter 7

Conclusion and Perspectives

7.1 Conclusion

This thesis addresses the joint modeling of visual and textual modalities for cross-modal problems. We approached this research topic by relying on learning a common representation space for both images and text, then derived representations that are more robust than the "naive" approach. The first aim consists in improving the quality of "matching" across modalities on the joint space. We also consider to reduce the computational cost of such multimedia representations for different tasks.

Our approach was to explicitly identify limitations of the common representation space which might hamper performance in multimedia tasks. In the first contribution described in Chapter 3, two such limitations have been identified. The first one concerns relevant data that is ignored in the joint space. In the case of cross modal retrieval, this data consists of words that are present in the reference database (or the queries) of the target task but absent from (or very rare in) the training database used to learn the joint space. The second limitation concerns the separation of projected data between visual and textual modalities on the common space. These projections tend to be grouped by modality rather than according to their semantics. This results in a limited quality of the matching across modalities. This quality has been quantified in terms of average distance between the projected points of both modalities, with regard to the intra-modality distances of these points.

Consequently, we proposed different models that work on the joint space of images and text to manage these identified limitations. The proposed models aim to explicitly improve the quality of matching across modalities and, consequently, enhance the performance of bi-modal and cross-modal tasks by reducing the gaps existing between the visual and textual modalities. Three contributions have been proposed to address this research problem:

The second contribution deals with the relevant information that is poorly represented on the joint space. In a retrieval task, such a piece of data, called *specific* information, corresponds to words that are present in the reference data but absent from (or very rare in) the training data. We put forward a model to combine them with *generic* information that is relatively well represented on the joint space. The proposed models support both uni-

modal and cross-modal retrieval tasks. Different experiments have been conducted on the challenging text illustration task. In this case, "specific" information mostly concerns names, trademarks or other very informative tags. Obtained results showed that by appropriately identifying and taking such "specific" information into account, the performance in cross-modal retrieval can be significantly improved. These results also show that in a realistic case where the difference between the training set used for common space learning and the reference set is important, our models are more effective with respect to the direct use of the projections onto the joint space.

The main work in our third contribution considers *cross-modal classification*. Our goal is to design a conceptual multimedia model at a higher semantic level by matching a given document to "more general concepts" resulting from a set of other documents. At the first step, we investigate the cross-modal classification consisting in training classifiers on data from one modality *e.g.* text and applying them to predict data from another modality *e.g.* image. To address this task, we proposed a method called Weighted Completion with Averaging (WCA) to build a robust representation accounting for both the visual and the textual information of a uni-modal document. At the core of this contribution we use a bi-modal dataset, called *auxiliary dataset*, that acts as a set of connections between the modalities within the joint space. We suggest to rely on this auxiliary dataset to find a complementary information in the missing modality of a uni-modal document. It leads to build a point representing the complementary modality of a given data point, thus we obtain a complete bi-modal WCA representation of a initially uni-modal data. Experiments have been conducted on well-known datasets [Hwang and Grauman 2012a; Chua et al. 2009] and showed that the WCA representation method significantly improves the cross-modal classification performance compared to the use of a latent space alone. Furthermore, WCA provides interesting results compared to recent cross-modal *retrieval* methods. It is important to note that the performance level we attain on cross-modal classification also compares well to state-of-the-art uni-modal and bi-modal classification results. Such a performance level makes our approach to cross-modal classification a convincing choice for real applications, such as learning classifiers from an existing large amount of annotated textual data and applying them to visual content.

Representation	$d_{\text{intra}}(I)$	$d_{\text{intra}}(T)$	$d_{\text{inter}}(\text{sample})$	$d_{\text{inter}}(\text{overall})$
KCCA (150)	1.18 ± 0.16	1.11 ± 0.19	1.39 ± 0.07	1.42 ± 0.06
WCA (150)	1.17 ± 0.15	1.16 ± 0.16	0.07 ± 0.15	1.17 ± 0.15
MACC (2400)	1.16 ± 0.13	1.15 ± 0.15	0.81 ± 0.13	1.16 ± 0.13

Table 7.1: Average Euclidean distances between image and text representations on Pascal VOC07 data. All representations are calculated on the 150-dimensional KCCA spaces. $k = 16, n = 5, \mu = 15$ for MACC and $\mu = 15$ for WCA.

In the fourth contribution, we put forward a robust representation method, called Multimedia Aggregated Correlated Components (MACC) that aggregates information provided by the projections of both visual and textual modalities on their joint space. MACC representation reduces the separation between the projections of visual and textual features by embedding them in a local context reflecting the data distribution in the common space. More precisely, we learn a codebook on the joint space using projections of all visual and textual features in the training dataset. For each bi-modal document, both visual and textual features are projected on this common space, then coded using the codebook and aggregated into a single MACC vector that is the multimedia representation of the document. This representation can be extended for uni-modal documents. In this case, the uni-modal completion processing relying on an auxiliary dataset (introduced in the second contribution) is performed to suggest the corresponding complementary information in the missing modality of the document. The extensive experimental evaluation on three challenging datasets [Hwang and Grauman 2012b; Rashtchian et al. 2010; Young et al. 2014] shows that our proposed MACC representation allows to reach state-of-the-art performance in various multimedia tasks such as bi-modal and cross-modal classification and image retrieval.

