Multi modal representation learning and cross-modal semantic matching

Wang, X.

MULTI MODAL REPRESENTATION LEARNING AND CROSS-MODAL SEMANTIC MATCHING

Proefschrift

ter verkrijging van
de graad van doctor aan de Universiteit Leiden,
op gezag van rector magnificus prof.dr.ir.H.Bijl,
volgens besluit van het college voor promoties
te verdedigen op vrijdag 24 juni 2022
klokke 11.15 uur

door

Xue WANG
geboren te Heilongjiang, China
in 1989

Promotor:
Prof. Dr. Ir. F. J. Verbeek

Co-promoters:
Dr. Y. Du (Xi'an Jiaotong University)
Dr. S. Verberne

Promotiecommissie:
Prof. Dr. A. Plaat
Prof. Dr. N. Mentens
Prof. Dr. M.S. Lew
Prof. Dr. H. Trautman (University of Munster)
Dr. Y. Guo (Chinese Academy of Sciences)

ISBN 978-94-6421-777-3

The research is financially supported by the Chinese Schoolarship Council (CSC No.201906280464).

CONTENTS

1

INTRODUCTION

1

Humans are very capable of processing a variety information to something that makes sense to them. The input of information is accomplished with the sensory systems through which the world is perceived. So, we acknowledge that humans perceive the real world through a spectrum of modes: vision, taste, hearing, smell, and touch. We see these as different channels of information or modals. For information processing, the brain integrates the information from these multiple modalities using a complex network of connected neurons. From information stored in these networks we are able to combine observations and recognize patterns. Making inferences over observation requires reasoning over the information which is an active form of using the brain, i.e. thinking. If a computer is able to reason over information that it assembles, this is referred to as a demonstration of intelligence. As opposed to the natural intelligence displayed by animals including humans, this computer intelligence is referred to as Artificial Intelligence (AI). With respect to natural intelligence, a process of reasoning is considered to leave a trace in the brain through the creation of connections in the brain; i.e. new connections between neurons. In artificial intelligence, these networks of connections are imitated by so called neural networks. Early versions of these neural networks were investigated and adapted in the early nineties of the previous century based on machine learning, e.g., long short-term memory (LSTM) [1] and LeNet 5 [2].

In recent years, with the development of AI, several neural networks have contributed to the status of the state-of-the-art in many research fields, such as computer vision (CV) [3, 4, 5], neural language processing (NLP)[6, 1, 7], etc.. These state-of-the-art neural networks, however, are based on a single modality. The aim of AI is to mimic the human way to learn and extract information for the purpose of advancing automated systems. Therefore, AI should be able to efficiently fuse information from different modalities.

In information processing a modality is considered to be a channel which is able to convey information. Other then natural information processing, in automated information processing, i.e. artificial intelligence, modalities are chosen for the type of information they represent rather then being connected with a sensory organ and perception. The information can represent an object or an event and so the modality channel thus represents image, text, video and audio, as the main types [8]. In this thesis we consider image and text as information carriers and investigate how reasoning on them can enhance the information we can analyze from a channel. In Fig. 1.1 this is depicted for the channels that are the subject of research in this thesis; i.e. the modalities figure (image) and figure caption (text). For humans it is easy to understand how the sentence in the figure caption represents the scene in the image as well as to locate the objects in the image that are represented by the words in the caption. The research presented in this thesis aim to develop methods to be able to learn from and reason over these two modalities. The methodology is based on principles and techniques common to Artificial Intelligence. We have the learn the computer to reason over these modalities and thus make inferences that are similar to those that humans make. In this man-

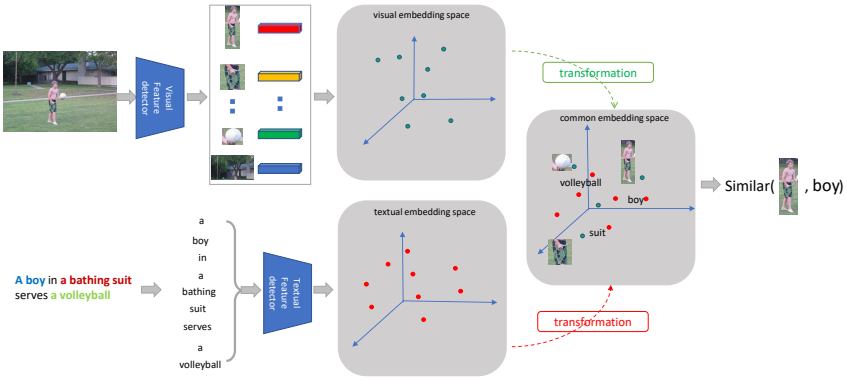


Figure 1.1: Cross-modal semantic matching is the task of matching between a textual description and an image. (a). Some methods learn different modality representations in embedding space respectively. The semantic matching performance is capped by the quality of the different modality representations. (b). The other methods directly transform the single modality representations into a common embedding space. This is significantly faster and also more accurate in inference.

ner we can generate figure captions that is well connected with image content. So, both captions and image need to be analyzed. The characterization is based on features that we can extract from these modalities. The set of features from one modality is referred to as an embedding. Each of the modalities has its own embedding space. And the embedding space is considered to be a representation of the modality, i.e. image or text. In order to be able to learn from these representations, one can build methods for each of the representations separately by extracting features and then learn from the different embedding spaces. However, another approach would be to build a common embedding space. This approach is referred to as multi-modal as it incorporates information from more than single modal [8]. This multi-modal approach is the focus of our research and the research questions that are formulated for the research presented in this thesis are based in this. It is clear that we consider a problem in AI that relates to the way humans process information. That is, information from more than one source is beneficial in further understanding of the complete picture and also helpful for better information retrieval. Consequently, the research areas of multi-modal data fusion and retrieval come into view for our research.

In interactive computing multi-modal systems are systems that offer interaction over multiple channels; that is several devices together constitute an interaction. For artificial intelligence a multi-modal system is characterized to contain more data modalities; so, the multi-modality is data driven. The system is considered as the entity within which the data are analyzed. The data from different modalities are often being paired based on the same meaning, in other words the paired object and word have the same semantics. Over recent years, there has been a rapid increase in multi-modal (data) systems; i.e., systems based on a com-

bination of images, text, video and audio. With respect to information retrieval, the classic approach is uni-modal. The embeddings, as mentioned earlier, are based one single information channel or modal. Within multi-modal systems, the modalities can be linked and thus we obtain a cross-modal system. The analysis of multi-modal systems focuses on cross-modal information embeddings. The principal task of cross-modal embeddings is to construct a common semantic space in which data points from different modalities but with the same semantics are close to each other [9]. The research on multi-modal systems and cross-modal embeddings provides important support for downstream computational tasks such as: information retrieval [10], image caption generation [11][12][13], visual grounding [14][15] and visual question answering (VQA) [11][16]. Visual grounding aims to localize those objects in an image with particular language information. In general, this language information should be one phrase or expression extracted from the image caption. Visual grounding is important for our tasks in computer vision. With respect to cross-modal grounding we can discern tasks with a different level of complexity, i.e., phrase grounding [17][18] and referring expression grounding [19]. Here the difference between phrase grounding and referring expression grounding is that the phrase should be extracted from the image caption, but the referring expressions are not.

It is very common to describe expressions as the relationship between two objects, however, a phrase only has the information for a single object [20]. Moreover, to date, most cross-modal benchmark datasets that are commonly used in research are bi-modal with both vision and text, e.g., MS COCO [21], Visual Genome [22] and Flickr30K [23]. In these datasets, the textual information is provided as the description of the visual information. The research and experiments presented in this thesis use this paired information to our advantage in order to build a cross-modality supervision learning strategy. This results in a fully or weakly supervised learning process for which the training data do not need explicit annotations. This means that the vision representation guides the improvement of the textual representation whereas the image captions supervise the object detection. The weakly supervised learning processes use phrase grounding without phrase location annotations and image retrieval. Cross-modal semantic matching is formally defined as the bridge of vision and language, i.e., image and text. In Fig. 1.2 the structure of this thesis is depicted. In this chapter we first give some insight in the datasets that we used in our experiments. Given these datasets we henceforth explain the major aspects addressed in this thesis. We investigate single modality representation learning using supervision from another modal supervision. And next, we elaborate how we construct a common information space. Both aspects share the importance of finding good semantic models from which can learn the representations (location or the vectors) of different modal data. In these representations points with the same semantics are closer to each other than points representing data that are more dissimilar.

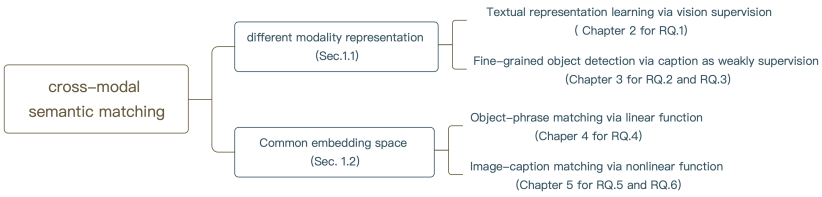


Figure 1.2: The structure of this thesis.

1.1. DATASETS FOR MULTI-MODALITY STUDIES

The availability of image–text paired data is limited due to significant manual efforts in collecting the annotations for special correspondences, e.g., phrase-object and expression-object, which is needed as data for visual grounding tasks. We should be aware of this limitation with repect to the dataset.

The research in this thesis is based on 3 commonly used datasets:

- The MS COCO dataset [21] is a large-scale dataset that addresses two tasks: object detection which uses either bounding box output or object segmentation output; and caption generation. There are 80 categories, 12 supercategories for 123,187 images, and each image corresponds to 5 captions.

- The Flickr30K dataset [24] and the Flickr30K Entities dataset [23]: The Flickr30k dataset has become a standard benchmark for sentence-based image description. The paper[23] of Flickr30K Entities dataset presents Flickr30k Entities, which augments the 158k captions from Flickr30k with 244k coreference chains, linking mentions of the same entities across different captions for the same image, and associating them with 276k manually annotated bounding boxes. Such annotations are essential for continued progress in automatic image description and grounded language understanding.

- The Visual Genome (VG) dataset [22] has a dense annotation of objects, attributes, and relationships within each image to learn object detection, attributes analysis and relationship extraction models. Specifically, the VG dataset contains over 108K images where each image has an average of 35 objects, 26 attributes, and 21 pairwise relationships between objects. The dataset can utilize the knowledge base semantic connection between the annotations of objects, attributes, relationships, and noun phrases in region descriptions and questions answer, as all annotation labels are match to WordNet synsets.

VG has more different object categories than MS COCO and Flickr30k without image descriptions, only with region description. There is a subset that con-

tains the same images (76,631 images) from the overlapping part between Visual Genome and MS COCO. The overlap set has annotations with captions, fine-grained category labels and coarse category labels. We use these data to train models based on the image-caption paired supervision for different cross-modal semantic matching tasks, i.e. phrase grounding, expression grounding and fine-grained visual representation learning.

1.2. IMPORTANCE OF CROSS-MODAL SUPERVISION FOR REPRESENTATION LEARNING

Vision and language are two important aspects of human intelligence to understand the real world, i.e., images and text are the most important expressions of vision and language. Nowadays, on the textual side, the representations of concepts can be extracted by word embeddings models such as Word2Vec [25, 26], GloVe [27], and BERT [28]. In recent years, Bidirectional Encoder Representations from Transformers (BERT)[28] set state-of-the-art performance on various sentence classification and sentence-pair regression tasks, by pretraining deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers. On the other hand, The features of visually represented concepts are often extracted by Convolutional Neural Networks (CNN) [4]. Visual representations are commonly derived by modeling the context and statistic distribution of one modality, e.g., words in documents [29] or objects in images [30] [31], instead of modeling the true visual semantics.

The data describing the semantics of text features or visual features are employed by visual information features or textual information features respectively. The unimodal representation models learned from the analysis of natural language or images do not usually reflect the true semantics of words, e.g. "apple" sometimes means fruit, sometimes means brand. For tasks at the intersection of vision and language, it seems prudent to model semantics as dictated by both text and vision. As a result, textual representation learning based on visual supervision will lead to the textual representation containing more spatial features, especially for relation words, i.e., the distance between "next to" and "near" will be close in the embedding space. The visual representation will be more detailed with fine-grained textual labels. Therefore, in this thesis, we are motivated to develop novel representation learning approaches: (1) to include cross-modal features (image features for text representation and text features for image representation); (2) to extract visual representations based on fine-grained categories object detection.

1.3. INTRODUCTION OF COMMON SEMANTIC SPACE REPRESENTATION LEARNING

Image–text matching is one of the fundamental topics in cross-modal research; it refers to measuring the semantic similarity between a sentence and an image or

phrases in sentence and objects in the image[32]. It has been widely adopted to various applications such as the retrieval of text descriptions from image queries or image search for textual queries. Research projects combining cross-modal, i.e., visual and language, semantic matching compatibility have attracted a lot of attention and accelerated the advancement of artificial intelligence in recent years.

During the last decade, research has gained great progress on image–text matching. Although in some circumstances a lot of progress has been achieved in image–text semantic matching research, it is still a challenging problem due to the huge visual semantic discrepancy and the semantic gap between vision and language. The huge visual semantic discrepancy is caused by visual classification based on pixel-level images usually lacking high-level semantic information as is present in the texts that are matched to it. For example, we describe a cat in an image, the caption could read somehting like: "a sleeping cat". However, in the image domain, it is very hard to train a model that can learn the visual representation to include the sleeping information. We propose two types of methods for estimating the similarity between the modalities, one is based on concept knowledge graphs, e.g., WordNet [33] and ConceptNet [34]; the other is embedding features to a common space based on an encoder [35].

In order to measure semantic matching, most of the prior work focuses on building a common semantic space in which comparisons between different modality features are performed with similarity metrics [36]. It is especially challenging as it requires a good representation of both the visual and textual domain and an effective way of linking them. Existing work can be roughly divided into two types. One is encoding words and vision features extracted from different modalities respectively into a unified vector space based on linear transformation [37][12]. The other is learning a non-linear model to map visual semantic features and textual features into a common semantic space [38]. Most works about linear projections for common space matching are based on Canonical Correlation Analysis (CCA) [37] and more recent work uses the attention mechanism [12]. CCA finds linear projections that maximize the correlation between projected vectors from the two modalities. The attention mechanism predicts an attention probability of each visual feature or textual feature, and picks out the weighted probability to calculate a context feature and subsequently feed it into a deep neural net (i.e. RNN [6] or CNN [4]) to learn the vectors in a common semantic space. Non-linear approaches map the visual and textual features into a common space, based on a non-linear projection function. Kernel CCA [39] can be seen as an extension of CCA with maximally correlated non-linear projections. Kernel CCA based methods have the ability of learning non-linear representations, but the learned representation is limited due to the fixed kernel. Therefor, we are motivated, in this thesis, to learn new models (1) based on attention mechanisms to reconstruct the representation of vision and text guided by inter- or intra-modality; (2) based on kernels to mixture mapping the semantic association between different modalities (image and text).

1

1.4. RESEARCH QUESTIONS AND PERSPECTIVES

RQ1: To what extent is it possible to improve the representation of visual features detected by CNNs or the representation of textual features embedding and reduce the semantic gap between visual and textual information?

Textual representation, i.e., the word embedding vector, can be learned by a deep learning model, e.g. LSTM, or shallow neural networks, e.g. Word2Vec or GloVe. Textual representation learning is independent and does not rely on the visual supervision. However, a cross-modal matching task should combine features from the two modalities together and analyze the connection or correlation between vision and text in these two modalities. We are, in particular, interested in how to use the visual information as a supervisor to improve textual representation. Inspired by the structure of Word2Vec, we want to design a word embedding model which can learn textual representations based on visual supervision and thereby reduce the semantic gap between text and vision. Another visual representation can be learned by classification based on CNN models, i.e., Fast-RCNN [3], Faster-RCNN [30] and Mask-RCNN [31]. These are object detection models; all regions will be marked by a category label. Therefore, the visual representations will be limited by the category label. The recolonization of fine-grained object categories by computer vision techniques has attracted extensive attention for cross-modal semantic matching research. The task is very challenging as it entails fine-grained representation learning and reigniting categories (e.g., "man", "woman" and "child" are all referring to "people"). Fine-grained visual representation learning can bridge the vision and text similarity in a textual vector space.

RQ2: How to utilize additional knowledge base to measure semantic matching?

Current knowledge bases such WordNet[33], Cyc [40], and ConceptNet [34] can offer a complementary and orthogonal source of evidence that helps in discovering highly confident facts from amongst the pool of all facts extracted from text. Knowledge bases are useful semantic matching tools for computational linguistics and natural language processing. Knowledge bases are typically represented as a directed graph which can seen as a hierarchical dictionary to obtain the semantic relations between words. For our analysis, we build a semantic map that provides a correspondence between the coarse labels and fine-grained label proposals coming from captions based on embedding techniques and knowledge bases.

RQ3: To what extent can curriculum learning measure the distribution of visual complexity and improve weak supervision for semantic matching?

Using external knowledge as the reasoning to replace the coarse object categories with fine-grained categories is a weakly supervised learning process. The main challenge is that the new categories will cause long-tail distributions of the visual labels. The motivation of curriculum learning (CL) [41] is to first learn from easy samples and then gradually learn from more difficult samples. This approach aims to reduce the negative effects brought by noisy data in early stages of training. It can also be applied for deciding the learning order of tasks. CL has been applied in many problems like image classification and object tracking. CL can assign samples with different transfer difficulties by metric of cross-media domain consistency. This is an iterative process with adaptive feedback, which gradually reduces the discrepancy between cross-media domains to enhance the robustness of model training and improves retrieval accuracy on cross-media target matching [42]. To address the problem of data distribution, we propose a novel learning process based on curriculum learning. First, we analyze the distribution of the data and rank the data according to different criteria. Second, we put the training data into the net by the order of the ranking. We are, in particular, interested in how much improvement can be achieved regarding the different ranking orders.

RQ4: How and with what quality can we model the semantic correlations between two different modalities?

Common embedding spaces are usually used for cross-modal matching tasks to measure the semantic correlation mining of heterogeneous media. We will research how to respectively transform different modality representations into a common embedding space.

Current research for linear mapping models, e.g., mixture of local linear mapping [43] cannot address the non-linearity of data distributions and cross-media correlations very well. To address the nonlinearity of data distributions and cross-media correlations, we transform the textual and visual data using a nonlinear transformation from the input spaces into two latent high-dimensional spaces by nonlinear feature space mapping functions, and then construct the mapping model between both modalities in the latent features spaces. The smoothness and sparseness of the parameters are introduced to enhance the generalization of models and the fitness between models and data. The parameters are estimated using kernel theory to avoid the explicit representation of both feature spaces.

RQ5: What is the effect of the attention mechanism to eliminate the different modal representations produced in the common embedding space?

Attention mechanisms were first used for transformer models and was initially shown to be effective in machine translation [12] and later for many other natural language processing tasks. In order to verify the attention mechanism is effectiveness for cross-modal semantic matching, in this thesis, we employ of the atten-

tion mechanism to contain the attention component of the network, which will learn the weight of each feature for self-modal or cross-modal learning. We investigate if a visual self-modality attention net can distinguish the important and relevant regions from the input image and assign higher weights to more important regions. For cross-modal semantic matching the query and key come from different modality (vision and text) to compute a weighted sum of uni-modal values (visual feature vector or textual feature vector). Consequently, we also need to build a model with cross-modal attention that can discover the latent alignment using both image regions and words in paired captions as context via attention across modalities, which produces more accurate image–text similarity for matching [44]. Therefore, based on the above analysis, we build a multi-modal model to analyze the self-modality context and cross-modality context at the same time, which can enhance the accuracy of the output prediction.

RQ6: How to employ the correspondence between images and text as supervision instead of the matching annotations to address the limited data issue?

Contrastive learning [13] was proposed for self-supervised and semi-supervised learning, which learns representations by maximizing agreement between differently augmented views of the same data example via a contrastive loss in the latent space. Contrastive Predictive Coding [35] extracts useful representations from high-dimensional data by using powerful autoregressive models. The probabilistic contrastive loss induces the latent space to capture information that is maximally useful to predict unseen samples. There has been a recent trend of exploring contrastive loss for weakly supervised cross-modal representation learning, which can maximize the mutual information between the visual and textual feature of a deep network. These are all based on a similar contrastive loss related to Noise Contrastive Estimation (NCE) [45]. To address the data limitation issue, in this thesis, we build our matching model based on contrastive learning framework, which consist of two branches, one is for visual modal and the other is for textual modal. By contrastive learning net we will got a similarity score between visual and text modality representation. The weight of contrastive learning model optimized by positive paired samples and negative paired samples. In our work, we define the paired image and caption as positive samples, and non-paired images and captions as negative samples. We use the NCE loss, which has also been explored for phrase grounding by Gupta et al. [18] (InfoGround), as learning objective for our model.

1.5. THESIS STRUCTURE

This thesis is structured along the research questions and perspectives presented in the previous paragraph.

Chapter 2 "Embedded Representation of Relation Words with Visual Supervision" presents a learning process to achieve the representation of the relation

words that are important in knowledge related tasks (related to the RQ.1). This model computes the visual features based on deep networks over an image patch that reflects a relation word and then establishes the visual similarity matrix for all relation words. We experiment with our implementation on publicly available datasets and demonstrate that our model really changes the distribution of word representation and achieves effective results in describing their true semantics.

Chapter 3 "Fine-Grained Label Learning in Object Detection with Weak Supervision of Captions" focuses on label inference curriculum network to fine-grained label learning by incorporating the coarse category labels and captions provided in public datasets (related to the RQ.2 and RQ.3). We build a semantic label map based on embedding techniques and a knowledge base to describe the correspondence between coarse labels and fine-grained label proposals. Then based on the label inference curriculum network with the consideration of the complexity of samples that describe the difficulty of fine-grained label learning. Experimental results demonstrate the effectiveness of our approach in the task of fine-grained label learning.

Chapter 4 "Kernel-Based Mixture Mapping for Image and Text Association" introduces a new approach called kernel-based mixture mapping to model the semantic correlations between images and text (related to RQ.4) . With this approach, we first construct latent high-dimensional feature spaces based on kernel theory to address the non-linearity of both the data distributions in the input spaces and the cross-model correlation. We present a probabilistic neighborhood model to describe the spatial locality of semantics by assuming that proximate examples in feature spaces generally have the same semantics and a conditional model to describe cross-modal conditional dependency. For optimization, we employ a hybird algorithm to find the solution of kernel-based mixture mapping based on expectation-maximization and sub-gradient ascent. The experimental results show that our approach outperforms the compared methods when modeling the relationships between images and text.

Chapter 5 "Visual Representation Contextualization Based on Constrastive Learning" focuses on a weakly supervised approach for visual contextualized representations, which is systematically learned by pooling object proposals to alleviate the suppression of each object feature (related to RQ.5 and RQ.6). We use visual-textual cross-modal attention to capture the correlation among object proposals of each image and generate the representation of each candidate incorporating the visual information of the other candidates. Visual-textual cross-modal attention represents the visual topic corresponding to each textual component in the caption in a cross-modal common space, guided by the attention of a word to object proposals. The experimental results show that our approach outperforms the compared methods when grounding words in caption to objects in image.

Chapter 6 "Conclusions and Discussion" summarizes our contribution for representation learning to measure cross-modal semantic matching, our proposed visualization framework, effect on some drawbacks of the framework. Future improvements of our methods are discussed guaranteeing the integrity of the pro-

posed framework.

2

EMBEDDED REPRESENTATION OF RELATION WORDS WITH VISUAL SUPERVISION

This chapter is based on the following publication:

Wang, X., Du, Y., Li, X., Cao, F., Su, C. (2019, February). Embedded representation of relation words with visual supervision. In 2019 Third IEEE International Conference on Robotic Computing (IRC) (pp. 409-412). IEEE.

CHAPTER SUMMARY

This chapter addresses RQ1.

RQ1: To what extent is it possible to improve the representation of visual features detected by CNNs or the representation of textual features embedding and reduce the semantic gap between visual and textual information?

Word representation learned from the analysis of natural language does not usually reflect the true semantics of words. This chapter proposes a new method, named Visually Supervised Word2Vec (VS-Word2Vec) model to achieve the representation of the relation words that are important in knowledge related tasks. Our method first computes the visual feature vector of relation words based on deep networks, and then achieve the visual similarity matrix for all relation words, which we think reflects their true semantics. VS-Word2Vec model then combines the visual similarity and the CBOW and builds an optimization problem to jointly learn the word vector representation. Therefore, VS-Word2Vec fuses the visual modality and natural language together. Experiments implemented over the public datasets demonstrate that VS-Word2Vec model really changes the distribution of word representation and achieves more effective results in describing their true semantics than CBOW model.

The distributed representation of words in a vector space is an important issue in natural language processing tasks. The representation is mainly derived by modeling the context and statistic distribution of words in language documents instead of modeling their true semantics [46]. However, the data to describe the semantics of words includes many other modalities, such as visual and auditory data. Human learns how to understand the world by fusing the multiple modalities.

Many methods have been proposed to incorporate morphological information into word representations [47, 48, 49, 50]. Recently, Mikolov et al. [26, 25] proposed a set of models, such as CBOW and skip-gram, based on word analogies that probe the structure of the word embedding space. Pennington et al. [27] introduced a new global log-bilinear regression model (GloVe) based on the global matrix factorization and local context window methods. Bojanowski [51] proposed an approach by considering the morphology of words based on the skip-gram model, in which each word was represented as a bag of character n-grams and the word vector was the sum of the n-gram representations. For tasks at the intersection of vision and language, it seems prudent to model semantics as dictated by both text and vision. Kottur et al. [52] presented the Visual Word2Vec model which learned visually grounded word embedding to capture visual notions of semantic relatedness. Lu et al. [53] proposed a model that used visual relationships to improve the previous work by leveraging language priors from semantic word embedding to fine-tune the likelihood of a predicted relationship.

Most of these methods represent each word based on a single modality, and ignore the joint learning of multiple modalities. The phenomenon is much distinct in the knowledge related problems. In a knowledge triple (Arg1, Predicate, Arg2), Arg1 and Arg2 are subject and object strings, respectively, and Predicate is the string of relation [54, 55]. The vector representation of many predicate words (also called relation words) doesn't reflect their true similarity. For example, "above" and "below" is similar to each other in the representation by CBOW model. However, in the visual media, we know that they mean the opposite relative positions and show much different semantics.

This chapter proposes a new method, named Visualla y Supervised Word2Vec (VS-Word2Vec) model, to learn the vector representation of relation words. In this method, we first compute the visual feature vector of relation words based on deep networks, and then achieve the visual similarity matrix as weight for all textual relation words, which we think reflects their true semantics. VS-Word2Vec model then combines the visual similarity weight matrix and CBOW, and builds an optimization problem to jointly learn the word vector representation over visual modality and text modality. Experiments implemented over the public datasets demonstrate that our approach really changes the distribution of word representation and achieves more effective results in describing their true semantics than CBOW model.

2.1. OUR MODEL

2.1.1. BASIC CBOW MODEL

Suppose we have a sequence of training words $D = (w_1, w_2, ..., w_t, ..., w_T)$, where T is the number of words in this sequence and each word w_t belongs to a vocabulary $W = \{w_{I,1}, w_{I,2}, ..., w_{I,|W|}\}$, $|W|$ is the number of words in vocabulary. We represent the vector representation of word $w \in W$ as $\mathbf{v}_w \in \mathbf{R}^m$. The objective of the CBOW model [25] is to maximize the log probability:

$$J_T = \sum_{w_t \in D} \log p(w_t | Cont(w_t)) \tag{2.1}$$

where $Cont(w_t)$ denotes the context of a word w_t, i.e., $Cont(w_t) = \{w_{t-c}, ..., w_{t-1}, w_{t+1}, ..., w_{t+c}\}$ with the context window size c.

The CBOW model uses a binary tree to represent all words in the vocabulary. The $|W|$ words in vocabulary are leaf units of the tree. For each leaf unit, there exists a unique path from the root to this unit, and this path is used to estimate the probability of the word corresponding to the leaf unit. The probability in Eq.2.1 can be formulated as follows:

$$\begin{aligned} p(w|Cont(w)) &= \prod_{j=2}^{l^w} p(d_j^w | \mathbf{x}_w, \theta_{j-1}^w) \\ &= \prod_{j=2}^{l^w} [\sigma(\mathbf{x}_w^T \theta_{j-1}^w)]^{d_j^w} \cdot [1 - \sigma(\mathbf{x}_w^T \theta_{j-1}^w)]^{d_j^w} \end{aligned} \tag{2.2}$$

where l^w denotes the length of the path from root unit to the leaf one corresponding to w, d_j^w indicates that the inner unit at the j-th level is the left or right child of the inner unit at the (j-1)-th level, $d_j^w = 1$ denotes the left child case and $d_j^w = 0$ otherwise, θ_{j-1}^w is the parameter corresponding to the (j-1)-th level inner unit, on the path. $\sigma(\cdot)$ denotes the sigmoid function, i.e.,

$$\sigma(\mathbf{x}_w^T \theta_{j-1}^w) = \frac{1}{1 + e^{-\mathbf{x}_w^T \theta_{j-1}^w}} \tag{2.3}$$

and $\mathbf{x}_w = \sum_{w_i \in Cont(w)} \mathbf{v}_{w_i}$.

2.1.2. VISUALLY SUPERVISED WORD2VEC MODEL

Besides the vocabulary W defined above, we have a relation word vocabulary $R = \{w_{r,1}, w_{r,2}, \cdots, w_{r,|R|}\} \subset W$, which often describes the relation of two objects in image, i.e., "behind" and "contain".

Images visually represent the concepts in terms of their appearance, motion and space relation, etc, instead of the abstraction used in natural language. Distributed representations of words in a vector space help learning algorithms to

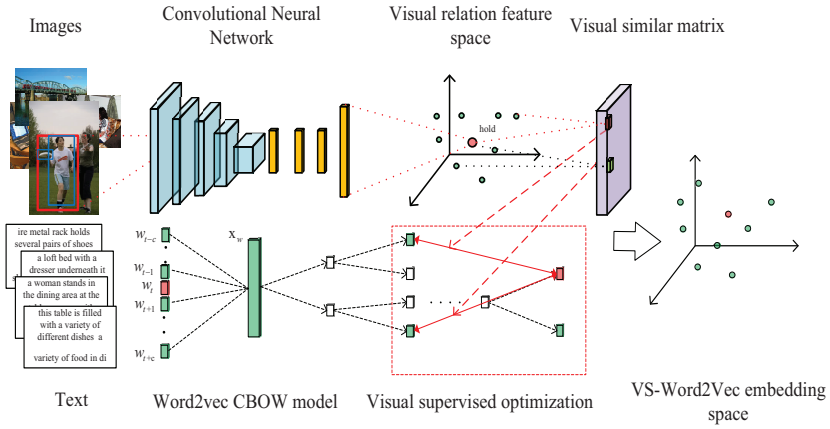


Figure 2.1: VS-Word2Vec Framework.

achieve better performance in natural language processing tasks by grouping similar word. However, we find that Word2Vec model cannot produce a good representation, especially for the relation words to represent the relationship of two entities in knowledge extraction task, in describing their semantic similarity. As shown in Fig.2.1, Our VS-Word2Vec model includes two parts: one is the basic CBOW, which learns the representation from natural language, and the second part is to compute the visual similarity of relation words based on the deep learning over image contents. Finally, our model jointly learns the representation by building an optimization problem over the above parts.

In the image data, each patch reflects a relation concept is bounded and is labeled by a relation word $w_{r,i} \in R$. The visual representation, denoted by $\mathbf{y}^q_{w_i}$, corresponding to $w_{r,i}$ can be computed over the bounded patch with Convolutional Neural Networks (CNNs)(16-layer VGG network), where the superscript q indicates the q-th patch example for the relation word. The visual representation for $w_{r,i}$ are given by the average of feature vector over its q visual examples:

$$\mathbf{y}_{w_i} = \frac{1}{Q_i} \sum_{q=1}^{Q_i} \mathbf{y}^q_{w_i} \tag{2.4}$$

where Q_i denotes the number of visual patch examples corresponding to $w_{r,i}$. The visual similarity two relation words $w_{r,i}$ and $w_{r,j}$ is defined as follows:

$$sim(w_{r,i}, w_{r,j}) = \frac{\mathbf{y}^T_{w_i} \mathbf{y}_{w_j}}{\|\mathbf{y}_{w_i}\| \cdot \|\mathbf{y}_{w_j}\|}. \tag{2.5}$$

For all $|R|$ relation words, we can obtain a similarity matrix $S_V = (s_{ij})_{|R|\times|R|}$ with $s_{ij} = sim(w_{r,i}, w_{r,j})$.

We aim to make the similarities of two relation words learned from natural language and from images consistent. Hence, we define the inconsistency of the similarities from two sources:

$$J_V(w_{r,i}, w_{r,j}) = (s_{ij} - \mathbf{x}_{w_{r,i}}^T \mathbf{v}_{w_{r,j}})^2 \tag{2.6}$$

where $\mathbf{x}_{w_{r,i}}$ has been defined in section 2.1.1. J_V is expected to be small to keep the consistency between the similarities derived from two sources. Our VS-Word2Vec can be formulated based on CBOW by the following optimization problem:

$$\max_{\{\mathbf{v}_{w_t}\}} J(\{\mathbf{v}_{w_t}\}), \tag{2.7}$$

where

$$J(\{\mathbf{v}_{w_t}\}) = \sum_{t=c+1}^{T-c} \big(\log p(w_t|Cont(w_t)) - \lambda \rho_t \sum_{w_{r,j} \in R} J_V(w_t, w_{r,j})\big) \tag{2.8}$$

where λ is a parameter to control the balance between two terms, $\rho_t = 0$ denotes that the word w_t is not the relation word defined in R, and $\rho_t = 1$ otherwise. The algorithm of VS-Word2Vec model is summarized in Algorithm 1.

2.2. EXPERIMENTAL RESULTS

2.2.1. DATASET AND EXPRERIMENT SETTINGS

We implement the experiments over the text8 dataset [1], which is often used as the test dataset for Word2Vec methods. We use the visual relationship detection data introduced by Lu et al. [53] to compute the similarity of relation words. The image relationship dataset contains 5000 images with 100 object categories and 70 predicates. In total, the dataset contains 37,993 relationships with 6,672 relationship types and 24.25 predicates per object category. In this chapter we focus on 22 predicates about 63 object categories. For both our model and the baseline method, we use the following parameters: the dimension of vector representation is 150, and the size of context window is $c = 2$. When building the word vocabulary, we keep the words those appear at least 3 times in the dataset. The learning rate is set to $\eta = 0.025$ in both CBOW and our model. The parameter λ in Eq.2.8 is experimentally set to 0.0025.

2.2.2. RESULTS AND ANALYSIS

Fig.2.2 illustrates the comparison of the similarity of 22 typical predicate words derived from images by deep networks, natural language by CBOW, and the multimodal data by our VS-Word2Vec, where all similarity values have been normalized into [0,1]. From the Figs.2.2(a) and 2.2(b), we observe that the similarity of relation words derived from image contents is different from that computed by CBOW

[1] http://mattmahoney.net/dc/text8.zip

Algorithm 1 VS-Word2Vec model.

Require: Word sequence D, word vocabulary W, relation word vocabulary R, images labeled with relation words.

Ensure: Vector representation $\{\mathbf{v}_{w_t}\}$;

Computer visual representation for each relation word based on Eq.2.4;
Compute visual similarity matrix S for all relation words with Eq.2.5;
Randomly initialize $\{\mathbf{v}_{w_t}\}$ and $\{\theta_{j-1}^{w_t}\}$;
for each $w_t \in D$ **do**
    $\mathbf{x}_{w_t} = \sum_{w_i \in Cont(w_t)} \mathbf{v}_{w_i}$; $\mathbf{e}_1 = \mathbf{0}$, $\mathbf{e}_2 = \mathbf{0}$;
    for $j = 2 : l^w$ **do**
        $g_1 = \eta(1 - d_j^{w_t} - \sigma(\mathbf{x}_{w_t}^T \theta_{j-1}^{w_t}))$;
        $\mathbf{e}_1 \leftarrow \mathbf{e}_1 + g_1 \theta_{j-1}^{w_t}$;
        $\theta_{j-1}^{w_t} \leftarrow \theta_{j-1}^{w_t} + g_1 \mathbf{x}_{w_t}$;
    end for
    if $w_t \in R$ **then**
        $\rho_t = 1$;
    else
        $\rho_t = 0$;
    end if
    for $j = 1 : q$ **do**
        $g_2 = \eta\lambda(\mathbf{x}_{w_t}^T \mathbf{v}_{w_j} - s_{tj})$;
        $\mathbf{e}_2 \leftarrow \mathbf{e}_2 + g_2 \mathbf{v}_{w_j}$;
    end for
    $\mathbf{v}_{w_t} \leftarrow \mathbf{v}_{w_t} + \mathbf{e}_1 + \rho_t \mathbf{e}_2$;
end for

over natural language. Fig.2.2(c) shows our results of VS-Word2Vec by considering the semantic similarity in images and the CBOW based natural language modeling together. The figure shows that our method really changes the similarity of relation words and tends to be close to the their true semantic similarity.

Fig.2.3 visualizes the location and distribution of 22 predicate relation words in 2-dimensional space. We observe that our approach really changes the distribution of the relation word representation compared the CBOW and describes their true semantics better. VS-Word2Vec model pushes away the location of relation words with different semantics and draw those words with similar semantics close. For example, “sit”, “lying” and “stand” represent three different kinds of motions and locate far away from each other in Fig.2.3(b) compared to Fig.2.3(a); “above” and “below” represent the opposite location of objects, and they are pushed away largely by our approach.

We also quantitatively evaluate the performance of our approach in measuring the word similarity or relatedness of semantics, which is also measured by the cosine distance shown in Eq.2.5. We choose SimVerb-3500 [56] as the ground

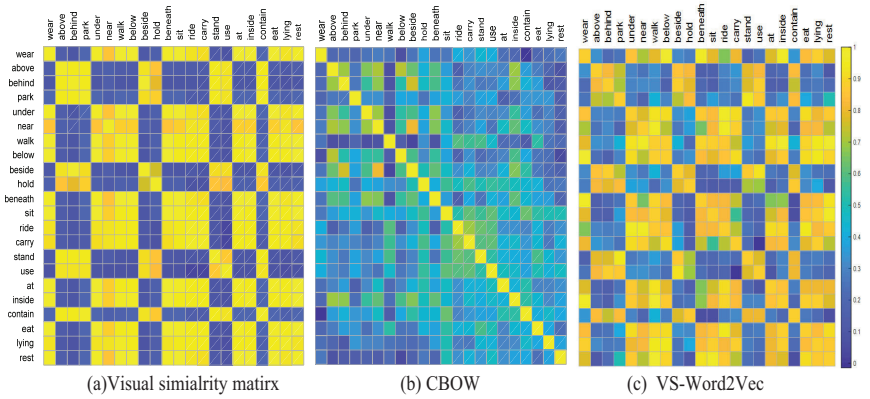


Figure 2.2: The comparison of the similarity of 22 typical relation words.

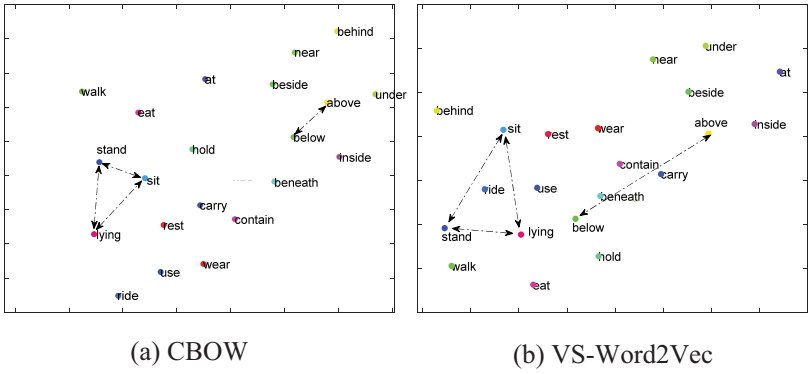


Figure 2.3: The visualization of the distribution of vector representation derived from the CBOW and our approach based on t-SNE.

truth, which gives the similarity of 3500 pairs of verbs. We normalize the similarity into [0,1] and use the following metric to evaluate the consistency of the similarity derived from our approach and the ground truth.

$$Con = \frac{\#(|S_g(p_i) - S_V(p_i)| < |S_g(p_i) - S_T(p_i)|)}{\#(p_i)} \tag{2.9}$$

where p_i denotes a pair of words, S_g, S_V and S_T are the similarity derived from ground truth, our VS-Word2Vec and CBOW Word2Vec, respectively, for the pair p_i. We choose 798 pairs that appear in both the Text8 and SimVerb-3500, and report the performance based on Eq.2.9 in Table 2.1. In this table, SYNONYMS, ANTONYMS, HYPER/HYPONYMS, COHYPONYMS and NONE are the different categories of pairs given in SimVerb-3500. We find that in three categories, the

Table 2.1: Confidence of Model

	CBOW	**Our Model**
SYNONYMS	**0.5618**	0.4492
ANTONYMS	0.3846	**0.6154**
HYPER/HYPONYMS	0.4192	**0.5808**
COHYPONYMS	**0.5231**	0.4769
NONE	0.1330	**0.8670**
Mean	0.4043	**0.5979**

similarity consistency of our approach is higher than CBOW Word2Vec model, and the average consistency of our approach is better.

2.3. CONCLUSION

In this chapter, we propose the VS-Word2Vec model to learn the vector representation of relation words by jointly compute over visual modality and natural language. In this method, we first compute the visual feature based on deep networks over an image patch that reflect a relation word, and then achieve the visual similarity matrix for all relation words. VS-Word2Vec model then resolve an optimization problem that consists of the terms related to the visual similarity and context in natural language. Experiments implemented demonstrate that our approach really changes the distribution of word representation and achieves more accurate similarity of words than CBOW model.

3

FINE-GRAINED LABEL LEARNING IN OBJECT DETECTION WITH WEAK SUPERVISION OF CAPTIONS

This chapter is based on the following publication:

Wang, X., Du, Y., Verberne, S. Verbeek, F.J. Fine-Grained Label Learning in Object Detectionwith Weak Supervision of Captions. Multimedia Tools and Applications. (under review)

CHAPTER SUMMARY

This chapter addresses RQ2 and RQ3.

RQ2: How to utilize additional knowledge base to measure semantic matching?
RQ3: To what extent can curriculum learning measure the distribution of visual complexity and improve weak supervision for semantic matching?

This chapter addresses the task of fine-grained label learning in object detection with the weak supervision of auxiliary information attached to images. Most of the recent work focused on the label prediction for objects in the same category space as in training data under the supervised learning framework and cannot be expanded to the learning of more fine-grained categories that have not been defined in training sets. In this chapter, we propose a new approach, called label inference curriculum network (LICN), to fine-grained label learning by incorporating the coarse category labels and captions provided in public datasets. First, we build a semantic label map based on embedding techniques and a knowledge base to describe the correspondence between coarse labels and fine-grained label proposals; second, we introduce the label inference curriculum network with the consideration of the complexity of samples that describes the difficulty of fine-grained label learning. To evaluate the performance of fine-grained label learning, we construct multiple datasets based on widely-used public datasets. Experimental results demonstrate the effectiveness of our approach in the task of fine-grained label learning.

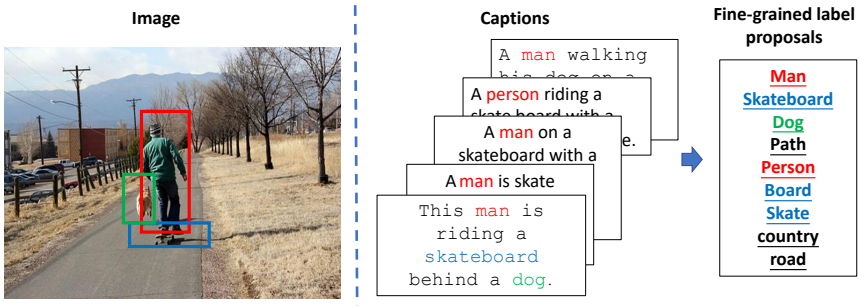


Figure 3.1: An illustration of the image-caption pair. For an image, the location of objects (bounding boxes), the corresponding coarse labels, and the attached captions are provided in the datasets for training. In general, the captions consist of a set of fine-grained label proposals for the objects in the image.

Visual object detection and classification is a fundamental problem in computer vision research and has a wide range of applications, such as face perception, autonomous vehicles and pedestrian detection. Since the renaissance of deep neural networks, object detection has been revolutionized by a series of groundbreaking works, including Faster-RCNN [30], Mask-RCNN [57] and YOLO [58].

Despite these achievements, most deep learning methods have an important limitation: they are trained with exhaustive and clean human annotations. These annotations are expensive as they require human to mark the label and the bounding boxes. Furthermore, labels provided by different annotators are possibly inconsistent. An alternative approach is to relax this requirement of exhaustively labeled data and to use web sources of annotated data, such as social media services like Flickr and Twitter, which have user-generated image tags or captions [59][60]. These data can be seen as natural annotations of the images, providing weak supervision of the collected data, which is a cheap way to increase the scale of datasets near-infinitely.

Weakly supervised object detection (WSOD) is training an object detection model without explicit bounding box annotations. The classic WSOD problem formulation [61][62] treats all object labels per image as a bag of proposals (image-level supervision), and learns to assign instance-level semantics to these proposals using multiple instance learning (MIL). The state-of-the-art model for weakly supervised object detection has reached 43.1% Mean Average Precision [63] on the Pascal VOC 2007 test set. However, there has a strong critical assumption of WSOD is that the image-level labels should be precise, indicating at least one proposal object in the image associated with one label in the image-level labels. This is always the case, especially not in real-world problems and real-world supervision.

A challenge of user-generated labeling (tags or captions) is that these anno-

tations have a lot of noisy labels: Past work has shown that weakly supervised learning algorithms can use these noisy labels [64][65]. However, captions lack information on minor objects or information that may be deemed unimportant, a phenomenon known as reporting bias [66][67]. For example, Fig.3.1 illustrates image captions that describe the same object (marked by a red bounding box) in the image but using different words (person and man) than the predefined category label (person). It is noteworthy that the word "man" is more fine-grained than "person" in describing this object. Also, references to objects may be ambiguous, for example in cases where there are multiple persons in the image.

In this chapter, we focus on a new problem called fine-grained label learning that is different from the traditional WSOD problems. Suppose we have a set of data that is paired image and captions, as shown in Fig.3.1, where the location and coarse labels are provided as ground truth in training sets. In this chapter, we aim to detection objects and learn the fine-grained labels under the joint supervision of the coarse label for an object and the captions for an image. The problem has the following two characteristics. First, the fine-grained labels need to be learned from captions, and thus the supervision of captions is considerably weak, noisy and ambiguous as analyzed above. Second, the uncertainty of noise and ambiguity in the supervision of captions results in different difficulties in the learning process for different examples, and thus the order of training data may affect the learning performance.