It is important noting that the proposed WCA and MACC representations have succeeded in narrowing the separation between modalities on the common representation space. To illustrate this fact, Table 7.1, 7.2 compares the average (intra-modal and inter-modal) Euclidean distances for different representation methods including the direct KCCA projection, WCA and MACC on Pascal VOC07 and FlickR 8K datasets. The average distances concerning on KCCA projections in these tables are extracted from the Table 3.1 (in Chapter 3) where we demonstrated the limitation about the separation between the

Representation	$d_{\text{intra}}(I)$	$d_{\text{intra}}(T)$	$d_{\text{inter}}(\text{sample})$	$d_{\text{inter}}(\text{overall})$
KCCA (50)	1.17 ± 0.13	0.75 ± 0.13	1.02 ± 0.12	1.28 ± 0.10
MACC (1600)	0.89 ± 0.08	0.81 ± 0.09	0.66 ± 0.11	0.87 ± 0.07

Table 7.2: Average Euclidean distances between image and text representations on 1000 FlickR 8K testing data. All representations are calculated on the 50-dimensional KCCA spaces. $k = n = 32, \mu = 64$ for MACC.

projections of corresponding visual and textual features on the joint space. Table 7.1 reports the results on 5011 Pascal VOC07 training data. On the 150-dimensional KCCA space, the average distance between an image projection and its corresponding textual projection is reduced (from 1.39 with KCCA projection to 0.07 with WCA representation). Also, WCA narrows the gap between the two clouds of visual and textual projections on their joint space (from 1.42 with KCCA projection to 1.17 with WCA). While KCCA and WCA representations reside in the same KCCA, MACC is represented in a different higher-dimensional space. The average distances reported on MACC are thus not directly comparable to those of KCCA and WCA methods. The average distances of projections within modalities and across modalities ($d_{inter}(overall)$) are quite similar, showing there is no separation between modalities. Furthermore, the average distance between visual and corresponding textual projection is smaller than the intra-modality and inter-modality ($d_{inter}(overall)$) distances. This indicates that MACC makes closer visual and textual representations of a document, resulting to a better quality of "matching" across these modalities and thus enhance the performance of classification tasks. The results on FlickR 8K data are shown in Table 7.2. Similarly, on the representation space of MACC, there is no particular separation across modalities *e.g.* the intra-modal and inter-modal distances are quite similar. More importantly, the inter-modal average distance ($d_{inter}(sample)$) between a visual point and its associated textual point is smaller than the intra-modal distances. This means that on the MACC representation space, the nearest neighbour of a visual point is its associated textual point and vice versa. This fact explains the improvement on retrieval performance obtained with the MACC representation on the FlickR 8K dataset.

7.2 Perspectives for future research

Three contributions were proposed in this dissertation for effectively and efficiently combining visual and textual modalities to address various bi-modal or cross-modal retrieval and classification problems. Without any doubt, these solutions can be improved in several ways. To conclude the thesis, this section discusses promising directions for future research related to the presented work. We divide these ideas into short-term and longer-term perspectives.

7.2.1 Short-term perspectives

Robust criterion for "specific" information identification. In the second contribution, the method we proposed aims to make better use of specific information that is poorly represented by a learned cross-modal model but nevertheless likely to be relevant for retrieval. In the text illustration experiments presented above, this information corresponds to words that are present in the test data but absent from the training data (or below the filtering threshold). This selection condition may appear still weak to distinguish specific information from noise. It is possible to make use of *e.g.* domain-specific rules to further improve this selection.

Multi-modal retrieval model for combining "specific" and "generic" information. Also in this second contribution, two retrieval models *e.g.* Eq 4.4 (uni-modal) and Eq 4.3 (cross-modal) were proposed for evaluating respectively query-caption and query-image similarities. These models consider only one media *e.g.* image or text of the reference bi-modal document containing at the same time image and its caption. Another interesting direction is to propose a model that exploits simultaneously visual and textual content of the reference document to enhance the retrieval performance. This can be done simply by performing a fusion between the results of two presented models or finding a suitable way to incorporate query-image and query-caption similarities on the joint space and then combine them with the similarity on their specific part.

MACC accounting "degree of similarity" with respect to codewords. In the MACC representation method, μ nearest codewords were used and contributed equally for the resulting bi-modal representation 6.3. However, one of the important aspects of soft coding is that it can take into consideration the "degree of similarity" of a projection point p with respect to different codewords. For further work, it is interesting to consider such information for MACC encoding *e.g.* by adding different weights for codewords in Eq 6.3.

Compact auxiliary set for uni-modal completion. In WCA and MACC method, we put forward a completion method to find the complementary information in the missing modality of a uni-modal data and then use it to build the final bi-modal representation. The completion process relies on an auxiliary bi-modal dataset that acts as a set of connections between the modalities within the joint space. In the presented work, we investigated different choices of such an auxiliary dataset for a target problem *e.g.* auxiliary data and target data belonging to the same dataset or totally different. Nevertheless, we usually employed the entire (training) set of a dataset as auxiliary data. An interesting perspective consists in the selection of a relevant sample of auxiliary data from this entire dataset. In other words, our goal is to restrict the number of data in the auxiliary set to reduce its size while maintaining the performance of target problems. To this end, the definition of a robust criterion for choosing data for such an auxiliary sample is crucial. A possibility is to rely on the quality of representation (mentioned in Section 3.3.1) of an auxiliary individual onto the common representation space for such a selection.

7.2.2 Longer-term perspectives

Generic methods for other cross-modal embedding techniques. An important and interesting direction related to this work concerns common representation space for visual and textual modalities. In this thesis, the joint spaces were built relying on (Kernel) Canonical Correlation Analysis (KCCA). However, the WCA and MACC representations are not specific to this particular type of joint space and we believe that they can be potentially used with any text-image embedding technique. It is thus interesting to investigate the effectiveness of our proposed models for other types of text-image joint

representations, particularly on recent embedding spaces issued from deep learning proposed by *e.g.* Yan and Mikolajczyk [2015]; Vukotić et al. [2016]. In the first step of this research, we investigated the MACC representation method on the image-text joint space issued from the Deep Canonical Correlation Analysis (DCCA) method [Andrew et al. 2013; Yan and Mikolajczyk 2015]. This latter consists of two deep networks corresponding to visual and textual modalities where the output layers (topmost layer of each network) are maximally correlated. Results obtained on this study can be found in Chami [2016].

Domain transfer. In several of the proposed methods we also considered the issue of a "universal resource". The aim is to learn a joint space for image and text on a large generic collection of multimedia documents and then apply this resource to address different target problems. However, it is challenging to directly apply the generic resource to a target problem since data from different collections (here, generic data and target data) has different distributions and properties. An interesting research direction consists in investigating the problem of "domain transfer" between such data. The goal is to better relate the generic resource and the target problem in order to enhance the performance of task in such a cross-domain context. A possible extension of this direction concerns the "domain transfer" problem between auxiliary data that we used to refine the image-text connection within the joint space and the target data and/or the generic data.

Extension for multi-lingual problems. In this thesis, we mainly consider the cross-modal problems between image and text. Nevertheless, the proposed approaches can be generic in multi-lingual settings. There, two different languages could be observed as two different modalities, and the entire cross-modal framework is exactly the same as proposed in our contributions. Particularly, it is possible to make the connection between the proposed cross-modal classification framework and the cross-lingual document classification scenario, which is also called cross-lingual knowledge transfer. Furthermore, it is interesting to study a unified framework for both cross-modal and cross-lingual classification settings which can be seen as zero-shot learning tasks.

Bibliography

Ahsan, U. and Essa, I. (2014). Clustering social event images using kernel canonical correlation analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 800–805.

Amir, A., Argill, J. O., Berg, M., fu Chang, S., Franz, M., Hsu, W., Iyengar, G., Kender, J. R., Kennedy, L., yung Lin, C., Naphade, M., (paul Natsev, A., Smith, J. R., Tesic, J., Wu, G., Zhang, D., Watson, I. T. J., and Watson, I. T. J. (2004). Ibm research trecvid-2004 video retrieval system. In *In Proc. of TREC Video Retrieval Evaluation*. Publications.

Andrew, G., Arora, R., Bilmes, J. A., and Livescu, K. (2013). Deep canonical correlation analysis. In *ICML (3)*, pages 1247–1255.

Avila, S., Thome, N., Cord, M., Valle, E., and AraúJo, A. D. A. (2013). Pooling in image representation: The visual codeword point of view. *Computer Vision and Image Understanding*, 117(5):453–465.

Barnard, K., Duygulu, P., Forsyth, D., de Freitas, N., Blei, D. M., and Jordan, M. I. (2003). Matching words and pictures. *J. Mach. Learn. Res.*, 3:1107–1135.

Blei, D. M. and Jordan, M. I. (2003). Modeling annotated data. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*, SIGIR '03, pages 127–134, New York, NY, USA. ACM.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Bosch, A., Zisserman, A., and Munoz, X. (2007). Representing shape with a spatial pyramid kernel. In *Proceedings of the 6th ACM International Conference on Image and Video Retrieval*, CIVR '07, pages 401–408, New York, NY, USA. ACM.

Bottou, L. (2010). Large-scale machine learning with stochastic gradient descent. In Lechevallier, Y. and Saporta, G., editors, *Proceedings of the 19th International Conference on Computational Statistics (COMPSTAT'2010)*, pages 177–187, Paris, France. Springer.

Boureau, Y.-L., Ponce, J., and Lecun, Y. (2010). A theoretical analysis of feature pooling in visual recognition. In *ICML*, Haifa, Israel.

Bruni, E., Tran, N. K., and Baroni, M. (2014). Multimodal distributional semantics. *J. Artif. Int. Res.*, 49(1):1–47.

Chami, I. (2016). Représentation commune des textes et des images. *Rapport de stage de master 1.*

Chatfield, K., Lempitsky, V., Vedaldi, A., and Zisserman, A. (2011). The devil is in the details: an evaluation of recent feature encoding methods. In *British Machine Vision Conference.*

Chatfield, K., Simonyan, K., Vedaldi, A., and Zisserman, A. (2014). Return of the devil in the details: Delving deep into convolutional nets. *arXiv preprint arXiv:1405.3531.*

Chen, X., Mu, Y., Yan, S., and Chua, T.-S. (2010). Efficient large-scale image annotation by probabilistic collaborative multi-label propagation. In *Proceedings of the 18th ACM International Conference on Multimedia*, pages 35–44, NY, USA.