To address the problem, this chapter formulates the task of fine-grained label learning with the joint supervision of coarse labels and captions and proposes a novel approach called label inference curriculum network (LICN).

First, we build a semantic mapping that provide a correspondence between the coarse labels and fine-grained label proposals coming from captions based on embedding techniques and a knowledge base. Furthermore, we design a curriculum learning process for the Faster R-CNN backbone, where a term called the complexity of samples (CoS) is defined to determine the order of training data in the curriculum learning process.

In summary, our contributions are four-fold. First, we introduce and formulate the problem of fine-grained label learning based on the joint supervision of the coarse category labels and captions. Second, we build a semantic mapping between the coarse labels and fine-grained label proposals coming from captions based on embedding techniques and a knowledge base. Third, we propose a novel approach called LICN and design the weakly supervised curriculum learning process for improving the learning performance, where the complexity of samples (CoS) is defined to determine the order of training data in the curriculum learning process. Finally, we construct the datasets consisting of both coarse and fine-grained labels based on MS COCO and Visual Genome for the evaluation of our approach, and the experimental results demonstrate the effectiveness of our approach.

The rest of this chapter is organized as follows. Section 3.1 presents a brief overview of related work. Section 3.2 formulates the problem of fine-grained label

learning and introduces our approach in details. Section 3.3 provides the experimental results and analysis, and Section 3.4 concludes the chapter.

3.1. RELATED WORK

The task of weakly supervised object detection involves the correlation of different media and the information distribution from images to the corresponding captions. Therefore, we review the related work in terms of lexico semantic analysis and weakly supervised entity localization.

The task in this chapter has some differences from the following related problems:

Learning from Text: Ye et al. [63] harvest detection models from free-form text and use a label inference module to amplify signals in the free-formed texts to supervise the learning of a multiple instance detection network. Fang et al. [65] use multiple instance learning to train visual detectors for words that commonly occur in captions. Most learning from text model does not use the same semantic word in text as new category.

Weakly Supervised Object Detection and Segmentation(WSOD) [68, 69, 70]: In general, this task aims to detect objects from images based on the supervision of a set of image-level labels. To the best of our knowledge, the existing WSOD methods have not involved captions, a type of weaker supervisory information than exact image-level labels, in object detection.

Fined-Grained Image Classification(FGIC) [71][72]: FGIC usually involves classifying the sub-classes of objects belonging to the same class. In each class, objects of different subclasses are both semantically and visually similar to each other.

3.1.1. LEXICO-SEMANTIC ANALYSIS

In the widely-used public image datasets, there is typically a semantic gap between the human-written captions and the categorical annotations of the objects in the images. For example,the annotation of the object in red box is “person” while the caption uses the word “man” in Fig 3.1. A variety of lexico-semantic methods have been proposed to bridge this semantic gap. These methods can be divided into two categories: knowledge-based methods and corpus-based methods [73][16]. Knowledge-based methods rely on external semantic resources (thesauri or lexical knowledge bases) to identify similarities between two words. For example, WordNet [74] and HowNet [75] [76] are used to measure semantic distance between a pair of words. Although these semantic metrics are interpretable and effective, they have as drawbacks that they lack context information and that the similarity can only be computed when both words are present in the lexicon.

Due to the knowledge-based methods limitations, corpus-based methods are then proposed to utilize context information around the center words. Current corpus-based methods train vector representations (called ‘embeddings’) based

on contexts of words in a large text collection. The word similarity study mostly uses a statistical description of the context [77][11]. The most used static word embedding model is Word2Vec [26][25], a highly efficient model proposed by Google. The model can simplify the processing the text context into a K-dimensional vector space, so we can use the spatial similarity to represent similarity in text semantics. Li et al. [78] provide a transferred vector approach, that utilizes a transferred vector for the representation of a word to reveal the word semantics better, not just relying on its own embedding. In our work, we use these two types of model, i.e., WordNet [74] and Word2Vec [26][25], to build a semantic map between the pre-annotated coarse labels and the fine-grained label proposals from captions.

3.1.2. WEAKLY SUPERVISED MULTIPLE INSTANCE LEARNING

Most weakly supervised methods for object detection formulate the task as a multiple instance learning (MIL) problem. In this problem, MIL addresses the data objects represented by a bag of instances and associated with a label (a set of labels) for each bag. If the image is labeled as containing an object, at least one of the label proposals will be responsible for providing the prediction of that object. The papers by Oquab et al. [79] and Zhou et al. [80] propose a Global Average (Max) Pooling layer to learn class activation maps. Bilen at al. [61] propose Weakly Supervised Deep Detection Networks (WSDDN) containing classification and detection data streams, where the detection stream weighs the results of the classification predictions. Kantorov et al. [81] improve WSDDN by considering context. Tang et al. [69][82] jointly train multiple refining models together with WSDDN, and show the final model benefits from the online iterative refinement. Diba et al. [57] and Wei et al. [83] apply a segmentation map and Wan et al. [62] incorporate saliency. Finally, Redmon et al. [58] introduce a min-entropy loss to reduce the randomness of the detection results.

Our work is similar to all the above since we also represent the proposals using a MIL weighted representation. However, we go one step further to successfully adopt a more challenging supervision scenario where the captions are utilized as the weak supervision for the learning fine-grained labels in the task of object detection.

3.1.3. CURRICULUM LEARNING

Curriculum learning[84] was proposed by Yoshua Bengio in 2009. It formalizes the learning process of humans and animals from easy cases to gradually more complex ones. In recent years, more and more weakly supervised learning methods based on curriculum learning have been proposed and obtain good performance [85][86]. CurriculumNet [41] designs a curriculum learning process by measuring the complexity of data using its distribution density in a feature space for the classification of large-scale weakly-supervised web images without human annotations, where the negative impact of noisy labels is reduced substantially. Wang et al. [87] address the object detection problem by learning an effective object

detector using weakly-annotated images with curriculum learning. Hacohen et al. [88] analyze the effect of curriculum learning, which involves the non-uniform sampling of mini-batches, on the training of deep networks. In this chapter, we design a curriculum learning process by defining a new measurement of the degree of difficulty in fine-grained label learning.

3.2. METHODOLOGY

3.2.1. OVERVIEW

In this chapter, we are given a pair consisting of an image and its captions. Formally, we have $\mathscr{D}_{tr} = \{(I_i, R_i, L_i^I, C_i)\}_{i=1}^{M_{tr}}$ and $\mathscr{D}_{te} = \{(I_i, L_i^I, C_i)\}_{i=1}^{M_{te}}$ as the training set and test set, respectively, where I_i and C_i denote the i-th image and caption, respectively, and $L_i^I = \{l_{i1}^I, l_{i2}^I, \cdots, l_{im_i}^I\}$ refers to the annotations of I_i, each considered as a coarse category label for one of the m_i visual object regions $R_i = \{r_{i1}, r_{i2}, \cdots, r_{im_i}\}$ segmented from this image. The caption C_i consists of a set of entities that generally provide more fine-grained category information than L_i^I for the visual object regions R_i and thus we extract them from captions as fine-grained label proposals, denoted by $L_i^C = \{l_{i1}^C, l_{i2}^C, \cdots, l_{in_i}^C\}$. In this manner, we have a coarse label vocabulary V_I and a fine-grained label proposal vocabulary V_C that consist of all coarse labels and fine-grained label proposals, respectively, where $l_{i\cdot}^I \in V_I$ and $l_{i\cdot}^C \in V_C$. Regarding the labels we make two observations: 1) the label proposals L_i^C from captions are generally more fine-grained than the annotations L_i^I preassigned to the image; 2) The correspondence at the granularity of instances (i.e., between a fine-grained label proposal $l_{i\cdot}^C$ and a visual object region $r_{i\cdot}$) is missing. An example can be seen in the second image of Fig. 3.2(a). It is in this image unknown which region corresponds to the fine-grained label "man" or "woman" as extracted from the captions.

We aim to learn and infer the fine-grained label $l_{i\cdot} \in V_I \cup V_C$ for each visual object region based on the supervision from the training data D_{tr}. As illustrated in Fig. 3.2, our framework includes two main processes: semantic mapping and curriculum learning-based fine-grained label learning. In the semantic mapping, we extract the entities from captions as the fine-grained label proposals $l_{i\cdot}^C \in V_C$ and measure the semantic similarity between the extracted label proposals and the coarse labels $l_{i\cdot}^I \in V_I$ based on the combination of the knowledge base WordNet and data-driven embedding techniques. To learn the fine-grained label for each object, we propose a curriculum learning-based method to train the model by adding data in an ascending order of example complexity.

3.2.2. SEMANTIC MAPPING

The purpose of the semantic map is to build the relationship between the coarse label $l_{i\cdot}^I$ and the fine-grained label proposals $l_{i\cdot}^C$ by measuring their similarity over the training set. We extract all nouns from captions with the CoreNLP toolkit [89]

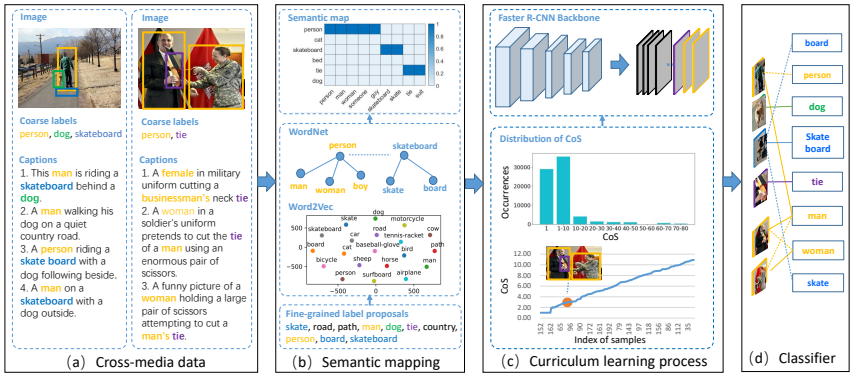


Figure 3.2: The framework of the proposed LICN approach. (a) The input data in the form of image-captions pairs, where the image, coarse labels and captions are provided in training sets. (b) Semantic mapping between the coarse labels and fine-grained label proposals based on embedding techniques and a knowledge base. (c) Curriculum learning process for the Faster R-CNN backbone, where the complexity of samples is defined to measure the degree of difficulty in learning the fine-grained labels. (d) The classifier for predicting fine-grained labels.

as the candidates for the fine-grained label proposals. In order to get a semantic map we pass three steps.

Semantic Mapping Based on Knowledge Base

We employ WordNet as the knowledge base to measure the semantic similarity between annotations and fine-grained label proposals. WordNet can represent relations between word senses with an ontology. For an annotation $l^I_{i.}$, we can obtain the synset $W_{kb}(l^I_{i.})$ from WordNet in the form of:

$$W_{kb}(l^I_{i.}) = \{H_{per}(l^I_{i.}), H_{pon}(l^I_{i.}), S_{non}(l^I_{i.})\}, \tag{3.1}$$

where $H_{per}(\cdot)$, $H_{pon}(\cdot)$ and $S_{non}(\cdot)$ refer to the hypernym, hyponym and synonym, respectively, for a given word in the WordNet.

Semantic Mapping Based on Embedding

We use Word2Vec as the embedding technique to measure the similarity of labels in $V_I \cup V_C$. We fine-tune the pre-trained Word2Vec model [26] on all captions in the data. In this chapter, we extract all words in captions to build a vocabulary. By our analysis, all coarse labels preassigned to images appear in this vocabulary, so that we can obtain the feature vector of each coarse label in embedding space. As the fine-grained labels are extracted from the captions, we can obtain the feature vectors of fine-grained labels as well. For a coarse label $l^I_{i.}$ and a fine-grained label proposal $l^C_{i.}$, we achieve their d_e-dimensional embedding vectors $\mathbf{l}^I_{i.}$ and $\mathbf{l}^C_{i.}$, respectively, with the Word2Vec model. The similarity between two vectors in the embedding space is measured by the cosine similarity $S(\cdot,\cdot)$.

Building the Semantic Map

As analyzed above, we build a semantic mapping between the annotations $l_{i\cdot}^{I}$ and the fine-grained label proposals $l_{i\cdot}^{C}$ with the following matrix:

$$W(l_{i\cdot}^{I}, l_{i\cdot}^{C}) = \begin{cases} 1, \ if \ l_{i\cdot}^{C} \in W_{kb}(l_{i\cdot}^{I}) \ and \ S(\mathbf{l}_{i\cdot}^{I}, \mathbf{l}_{i\cdot}^{C}) > \varepsilon \\ 0, \ otherwise, \end{cases} \tag{3.2}$$

where ε is a threshold in $[0,1]$. With Eq.3.2, we can find one or multiple fine-grained label proposals that are semantically similar with the given annotation. Since a visual object region strictly corresponds to an annotation in the dataset, we can achieve a weak correspondence between visual object regions and fine-grained label proposals.

3.2.3. FINE-GRAINED LABEL LEARNING BASED ON CURRICULUM LEARNING

Curriculum learning is an effective learning framework that imposes structure on the training set relying on a notion of "easy" and "hard" examples [84]. In the following subsection, we will find that the examples are of different difficulties to learn and infer the fine-grained labels. Therefore, we perform the fine-grained object label learning based on the curriculum learning framework.

Backbone for Object Detection

Based on the semantic mapping introduced in Subsection 3.2.2, we have achieved the correspondence between each visual object region $r_{i\cdot}$ in the i-th image and its fine-grained label proposals (a subset of L_i^C). Without ambiguity, we redenote them by r_k and $\widetilde{L}_k^C$ by removing the subscript i (the index of images), where k is the index of a visual object region in the dataset, $r_k \in R_i$ and $\widetilde{L}_k^C \subset L_i^C$. Thus, our objective is to localize the visual object and learn its fine-grained label with the weak supervision of a set of fine-grained label proposals $\widetilde{L}_k^C$ to the visual object region r_k.

We use the Faster R-CNN model [30], denoted by $F_{det}(I_i)$, as the backbone of our work. The Faster R-CNN consists of three modules: a convolutional neural network for generating the feature map of an image, an RPN (region proposal network) for generating a set of rectangular object proposals performed on the feature map, and a classifier for learning the category label of each region. The output of the backbone can be described as follows:

$$(\mathbf{P}_i, R_i) = F_{det}(I_i), \tag{3.3}$$

where $R_i = \{r_{ij}\}_{j=1}^{m_i}$ denotes the set of m_i visual object regions extracted from the image I_i, where the location of each region is described by four coordinates of the bounding box, and $\mathbf{P}_i = [\mathbf{p}_{i,1}, \mathbf{p}_{i,2}, \cdots, \mathbf{p}_{i,m_i}]$ denotes the probabilities that all object regions in R_i are predicted to categories. Without ambiguity, we rewrite

$\mathbf{p}_{i,j}$ as $\mathbf{p}_k = [p_{k,1}, p_{k,2}, \cdots, p_{k,C_C}]^T$ by removing the index of images, where $p_{k,c}$ denotes the probability that a visual object region r_k is categorized into the c-th class and C_C denotes the cardinality of V_C (the same as the cardinality of $V_I \cup V_C$ as all annotations appear in the fine-grained label proposals). In our work, we define the space of categories with the fine-grained label proposals, i.e., V_C.

The Complexity of Samples

Different samples have different difficulty in the learning of fine-grained labels. For example, if there is only an object region annotated by "person" in an image and only an fine-grained label proposal "man" in the caption is related to the annotation according to the semantic mapping in Eq.3.2, it is easy to infer the fined-grained label for the object region. In contrast, if there are multiple fined-grained label proposals corresponding to the annotation according to the semantic mapping, it is much difficult to discriminate which one is the true fine-grained label of the object region in the image. We introduce a term called *the complexity of samples* (CoS) to describe the difficulty in the task. We define the CoS of a sample $D_i \in \mathscr{D}$ as follows:

$$H_{CoS}(D_i) = -\sum_{l^I_{i\cdot}} \sum_{l^C_{i\cdot}} \Pr(l^C_{i\cdot}|l^I_{i\cdot}) \log(\Pr(l^C_{i\cdot}|l^I_{i\cdot})), \tag{3.4}$$

where $\Pr(l^C_{i\cdot}|l^I_{i\cdot})$ is the conditional probability of the fine-grained label proposal $l^C_{i\cdot}$ given the annotation $l_{i\cdot}$ and can be achieved by:

$$\Pr(l^C_{i\cdot}|l^I_{i\cdot}) = \frac{W(l^I_{i\cdot}, l^C_{i\cdot})}{\sum_{l^C_{i\cdot} \sim l^I_{i\cdot}} W(l^I_{i\cdot}, l^C_{i\cdot})}, \tag{3.5}$$

where $l^C_{i\cdot} \sim l^I_{i\cdot}$ denotes all fine-grained label proposals $l^C_{i\cdot}$ related to the annotation $l^I_{i\cdot}$ according to Eq.3.2. As shown in Eq.3.4, CoS is defined based on the Shannon's Entropy that is mainly used to measure the uncertainty of a discrete random variable. In this chapter, we consider $l^I_{i\cdot}$ as the random variable and $l^C_{i\cdot}$ as its values with non-zero probability. If more fine-grained label proposals are related to the annotation, the correspondence between them is more uncertain and the label proposal is thus more intractable. Moreover, if there are multiple visual objects detected in an image, the CoS tends to be a larger value accordingly based on Eq.3.4.

Curriculum Learning Process

Based on the semantic mapping, we have obtained the fine-grained label proposals $\widetilde{L}^C_k$ for each visual object region r_k. Here we transform $\widetilde{L}^C_k$ to a binary vector $\mathbf{y}_k = [y_{k,1}, y_{k,2}, \cdots, y_{k,C_C}]^T \in \{0,1\}^{C_C}$. $y_{k,c} = 1$ ($y_{k,c} = 0$) means the c-th fine-grained label proposal of V_C is present (absent) in $\widetilde{L}^C_k$.

In the curriculum learning process, the training data are fed to the Faster R-CNN in the order of easy samples (with low CoS) to hard samples (with high CoS). The loss for the learning of fine-grained labels is defined as follows:

$$L_{ws}^{k} = \sum_{c=1}^{C_C} y_{k,c} \cdot \log p_{k,c} + (1 - y_{k,c}) \cdot (1 - \log p_{k,c}), \tag{3.6}$$

where L_{ws}^{k} refers to the weakly supervised loss. Different from the original Faster R-CNN, the ground truth of label vector, i.e., $\mathbf{y}_k$, may consist of multiple ones corresponding to multiple fine-grained label proposals, rather than being a one-hot vector.

3.3. EXPERIMENTAL RESULTS AND DISCUSSION

In this section, we evaluate the effectiveness of the proposed model LICN by answering the following two questions. Q1: What is the quality of the learnt fine-grained label proposals semantic map reasonable for this weakly supervised object detection model? Q2: How effective the proposed LICN approach is in terms of the fine-grained label learning based on weakly supervised paradigm learning?

3.3.1. EXPERIMENTAL SETUP

For the experimental setup, we first describe the dataset and then the implementation details.

Datasets

The experiments are conducted on the MS COCO 2017 dataset, Visual Genome, the Pascal VOC 2007 test dataset, and our constructed datasets based on the three datasets. Table 3.1 shows an overview of these datasets.

- The *MS COCO 2017* dataset contains 118,287 training images and 5,000 validation images. It provides 5 human-annotated captions per image and a

Table 3.1: An overview of the datasets.

datasets	# of images	# of categories	# of objects
Visual Genome	107,228	80,138	3,909,697
MS COCO	118,287	80	860,001
FG-COCO	118,287	169	860,001
sCOCO training	76,631	69	200,962
FG-sCOCO training	76,631	150	200,962
FG-sCOCO test	13,175	150	29,169
FG-sCOCO val.	2,000	150	14,090
Visual Genome test	54,212	150	496,809

total of 80 category labels for the object regions segmented from all the images. The category labels play the role of the annotations L_i^I and the captions are used for the building of the semantic map and the extraction of fine-grained label proposals L_i^C for image I_i.

- *Visual Genome* contains 107,228 images, 3,909,697 objects from 80,138 categories, and other information such as the relationships between objects. The categories in Visual Genome are much more fine-grained than those in MS COCO, and thus we use this dataset for testing the performance of fine-grained label inference and the category labels as the ground truth.

- The *Pascal VOC 2007 test* dataset has 4,952 images and 20 categories of objects. It is utilized as as the test dataset.

Based on the above datasets, We construct the following datasets for training and testing our approach from different aspects:

- *FG-COCO*: We replace the coarse category labels of the objects in each image in MS COCO by the fine-grained label proposals appearing in the corresponding caption based on the semantic map and thus obtain FG-COCO. A total of 169 category labels (including the original coarse labels from MS COCO and new fine-grained category labels) are generated for the objects in the dataset.

- *FG-sCOCO test dataset*: There are a set of images appearing both in MS COCO and in Visual Genome. For an image in the set, if the Intersection over Union (IoU) between a bounding box from MS COCO and a bounding box from Visual Genome is larger than 0.90, we keep the image as an image example, and the bounding box from MS COCO and category labels from Visual Genome (must appear in the corresponding caption from MS COCO as well) as the ground truth of the location and fine-grained label for an object, respectively. We randomly choose 2000 images from the set for validation (called FG-sCOCO val. as shown in Table 3.1), and the rest is for test. As a result, the FG-sCOCO test dataset consists of 13,175 images and 29,169 objects with 150 category labels (including the original coarse labels from MS COCO and new fine-grained category labels). In the experiments, we adopt the FG-sCOCO test dataset to evaluate the performance of fine-grained label learning and inference.

- *FG-sCOCO training dataset*: It is a subset of FG-COCO, which excludes all the images appearing in the FG-sCOCO test and FG-sCOCO val. dataset. This dataset consists of 76,631 images and 200,962 objects with 150 category labels (including the original coarse labels from MS COCO and new fine-grained category label proposals from the semantic map). To make the learning robust, we keep only the categories consisting of more than 200 examples of object regions in the dataset.

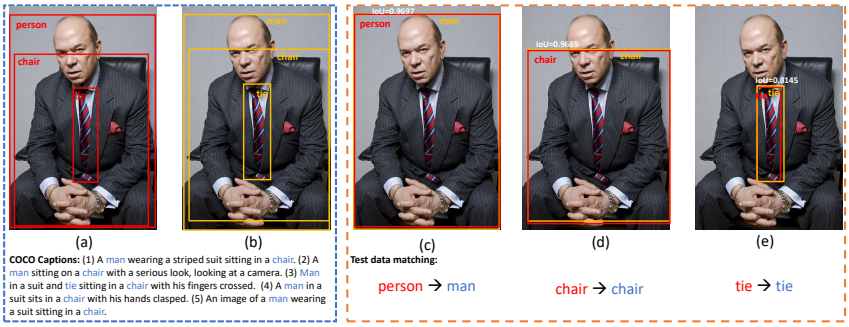


Figure 3.3: Test data example: (a) shows an example from MS COCO with object bounding boxes and the associated category labels (red color); (b) shows the same image in the Visual Genome dataset with object bounding boxes and the associated category labels (blue color); (c), (d) and (e) show the matching between the object regions from MS COCO and Visual Genome with an IoU value larger than 0.90. We see that "person" matches to "man", "chair" to "chair" and "tie" to "tie".