Chen, X. and Yuille, A. (2014). Articulated pose estimation by a graphical model with image dependent pairwise relations. In *Advances in Neural Information Processing Systems (NIPS).*

Chen, X. and Zitnick, L. C. (2015). Mind's eye: A recurrent visual representation for image caption generation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*

Chua, T.-S., Tang, J., Hong, R., Li, H., Luo, Z., and Zheng, Y.-T. (July 8-10, 2009). NUS-WIDE: A real-world web image database from National University of Singapore. In *Proc. of ACM Conference on Image and Video Retrieval (CIVR'09)*, Santorini, Greece.

Coelho, F. and Ribeiro, C. (2011). Automatic illustration with cross-media retrieval in large-scale collections. In *Content-Based Multimedia Indexing (CBMI), 2011 9th International Workshop on*, pages 25–30. IEEE.

Costa Pereira, J., Coviello, E., Doyle, G., Rasiwasia, N., Lanckriet, G., Levy, R., and Vasconcelos, N. (2014). On the role of correlation and abstraction in cross-modal multimedia retrieval. *TPAMI*, 36(3):521–535.

Csurka, G., Bray, C., Dance, C., and Fan, L. (2004a). Visual categorization with bags of keypoints. *Workshop on Statistical Learning in Computer Vision, ECCV*, pages 1–22.

Csurka, G., Dance, C. R., Fan, L., Willamowski, J., and Bray, C. (2004b). Visual categorization with bags of keypoints. In *In Workshop on Statistical Learning in Computer Vision, ECCV*, pages 1–22.

Datta, R., Joshi, D., Li, J., and Wang, J. Z. (2008). Image retrieval: Ideas, influences, and trends of the new age. *ACM Computing Surveys (CSUR)*, 40(2):5.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990). Indexing by latent semantic analysis. *JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE*, 41(6):391–407.

Delgado, D., Magalhaes, J., and Correia, N. (2010). Assisted news reading with automated illustration. In *Proceedings of the 18th ACM International Conference on Multimedia*, MM '10, pages 1647–1650, New York, NY, USA. ACM.

Dong, J., Xia, W., Chen, Q., Feng, J., Huang, Z., and Yan, S. (2013). Subcategory-aware object classification. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 827–834. IEEE.

Dosovitskiy, A., Fischer, P., Ilg, E., Hausser, P., Hazirbas, C., Golkov, V., van der Smagt,

P., Cremers, D., and Brox, T. (2015). Flownet: Learning optical flow with convolutional networks. In *The IEEE International Conference on Computer Vision (ICCV)*.

Everingham, M., Van Gool, L., Williams, C. K., Winn, J., and Zisserman, A. (2010). The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338.

Feng, F., Wang, X., and Li, R. (2014). Cross-modal retrieval with correspondence autoencoder. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 7–16. ACM.

Feng, Y. and Lapata, M. (2010). Topic models for image annotation and text illustration. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 831–839, Stroudsburg, PA, USA. Association for Computational Linguistics.

Fischer, P., Dosovitskiy, A., and Brox, T. (2014). Descriptor matching with convolutional neural networks: a comparison to SIFT. *CoRR*, abs/1405.5769.

Frome, A., Corrado, G. S., Shlens, J., Bengio, S., Dean, J., Mikolov, T., et al. (2013). Devise: A deep visual-semantic embedding model. In *Advances in neural information processing systems*, pages 2121–2129.

Gabrilovich, E. and Markovitch, S. (2007). Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *IJcAI*, volume 7, pages 1606–1611.

Girshick, R. B., Donahue, J., Darrell, T., and Malik, J. (2013). Rich feature hierarchies for accurate object detection and semantic segmentation. *CoRR*, abs/1311.2524.

Gong, Y., Ke, Q., Isard, M., and Lazebnik, S. (2014). A multi-view embedding space for modeling internet images, tags, and their semantics. *IJCV*, 106(2):210–233.

Hardoon, D. R., Szedmak, S. R., and Shawe-Taylor, J. R. (2004). Canonical correlation analysis: An overview with application to learning methods. *Neural Comput.*, 16(12):2639–2664.

He, K., Zhang, X., Ren, S., and Sun, J. (2015a). Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *IEEE International Conference on Computer Vision (ICCV 2015)*.

He, K., Zhang, X., Ren, S., and Sun, J. (2015b). Spatial pyramid pooling in deep convolutional networks for visual recognition. *TPAMI*, 37(9):1904–1916.

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

Hodosh, M., Young, P., and Hockenmaier, J. (2013). Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47:853–899.

Hofmann, T. (1999). Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 50–57. ACM.

Hofmann, T. (2001). Unsupervised learning by probabilistic latent semantic analysis. *Mach. Learn.*, 42(1-2):177–196.

Hotelling, H. (1936). Relations between two sets of variates. *Biometrika*, 28(3/4):321–377.