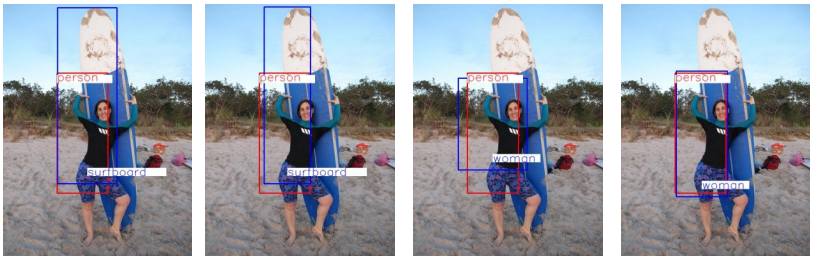


Figure 3.4: IoU example: The red color box and blue color box come from MS COCO and Visual Genome, respectively, and the IoU value of two different boxes of the same object should be high.

- *sCOCO training dataset*: As a subset of MS COCO, it consists of all the images in FG-sCOCO training dataset, and its bounding boxes and category labels are from MS COCO. As a result, the dataset consists of 76,631 images and 69 category labels for 200,962 objects.

- *Visual Genome test dataset*: Different from FG-sCOCO test dataset, Visual Genome test dataset is the subset of Visual Genome that excludes all the images appearing in MS COCO. In this dataset, we only keep those objects whose category labels appear in the FG-sCOCO training dataset. As a result, the dataset consists of 54,212 images and 496,809 objects with 150 category labels.

The following is an analysis of the building of the FG-sCOCO test dataset. We assume that the IoU value is high for a paired bounding boxes of the same object

Table 3.2: The characteristics of the interaction of MS COCO and Visual Genome with different IoU threshold values.

IoU	# of images	# of objects	# of categories in MS COCO	# of categories in Visual Genome
0.50	30,983	96,529	79	2,004
0.55	29,337	85,468	79	1,680
0.60	27,621	75,772	79	1,407
0.65	26,118	67,248	79	1,143
0.70	24,890	59,503	78	940
0.75	23,591	51,222	77	787
0.80	21,958	41,848	76	654
0.85	19,603	31,303	76	537
0.90	15,529	19,702	74	413
0.95	7,306	7,957	72	281

in the same image from the overlapping part between Visual Genome and MS COCO. As shown in Fig.3.3, the paired bounding boxes with high IoU value has the same semantics, but may have different object labels. As Visual Genome has 80K category labels which contain all fine-grained categories, we use these object labels as ground truth label to evaluate the semantic map (Q1). We illustrate the role of the threshold on the IoU value in Fig. 3.4. Table 3.2 shows the effect of different IoU threshold values on the data. For example, for the IoU of 0.90, there are 19,702 paired objects with an IoU larger than 0.90 from 15,529 images, and these objects belong to 74 categories in MS COCO and 413 categories in Visual Genome. Considering the count of test data and the count of the object categories, we will evaluate our model on the FG-sCOCO test dataset with IoU in $[0.90, 1]$. For the object detection (Q2), we found that size of images are a little different between MS COCO and Visual Genome for the image with same id. We resize the size of Visual Genome images to make them equal to the size of same image in MS COCO.

Implementation Details

We train the proposed models on two different datasets: FG-COCO and FG-sCOCO, and thus generate the following four configurations:

- LICN-E2C$_{FG-COCO}$: learned on the FG-COCO dataset by feeding training examples from easy to complex;
- LICN-C2E$_{FG-COCO}$: learned on the FG-COCO dataset by feeding training examples from complex to easy;
- LICN-E2C$_{FG-sCOCO}$: learned on the FG-sCOCO training dataset by feeding training examples from easy to complex;

Figure 3.5: The effect of Word2Vec similarity parameter ε in Eq.2 on the performance (weighted SMJI) of semantic mapping for the FG-sCOCO validation set.

- LICN-C2E$_{FG-sCOCO}$: learned on the FG-sCOCO training dataset by feeding training examples from complex to easy.

We use Faster R-CNN with a backbone of VGG-16 as the basic framework of our work. The VGG-16 backbone is pre-trained on ImageNet and then fine-tuned on our training datasets. In the process of fine-grained label learning, we use the stochastic gradient descent (SGD) optimizer with a momentum of 0.9 and a learning rate of 0.01. We set the maximum epoch to 20 for the convergence of learning process. The minibatch size is set to 1 for the flexible feeding of the examples of different complexity. All the experiments are conducted on a platform of 8 Nvidia Titan V GPUs with Pytorch.

3.3.2. EVALUATION METRICS

Semantic Mapping

We define a weighted semantic map Jaccard index (SMJI) for measuring the closeness between the fine-grained labels mined by semantic mapping and the fine-grained label ground truth provided in the FG-sCOCO validation set. The weighted SMJI is defined as follows:

$$W\_SMJI = \sum_k W_k \cdot \frac{L_k^{SM} \cap L_k^{GT}}{L_k^{SM} \cup L_k^{GT}}, \tag{3.7}$$

where L_k^{SM} and L_k^{GT} denote the sets of fine-grained labels mined by semantic mapping and the fine-grained label ground truth provided in the FG-sCOCO validation set, respectively, corresponding to the k-th coarse category label, and the operators $\cup$ and $\cap$ denote the union and intersection of two sets, respectively. For example, for the coarse category label of "person", L_k^{SM}= {"guy", "man", "person",

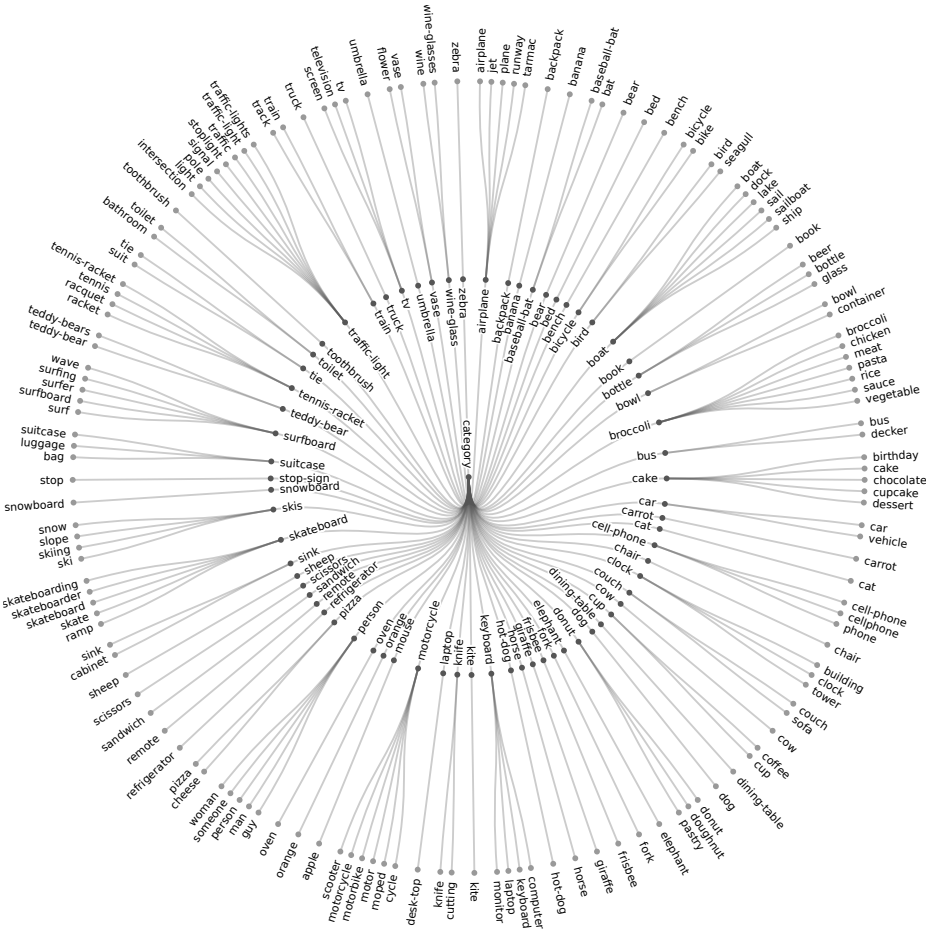


Figure 3.6: The illustration of the semantic map that consists of 69 coarse category labels (points on the inner circle) and 81 fine-grained category labels (points on the outer circle) appearing in the FG-sCOCO validation set.

"woman", "someone"} and L_k^{GT}= {"guy", "man", "person", "skateboarder", "surfer", "woman"}. The weight W_k in Eq. 3.7 is defined as follows:

$$W_k = \frac{|\, L_k^{GT} \,|}{\sum_k |\, L_k^{GT} \,|} \tag{3.8}$$

where $|\cdot|$ denotes the cardinality of a set. Fig. 3.5 reports the weighted SMJI on the FG-sCOCO validation set as the threshold ε changes. From the figure, we observe that the performance of semantic mapping in mining the fine-grained labels is optimal when $\varepsilon = 0.72$. Thus, we choose $\varepsilon = 0.72$ in the following exper-

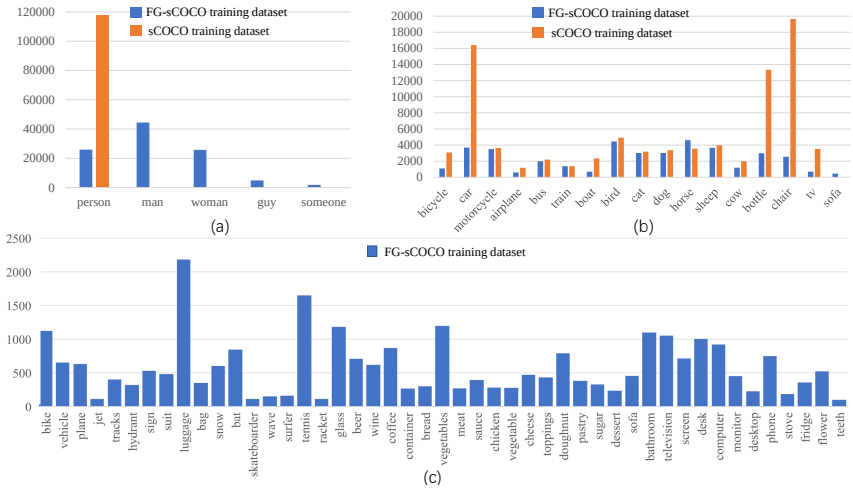


Figure 3.7: The comparison of occurrence frequencies of category labels between before and after semantic mapping, where the orange bars indicate the occurrence frequencies of the coarse labels in sCOCO training dataset and blue bars indicate the occurrence frequencies of the labels (either the original coarse labels or the generated fine-grained label proposals) in the our constructed FG-sCOCO training dataset after semantic mapping. a) Comparison between the coarse label of category "person" and the corresponding fine-grained labels, b) comparison on 17 coarse categories, and c) comparison on the generated fine-grained categories.

iments. Fig. 3.6 illustrates the semantic map that consists of 69 coarse category labels and 81 fine-grained category labels appearing in the FG-sCOCO validation set. From the figure, we observe that most fine-grained label proposals extracting from captions are semantically similar with the coarse labels, while a few noises are introduced by the semantic mapping. For example, the generated "chicken", "meat", "pasta", "rice" and "sauce" are not semantically similar with the coarse label "broccoli". These noises will be reduced with the curriculum learning process.

In Fig. 3.7, we illustrate of the occurrence frequencies of the category labels (including the coarse and fine-grained labels) in the FG-sCOCO training dataset and the sCOCO training dataset, which correspond to the data with and without semantic mapping, respectively. Due to the large difference in the occurrence frequencies of these categories, we report the results separately in three subfigures. From the figure, we find that a large amount of fine-grained label proposals are generated with semantic mapping.

Object Detection

We utilize a widely-used metric, namely average precision (AP), to evaluate the performance of object detection. AP is defined as the average detection precision under different recalls and usually evaluates the performance in a category specific manner. To compare performance over all object categories, the mean AP

Table 3.3: Average precision (AP) (%) results of LICNs trained on FG-sCOCO training dataset. The results are reported on the FG-sCOCO test dataset.

Method	Avg. Precision, IoU			Avg. Precision, Area		
	0.5:0.95	0.5	0.75	S	M	L
LICN-C2E	21.90	37.00	22.80	15.40	16.80	24.00
LICN-E2C	23.60	37.40	25.40	13.10	19.10	25.30

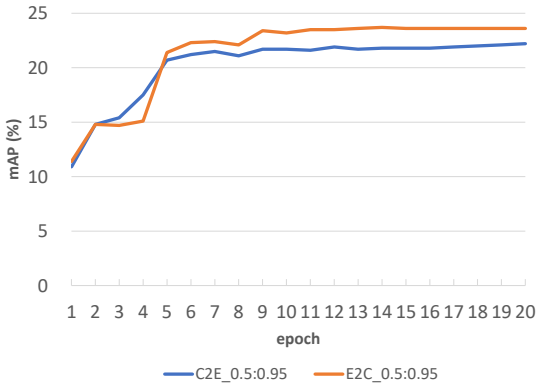
(mAP) averaged over all object categories is usually used as the final metric of performance. To measure the object localization accuracy, the IoU is used to check whether the IoU between the predicted box and the ground truth bounding box is greater than a predefined threshold 0.5. Instead of using a fixed IoU threshold, based on MS COCO AP is averaged over multiple IoU thresholds between 0.5 (coarse localization) and 0.95 (perfect localization).

3.3.3. PERFORMANCE AND ANALYSIS

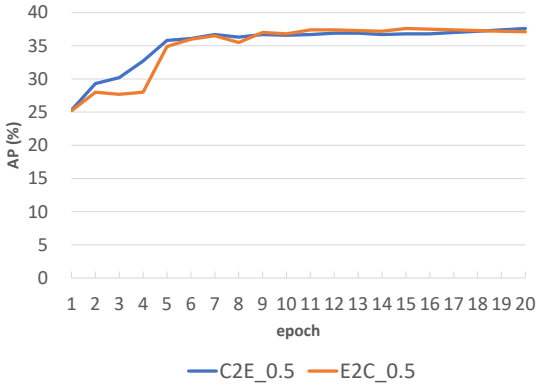
FG-sCOCO

We first evaluate our method on the FG-sCOCO validation dataset to analyze the importance of curriculum learning, where the proposed LICN models are trained on the FG-sCOCO training dataset.

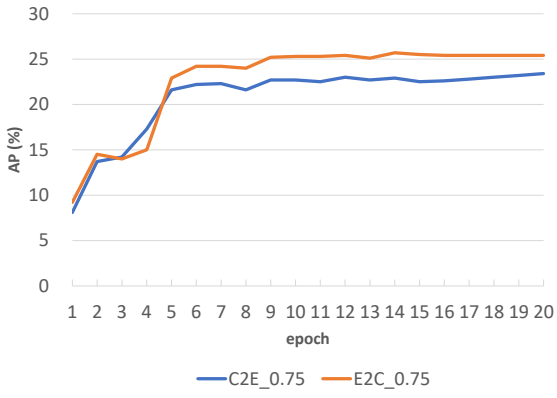
Fig.3.8 shows the results of the LICN models for the FG-sCOCO validation dataset. We find that the E2C version of LICN improves the performance of fine-grained label learning. As shown in Fig.3.8(a), in terms of the mean AP of 0.5:0.05:0.95 IoU, LICN-E2C performs approximately 0.02 AP improvement better than the LICN-C2E model. However, Fig.3.8(b) for the 0.50 IoU AP, after 7 epochs there is not a large difference between the LICN-E2C and LICN-C2E model. Fig.3.8(c) shows the 0.75 IoU AP, for which LICN-E2C performs approximately 0.03 AP better than the LICN-C2E model. LICN-E2C improves the performance for the predictions of the 0.75 IoU. As IoU means the object location accuracy, IoU close to 1 means that the predicted object location is close to the ground truth. We observe that the improvement is brought by complexity ranked as the IoU increases. This can be explained by the fact that LICN object detection is able to compute the complexity of the images, and thus it is prone to make fine-grained label predictions for the same object. Table 3.3 shows a more detailed experimental result on the FG-sCOCO test dataset. In Table 3.3, " Avg. Precision, Area S M L" means the average precisions for small ($area < 32^2$), medium ($32^2 < area < 96^2$), and large ($area > 96^2$) objects, respectively, where the area is measured as the number of pixels in the segmentation mask. The table shows that in the case of 0.75 IoU, the E2C version improves the performance by 2.6% compared with the C2E version, which demonstrates that it is better to learn the fine-grained labels with the consideration of the complexity of samples defined in the "Methodology" section.



(a)



(b)



(c)

Figure 3.8: Results of LICN-E2C and LICN-C2E on the FG-sCOCO validation set for different training epochs. (a) Mean AP of 0.50:0.95 in steps of 0.05, (b) AP of 0.5, and (c) AP of 0.75.

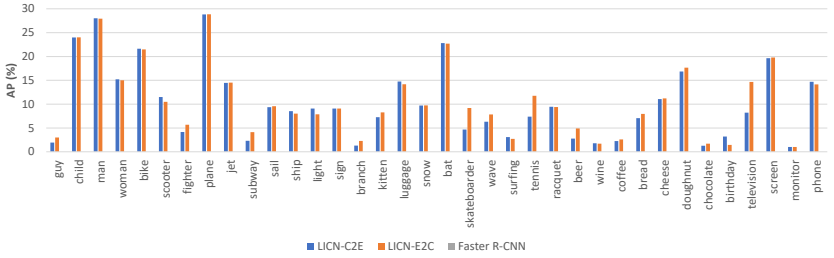
Table 3.4: Average precision (AP) (%) results for all the 20 categories of the Pascal VOC 2007 test datatet. Faster R-CNN was trained on MS COCO and the sCOCO training dataset consisting of 80 coarse labels and 69 coarse labels, respectively, and LICN was trained on the FG-COCO dataset and FG-sCOCO training dataset with the expanded labels consisting of both the coarse and fine-grained labels.

	Method	aero	bicycle	bird	boat	bottle	bus	car	cat	chair	cow	diningtable	dog	horse	motorbike	person	pottedplant	sheep	sofa	train	tvmonitor	**mean**
FG-COCO	ratio	0.48	0.89	0.98	0.90	0.96	0.96	1.00	0.99	1.00	0.99	1.00	1.00	1.00	0.94	0.69	1.00	0.95	1.00	0.96	0.92	
	Faster R-CNN[30]	**84.0**	**83.1**	**76.5**	**58.9**	**67.7**	**87.4**	77.1	85.6	61.0	**83.9**	**66.3**	78.4	86.3	86.6	**86.2**	50.9	81.7	68.1	86.1	**78.8**	**76.7**
	LICN-C2E	76.8	71.7	74.3	52.7	62.4	87.2	79.7	85.3	60.6	82.5	65.0	79.3	85.5	85.7	70.5	50.2	81.4	68.1	86.5	74.2	74.0
	LICN-E2C	71.6	69.9	75.2	52.9	64.1	87.0	**80.0**	**86.5**	**62.0**	83.6	65.6	**81.0**	**86.4**	**86.7**	68.2	**54.2**	**83.2**	**70.7**	**86.8**	76.9	74.6
FG-sCOCO	ratio	0.26	0.51	0.97	0.67	0.38	0.72	1.00	1.00	1.00	1.00	-	1.00	1.00	0.86	0.11	-	0.93	1.00	0.91	0.59	
	Faster R-CNN[30]	**77.4**	**79.7**	71.5	**58.9**	**52.3**	**85.2**	74.4	**86.3**	38.4	77.5	-	80.4	85.6	81.9	**83.9**	-	81.2	**64.1**	85.2	**64.3**	**73.8**
	LICN-C2E	69.9	76.1	68.6	50.9	41.8	81.4	73.4	85.9	37.3	74.4	-	78.3	84.0	81.2	46.3	-	76.1	63.7	84.1	59.6	68.5
	LICN-E2C	71.7	77.3	**73.8**	48.0	42.7	79.3	**75.4**	86.2	**39.4**	**79.5**	-	**80.5**	**86.2**	**82.7**	47.1	-	**81.4**	63.7	**86.1**	61.7	70.1

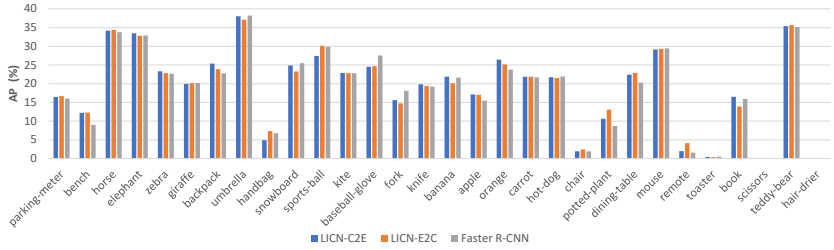
VOC 2007

We train our model on FG-COCO and the FG-sCOCO training dataset and test the learned models on the VOC 2007 test dataset to evaluate the object detection performance. Correspondingly, the Faster R-CNN baseline is trained on MS COCO and the sCOCO training dataset. Table 3.4 shows the experimental results for the 20 coarse categories in the VOC 2007 test dataset, where only 18 categories are shown for our model learned on the FG-sCOCO training dataset as the categories of "diningtable" and "pottedplant" do not appear in the training set. The table shows a term called *ratio*, which is defined as the ratio of the number of occurrences for a category in the training set FG-COCO (FG-sCOCO training) to that in the training set MS COCO (sCOCO training) and describes the degree of how many objects in a coarse category of MS COCO (sCOCO training) have not been re-assigned to a corresponding fine-grained category of FG-COCO (FG-sCOCO training) with the semantic mapping. The ratio equal to 1 means that no object in MS COCO (sCOCO training) is re-assigned to a fine-grained category and its coarse label is kept in constructing FG-COCO (FG-sCOCO training). From the table, we observe that for most of the categories with the ratio close to 1, such as "car", "chair", "dog" and "train", the detection result of our proposed LICN-E2C version has better performance than the Faster R-CNN baseline. For these categories, the training examples are almost the same between FG-COCO (FG-sCOCO training) and MS-COCO (sCOCO training). The result demonstrates that our approach improves the label inference performance in the image detection problem. For the categories with the ratio much lower than 1, such as "aero" and "person", LICN has a lower performance than Faster-RCNN. We note that in this case, there is a large difference between the training sets for LICN and Faster R-CNN: FG-COCO (FG-sCOCO training) has a much larger label space and less training examples for many categories than MS COCO (sCOCO training), which significantly increases the difficulty of label learning and inference and thus results in the the

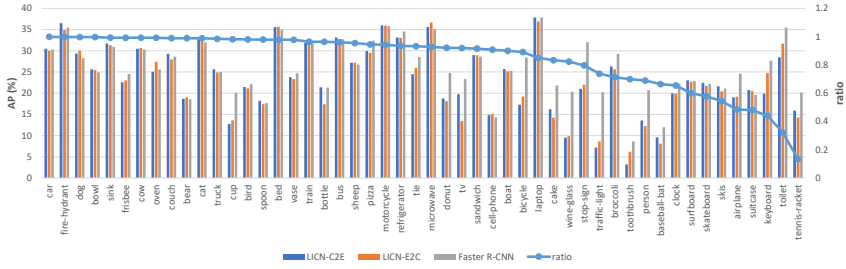
drop of AP of LICN. It is noteworthy that our LICN-E2C achieves improvements of 0.6% and 1.6% compared with LICN-C2E with the training on FG-COCO and the FG-sCOCO training dataset, respectively. The results indicate that it is important to train the model in an ascending order of example complexity in improving the object detection performance.



(a)



(b)



(c)

Figure 3.9: The comparison of LICNs and Faster R-CNN, where the former is trained on FG-COCO and the latter on MS COCO. As introduced in Subsection 4.1.1, both datasets consist of the same images. The testing results are reported for the Visual Genome test dataset. (a) shows the results for the fine-grained categories whose labels are not appearing in MS COCO. (b) shows results for the coarse categories that have no corresponding fine-grained labels in the semantic map, i.e., $ratio = 1$. (c) shows the results for the coarse categories where different proportions of object samples are re-labeled by new fine-grained labels with semantic mapping, i.e., $ratio \in (0,1)$.

Visual Genome

In this subsection, we evaluate the performance of our approach on the Visual Genome test dataset, where LICNs and Faster R-CNN are trained on FG-COCO and MS COCO, respectively.

Fig. 3.9 reports the comparison results of different methods on the test dataset in three cases: a) Fig. 3.9(a) shows the results for the fine-grained categories that do not appear in MS COCO and do come from the semantic mapping; b) Fig. 3.9(b) is for the coarse categories that have no corresponding fine-grained labels in the semantic map, i.e., the information for these categories in the training set MS COCO is the same as that in FG-COCO, and $ratio = 1$; and c) Fig. 3.9(c) is for the coarse categories, where different proportions of object samples with these category labels in training set MS COCO are re-labeled by new fine-grained labels with semantic mapping in building FG-COCO, i.e., $ratio \in (0,1)$. In Fig. 3.9(a), we see that the proposed LICN-E2C performs better than LICN-C2E for some fine-grained categories, such as “guy", “fighter", “subway", “branch", “skateboarder", “wave", “tennis" , “bear" and “television". The mean AP of the LICN-E2C model over all categories in Fig. 3.9(a) is 10.73, which achieves 0.61 mAP improvement over LICN-C2E (10.12). However, Faster R-CNN baseline training on the coarse categories cannot detect the new fine-grained categories. So its $AP = 0$ for these categories (the gray bars are not visible for that reason). From Fig. 3.9(b), we can see that for those coarse categories that have not been re-annotated with fine-grained category labels, there are no obvious differences between these three models. As shown in Fig. 3.9(c), for each coarse category in which a proportion of object samples have been re-annotated with fine-grained labels from captions by semantic mapping, Faster R-CNN has a better performance because it's training dataset, i.e., MS COCO, consists of less categories and more examples in each of these categories than the training set of LICN. With the ratio decreases, Faster R-CNN tends to increase the improvement because the number of objects re-assigned from the coarse categories to the fine-grained categories increases continuously. But our LICN model also achieves a performance close to Faster R-CNN for the categories with the ratio close to 1.

Actually, the problem of fine-grained label learning with the weak supervision of captions resolved by our approach is more challenging than the object detection and label inference resolved by the compared method, i.e., Faster R-CNN. The main reason is that the category space coming from captions in our problem (e.g., 150-dim as shown in Table 3.1) is much larger and consists of much more labeling noise than that in the latter problem.

Example Illustrations

Fig. 3.10 shows 5 fine-grained categories, namely “man”, “woman”, “plane”, “bike" and “bat”, predicted in object detection with our approach. For each category, we show 4 representative images with top confidence of category prediction. The illustration shows that our LICN approach can truly predict fine-grain category label with the weak supervision of captions.

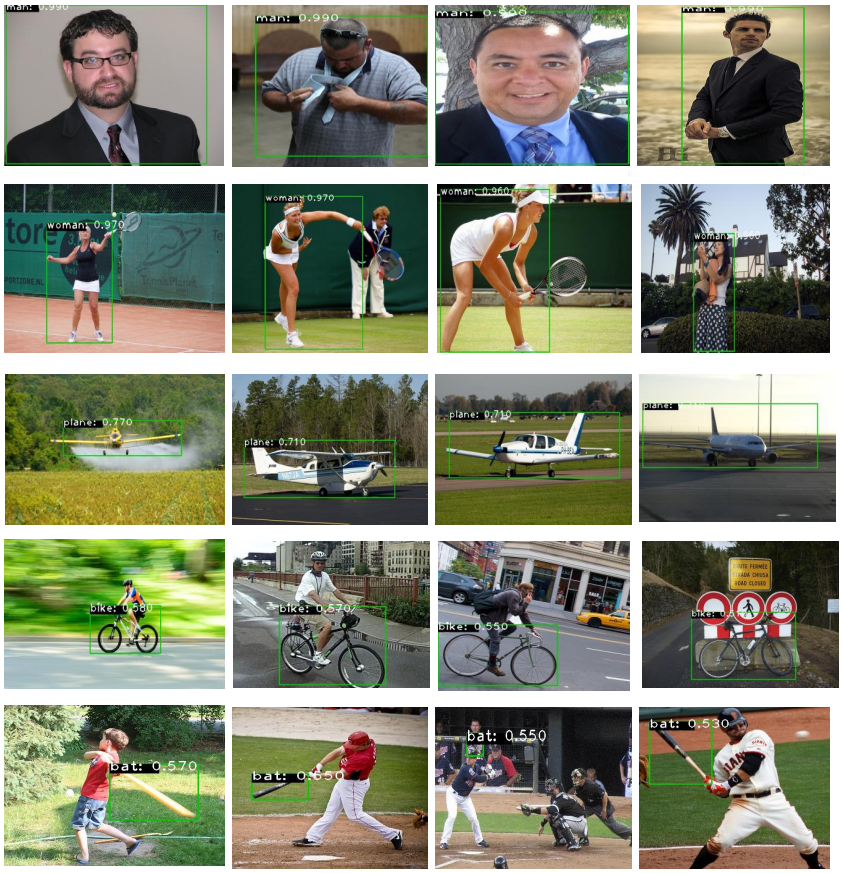


Figure 3.10: Example illustration of 5 fine-grained categories: “man”, “woman”, “bike”, “plane” and “bat”, which correspond to the coarse categories: “person”, “person”, “airplane”, “bicycle” and “baseball bat”, respectively. The values next to bounding boxes indicate the confidences of fine-grained label prediction.

3.4. CONCLUSION AND FUTURE WORK

This chapter seeks to answer the question of how to learn the fine-grained object labels in object detection with the help of auxiliary information attached to images. In this chapter, we propose a novel approach called label inference curriculum network (LICN) to the problem of fine-grained object label learning with the weak supervision of captions. First, we construct a semantic map that builds a correspondence between the coarse category labels provided by public datasets and the fine-grained category labels extracted from captions based on the combination of embedding techniques and knowledge bases. Second, we present the label inference curriculum network with the consideration of the complexity of samples that describes the difficulty of fine-grained label learning. To evaluate the performance of fine-grained object label learning in different aspects, we construct multiple datasets based on widely-used public datasets. Experimental results implemented on the public datasets and our constructed datasets demonstrate the effectiveness of our approach and show that it is helpful to structure the training process in the order of easy samples to hard samples in the task under the framework of curriculum learning.

4

KERNEL-BASED MIXTURE MAPPING FOR IMAGE AND TEXT ASSOCIATION

This chapter is based on the following publication:

Du, Y., Wang, X., Cui, Y., Wang, H., Su, C. (2019). Kernel-Based Mixture Mapping for Image and Text Association. IEEE Transactions on Multimedia, 22(2), 365-379.

CHAPTER SUMMARY

This chapter addresses RQ4.

RQ4: How and with what quality can we model the semantic correlations between two different modalities?

Modeling the relationship between multi modal media, including images, videos, and text, can reduce the gap between the modalities and promote cross-media retrieval, image annotation, etc. In this chapter, we propose a new approach called kernel-based mixture mapping (KMM) to model the semantic correlations between web images and text. With this approach, we first construct latent high-dimensional feature spaces based on kernel theory to address the non-linearity of both the data distributions in the input spaces and the cross-model correlation. Second, we present a probabilistic neighborhood model to describe the spatial locality of semantics by assuming that proximate examples in feature spaces generally have the same semantics and a conditional model to describe cross-modal conditional dependency. Finally, we build a probabilistic mixture model to jointly model the spatial locality of semantics and the conditional dependency between different modalities. By combining nonlinear transformation and probabilistic models, KMM can address the non-linearity of cross-modal correlation, the complexity of the semantic distributions at the global scale, and the continuity of semantic distributions at the local scale. We present a hybrid optimization algorithm to find the solution of KMM based on expectation-maximization and sub gradient ascent; this algorithm avoids estimating the parameters of KMM in high-dimensional feature spaces and is proved to converge to an (local) optimal solution. We demonstrate the performance of KMM using for public datasets.The experimental results show that our approach outperforms the compared methods when modeling the relationships between image and text.

With the rapid development of the Internet, there has been a massive explosion of multimedia content, such as text, image, audio and videos, on the web. These types of content usually coexist in a multimedia document and complement each other to express similar semantics. For example, an image provides a visual description of a concept, yet this description is usually incomplete. In contrast, text can accurately describe the abstraction of a concept, but it is not intuitive. Consequently, joint exploitation of the full information from different modalities could facilitate accurate content interpretation. Currently, many real-world internet applications, such as cross-media retrieval [90, 91, 92], image caption or summary generation [93], image annotation [94, 95, 96] and information recommendation [97], involve multimodal data. For these applications, the relationship between modalities needs to be considered. Many previous studies focused on the understanding of the unimodal scenario, in which data are homogeneously represented and similarity is measured in a single feature spaces. However, different data modalities are associated with different metric spaces, and thus similarity cannot be measured directly between heterogeneous modalities. The vastly different representations derived from heterogeneous modalities make it very challenging to associate signals across these modalities.

The work related to the semantic correlation mining of heterogeneous media can be categorized into the following four main classes: 1) linear/non-linear mapping [98], [99], such as canonical correlation analysis (CCA), 2) probabilistic models, such as probabilistic latent semantic analysis (PLSA) [100], 3) graph-based correlation propagation methods [92], [101], and 4) deep learning-based methods [102],[103]. In [43], the authors presented an approach called mixture of local linear mapping (MLLM) to cross-modal semantic correlation modeling. MLLM considers that close examples in a local region generally represent a uniform concept and are supposed to be mapped to another modality based on a linear model, and then combines multiple linear mapping models to represent the relationships between different modalities on the whole data distribution. However, MLLM cannot address the non-linearity of data distributions and cross-media correlations very well.

In this chapter, we first analyze the ineffectiveness of linear mapping models and then propose a novel approach, called kernel-based mixture mapping (KMM), to model the semantic association between text and images. Similar to our previous method MLLM, KMM considers that the data in a local region of the input spaces follow a local mapping model and uses a mixture of local mapping models to substitute a more complex nonlinear mapping. In KMM, we introduce a probabilistic neighborhood model to accurately describe how data in a local region follow the corresponding local mapping model. To address the nonlinearity of data distributions and cross-media correlations, KMM first transforms the textual and visual data from the input spaces into two latent high-dimensional spaces by nonlinear feature space mapping functions, and then constructs the mapping model between both modalities in the latent features spaces. The smoothness and sparseness of the parameters are introduced to enhance the generaliza-

tion of models and the fitness between models and data. We present a hybrid optimization algorithm based on expectation-maximization (EM) and subgradient ascent to find the solution of KMM; the parameters are estimated using kernel theory to avoid the explicit representation of both feature spaces.

In summary, our contributions are three-fold: 1) We analyze the ineffectiveness of linear models and reveal that linear models' prediction is close to a zero vector for cross-media retrieval due to the linear uncorrelation between images and text at the global scale in feature space. 2) We present a parameterized model-driven approach, called KMM, to model cross-modal association. KMM provides a kernel-based probabilistic mixture model to describe the distribution that cross-modal data need to follow and addresses the complexity of the semantic distribution at the global scale, its continuity at the local scale, and the non-linearity in the mapping of different modalities. 3) We introduce a hybrid optimization algorithm based on the frameworks of EM and subgradient ascent and prove its convergence to an (local) optimum. The optimization algorithm overcomes the difficulty in estimating the parameters of the KMM model because our model does not consist of explicit inner products for being replaced directly by kernel functions.

The rest of this chapter is organized as follows. Section 4.1 presents a brief overview of related work. Section 4.2 breifly analyzes the ineffectiveness of linear mapping models. Section 4.3 describes our KMM approach to the modeling of image and text association. Section 4.4 presents the optimization, algorithm and analysis for KMM. Section 4.5 provides the experimental results, and Section 4.6 concludes the chapter.

4.1. RELATED WORK

Studies related to cross-media modeling can be divided into four main classes.

(1) **Linear or nonlinear mapping**. This class of methods builds a linear or nonlinear (closed-form) transformation model between heterogeneous input spaces or from both input spaces to a latent semantic space where similarity is measured. Grangier and Bengio [104] proposed a linear discriminant approach for cross-modal retrieval by linearly transforming one modality to the other and extended the linear model to a nonlinear one through the kernel trick. Jiang and Tan [105] presented a vague linear transformation to measure the information similarity between visual and textual modalities through a set of predefined domain-specific information categories. There are some other approaches that transform both modalities into a common space, which can be constructed based on CCA [90, 106, 107], matrix factorization [108, 109], or by preserving a certain structure of data [110]. The similarity between multiple modalities can be measured in the common space. Tang et al. [111] presented a cross-space affinity model that was learned with an optimization problem, where the restriction of exact correspondences between different modalities was relaxed to their relative similarities. In addition, some researchers proposed mixture models to describe the relationship between two sources of data. In [112], Deleforge et al. introduced a model

called Gaussian locally-linear mapping (GLLiM) for high-dimensional regression. Different from [43], GLLiM model aims to solve the inverse regression problem. Hannah et al. [113] presented a more general regression model by introducing generalized linear models. To handle diverse content more appropriately, Hua et al. [114] presented a method called TINA that built a set of local linear projections for each modality and then measured the relations of pairs of local models for different modalities. To address nonlinearity of data distributions, Zhang et al. [115] and Xu et al. [116] introduced kernel mapping in data representation.

(2) **Probabilistic methods**. Probabilistic methods generally aim to maximize the probabilities that the data of one modality can be generated for the given inputs of the other modality. Jeon et al. [117] proposed an approach to annotating and retrieving images that directly modeled the joint distributions over blobs in images and words in text. Different from [117], Monay and Pere [100] computed the joint distributions over images and text based on PLSA by introducing a latent semantic variable. Feng and Lapata [118] proposed an approach to image captioning based on the latent Dirichlet allocation (LDA) model and generated the keywords of captions by maximizing the posterior probabilities given the image and its corresponding textual documents. Zhang et al. [119] supposed that features of images and text are independently generated by a certain concept and modeled cross-media relationships under the Bayesian framework. To improve learning performance, Wu et al. [120] incorporated unlabeled data in the training process of image retrieval and learned the model by maximizing the joint probabilities of labeled and unlabeled data with the discriminant-EM algorithm. To relax the restriction discussed in many studies regarding full correspondence between modalities, Jia et al. [121] proposed a method for analyzing the semantic correlation between modalities based on a Markov random field of topic models for realistic scenarios, where a narrative text is only loosely related to an image. Different from the above studies, Pham et al. [122] presented a method for learning fine-grained relationships between images and text, i.e., the correspondences between the keywords in text and the visual regions in images, based on EM algorithm.

(3) **Graph-based correlation propagation**. Generally, graph-based methods model multimedia with each document as a vertex and the relationship between documents as an edge, and propagate the correlation information to learn the cross-modality similarity over the graph [92]. Zhai et al. [101] constructed a kNN graph for each modality and performed cross-media retrieval by determining whether the examples from different modalities have the same label or not. Lin et al. [123] presented a PLSA-based aspect model to measure the inter-correlation between different modalities and intra-correlation in the same modality, and then constructed a multi-modal propagation network for cross-media retrieval. Lazaridis et al. [124] presented a novel framework based on kNN graphs for multimodal search of rich media objects, in which Laplacian eigenmaps were employed to merge low-level descriptors and create a new low-dimensional multimodal feature space. Xue et al. [125] proposed a graph-based approach that contained

two processes of semantic correlation computing for modeling the semantic correlation between web images and text. In the work, information propagation was jointly driven by the local semantics of visual blobs or words and the global semantics of documents. In [126], a multiple graph-based multi-label learning framework was proposed for image annotation problem, in which the visual content of images, semantic correlation of tags and the prior information provided by users were simultaneously considered. The multi-graph strategy was also used in [127], where the authors jointly modeled the intra-modal local topology structures of each graph constructed on one modality and the inter-modal local topology structures to obtain the final common embedding space for multiple modalities.

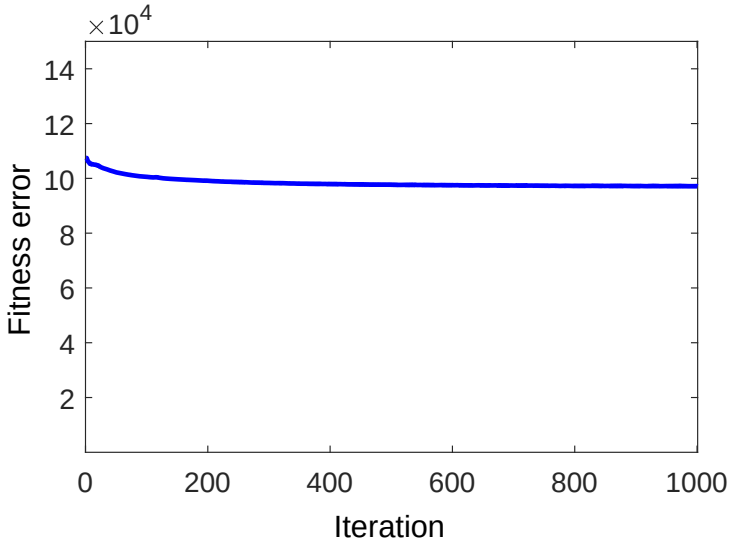
(4) **Deep learning-based methods**. In general, deep learning-based methods jointly map different modalities into an embedding space using deep networks and measure similarity in this space. Deep CCA [128] is a representative approach to cross-media correlation modeling, which represents each modality with a deep network and measures the similarity based on CCA. Different from deep CCA, Wang et al. [129, 103] measured the similarity between different modalities based on cross-view ranking constraints or the element-wise product. Eisenschtat and Wolf [102] presented a bidirectional neural network architecture for matching images and text, in which two tied neural network channels were used to project both views into a common, maximally correlated space using Euclidean loss. To make cross-modal correlation modeling more precise, Peng et al. [130] fused coarse-grained instances and fine-grained patches and learned the relationships between images and text based on the constraints of the intra-modality semantic category and the inter-modality pairwise similarity. To make an efficient retrieval, Hong et al. [131] presented a novel joint semantic-visual space by leveraging visual descriptors to narrow the semantic gap and provided an efficient online multimedia service. In addition to image-text association modeling, Wang et al. [132] focused on making correlations between movies and text and proposed a novel model called layered memory network, which can encode the temporal alignment between sentences and frames inside movie clips. Most of the deep learning-based methods model the relationships between different modalities by parameter tuning in the representation process. Different from these methods, we build an explicit probabilistic model to describe the cross-modality relation based on the representation from deep networks.

4.2. LINEAR MODELS AND THE INEFFECTIVENESS

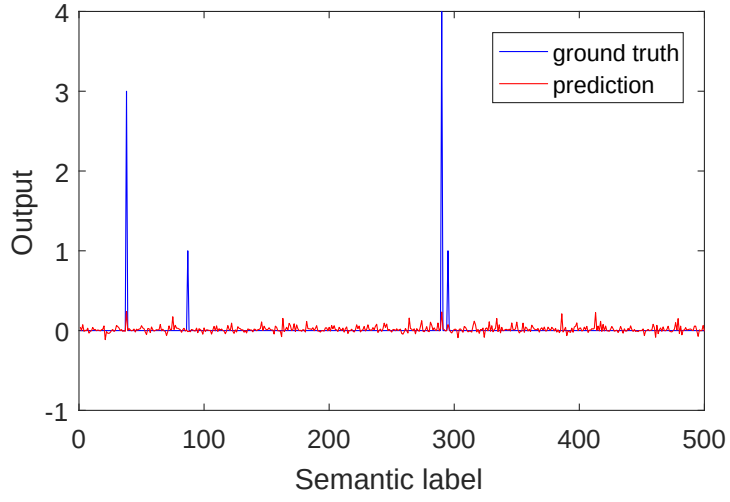
In general, there is a natural correspondence between visual space and textual space. Let

$$\mathscr{M} : \mathbf{R}^T \rightarrow \mathbf{R}^I \tag{4.1}$$

be an invertible map from the textual space to the visual space. Similarly, given a query of visual image $\mathbf{x}^I \in \mathbf{R}^I$, its corresponding textual sample in textual space can be achieved with the inverse of $\mathscr{M}$, i.e., $\mathscr{M}^{-1}(\mathbf{x}^I)$.



(a)



(b)

Figure 4.1: An example of linear transformation from textual spaces to visual spaces using the Corel5K dataset: (a) the decrease of fitness error with iteration and (b) the comparison between the prediction (red curve) and the ground-truth image (blue curve) for a certain textual input. In Fig. 4.1(b), images are represented by the bag of visual words (BoVW) model (500 visual words), and the y-axis denotes the value of each entry of feature vectors for the prediction or corresponding ground truth given a textual input.

Many previous models for mapping the heterogeneous modalities are constructed as linear models [90, 106, 105]. Jiang and Tan [105] transformed text (or images) to the other modality with a linear model and computed the similarity between the ground truth and the corresponding prediction:

$$\hat{\mathbf{x}}^I = \mathbf{M}_{CI}\mathbf{M}_{TC}\mathbf{x}^T \ , \tag{4.2}$$

where $\hat{\mathbf{x}}^I$ is the prediction in the visual space, and $\mathbf{M}_{TC}$ and $\mathbf{M}_{CI}$ are the transformation matrices from textual spaces to concept spaces and from concept spaces to visual spaces, respectively. The similarity can be measured using Euclidean distance in both spaces:

$$d_F(\mathbf{x}^I, \mathbf{x}^T) = \|\mathbf{x}^I - \mathbf{M}_{CI}\mathbf{M}_{TC}\mathbf{x}^T\|_2 \ . \tag{4.3}$$

4

Actually, images and text originate from two completely different systems. Furthermore, visual features and textual features are complicated and nonlinearly distributed in their respective spaces. Consequently, constructing a map between both spaces with a linear model is intuitively inaccurate. Fig. 4.1 illustrates an example of linear transformation from textual spaces to visual spaces using the Corel5K dataset. From Fig. 4.1(a), we find that the fitness error decreases by only approximately 10% through iterative optimization. Fig. 4.1(b) shows that the prediction result in visual space is similar to a random noise around 0 along the semantic label dimension and has a large difference from the ground truth. Theorem 1 provides a theoretical analysis.