Huang, Y., Wu, Z., Wang, L., and Tan, T. (2014). Feature coding in image classification: A comprehensive study. *TPAMI*, 36(3):493–506.

Hwang, S. J. and Grauman, K. (2012a). Learning the relative importance of objects from tagged images for retrieval and cross-modal search. *IJCV*, 100(2):134–153.

Hwang, S. J. and Grauman, K. (2012b). Reading between the lines: Object localization using implicit cues from image tags. *TPAMI*, 34(6):1145–1158.

Ioffe, S. and Szegedy, C. (2015). Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *JMLR Proceedings*, pages 448–456. JMLR.org.

Jégou, H., Douze, M., Schmid, C., and Pérez, P. (2010). Aggregating local descriptors into a compact image representation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2010)*, pages 3304–3311. IEEE.

Jegou, H., Perronnin, F., Douze, M., Sánchez, J., Perez, P., and Schmid, C. (2012). Aggregating local image descriptors into compact codes. *IEEE Trans. Pattern Anal. Mach. Intell.*, 34(9):1704–1716.

Jia, Y., Salzmann, M., and Darrell, T. (2011). Learning cross-modality similarity for multinomial data. In *2011 International Conference on Computer Vision*, pages 2407–2414. IEEE.

Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., and Darrell, T. (2014). Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the 22Nd ACM International Conference on Multimedia*, MM '14, pages 675–678, New York, NY, USA. ACM.

Joly, A. and Buisson, O. (2011). Random maximum margin hashing. In *The 24th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2011, Colorado Springs, CO, USA, 20-25 June 2011*, pages 873–880. IEEE.

Joshi, D., Wang, J. Z., and Li, J. (2004). The story picturing engine: finding elite images to illustrate a story using mutual reinforcement. In *Proceedings of the 6th ACM SIGMM international workshop on Multimedia information retrieval*, pages 119–126. ACM.

Joshi, D., Wang, J. Z., and Li, J. (2006). The story picturing engine—a system for automatic text illustration. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 2(1):68–89.

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*, pages 3128–3137.

Karpathy, A., Joulin, A., and Li, F. F. F. (2014). Deep fragment embeddings for bidirectional image sentence mapping. In *Advances in neural information processing systems*, pages 1889–1897.

Klein, B., Lev, G., Sadeh, G., and Wolf, L. (2015). Associating neural word embeddings with deep image representations using fisher vectors. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States.*, pages 1106–1114.

Lazebnik, S., Schmid, C., and Ponce, J. (2006). Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2*, CVPR '06, pages 2169–2178, Washington, DC, USA. IEEE Computer Society.

Li, A., Shan, S., Chen, X., and Gao, W. (2011). Face recognition based on non-corresponding region matching. In *2011 International Conference on Computer Vision*, pages 1060–1067. IEEE.

Li, H., Li, Y., and Porikli, F. (2014). Robust online visual tracking with an single convolutional neural network. In *Asian Conference on Computer Vision (ACCV)*, pages 1–16, Singapore.

Li, Y., Crandall, D. J., and Huttenlocher, D. P. (2009). Landmark classification in large-scale image collections. In *IEEE 12th International Conference on Computer Vision, ICCV 2009, Kyoto, Japan, September 27 - October 4, 2009*, pages 1957–1964.

Liu, L., Wang, L., and Liu, X. (2011). In defense of soft-assignment coding. In *Proceedings of the 2011 International Conference on Computer Vision*, ICCV '11, pages 2486–2493, Washington, DC, USA. IEEE Computer Society.

Liu, N., DellandréA, E., Chen, L., Zhu, C., Zhang, Y., Bichot, C.-E., Bres, S., and Tellez, B. (2013). Multimodal recognition of visual concepts using histograms of textual concepts and selective weighted late fusion scheme. *Comput. Vis. Image Underst.*, 117(5):493–512.

Liu, Y., Zhang, D., Lu, G., and Ma, W.-Y. (2007). A survey of content-based image retrieval with high-level semantics. *Pattern Recogn.*, 40(1):262–282.

Long, J., Shelhamer, E., and Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60(2):91–110.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

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

Monay, F. and Gatica-Perez, D. (2007). Modeling semantic aspects for cross-media image indexing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(10):1802–1817.

Müller, H., Clough, P., Deselaers, T., and Caputo, B. (2010). *ImageCLEF: Experimental Evaluation in Visual Information Retrieval*. Springer Publishing Company, Incorporated, 1st edition.

Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., and Ng, A. Y. (2011). Multimodal deep learning. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 689–696.

Novak, D., Batko, M., and Zezula, P. (2015). Large-scale image retrieval using neural net descriptors. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, Santiago, Chile, August 9-13, 2015*, pages 1039–1040.

Over, P., Fiscus, J., Sanders, G., Joy, D., Michel, M., Awad, G., Smeaton, A., Kraaij, W., and Quénot, G. (2014). Trecvid 2014–an overview of the goals, tasks, data, evaluation mechanisms and metrics. In *Proceedings of TRECVID*, page 52.

Pepik, B., Benenson, R., Ritschel, T., and Schiele, B. (2015). What is holding back convnets for detection? *CoRR*, abs/1508.02844.