Theorem 1. *If $\mathbf{x}^T$ and $\mathbf{x}^I$ are linearly uncorrelated, the solution to Eq.4.2 with the minimization of the distance shown in Eq.4.3 over all data is a zero vector that is independent of the distribution of $\mathbf{x}^I$.*

Proof. Let $\mathbf{x}^I = (\mathbf{x}_1^I, \mathbf{x}_2^I, \cdots, \mathbf{x}_N^I)$ and $\mathbf{x}^T = (\mathbf{x}_1^T, \mathbf{x}_2^T, \cdots, \mathbf{x}_N^T)$ be the data matrices for images and text, respectively. Without loss of generality, we assume that $\mathbf{x}^I$ and $\mathbf{x}^T$ have a mean of zero. When $\mathbf{x}^T$ and $\mathbf{x}^I$ are linearly uncorrelated, the correlation coefficient can be computed as follows:

$$\begin{aligned} \rho &= \frac{tr(\mathbf{C}_{TI}\mathbf{C}_{TI}^T)}{\sqrt{tr(\mathbf{C}_{TT}\mathbf{C}_{TT}^T)tr(\mathbf{C}_{II}\mathbf{C}_{II}^T)}} \\ &= 0 \ , \end{aligned} \tag{4.4}$$

where $\mathbf{C}_{TI} = \mathbf{X}^I\mathbf{x}^{T\prime}$, $\mathbf{C}_{TT} = \mathbf{X}^T\mathbf{x}^{T\prime}$ and $\mathbf{C}_{II} = \mathbf{X}^I\mathbf{x}^{I^T}$. In this chapter, $\mathbf{x}^{T\prime}$ means transpose of the matrix $\mathbf{x}^T$. Thus, $tr(\mathbf{C}_{TI}\mathbf{C}_{TI}^T) = \|\mathbf{C}_{TI}\|_F^2 = 0$. By minimizing the distance in Eq.4.3, the following is attained:

$$\begin{aligned} \hat{\mathbf{x}}^I &= \mathbf{M}_{CI}\mathbf{M}_{TC}\mathbf{x}^T \\ &= \mathbf{X}^I\mathbf{x}^{T\prime}(\mathbf{x}^T\mathbf{x}^{T\prime})^{-1}\mathbf{x}^T \\ &= \mathbf{C}_{TI}\mathbf{C}_{TT}^{-1}\mathbf{x}^T \ . \end{aligned} \tag{4.5}$$

Then, $\|\hat{\mathbf{x}}^I\| \leq \|\mathbf{C}_{TI}\|_F\|\mathbf{C}_{TT}^{-1}\mathbf{x}^T\| = 0$, and thus $\hat{\mathbf{x}}^I = 0$. □

4.3. PROPOSED MODEL

4.3.1. LOCAL LINEAR MAPPING

In this approach, we write the map as $\mathscr{M} : X \rightarrow Y$. Without loss of generality, we let $X = \mathbf{R}^T$ and $Y = \mathbf{R}^I$. We consider that the map from a local region of X to Y can be described by a linear model due to the simplicity of the local data distribution. We characterize the linear mapping model $\mathscr{M}$ over the local region by the concatenation of two matrices as follows:

$$\begin{aligned} \mathbf{y}_i &= \hat{\mathbf{y}}_i + \varepsilon_i \\ &= \mathbf{W} \cdot \mathbf{V}\mathbf{x}_i + \varepsilon_i \ , \end{aligned} \tag{4.6}$$

where $\mathbf{x}_i \in X$ denotes an input (or a query), $\mathbf{y}_i, \hat{\mathbf{y}}_i \in Y$ are the corresponding ground-truth output and the prediction in the other modality, respectively, $\mathbf{V}$ is the transformation matrix from the input space to a latent semantic space, $\mathbf{W}$ is the transformation matrix from the semantic space to the output space, and ε_i denotes the fitness error. In our work, we assume the fitness error ε_i follows a normal distribution with zero mean and covariance matrix Σ. Given the model $\mathscr{M}$ and an input $\mathbf{x}_i$, the probability distribution of the ground-truth output $\mathbf{y}_i$ is formulated as follows:

$$\Pr(\mathbf{y}_i|\mathbf{x}_i, \mathscr{M}) = \frac{1}{\sqrt{(2\pi)^{d_y}|\Sigma|}} e^{-\frac{1}{2}d(\mathbf{y}_i, \mathbf{x}_i)} \ , \tag{4.7}$$

where $d(\mathbf{y}_i, \mathbf{x}_i) = (\mathbf{y}_i - \mathbf{W}\mathbf{V}\mathbf{x}_i)^T \Sigma^{-1} (\mathbf{y}_i - \mathbf{W}\mathbf{V}\mathbf{x}_i)$.

As analyzed above, we consider that a set $\mathscr{R}_m$ of close examples in a local region indexed by m has uniform semantics and approximately follows one cross-media mapping model. Intuitively, the data near the centroid of $\mathscr{R}_m$ follow the mapping model with high confidence, and those far from the centroid follow with low confidence. We then characterize the confidence with a neighborhood model $K_{\mathbf{H}}(\mathbf{x} - \mu)$ with a symmetric positive definite $d_x \times d_x$ bandwidth matrix $\mathbf{H}$, where μ is the centroid of the local region and

$$K_{\mathbf{H}}(\mathbf{x} - \mu) = |\mathbf{H}|^{-1/2} K(\mathbf{H}^{-1/2}(\mathbf{x} - \mu)) \ . \tag{4.8}$$

$K(\mathbf{x})$ is a bounded function with compact support satisfying [133]

$$\begin{aligned} &\int_{\mathbf{R}^{d_x}} K(\mathbf{x}) d\mathbf{x} = 1 \quad \lim_{||\mathbf{x}|| \rightarrow \infty} ||\mathbf{x}||^{d_x} K(\mathbf{x}) d\mathbf{x} = 0 \\ &\int_{\mathbf{R}^{d_x}} \mathbf{x} K(\mathbf{x}) d\mathbf{x} = 0 \quad \int_{\mathbf{R}^{d_x}} \mathbf{x}\mathbf{x}^T K(\mathbf{x}) d\mathbf{x} = c_K \mathbf{I} \ , \end{aligned} \tag{4.9}$$

where c_K is a constant. A Gaussian function with a zero-mean vector and an identity covariance matrix satisfies such constraints in Eq.4.9. We use $K_{\mathbf{H}_m}(\mathbf{x} - \mu_m)$ to describe the probability or confidence of the data that follow the mapping model $\mathscr{M}_m$ over $\mathscr{R}_m$, i.e., $\Pr(\mathbf{x}_i|\mathscr{M}_m)$.

The joint probability of the pair $(\mathbf{x}_i, \mathbf{y}_i)$ generated by the model $\mathscr{M}_m$ is:

$$\begin{aligned}\Pr(\mathbf{x}_i, \mathbf{y}_i|\mathscr{M}_m) &= \Pr(\mathbf{x}_i|\mathscr{M}_m)\Pr(\mathbf{y}_i|\mathbf{x}_i, \mathscr{M}_m) \\ &= K_{\mathbf{H}_m}(\mathbf{x}_i - \mu_m)\Pr(\mathbf{y}_i|\mathbf{x}_i, \mathscr{M}_m)\ ,\end{aligned} \tag{4.10}$$

where μ_m and $\mathbf{H}_m$ denote the centroid (replaced by the mean vector in computing) and bandwidth matrix, respectively, of the local region $\mathscr{R}_m$ that $\mathbf{x}_i$ belongs to.

An alternative factorization of the joint probability shown in Eq.4.10 can be performed as follows:

$$\Pr(\mathbf{x}_i, \mathbf{y}_i|\mathscr{M}'_m) = \Pr(\mathbf{y}_i|\mathscr{M}'_m)\Pr(\mathbf{x}_i|\mathbf{y}_i, \mathscr{M}'_m)\ ,$$

where $\mathscr{M}'_m$ denotes a mapping model from Y to X. This factorization is related to the inverse regression [112], where $\mathbf{y}_i$ is considered as a regressor. In this case, we need to define a neighborhood model $K_{\mathbf{H}}(\mathbf{y} - \mu)$ to describe the probability of $\mathbf{y}_i$ that follows the mapping model, i.e., $\Pr(\mathbf{y}_i|\mathscr{M}'_m)$. Compared with Eq.4.10, this factorization will result in a high computational complexity because we need to compute $\Pr(\mathbf{y}_i|\mathscr{M}'_m)$ for all $\mathbf{y}_i$ in a dataset given a certain query $\mathbf{x}_i$.

4.3.2. KERNEL-BASED MIXTURE MAPPING

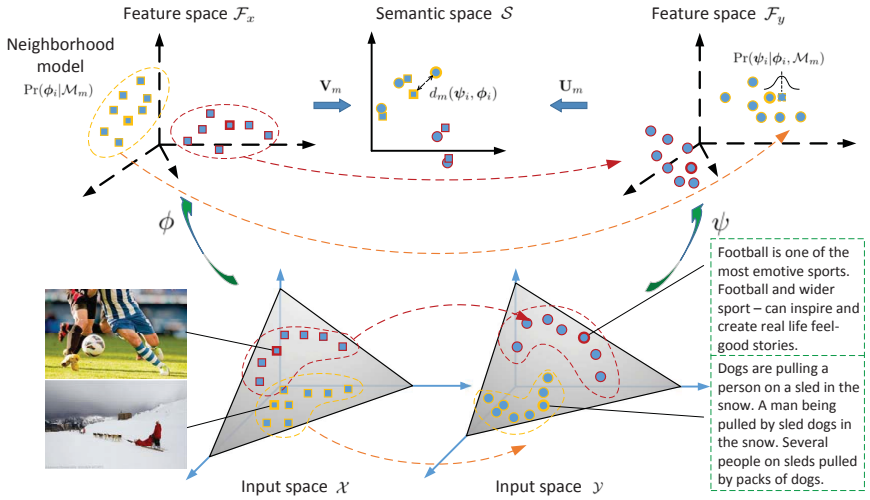
Due to the complexity of the data distribution, the map between two modalities may not follow the linear model in the original input space, and the local region in the input space cannot be depicted well by the expected Gaussian neighborhood model. Therefore, we formulate this problem in a high-dimensional latent feature space based on kernel theory. Let us consider $\phi : X \to \mathscr{F}_x$ and $\psi : Y \to \mathscr{F}_y$ that map the original input spaces into two feature spaces of dimensions d_ϕ and d_ψ, respectively, where both $\mathscr{F}_x$ and $\mathscr{F}_y$ are inner product spaces. Here, as shown in Fig. 4.2(a), we build a d_s-dimensional semantic space S by the linear transformation over both feature spaces. In the semantic space, it is easier to introduce the kernel theory and measure the similarity of two modalities. Similar to Eq.4.7, the map between two modalities can be represented by the following probabilistic model in the semantic space:

$$\Pr(\psi_i|\phi_i, \mathscr{M}_m) = \frac{1}{\sqrt{(2\pi)^{d_s}|\Sigma_m|}} e^{-\frac{1}{2}d_m(\psi_i, \phi_i)}\ , \tag{4.11}$$

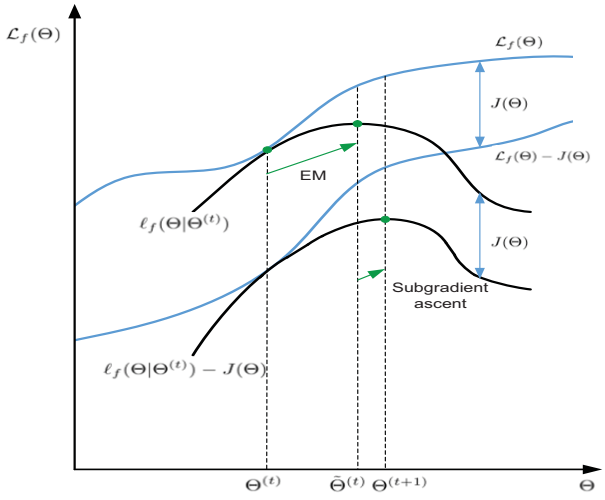
where $\phi_i \triangleq \phi(\mathbf{x}_i)$, $\psi_i \triangleq \psi(\mathbf{y}_i)$, and

$$d_m(\psi_i, \phi_i) = (\mathbf{U}_m\psi_i - \mathbf{V}_m\phi_i)^T\Sigma_m^{-1}(\mathbf{U}_m\psi_i - \mathbf{V}_m\phi_i). \tag{4.12}$$

where Σ_m denotes the covariance matrix of data points $\{\mathbf{U}_m\psi_i - \mathbf{V}_m\phi_i\}$ associated with the model $\mathscr{M}_m$ in semantic space and Σ_m^{-1} denotes the inverse. The distance $d_m(\cdot,\cdot)$ measured in S is achieved by the combination of features with matrices $\mathbf{U}_m$ and $\mathbf{V}_m$. Generally, the rows of $\mathbf{U}_m$ and $\mathbf{V}_m$ are located in the space spanned by the



(a)



(b)

Figure 4.2: Our approach. (a) The framework. The small squares and circles denote examples of images and text, respectively, located in input or feature spaces, and different colors indicate different local regions. In the input space, the local regions are supposed to follow a Gaussian neighborhood model in the feature space, while they do not follow this model in the input space. (b) Convergence analysis of hybrid optimization, which is introduced in Section 4.4 in detail.

columns of $\Psi = (\psi_i)$ and $\Phi = (\phi_i)$, respectively, i.e., $\mathbf{U}_m = \mathbf{A}_m \Psi^T$ and $\mathbf{V}_m = \mathbf{B}_m \Phi^T$. Eq.4.12 can be rewritten as:

$$d_m(\psi_i, \phi_i) = (\mathbf{A}_m K_{y,i} - \mathbf{B}_m K_{x,i})^T \Sigma_m^{-1} (\mathbf{A}_m K_{y,i} - \mathbf{B}_m K_{x,i}), \tag{4.13}$$

where each row of $\mathbf{A}_m$ and $\mathbf{B}_m$ denotes the coefficients with which the rows of $\mathbf{U}_m$ and $\mathbf{V}_m$ can be linearly reconstructed by the data points $\{\psi_i\}$ and $\{\phi_i\}$, respectively, and $K_{x,i}$ and $K_{y,i}$ denote the i-th column of kernel matrices $\mathbf{K}_x = (\phi_k \cdot \phi_l)$ and $\mathbf{K}_y = (\psi_k \cdot \psi_l)$, respectively. Based on the kernel theory [134], we can choose nonlinear kernel functions $f_\phi : X \times X \to \mathbf{R}$ and $f_\psi : Y \times Y \to \mathbf{R}$, which should follow Mercer's condition, to satisfy $f_\phi(\mathbf{x}_k, \mathbf{x}_l) = \phi_k \cdot \phi_l$ and $f_\psi(\mathbf{y}_k, \mathbf{y}_l) = \psi_k \cdot \psi_l$. Therefore, we can achieve the kernel matrix in input space instead of in feature space by choosing appropriate kernel functions.

4

The neighborhood model in the feature space $\mathscr{F}_x$ can be rewritten in the following form:

$$\Pr(\phi_i | \mathscr{M}_m) = \frac{1}{\sqrt{(2\pi)^{d_\phi} |\mathbf{H}_m|}} e^{-\frac{1}{2} \tilde{\phi}_{mi}^T \mathbf{H}_m^{-1} \tilde{\phi}_{mi}}, \tag{4.14}$$

where $\tilde{\phi}_{mi} = \phi_i - \mu_m$. It is worth noting that the neighborhood model could not have been computed in the input space X so far. We will introduce its solution method in the next section.

Due to the complicated data distribution and the nonlinear mapping between the textual and visual spaces, a single mapping model is insufficient in modeling the relationship between different media. To this end, we develop a probabilistic mixture model to characterize the cross-media mapping. Given the model, a log-likelihood function is defined based on the joint probability of N cross-media data pairs $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^N$ as follows:

$$\begin{aligned} \mathscr{L}_f &= \ln \prod_{i=1}^{N} \Pr(\phi_i, \psi_i) \\ &= \ln \prod_{i=1}^{N} \sum_{m=1}^{M} \omega_m \Pr(\phi_i | \mathscr{M}_m) \Pr(\psi_i | \phi_i, \mathscr{M}_m), \end{aligned} \tag{4.15}$$

where M is the number of components in the mixture model, and ω_m is the weight of the m-th component $\mathscr{M}_m$ with $\sum_{m=1}^{M} \omega_m = 1$ and $\omega_m \geq 0$. In the mixture model, the first probabilistic term aims to make close points share the same component $\mathscr{M}_m$, and the second term focuses on modeling the relationship between two modalities.

4.3.3. CONSTRAINTS IN THE MODEL

According to Eq.4.13, $\mathbf{A}_m$ and $\mathbf{B}_m$ are the coefficient matrices used to reconstruct $\mathbf{U}_m$ and $\mathbf{V}_m$ based on the kernel matrices $\mathbf{K}_y$ and $\mathbf{K}_x$, respectively. Here, we consider two extra constraints.

Smoothness

In general, two close data points in the input spaces X and Y are expected to have images close together in the latent semantic space S. To this end, we introduce a smoothness constraint that is defined as follows:

$$\begin{aligned} J_A &= \sum_{i\sim j} \|\mathbf{A}(K_{y,i} - K_{y,j})\|_2^2 \\ &\leq \|\mathbf{P}\|_F^2 \|\mathbf{A}\|_F^2 \ , \\ J_B &= \sum_{i\sim j} \|\mathbf{B}(K_{x,i} - K_{x,j})\|_2^2 \\ &\leq \|\mathbf{Q}\|_F^2 \|\mathbf{B}\|_F^2 \ , \end{aligned} \tag{4.16}$$

where $i \sim j$ denotes that the i-th and j-th data points are close, and $\mathbf{P}$ and $\mathbf{Q}$ are the matrices whose columns are the vectors $\{(K_{y,i} - K_{y,j})\}$ and $\{(K_{x,i} - K_{x,j})\}$, respectively, in a certain order for all $i \sim j$. Thus, we characterize the smoothness of the cross-media mapping based on Eq.4.16 as follows:

$$J_{sm}(\mathbf{A}_m, \mathbf{B}_m) = \lambda_{A,m}\|\mathbf{A}_m\|_F^2 + \lambda_{B,m}\|\mathbf{B}_m\|_F^2 \ , \tag{4.17}$$

where for simplicity, we use parameters $\lambda_{A,m}$ and $\lambda_{B,m}$ to replace the exact Frobenius norm of $\mathbf{P}$ and $\mathbf{Q}$, respectively. In our work, we let $\lambda_{A,m} = \lambda_{B,m} = 1$ and use a single λ_1 to control the importance of the smoothness term. To obtain a smooth mapping model, $J_{sm}(\mathbf{A}, \mathbf{B})$ needs to be constrained to a small value.

Sparseness

The rows of $\mathbf{U}_m$ and $\mathbf{V}_m$ can be considered as a new basis (possibly nonorthonormal) for the projection of examples $\{\psi_i\}$ and $\{\phi_i\}$, respectively, and can be linearly reconstructed by these examples. To make each basis tend to represent some specific semantics held by a subset of examples, we expect to reconstruct the rows of $\mathbf{U}_m$ and $\mathbf{V}_m$ using a few examples by enforcing each row of $\mathbf{A}_m$ and $\mathbf{B}_m$ to have a few non-zero elements. We call the characteristics sparseness and formulate it using the L_1-norm as follows:

$$J_{sp}(\mathbf{A}_m, \mathbf{B}_m) = \|\mathbf{A}_m\|_1 + \|\mathbf{B}_m\|_1 \ . \tag{4.18}$$

Incorporating both constraints into our problem, we have the final optimization problem to compute cross-media correlation:

$$\max_{\Theta} \mathscr{L}_f - \sum_{m=1}^{M} \left(\lambda_1 J_{sm}(\mathbf{A}_m, \mathbf{B}_m) + \lambda_2 J_{sp}(\mathbf{A}_m, \mathbf{B}_m)\right) \ , \tag{4.19}$$

where $\Theta = \{\omega_m, \mu_m, \mathbf{H}_m, \mathbf{A}_m, \mathbf{B}_m, \Sigma_m\}_{m=1}^{M}$ is the parameter set, and λ_1 and λ_2 are used to control the balance between the terms.

4.4. OPTIMIZATION, ALGORITHM AND ANALYSIS

4.4.1. OPTIMIZATION AND ALGORITHM

Similar to Wang et al.'s work [135], we define the following notations as shown in Table 4.1. The first five rows in this table formulate the traditional estimation of Gaussian mixture models in the input space based on EM [136]. Here, the superscript (t) refers to the t-th iteration.

Table 4.1: Notations.

$p_{mi}^{(t)} = \Pr(\mathscr{M}_m \vert \phi_i, \psi_i, \Theta^{(t)})$
$w_{mi}^{(t)} = \sqrt{p_{mi}^{(t)} / \sum_{j=1}^{N} p_{mj}^{(t)}}$
$\mu_m^{(t)} = \sum_{i=1}^{N} (w_{mi}^{(t)})^2 \phi_i$
$\tilde{\phi}_{mi}^{(t)} = \phi_i - \mu_m^{(t)}$
$\mathbf{H}_m^{(t)} = \sum_{i=1}^{N} (w_{mi}^{(t)})^2 \tilde{\phi}_{mi} \tilde{\phi}_{mi}^T$
$(\mathbf{K}_x)_{ij} = \phi_i \cdot \phi_j = f_\phi(\mathbf{x}_i, \mathbf{x}_j)$
$(\mathbf{K}_{x,m})_{ij}^{(t)} = w_{mi}^{(t)} \phi_i \cdot w_{mj}^{(t)} \phi_j$
$(\tilde{\mathbf{K}}_{x,m})_{ij}^{(t)} = w_{mi}^{(t)} \tilde{\phi}_{mi} \cdot w_{mj}^{(t)} \tilde{\phi}_{mj}$
$(\mathbf{K}'_{x,m})_{ij}^{(t)} = \phi_i \cdot w_{mj}^{(t)} \phi_j$
$(\tilde{\mathbf{K}}'_{x,m})_{ij}^{(t)} = \tilde{\phi}_{mi} \cdot w_{mj}^{(t)} \tilde{\phi}_{mj}$

The optimization problem Eq. 4.19 is different from previous regularization-based learning problems because it contains the hidden information. More specifically, we do not know which component $\mathscr{M}_m$ "generates" each pair $(\mathbf{x}_i, \mathbf{y}_i)$. To solve the optimization problem, we present a hybrid optimization algorithm based on EM and subgradient ascent. The parameters of the proposed model are $\Theta = \{\omega_m, \mu_m, \mathbf{H}_m, \mathbf{A}_m, \mathbf{B}_m, \Sigma_m\}_{m=1}^{M}$, where the first three parameters describe the neighborhood model, and the rest are for cross-media mapping.

Based on the EM algorithm, we define the following function $\tilde{\mathscr{L}}_f$ in the expectation step to help optimize problem Eq. 4.19.

$$\begin{aligned} \tilde{\mathscr{L}}_f &= \sum_{m=1}^{M} \sum_{i=1}^{N} p_{mi} \ln \left(\omega_m \Pr(\phi_i \vert \mathscr{M}_m) \Pr(\psi_i \vert \phi_i, \mathscr{M}_m) \right) \\ &= \sum_{m=1}^{M} \sum_{i=1}^{N} p_{mi} \left(\ln \omega_m + \ln \Pr(\phi_i \vert \mathscr{M}_m) + \ln \Pr(\psi_i \vert \phi_i, \mathscr{M}_m) \right). \end{aligned} \tag{4.20}$$

According to EM, the growth of $\tilde{\mathscr{L}}_f$ can increase $\mathscr{L}_f$ shown in problem Eq. 4.19.

By setting the partial derivative of $\tilde{\mathscr{L}}_f$ to zero, we can easily achieve

$$\omega_m^{(t)} = \frac{1}{N}\sum_{i=1}^{N} p_{mi}^{(t)} \ , \tag{4.21}$$

where $p_{mi}^{(t)}$ is defined in Table 4.1 and can be expanded as:

$$p_{mi}^{(t)} = \frac{\omega_m^{(t-1)} \Pr(\phi_i|\mathscr{M}_m, \Theta^{(t-1)}) \Pr(\psi_i|\phi_i, \mathscr{M}_m, \Theta^{(t-1)})}{\sum_{k=1}^{M} \omega_k^{(t-1)} \Pr(\phi_i|\mathscr{M}_k, \Theta^{(t-1)}) \Pr(\psi_i|\phi_i, \mathscr{M}_k, \Theta^{(t-1)})}. \tag{4.22}$$

The feature space $\mathscr{F}_x$ is usually of high dimension and cannot be represented explicitly. Hence, we do not directly compute the distribution $\Pr(\phi_i|\mathscr{M}_m, \Theta^{(t-1)})$ in Eq.4.15 and Eq.4.21 (sometimes $\Theta^{(t-1)}$ may be omitted to save space) and estimate the parameters $\{\mu_m, \mathbf{H}_m\}_{m=1}^{M}$ in the feature space. Instead, we may estimate the distribution in the input space with the kernel trick. First, based on the work in [137], we can rewrite the exponent term in Eq.4.14 as:

$$\begin{aligned} \tilde{\phi}_{mi}^T \mathbf{H}_m^{-1} \tilde{\phi}_{mi} =& \tilde{\phi}_{mi}^T \mathbf{V} \Lambda^{-1} \mathbf{V}^{\mathrm{T}} \tilde{\phi}_{mi} \\ =& \sum_{j=1}^{d_\phi} y_j^2 / \lambda_j \ , \end{aligned} \tag{4.23}$$

where $\mathbf{V}$ and $\Lambda^{-1} = diag(\lambda_1^{-1}, \lambda_2^{-1}, \cdots, \lambda_{d_\phi}^{-1})$ denote the matrices of the eigenvectors and eigenvalues of $\mathbf{H}_m^{-1}$, respectively, and $y_j = \tilde{\phi}_{mi}^T \mathbf{V}_j$ is the projection of $\tilde{\phi}_{mi}^T$ over the j-th eigenvector $\mathbf{V}_j$. We note that $\tilde{\mathbf{K}}_{x,m}$ and the bandwidth matrix $\mathbf{H}_m$ have the same nonzero eigenvalues $\{\lambda_j\}$. It was proved in [135] that

$$y_j = \beta_j^T \Gamma_{\cdot,i} \ , \tag{4.24}$$

where β_j is the eigenvector of $\tilde{\mathbf{K}}_{x,m}$ corresponding to the eigenvalue λ_j, and $\Gamma_{\cdot,i}$ is the column of $\tilde{\mathbf{K}}'_{x,m}$ corresponding to $\mathbf{x}_i$. Note that d_ϕ is unknown due to the implicit feature map ϕ, and we approximately estimate the distribution as the marginal density function by keeping d'_ϕ ($d'_\phi < d_\phi$) principal components that correspond to the d'_ϕ largest nonzero eigenvalues and discard the rest in Eq.4.23. Moreover, the factor $(2\pi)^{d_\phi/2}\mathbf{H}_m^{1/2}$ in Eq.4.14 can be replaced by $(2\pi)^{d'_\phi/2}\prod_{j=1}^{d'_\phi}\lambda_j^{1/2}$. Then, we can iteratively estimate $\Pr(\phi_i|\mathscr{M}_m)$ in Eq.4.20 by updating the kernel-matrix parameters $\tilde{\mathbf{K}}_{x,m}$ and $\tilde{\mathbf{K}}'_{x,m}$ shown in Table 4.1 in the input space, instead of μ_m and $\mathbf{H}_m$ in the feature space, since both sets of parameters describe the same distribution. Note that $\tilde{\mathbf{K}}_{x,m}$ and $\tilde{\mathbf{K}}'_{x,m}$ can be easily computed as the centralized versions of $\mathbf{K}_{x,m}$ and $\mathbf{K}'_{x,m}$, respectively.

For the update of the parameters $\{\mathbf{A}_m, \mathbf{B}_m, \Sigma_m | m = 1, 2, \cdots, M\}$, we build a new optimization function based on problem Eq. 4.19 and Eq.4.20:

$$\mathscr{Q} = \sum_{m=1}^{M} \left(\sum_{i}^{N} p_{mi} \log \Pr(\psi_i|\phi_i, \mathscr{M}_m) - J(\mathbf{A}_m, \mathbf{B}_m) \right), \tag{4.25}$$

where $J(\cdot,\cdot) = \lambda_1 J_{sm}(\cdot,\cdot) + \lambda_2 J_{sp}(\cdot,\cdot)$. $\mathscr{Q}$ includes the non-differentiable terms of $\|\cdot\|_1$ for the sparseness constraint, and thus a closed-form solution cannot be obtained by directly taking the derivative. We use the subgradient ascent scheme to iteratively maximize $\mathscr{Q}$. At each time t, we compute the subgradients as

$$\nabla_{\Sigma_m}\mathscr{Q} = -\frac{1}{2}\sum_i p_{mi}(\Sigma_m^{-1} - D_{m,i}\Sigma_m^{-2}D_{m,i}^T) ,$$
$$\nabla_{\mathbf{A}_m}\mathscr{Q} = -\sum_i p_{mi}\Sigma_m^{-1}\mathbf{A}_m D_{m,i}^T - \lambda_1\mathbf{A}_m - \lambda_2\Delta_{\mathbf{A}_m} ,$$
$$\nabla_{\mathbf{B}_m}\mathscr{Q} = \sum_i p_{mi}\Sigma_m^{-1}\mathbf{B}_m D_{m,i}^T - \lambda_1\mathbf{B}_m - \lambda_2\Delta_{\mathbf{B}_m} ,$$

where $D_{m,i} = \mathbf{A}_m K_{x,i} - \mathbf{B}_m K_{y,i}$, and Δ is defined as

$$(\Delta_{\mathbf{A}_m})_{ij} = \mathrm{sgn}((\mathbf{A}_m)_{ij}),\ (\Delta_{\mathbf{B}_m})_{ij} = \mathrm{sgn}((\mathbf{B}_m)_{ij}) .$$

Here, sgn(z) outputs 1 when $z > 0$, 0 when $z < 0$, and a random value uniformly distributed in [-1,1] when $z = 0$. Given the subgradients, we update the solution for Σ_m, $\mathbf{A}_m$ and $\mathbf{B}_m$ to maximize $\mathscr{Q}$ as follows:

$$\begin{aligned}\Sigma_m^{(t+1)} &= \Sigma_m^{(t)} + \eta_{\Sigma_m}^{(t)} \cdot \nabla_{\Sigma_m}\mathscr{Q} ,\\ \mathbf{A}_m^{(t+1)} &= \mathbf{A}_m^{(t)} + \eta_{\mathbf{A}_m}^{(t)} \cdot \nabla_{\mathbf{A}_m}\mathscr{Q} ,\\ \mathbf{B}_m^{(t+1)} &= \mathbf{B}_m^{(t)} + \eta_{\mathbf{B}_m}^{(t)} \cdot \nabla_{\mathbf{B}_m}\mathscr{Q} ,\end{aligned} \tag{4.26}$$

where $\eta_{\Sigma_m}^{(t)}$, $\eta_{\mathbf{A}_m}^{(t)}$ and $\eta_{\mathbf{B}_m}^{(t)}$ are the step sizes at time t. In the experiment, we set the step size to $1/t$.

Eigenvalue decomposition of large-scale matrices

When an example $\mathbf{x}_i$ is far from the center of the Gaussian component $\mathscr{M}_m$, it belongs to this component with low probability p_{mi}. Hence, we can set the corresponding columns and rows of $\tilde{\mathbf{K}}_{x,m}$ to zero to obtain an approximation, and perform eigenvalue decomposition on a smaller matrix after the elementary transformation of the matrices. Let $p_m = \max_j p_{mj}$, and in the experiments, we set the i-th columns and rows of $\tilde{\mathbf{K}}_{x,m}$ to zero if $p_{mi} < 0.01 p_m$.

Parameter initialization

First, we use the K-means clustering algorithm with the training data $\{\mathbf{x}_i\}_{i=1}^N$ and achieve M clusters in the input space. Then, we compute the values of p_{mi}, $\mathbf{K}_{x,m}$, $\tilde{\mathbf{K}}_{x,m}$, $\mathbf{K}'_{x,m}$ and $\tilde{\mathbf{K}}'_{x,m}$ over the hard partitions as the parameters at the time $t = 0$, $\Sigma_m^{(0)}$, $\mathbf{A}_m^{(0)}$ and $\mathbf{B}_m^{(0)}$ are set randomly.

4.4.2. CONVERGENCE ANALYSIS

Our model includes a hidden variable, i.e., $\mathscr{M}_m$, to indicate which local model that a pair of data points follows. The EM algorithm is a powerful tool for solving

Algorithm 2 KMM parameter estimation algorithm

Require: Image-text paired documents $D = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{N}$, each including an image and a textual document with the same semantics, kernel functions k_ϕ and k_ψ, parameters M, λ_1, λ_2, d_s, step sizes $\eta_{\Sigma_m}^{(t)}$, $\eta_{\mathbf{A}_m}^{(t)}$ and $\eta_{\mathbf{B}_m}^{(t)}$, and the maximum number of iterations T;

Ensure: The estimated parameter set $\Theta = \{\omega_m, \tilde{\mathbf{K}}_{x,m}, \tilde{\mathbf{K}}'_{x,m}, \mathbf{A}_m, \mathbf{B}_m, \Sigma_m\}_{m=1}^{M}$;

1: Initialize parameter set $\Theta^{(0)} = \{\omega_m^{(0)}, \tilde{\mathbf{K}}_{x,m}^{(0)}, \tilde{\mathbf{K}}_{x,m}^{\prime(0)}, \mathbf{A}_m^{(0)}, \mathbf{B}_m^{(0)}, \Sigma_m^{(0)}\}_{m=1}^{M}$, $t = 0$;
2: **repeat**
3: $t = t + 1$, and step sizes $\eta_{\Sigma_m}^{(t)}, \eta_{\mathbf{A}_m}^{(t)}, \eta_{\mathbf{B}_m}^{(t)} = 1/t$;
4: Compute $p_{mi}^{(t)}$ with Eq.4.22 based on $\Theta^{(t-1)}$;
5: Update $\omega_m^{(t)}$ with Eq.4.21, $\tilde{\mathbf{K}}_{x,m}^{(t)}$ and $\tilde{\mathbf{K}}_{x,m}^{\prime(t)}$ based on Table 4.1, and then get $\tilde{\Theta}^{(t)}$;
6: Update $\Sigma_m^{(t)}$, $\mathbf{A}_m^{(t)}$ and $\mathbf{B}_m^{(t)}$ with Eq.4.26, and then get $\Theta^{(t+1)}$;
7: **until** $t \geq T$.

such problems and can generally guarantee that the iterative optimization converges to a local optimal solution. In our work, we present a hybrid optimization algorithm based on the combination of EM and subgradient ascent. In this subsection, we introduce two notations: X denotes the observed data and z hidden states commonly used in the EM algorithm, which correspond to $\{(\phi_i, \psi_i)\}$ and $\{\mathscr{M}_m\}$, respectively, in our model. Here, we rewrite $\mathscr{L}_f = \ln \prod_{i=1}^{N} \Pr(\phi_i, \psi_i)$ in Eq.4.15, i.e., $\ln \Pr(X|\Theta)$, as $\mathscr{L}_f(\Theta^{(t)})$ to emphasize the parameters at a particular time t, and we define the following variable:

$$\begin{aligned}\ell_f(\Theta|\Theta^{(t)}) =& \mathscr{L}_f(\Theta^{(t)}) + \sum_z \Pr(z|X,\Theta^{(t)}) \ln\left(\frac{\Pr(X,z|\Theta)}{\Pr(X,z|\Theta^{(t)})}\right) \\ =& \mathscr{L}_f(\Theta^{(t)}) + l_1(\Theta|\Theta^{(t)}) + l_2(\Theta|\Theta^{(t)}) \\ & - \sum_z \Pr(z|X,\Theta^{(t)}) \ln \Pr(X,z|\Theta^{(t)}) \ , \end{aligned} \tag{4.27}$$

where

$$\begin{aligned} l_1(\Theta|\Theta^{(t)}) &= \sum_z \Pr(z|X,\Theta^{(t)}) \ln P_1(X,z|\Theta) \ , \\ l_2(\Theta|\Theta^{(t)}) &= \sum_z \Pr(z|X,\Theta^{(t)}) \ln P_2(X,z|\Theta) \ , \end{aligned}$$

and $\Theta^{(t)}$ denotes the current parameters at time t. In this subsection, $P_1(\cdot)$ and $P_2(\cdot)$ correspond to $\omega_m \Pr(\phi_i|\mathscr{M}_m)$ and $\Pr(\psi_i|\phi_i, \mathscr{M}_m)$, respectively, in Eq.4.16. Based on Jensens inequality, we have that $\mathscr{L}_f(\Theta) \geq \ell_f(\Theta|\Theta^{(t)})$, and then that

$$\mathscr{L}_f(\Theta) - J(\Theta) \geq \ell_f(\Theta|\Theta^{(t)}) - J(\Theta) \ . \tag{4.28}$$

where $J(\Theta)$ represents the regularization terms in problem Eq. 4.19 and Eq.4.25, i.e., $\sum_{m=1}^{M} J(\mathbf{A}_m, \mathbf{B}_m)$. In Eq.4.28, the equality holds if and only if $\Theta = \Theta^{(t)}$.

In the optimization process shown in Algorithm 1, we have the following two steps in each iteration. 1) We update ω_m, $\tilde{\mathbf{K}}_{x,m}$ and $\tilde{\mathbf{K}}'_{x,m}$ based on the EM algorithm to maximize $\sum_{m,i} p_{mi}\big(\ln\omega_m + \ln \Pr(\phi_i|\mathscr{M}_m)\big)$ in Eq.4.21, i.e., $l_1(\Theta|\Theta^{(t)})$, and obtain the parameter denoted by $\tilde{\Theta}^{(t)}$; 2) By fixing the updated parameters in step 1, we update parameters Σ_m, $\mathbf{A}_m$ and $\mathbf{B}_m$ via Eq.4.26 to maximize $\mathscr{Q}$, i.e., $l_2(\Theta|\Theta^{(t)}) - J(\Theta)$. Consequently, based on the above steps, we increase the right side of Eq.4.28 and obtain the updated parameters, denoted by $\Theta^{(t+1)}$. According to Eq.4.28, $\mathscr{L}_f(\Theta^{(t+1)}) - J(\Theta^{(t+1)}) > \mathscr{L}_f(\Theta^{(t)}) - J(\Theta^{(t)})$. Consequently, our algorithm will converge to an (local) optimal solution. Fig. 4.2(b) illustrates the optimization process.

4

4.4.3. COMPLEXITY ANALYSIS

The computational complexity of parameter estimation is mainly derived from the update of kernel matrices, e.g., $\tilde{\mathbf{K}}_{x,m}$, the eigenvalue decomposition of $\tilde{\mathbf{K}}_{x,m}$, the subgradient computation and the update in Eq.4.26. Suppose the numbers of examples handled by each component $\mathscr{M}_m$, denoted by N_m, are the same and do not change as the amount of data increases; the proposed algorithm includes the following four main parts: 1) computing $\tilde{\mathbf{K}}_{x,m}$ in $O(N_m^2 M)$ time, 2) performing the eigenvalue decomposition of $\tilde{\mathbf{K}}_{x,m}$ in $O(N_m^3 M)$ time, 3) computing the subgradients with respect to Σ_m and $\mathbf{A}_m$ ($\mathbf{B}_m$) in $O(d_s^3 M + d_s N_m^2 M)$ time and $O(d_s^3 M + d_s^2 NM + d_s N_m^2 M + d_s NM)$ time, respectively, and 4) updating Σ_m and $\mathbf{A}_m$ ($\mathbf{B}_m$) in $O(d_s^2 M)$ time and $O(d_s N)$ time, respectively. Suppose the algorithm converges in T iterations; the total computational complexity is $O(NMT)$ by keeping the higher-order terms in the above analysis.

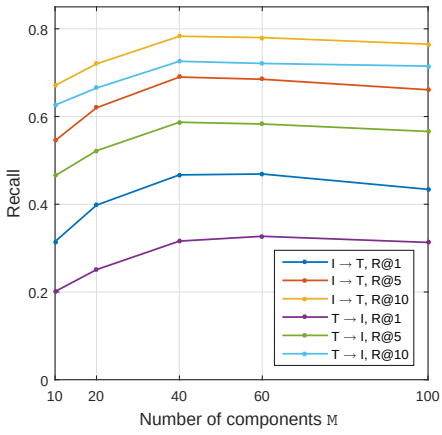
4.5. EXPERIMENTAL RESULTS

4.5.1. DATASET AND EXPERIMENTAL SETTING

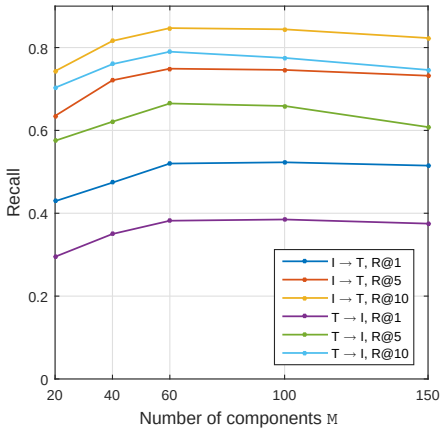
Four public real-world datasets are used in our experiments.

- *Flickr8K* dataset [138] consists of 8,000 images from the Flickr.com website, which focuses on people or animals performing actions. For each image, five captions were generated by different annotators using a crowdsourcing service. The dataset is split into disjoint training, validation, and test sets with 6,000, 1,000, and 1,000 pairs, respectively.

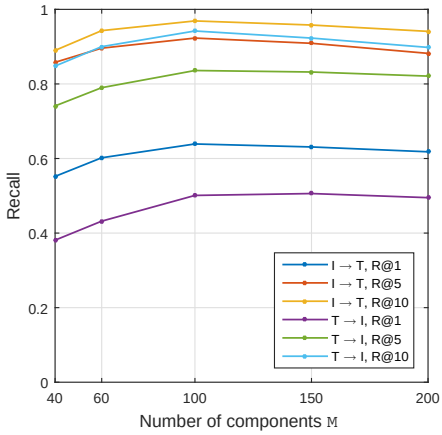
- *Flickr30K* dataset [24] extends the Flicker8K and consists of 31,783 images of everyday activities, events and scenes, each paired with five captions, i.e., a total of 158,915 captions. The captions were annotated in a similar style as in Flicker8K. We use 1,000 examples for testing, 1,000 examples for validation and the rest for training.



(a) Flickr8K



(b) Flickr30K



(c) MSCOCO

Figure 4.3: The performance of cross-media retrieval versus the number of components M in the case of "1 image vs. 1 caption" on three datasets. "I → T" and "T → I" denote "image → text" and "text → image", respectively.

- *MSCOCO* dataset [21] contains 123,287 images, each corresponding to 5 captions. Similar to [139], we randomly generate the splits that contain 5,000 images with corresponding captions for both validation and testing, and the rest of the images are used for training. The results are reported on a subset of 1,000 testing images.

- *NUS-WIDE-10K* dataset [140] has 10,000 image/text pairs in total, selected evenly from the 10 largest categories of the NUS-WIDE dataset. The dataset is split into three subsets following [130]: training set with 8,000 pairs, testing set with 1,000 pairs and validation set with 1,000 pairs.

We implement 5 independent experiments to alleviate the variation caused by random splits of datasets.

The data are represented as follows.

- *Image representation*: In the experiments, we employ two pre-trained deep networks on ImageNet, i.e., VGG-16 networks [4] and ResNet-152 [5]. We use the images resized to 224×224 as the input for both networks, and achieve 4096-dimensional feature vectors from VGG and 2048-dimensional vectors from ResNet.

- *Text representation*: We extract textual features based on Word2vec [141]. We represent every word in a commonly used 150-dimensional embedding space, and then cluster them into K groups. Finally, we employ a bag-of-words representation to describe an text instance based on the feature vectors of words. In the experiments, we let $K = 500$.

Cross-media retrieval includes two tasks: text retrieval given a query of an image and image retrieval given a textual query, which are denoted by "image → text" and "text → image", respectively. We evaluate the performance with R@r that denotes the recall at r for both tasks. Since Flickr8k, Flickr30k and MSCOCO contain 5 captions per image, we evaluate the proposed approach in two cases: 1) "1 image vs. 1 caption", in which each caption is considered as a response or a query in the retrieval, and the recall at r for "image → text" task is computed based on whether at least one of the correct captions is among the first r retrieved ones [142], and 2) "1 image vs. 5 captions", in which the 5 captions corresponding to an image are concatenated as a response or a query [128]. In the cases of "1 image vs. 1 caption" and "1 image vs. 5 captions", we train KMM based on the pair of an image and each of its 5 captions [103] and the pair of an image and its concatenated captions [128], respectively. Regarding NUS-WIDE-10K, like [130], we consider the set of multiple tags for an image as a text instance in both retrieval tasks and evaluate the performance with the mean average precision (mAP) score.

In the experiment, we compare the proposed approach with the following state-of-the-art methods.

- Deep CCA [128]: representing images and captions using deep neural networks and then correlating them by CCA.

- HGLMM and GMM+HGLMM [139]: combining Gaussian and Laplacian distributions into one hybrid distribution model that can benefit from the properties of the two distributions.
- MLLM [43]: a mixture of local linear mapping model with VGG-16-based visual representation and Word2vec-based text representation.
- 2-Way Net [102]: employing two tied neural network channels that project the two views into a common, maximally correlated space using Euclidean loss.
- Embedding Networks [103]: learning a shared latent embedding space based on two-way networks with a maximum-margin ranking loss and neighborhood constraints.
- DVSA [143]: an alignment model based on the combination of CNNs over image regions and bidirectional recurrent neural networks over sentences.
- OrderEmbedding [144]: learning the embeddings of images and captions by defining a loss function that encourages the order-violation penalty for ground truth caption-image pairs to be lower than that for all other pairs, by a margin.
- DSvEL [142]: a new two-path neural network with a visual path that leverages recent space-aware pooling mechanisms.
- CSE [145]: using CNNs to represent images and sentences and combining mid-level representations and global semantic learning.
- CCL [130]: fusing multi-grained features and learning the correlation based on the constraints of the intra-modality semantic category and the inter-modality pairwise similarity.
- RRF-Net [146]: a model that adapts the recurrent mechanism to residual learning and integrates the intermediate recurrent outputs.

4.5.2. PARAMETER TUNING AND ANALYSIS

The key parameters of KMM include the number of components M in Eq.4.15, the balance control parameters λ_1 and λ_2 in problem Eq. 4.19, and the dimension, d_s of semantic space S. To maximize the performance over validation sets, we determine the parameters by searching on the following grids: $\lambda_1, \lambda_2 \in \{10^2, 10^1 \cdots, 10^{-3}\}$, $M \in \{10, 20, 40, 60, 100, 150, 200\}$, and $d_s \in \{20, 50, 100, 150, 200\}$. In the experiment, we set $d_s = 50$ for Flickr8K, Flickr30K and NUS-WIDE-10K, and $d_s = 100$ for MSCOCO. To avoid inner product computation in the implicit feature spaces $\mathscr{F}_x$ and $\mathscr{F}_y$, we introduce the kernel function in Section 4.3.2 to achieve the computational results in the input spaces. We choose a polynomial kernel of degree 2, i.e., $f_\phi(\mathbf{x}_k, \mathbf{x}_l) = (\mathbf{x}_k \cdot \mathbf{x}_l + 1)^2$, $f_\psi(\mathbf{y}_k, \mathbf{y}_l) = (\mathbf{y}_k \cdot \mathbf{y}_l + 1)^2$, via experimentation.