Perronnin, F. and Dance, C. (2007). Fisher kernels on visual vocabularies for image categorization. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE.

Perronnin, F. and Larlus, D. (2015). Fisher vectors meet neural networks: A hybrid classification architecture. In *CVPR*.

Perronnin, F., Sánchez, J., and Liu, Y. (2010a). Large-scale image categorization with explicit data embedding. pages 2297–2304.

Perronnin, F., Sánchez, J., and Mensink, T. (2010b). Improving the fisher kernel for large-scale image classification. In *Proceedings of the 11th European Conference on Computer Vision: Part IV*, ECCV'10, pages 143–156, Berlin, Heidelberg. Springer-Verlag.

Putthividhy, D., Attias, H. T., and Nagarajan, S. S. (2010). Topic regression multi-modal latent dirichlet allocation for image annotation. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 3408–3415. IEEE.

Ranjan, V., Rasiwasia, N., and Jawahar, C. (2015). Multi-label cross-modal retrieval. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4094–4102.

Rashtchian, C., Young, P., Hodosh, M., and Hockenmaier, J. (2010). Collecting image annotations using amazon's mechanical turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, CSLDAMT '10, pages 139–147, Stroudsburg, PA, USA. Association for Computational Linguistics.

Rasiwasia, N., Mahajan, D., Mahadevan, V., and Aggarwal, G. (2014). Cluster canonical correlation analysis. In *AISTATS*, pages 823–831.

Rasiwasia, N. and Vasconcelos, N. (2008). Scene classification with low-dimensional semantic spaces and weak supervision. In *CVPR*, pages 1–6.

Rasiwasia, N. and Vasconcelos, N. (2009). Holistic context modeling using semantic co-occurrences. In *CVPR*, volume 0, pages 1889–1895, Los Alamitos, CA, USA.

Razavian, A. S., Azizpour, H., Sullivan, J., and Carlsson, S. (2014). Cnn features off-the-shelf: an astounding baseline for recognition. *CoRR*, abs/1403.6382.

Robertson, S., Zaragoza, H., et al. (2009). The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. (2015). ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252.

Salton, G. and McGill, M. J. (1986). *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc., New York, NY, USA.

Sánchez, J., Perronnin, F., Mensink, T., and Verbeek, J. (2013). Image classification with the fisher vector: Theory and practice. *IJCV*, 105(3):222–245.

Schroff, F., Kalenichenko, D., and Philbin, J. (2015). Facenet: A unified embedding for face recognition and clustering. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., and LeCun, Y. (2013). Overfeat: Integrated recognition, localization and detection using convolutional networks. In *International Conference on Learning Representations (ICLR 2014)*, page 16. CBLS.

Sharma, A., Kumar, A., Daume, H., and Jacobs, D. W. (2012). Generalized multiview analysis: A discriminative latent space. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2160–2167. IEEE.

Shen, H. T., Ooi, B. C., and Tan, K.-L. (2000). Giving meanings to www images. In *Proceedings of the eighth ACM international conference on Multimedia*, pages 39–47. ACM.

Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556.

Sivic, J. and Zisserman, A. (2003). Video Google: A text retrieval approach to object matching in videos. In *Proceedings of the International Conference on Computer Vision*, volume 2, pages 1470–1477.

Smeaton, A. F., Over, P., and Kraaij, W. (2006). Evaluation campaigns and trecvid. In *Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval*, MIR '06, pages 321–330, New York, NY, USA. ACM.

Smeulders, A. W. M., Worring, M., Santini, S., Gupta, A., and Jain, R. (2000). Content-based image retrieval at the end of the early years. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(12):1349–1380.

Snoek, C. G., Worring, M., and Smeulders, A. W. (2005). Early versus late fusion in semantic video analysis. In *Proceedings of the 13th annual ACM international conference on Multimedia*, pages 399–402. ACM.

Socher, R., Karpathy, A., Le, Q. V., Manning, C. D., and Ng, A. Y. (2014). Grounded compositional semantics for finding and describing images with sentences. *Transactions of the Association for Computational Linguistics*, 2:207–218.

Srihari, R. K., Zhang, Z., and Rao, A. (2000). Intelligent indexing and semantic retrieval of multimodal documents. *Information Retrieval*, 2(2-3):245–275.

Srivastava, N. and Salakhutdinov, R. R. (2012). Multimodal learning with deep boltzmann machines. In *Advances in neural information processing systems*, pages 2222–2230.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9.

Tamaazousti, Y., Le Borgne, H., Popescu, A., Gadeski, E., Ginsca, A., and Hudelot, C. (2017). Vision-language integration using constrained local semantic features. *Journal of Computer Vision and Image Understanding*. Pending revision.

Tammazousti, Y., Le Borgne, H., and Popescu, A. (2016). Constrained local enhancement of semantic features by content-based sparsity. In *International Conference on Multimedia Retrieval*.

Tran, T. Q. N., Le Borgne, H., and Crucianu, M. (2015). Combining generic and specific information for cross-modal retrieval. In *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*, ICMR '15, pages 551–554, New York, NY, USA. ACM.

Tran, T. Q. N., Le Borgne, H., and Crucianu, M. (2016a). Aggregating image and text quantized correlated components. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Tran, T. Q. N., Le Borgne, H., and Crucianu, M. (2016b). Cross-modal classification by completing unimodal representations. In *Proceedings of the 2016 ACM Workshop on Vision and Language Integration Meets Multimedia Fusion*, iV&L-MM '16, pages 17–25, New York, NY, USA. ACM.

Udupa, R. and Khapra, M. (2010). Improving the multilingual user experience of wikipedia using cross-language name search. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 492–500, Stroudsburg, PA, USA. Association for Computational Linguistics.

van Gemert, J. C., Veenman, C. J., Smeulders, A. W. M., and Geusebroek, J.-M. (2010). Visual word ambiguity. *IEEE Trans. Pattern Anal. Mach. Intell.*, 32(7):1271–1283.

Vasconcelos, N. (2004). Minimum probability of error image retrieval. *IEEE Transactions on Signal Processing*, 52(8):2322–2336.

Villegas, M., Paredes, R., and Thomee, B. (2013). Overview of the imageclef 2013 scalable concept image annotation subtask.

Vukotić, V., Raymond, C., and Gravier, G. (2016). Bidirectional joint representation learning with symmetrical deep neural networks for multimodal and crossmodal applications.

In *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*, ICMR '16, pages 343–346, New York, NY, USA. ACM.

Wang, G., Hoiem, D., and Forsyth, D. (2009). Building text features for object image classification. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1367–1374. IEEE.

Wang, J., Yang, J., Yu, K., Lv, F., Huang, T., and Gong, Y. (2010). Locality-constrained linear coding for image classification. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 3360–3367. IEEE.

Wang, K., He, R., Wang, L., Wang, W., and Tan, T. (2015). Joint feature selection and subspace learning for cross-modal retrieval.

Wang, K., Yin, Q., Wang, W., Wu, S., and Wang, L. (2016a). A comprehensive survey on cross-modal retrieval. *CoRR*, abs/1607.06215.

Wang, L., Li, Y., and Lazebnik, S. (2016b). Learning deep structure-preserving image-text embeddings. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Wang, Y., Wu, F., Song, J., Li, X., and Zhuang, Y. (2014). Multi-modal mutual topic reinforce modeling for cross-media retrieval. In *Proceedings of the 22Nd ACM International Conference on Multimedia*, MM '14, pages 307–316, New York, NY, USA. ACM.

Wei, Y., Xia, W., Huang, J., Ni, B., Dong, J., Zhao, Y., and Yan, S. (2014). Cnn: Single-label to multi-label. *arXiv preprint arXiv:1406.5726*.

Weston, J., Bengio, S., and Usunier, N. (2011). Wsabie: Scaling up to large vocabulary image annotation. In *IJCAI*, volume 11, pages 2764–2770.

Williams, D. and Hinton, G. (1986). Learning representations by back-propagating errors. *Nature*, 323:533–536.

Xie, S. and Tu, Z. (2015). Holistically-nested edge detection. In *The IEEE International Conference on Computer Vision (ICCV)*.

Yan, F. and Mikolajczyk, K. (2015). Deep correlation for matching images and text. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yang, J., Yu, K., Gong, Y., and Huang, T. (2009). Linear spatial pyramid matching using sparse coding for image classification. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1794–1801. IEEE.

Yao, T., Mei, T., and Ngo, C.-W. (2015). Learning query and image similarities with ranking canonical correlation analysis. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 28–36. IEEE.

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. *TACL*, 2:67–78.

Zbontar, J. and LeCun, Y. (2015). Computing the stereo matching cost with a convolutional neural network. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zeiler, M. D. and Fergus, R. (2013). Visualizing and understanding convolutional networks. *CoRR*, abs/1311.2901.

Zhou, B., Lapedriza, A., Xiao, J., Torralba, A., and Oliva, A. (2014). Learning deep features for scene recognition using places database. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 27*, pages 487–495. Curran Associates, Inc.

Zhou, X. S. and Huang, T. S. (2000). Cbir: from low-level features to high-level semantics.

Znaidia, A. (2014). *Handling imperfections for multimodal image annotation*. PhD thesis, Châtenay-Malabry, Ecole centrale de Paris.

Znaidia, A., Shabou, A., Le Borgne, H., Hudelot, C., and Paragios, N. (2012). Bag-of-multimedia-words for image classification. In *ICPR*, pages 1509–1512. IEEE.

Appendices

Appendix A

Tags cleaning

Tags provided by users remain sometimes noisy, even after several filter process proposed by the authors of dataset. For instance, in Nus WIDE dataset, some tags are concatenated and result into a unique and non-existing word, *e.g. sunsetoverthesea*. This fact infers a shortcoming in textual feature extraction. For example, the term "*sunsetoverthesea*" is naturally absent from the Word2Vec vocabulary; hence the Word2Vec model will not able to correctly represent it. To improve the quality of textual features, we automatically separate the words (producing *e.g. sunset over the sea*) before employing techniques of textual features extraction. For this, each tag is matched to a dictionary of existing words (*e.g.* the tag dictionary of 5,018 terms for Nus WIDE) and we then retain only the valid largest sub-strings. The proposed process is described in the following Python code.

```
dict_file = 'dictionary.txt'
tags_file = 'allTags_before.txt'
final_tags_file = 'allTags_after.txt'

# collect words in dictionary
dict = []
with open(dict_file, "r") as file:
    for line in file:
        dict.append(line.strip())
file.close()

# clean each row of tags
with open(tags_file, "r") as file,
open(final_tags_file,"w") as outfile:

    for line in file:
        tags = line.strip().split(' ')
        final_tags = []
        for t in tags:
            # for each tag string, candidate list contains
            # all possible words (from dict) that may be
            # present in tag string
            candidate = []
            for d in dict:
                if d in t:
                    candidate.append(d)
            # we take only the longest word among possible
            # words in candidate
            # (for ex: to avoid the case "sunglasses"
            # decomposing into "sun", "glasses", "sunglasses")
            for c in candidate:
                if any((c in d) and (len(c)!=len(d)) for d in
                   candidate):
                    print()
                else:
                    final_tags.append(c)

        for f in (set(final_tags)):
            outfile.write('%s %f)
        outfile.write('\n')
file.close()
outfile.close()
```

Figure A.1: Python program to clean the tags

Appendix B

List of publications

International Conferences with Peer-review

T.Q.N. Tran, H. Le Borgne, M. Crucianu. 2015. *Combining Generic and Specific Information for Cross-modal Retrieval.* In Proceedings of the 5th ACM on International Conference on Multimedia Retrieval (ICMR '15). ACM, New York, NY, USA, pages 551-554.

T.Q.N. Tran, H. Le Borgne, M. Crucianu. 2016. *Aggregating image and text quantized correlated components.* In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR '16), pages 2046-2054.

T.Q.N. Tran, H. Le Borgne, M. Crucianu. 2016. *Cross-modal Classification by Completing Unimodal Representations.* In Proceedings of the 2016 ACM workshop on Vision and Language Integration Meets Multimedia Fusion (iV&L-MM '16). ACM, New York, NY, USA, pages 17-25.

Evaluation Campaign Participations

H. Le Borgne, E. Gadeski, I. Chami, T.Q.N. Tran, Y. Tamaazousti, A. Gînsca, A. Popescu: *Image Annotation and Two Paths to Text Illustration.* CLEF (Working Notes) 2016: 322-333, CEUR Workshop proceedings, Évora, Portugal, 5-8 September 2016.

Patent

T.Q.N. Tran, H. Le Borgne, M. Crucianu. *Procédé de description de documents multimedia par traduction intermodalités, système et programme d'ordinateur as- sociés.* France, INPI number 1651591. Filing date: 26 February 2016. (In progress)

Thi Quynh Nhi TRAN

Robust and comprehensive joint image-text representations

Abstract :
This thesis investigates the joint modeling of visual and textual content of multimedia documents to address cross-modal problems. A common representation space on which images and text can be both represented and directly compared is a generally adopted solution. Such a joint space still suffers from several deficiencies. The first limitation concerns significant information yet poorly represented on the common space. The second limitation consists in a separation between modalities on the common space. To deal with the first limitation, we put forward a model that combine such poorly-represented data with one that is relatively well-represented on the joint space. To cope with the separation between modalities on the joint space, we propose two representation methods that aggregate information from both the visual and textual modalities projected on the joint space. Specifically, for uni-modal documents we suggest a completion process relying on an auxiliary dataset to find the corresponding information in the absent modality and then use such information to build a final bi-modal representation for a uni-modal document. Evaluations show that our approaches achieve state-of-the-art results on several standard and challenging datasets for cross-modal retrieval or bi-modal and cross-modal classification.
Keywords :
common representation, cross-modal retrieval, cross-modal classification, (kernel) canonical correlation analysis, multi-modal representation, image and text.

Résumé :
La présente thèse étudie la modélisation conjointe des contenus visuels et textuels extraits à partir des documents multimédia pour résoudre les problèmes intermodaux. Un espace de représentation commun est la solution généralement adoptée. Sur cet espace, images et textes peuvent être représentés par des vecteurs de même type sur lesquels la comparaison intermodale peut se faire directement. Un tel espace commun souffre de plusieurs insuffisances. La première concerne les informations très importantes mais qui sont mal représentées sur cet espace. La deuxième insuffisance porte sur la séparation entre les projections des différentes modalités sur l'espace commun. Pour faire face au premier problème, nous avons proposé un modèle qui combine ces informations avec des données relativement bien représentées sur l'espace commun. Nous mettons en avant deux méthodes de représentation pour les documents bi-modaux ou uni-modaux qui regroupent à la fois des informations visuelles et textuelles projetées sur l'espace commun. Pour les documents uni-modaux, nous proposons un processus de *complétion* basé sur un ensemble de données auxiliaires pour trouver les informations correspondantes dans la modalité absente. Ces informations complémentaires sont ensuite employées pour construire une représentation bi-modale d'un document uni-modal. Nos approches permettent d'améliorer l'état de l'art pour la recherche intermodale ou la classification bi-modale et intermodale.
Mots clés :
espace commun de représentation, analyse canonique des corrélations, recherche intermodale, classification intermodale, représentation multimodale, image et texte.