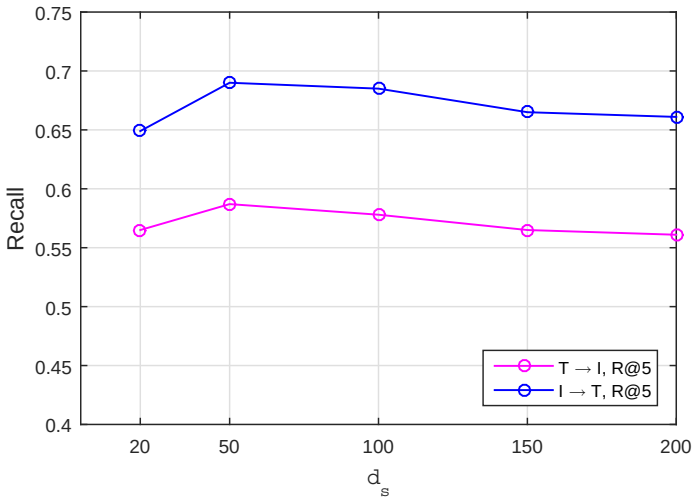


Figure 4.4: The effect of the dimension, d_s, of semantic space on the retrieval performance (R@5) of KMM with $M = 40$ in the case of "1 image vs 1 caption" on Flickr8K. "I → T" and "T → I" denote "image → text" and "text → image", respectively.

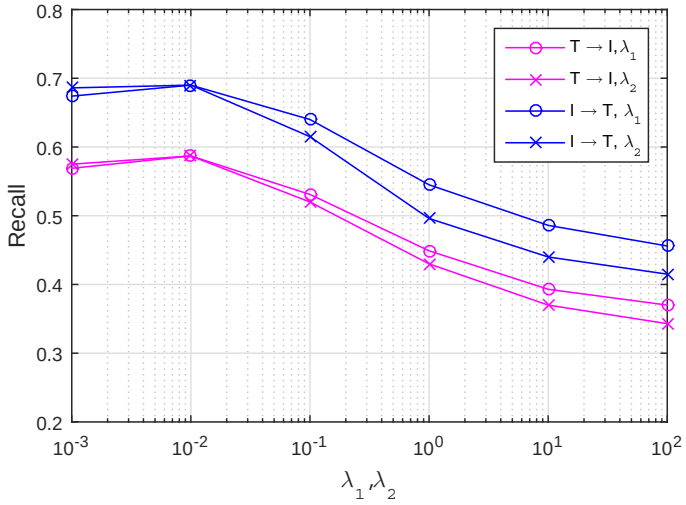


Figure 4.5: The effect of λ_1 and λ_2 on the retrieval performance (R@5) of KMM with $M = 40$ in the case of "1 image vs 1 caption" on Flickr8K. "I → T" and "T → I" denote "image → text" and "text → image", respectively. We change one parameter by setting the other to the optimal value, i.e., 10^{-2}.

Table 4.2: Performance (percent) comparison of bi-directional retrieval on Flickr8K. The top two parts in the body of the table correspond to the case of "1 image vs. 1 caption" and the bottom two parts correspond to the case of "1 image vs. 5 captions".

Approaches	Visual Backend	Flickr8K					
		Image → Text			Text → Image		
		R@1	R@5	R@10	R@1	R@5	R@10
HGLMM [139]	VGG	28.5	58.4	71.7	20.6	49.4	64.0
GMM+HGLMM [139]	VGG	31.0	59.3	73.7	21.2	50.0	64.8
2-Way Net [102]	VGG	43.4	63.2	-	29.3	49.7	-
MLLM [43]	VGG	34.2	59.9	70.8	26.9	54.0	67.9
KMM (without sparseness)	ResNet	45.7	68.5	77.3	31.0	57.1	71.2
KMM (without smoothness)	ResNet	44.6	67.2	76.3	30.1	55.8	70.6
KMM	VGG	43.4	65.9	74.5	29.0	56.1	69.3
KMM	ResNet	**46.7**	**69.0**	**78.3**	**31.6**	**58.7**	**72.6**
Deep CCA [128] (1 ima. vs. 5 cap.)	AlexNet	28.2	56.1	69.8	26.3	54.0	67.5
MLLM [43] (1 ima. vs. 5 cap.)	VGG	32.5	59.2	70.3	27.8	54.7	68.9
KMM (1 ima. vs. 5 cap.)	VGG	42.9	65.6	74.6	29.8	56.3	70.2
KMM (1 ima. vs. 5 cap.)	ResNet	**46.1**	**68.9**	**78.5**	**32.5**	**59.1**	**73.3**

Fig. 4.3 illustrates the effect of parameter M on the performance of cross-media retrieval for three datasets in the case of "1 image vs. 1 caption". On the whole, we observe that the recalls reach the highest values at $M = 40$ and 60 for Flickr8K and Flickr30K, respectively. For the more complex MSCOCO dataset, a larger value, $M = 100$, can produce better performance than the other values of M. The phenomenon is consistent with our intuition. That is, a model of larger capacity, i.e., the one with larger M in this work, is required for modeling a more complex dataset. In addition, the performances measured by different metrics on a specific dataset likely do not reach the highest value at the same M. For example, for Flickr8K, the recall R@1 in task "text→ image" is 31.6% at $M = 40$, which is slightly lower than 32.7% at $M = 60$. We also notice that a value of M that is too large may decrease the size of the local region $\mathscr{R}_m$ that supports local model $\mathscr{M}_m$, which tends to cause over-fitting in the learning of parameters $\mathbf{A}_m$, $\mathbf{B}_m$ and Σ_m and affects the performance of cross-media retrieval.

Fig. 4.4 shows the effect of the dimension d_s of semantic spaces on the retrieval performance (R@5) of KMM with $M = 40$ in the case of "1 image vs. 1 caption" on Flickr8K. As seen, the dimension of semantic spaces has effects on the performance. More specifically, the recall reaches the peak at $d_s = 50$ and then begins to degrade. An appropriate dimension for a latent semantic space depends on the complexity of semantics contained in datasets. A lower dimensional semantic space may result in an insufficient capacity to represent the distribution of semantics, while a higher dimension may cause a looser distribution of semantics as well as larger sizes of transformation matrices $\mathbf{U}_m$ and $\mathbf{V}_m$.

Fig. 4.5 illustrates the effect of parameters λ_1 and λ_2 on the retrieval performance (R@5) of KMM with $M = 40$ in the case of "1 image vs. 1 caption" on Flickr8K. In the figure, we show the effect of one parameter while setting the other to the optimal value. By experiments, we find that the retrieval performance peaks at $\lambda_1 = \lambda_2 = 10^{-2}$ and then retrieval performance begins to degrade as λ_1 or λ_2 continues to be added. In general, we find λ_1 leads a faster increase and slower degradation of performance than λ_2 as parameters are added. The results indicate that the smoothness term plays a more important role than the sparseness term in maintaining a good retrieval performance. In the experiments, we set λ_1 or λ_2 to 10^{-2} for all datasets. We conduct a further analysis for λ_1 and λ_2 by an ablation study in the next subsection.

4.5.3. PERFORMANCE ON CROSS-MEDIA RETRIEVAL

Ablation study

To further reveal the contribution of the two constraints in problem Eq. 4.19, we test the performance of KMM with three configurations. The variants include: 1) KMM (without smoothness), which is obtained by removing the smoothness term, 2) KMM (without sparseness), which ignores the L1-norm regularization that constrains the sparseness of learning results, and 3) KMM, which is the full version formulated in problem Eq. 4.19. Table 4.2, Table 4.3 and Table 4.4 show

Table 4.3: Performance (percent) comparison of bi-directional retrieval on Flickr30K. The top two parts in the body of the table correspond to the case of "1 image vs. 1 caption" and the bottom two parts correspond to the case of "1 image vs. 5 captions".

Approaches	Visual Backend	Flickr30K					
		Image → Text			Text → Image		
		R@1	R@5	R@10	R@1	R@5	R@10
HGLMM [139]	VGG	34.4	61.0	72.3	24.4	52.1	65.6
GMM+HGLMM [139]	VGG	35.0	62.0	73.8	25.0	52.7	66.0
2-Way Net [102]	VGG	49.8	67.5	-	36.0	55.6	-
MLLM [43]	VGG	44.5	62.8	73.4	28.1	53.6	65.8
DSvEL [142]	ResNet	46.5	72.0	82.2	34.9	62.4	73.5
CSE [145]	ResNet	44.6	74.3	83.8	36.9	**69.1**	**79.6**
Embedding Networks [103]	VGG	43.2	71.6	79.8	31.7	61.3	72.4
KMM (without sparseness)	ResNet	50.9	74.0	84.2	38.0	64.9	77.4
KMM (without smoothness)	ResNet	50.3	71.6	83.4	37.2	63.9	76.6
KMM	VGG	49.1	72.2	81.5	35.1	60.1	72.8
KMM	ResNet	**52.0**	**74.9**	**84.7**	**38.2**	66.5	79.0
Deep CCA [128] (1 ima. vs. 5 cap.)	AlexNet	27.9	56.9	68.2	26.8	52.9	66.9
MLLM [43] (1 ima. vs. 5 cap.)	VGG	43.7	62.2	73.5	28.5	54.1	66.1
CCL [130] (1 ima. vs. 5 cap.)	VGG	37.7	69.4	81.1	37.3	**68.4**	**80.0**
KMM (1 ima. vs. 5 cap.)	VGG	48.5	71.9	81.2	36.2	60.9	73.2
KMM (1 ima. vs. 5 cap.)	ResNet	**51.6**	**75.4**	**85.3**	**39.4**	66.9	79.5

the comparison results of the variants in the case of "1 image vs. 1 caption" on Flickr8K, Flickr30K and MSCOCO. From the tables, we observe that the variants KMM (without smoothness) and KMM (without sparseness) generally perform slightly worse than KMM (with visual representation using ResNet). Both variants have a degradation of 1.0 ~ 3.3% on the whole compared with KMM. From the figures, we observe that the smoothness term plays a more important role than the sparseness term in improving performance. The main cause is that smoothness may enforce two similar examples to be close together in the latent space.

Performance comparison

First, we evaluate and analyze the performance of the proposed approach in the case of "1 image vs. 1 caption" (i.e., the top two parts of Table 4.2, Table 4.3 and Table 4.4). Table 4.2 and Table 4.3 show the bi-directional retrieval results for the Flickr8K and Flickr30K datasets. We implement KMM with two visual representations: VGG-based and ResNet-based. It is known that ResNet generally performs better than VGG in many tasks. As expected, KMM with ResNet-based visual representation achieves better performance than KMM with VGG-based visual representation and has an increase of 2.6 ~ 6.4%. For Flickr8K and Flickr30K, we compare our approach with 4 and 7 state-of-the-art methods, respectively. The table shows that our approach achieves better performance than the compared methods in most cases. In the task of "Text → Image", CSE achieves better results than ours in terms of the metrics R@5 and R@10. Compared with our previous work MLLM, which can be considered as a simple version of KMM that does not introduce kernel mapping, we find that KMM achieves a large improvement. The results mean that the kernel mapping may lead to better modeling for nonlinear data distributions and nonlinear relationships between modalities. Table 4.4 shows the comparison between KMM and 9 state-of-the-art methods for the MSCOCO dataset. From the table, we find that the performance of our approach is better than or close to those of the compared methods. Our approach achieves the best performance for the metrics R@5 and R@10 in the task of "Image → Text" and the metric R@10 in the task of "Text → Image", while DSvEL obtains better results than ours in the other cases.

We also evaluate our approach in the case of "1 image vs. 5 captions" and report the results in Table 4.2, Table 4.3 and Table 4.4 (i.e., the bottom two parts of the tables). The tables show that KMM is superior to MLLM and Deep CCA. Regarding Flickr30K, we find that CCL achieves better performance than our approach for the metric R@5 and R@10 in the task of "Text → Image". Comparing the case of "1 image vs. 1 caption" with "1 image vs. 5 captions", we notice that the former has a change of $-1.2 \sim +0.7$% in terms of the three recalls when KMM works with the ResNet-based visual representation. More specifically, for the task of "Text → Image", the former has a slight decline in terms of all metrics; for the task of "Image → Text", the former tends to achieve a higher recall in terms of R@1 and a lower recall in terms of R@10. We consider that this change may derive from the richness of association information between different modalities and the way

Table 4.4: Performance (percent) comparison of bi-directional retrieval on MSCOCO. The top two parts in the body of the table correspond to the case of "1 image vs. 1 caption" and the bottom two parts correspond to the case of "1 image vs. 5 captions".

Approaches	Visual Backend	Image → Text			Text → Image		
		R@1	R@5	R@10	R@1	R@5	R@10
HGLMM [139]	VGG	37.7	66.6	79.1	24.9	58.8	76.5
GMM+HGLMM [139]	VGG	39.4	67.9	80.9	25.1	59.8	76.6
MLLM [43]	VGG	47.6	76.3	85.2	35.9	64.3	81.6
OrderEmbedding [144]	VGG	46.7	-	88.9	37.9	-	85.9
DVSA [143]	VGG	38.4	69.9	80.5	27.4	60.2	74.8
2-Way Net [102]	VGG	55.8	75.2	-	39.7	63.3	-
DSvEL [142]	ResNet	**69.8**	91.9	96.6	**55.9**	**86.9**	94.0
CSE [145]	ResNet	56.3	84.4	92.2	45.7	81.2	90.6
Embedding Networks [103]	VGG	54.0	84.0	91.2	43.3	76.8	87.6
KMM (without sparseness)	ResNet	63.1	91.5	96.5	49.5	82.9	93.4
KMM (without smoothness)	ResNet	62.1	90.6	94.9	48.4	81.5	92.0
KMM	VGG	56.9	85.5	92.1	42.0	75.6	88.5
KMM	ResNet	63.9	**92.3**	**96.9**	50.1	83.6	**94.2**
MLLM [43] (1 ima. vs. 5 cap.)	VGG	47.3	76.1	85.6	36.8	65.5	82.3
KMM (1 ima. vs. 5 cap.)	VGG	56.3	85.1	91.8	42.9	76.1	88.7
KMM (1 ima. vs. 5 cap.)	ResNet	**63.2**	**92.0**	**97.1**	**50.8**	**84.2**	**94.6**

Table 4.5: MAP scores (percent) of bi-directional retrieval on NUS-WIDE-10K.

Approaches	Image → Text	Text → Image	Average
Deep CCA [128]	40.7	41.6	41.2
GMM+HGLMM [139]	44.0	45.3	44.7
MLLM [43]	49.7	48.1	48.9
CCL [130]	50.6	53.5	52.1
KMM (VGG)	51.7	51.6	51.7
KMM (ResNet)	**54.8**	**54.4**	**54.6**

of retrieval. Intuitively, the case of "1 image vs. 1 caption" has less association information than "1 image vs. 5 captions" due to its shorter text, hence it may result in a slightly lower recall on the whole in the image retrieval given a textual query; while for the task of "Image → Text", all 5 correct captions can be used as the candidates to match a given image query and increase the possibility of a correct one among the first r responses, especially for the metric R@1.

From Table 4.2, Table 4.3 and Table 4.4, we observe that ResNet-based representation generally leads to better performance than VGG- and AlexNet-based representations due to its better abstraction of visual semantics using the structure of more layers. Regarding the superiority of CSE, DSvEL and CCL to our approach in some cases, we consider that there are two main causes. One is the visual localization. For example, DSvEL introduces a localization mechanism to emphasize the visual concepts associated with the corresponding text. CCL uses local visual patches as well as whole images as the input of model. CSE adds the consistency constraints on the intermediate regional features. The fine grained information may help capture accurate mapping between modalities. In addition, the multi-layered association in the feature extraction via deep networks may cause the improvement. Both CCL and CSE introduce consistency constraints for images and text at different layers of deep networks, which truely reinforce the association of heterogeneous modalities.

In Table 4.5, we report the performance of bi-directional retrieval on the NUS-WIDE-10K dataset in terms of the mAP metric. Since NUS-WIDE-10K has class labels, we can compute the mAP for the retrieval task. In the experiment, we compare our approach with 4 state-of-the-art methods. As shown from the table, KMM (with ResNet-based visual representation) maintains an advantage with all 4 compared methods and KMM (with VGG-based visual representation) obtains similar results with CCL.

Performance on cross-dataset evaluation

Following RRF-Net [146] and CSE [145], we also evaluate the performance of our approach in terms of cross-dataset generalization. In this experiment, we em-

Table 4.6: Performance (percent) of bi-directional retrieval on cross-dataset in the case of "1 image vs. 1 caption".

Data Setting	Approaches	Image → Text			Text → Image		
		R@1	R@5	R@10	R@1	R@5	R@10
Train: Flickr30K, Test: MSCOCO	RRF-Net [146]	24.8	**53.0**	64.8	18.8	44.1	58.5
	CSE [145]	24.6	49.2	62.5	**19.1**	44.4	58.6
	KMM (ResNet)	**25.4**	52.5	**65.4**	**19.1**	**44.8**	**58.9**
Train: MSCOCO, Test: Flickr30K	RRF-Net [146]	28.8	53.8	66.4	21.3	42.7	53.7
	CSE [145]	30.6	59.3	71.0	26.0	52.1	**64.3**
	KMM (ResNet)	**32.7**	**60.1**	**71.6**	**26.6**	**52.4**	63.7

ploy the model trained on Flickr30K or MSCOCO to evaluate the test set of the other dataset. Table 4.6 reports the results of bi-directional retrieval in the case of "1 image vs. 1 caption" for the cross-dataset. The performance of the generalization is similar to and positively correlated with the performance in Table 4.2, Table 4.3 and Table 4.4. The table also shows that it is easier to transfer a model trained on a large dataset to a small one than the converse case. From the table, we observe that, on the whole, our approach achieves better performance on the cross-dataset evaluation. We consider that this may be caused by two reasons. 1) In the training process, the deep networks pre-trained on ImageNet are changeless and the feature space is uniform for different datasets. In this case, the KMM model trained in the feature space that is independent of datasets can transfer the association knowledge across datasets more stably. 2) As a model-driven approach, KMM introduces an explicit probabilistic model to describe both the data distribution and relationship distribution, which can be considered as prior information from the Bayesian viewpoint, and can generally improve generalizability.

Example illustration

Fig. 4.6 shows some examples of cross-media retrieval results in the cases of "1 image vs. 1 caption" (top two rows) and "1 image vs. 5 captions" (bottom four rows) for the MSCOCO test data. All retrieval algorithms encourage the ground truth associated with queries to be located as close to the front of the response as possible. In the first case, we find a response (in the 2nd row) that is not the ground truth associated with the query appears in front of a correct caption; in the second case, we show two examples (in the 4th and 6th rows) in which the ground truth does not appear at the 1st position in the retrieval results. We find that the retrieval results at the 1st position are truely similar with the queries. For example, in the 4th row, although the returned image at the 1st position is not the ground truth associated with the query, it consists of the same objects, such as "plane" and "runway", as the ground truth and highly matches the query.

4.6. CONCLUSIONS

In this chapter, we present a kernel-based probabilistic mixture model, called KMM, for modeling the semantic correlation between web images and text. KMM assumes that the relationship between different modalities follows multiple basic transformations, each working over a local region described by a neighborhood model in the input space. We employ kernel theory to address the nonlinearity of the data distribution and cross-modal mapping. We present a hybrid optimization algorithm based on EM and subgradient ascent to estimate the parameters of KMM and prove that the algorithm can converge to an (local) optimal solution. By combining nonlinear transformation and probabilistic models, KMM addresses the complexity of the semantic distribution over the global input space, its continuity at the local scale, and the nonlinearity in the mapping of different modalities. The experimental results demonstrate the superiority of our

Query	Retrieval results				
Three large elephants and one small elephant walking through a dusty field.					
	A door for exiting and entering the house in the kitchen.	The kitchen has a white door with a window	The back door with a window in the kitchen	A kitchen with a dishwasher double door panty and a back door	A kitchen door next to a kitchen sing and counter top
Kids playing a game of base ball while people watch. Parents watching Young boys playing baseball in the sun a young boy is at home ...					
An Aer Lingus plane touches down on an airport runway. Passenger airliner at the end of a runway waiting to take off...					
	A display in a store filled with ripe bananas. A store display that has a lot of bananas on display for sale...	A pile of oranges in crates topped with yellow bananas. There are bananas, pineapples, oranges...	A planter filled with lots of yellow and red green leafed flowers. a group of flowers sitting in a vace...	Burger with broccoli, pickle, and fork on orange plate On a plate is kept a burger and a bowl of broccoli...	A bunch of bananas sitting on top of a wooden table.A closeup of a group of bananas on a table...
	A group of people fly kites into the air on a large grassy field. Group of peo - ple outdoors flying kites...	A field full of people standing on top of a grassy field flying kites. Group of many people on a field...	A group of umbrellas together in a plaza near the Eiffel Tower. The Eiffel Tower is shown in all...	The sky is cloudy over a stop sign. A traffic sign near a high grass field near a road...	People in the water and Parachutes overhead. Many Different Sails flying over a large ...

Figure 4.6: Example cross-media retrieval results over MSCOCO test data. The top two rows correspond to the case of "1 image vs. 1 caption" and the bottom four rows correspond to the case of "1 image vs. 5 captions". Images surrounded by blue boxes and blue-colored text are ground truth. Retrieval results are arranged in decreasing order of similarity.

approach over representative state-of-the-art methods of modeling the relationships between images and text.

5

VISUAL REPRESENTATION CONTEXTUALIZATION BASED ON CONSTRASTIVE LEARNING

This chapter is based on the following publication:

Wang, X., Du, Y., Verberne, S. Verbeek, F.J. Improving Weakly Supervised Phrase Grounding via Visual Representation Contextualization with Contrastive Learning. Applied Intelligence. (under review)

CHAPTER SUMMARY

This chapter addresses RQ5 and RQ6.

RQ5: What is the effect of the attention mechanism to eliminate the different modal representations produced in the common embedding space?
RQ6: How to employ the correspondence between images and text as supervision instead of the matching annotations to address the limited data issue?

Weakly supervised phrase grounding aims to map the phrases in an image caption to the objects appearing in the image under the supervision of image-caption correspondences. We observe that the current studies are insufficient to model the complicated interactions between visual components (i.e., visual regions) and between visual and textual components (i.e., phrases). Therefore, this chapter presents a novel weakly supervised learning approach to phrase grounding in which we systematically model the visual contextualized representation with three modules: (1) object proposals pooling (OPP), (2) visual self-attention (VSA) and (3) visual-textual cross-modal attention (VTCA). OPP alleviates the suppression of object proposals and benefits the visual representation in terms of trading off the richness of visual components and the computational efficiency. VSA aims to capture the correlation among the object proposals and generate the representation of each proposal by incorporating the visual information of the others. In order to measure the cross-modal compatibility in terms of topics, we introduce the VTCA module to represent the visual topic corresponding to each textual component in a cross-modal common vector space. In the training process, we build a mixed contrastive loss function by considering both the cross-modal compatibility and the difference of visual representations in the VSA module. Compared to the state-of-the-art methods, the proposed approach improves the performance by 3.88% point and 1.24% point on $R@1$, and by 2.23% point and 0.26% point on Pt_Acc, when trained on the MS COCO and Flickr30K Entities training set, respectively. We have made our code available for follow-up research.

Tasks combining cross-modal (visual-and-language) compatibility have attracted a lot of attention and contributed to the advancement of artificial intelligence in recent years. Examples of cross-modal tasks are image caption generation [147], visual question answering (VQA) [148], visual reasoning [149, 150], and phrase grounding [23]. Phrase grounding aims to localize the objects in images and at the same time, based on paired images and captions, maps them to the phrases in captions. Phrase grounding requires a model to understand the fine-grained correspondence between images and language. A large part of previous works plummer2017phrase,fukui2016multimodal,wang2018learning are based on supervised learning, i.e., with supervision of the correspondence between visual regions and phrases. However, the availability of this kind of labelled data is limited due to significant manual efforts in collecting the annotations for region-phrase correspondences.

To address the issue of limited availability of data, researchers have proposed a few weakly supervised phrase grounding methods, which only employ the correspondence between images and text as supervision instead of the matching annotations of visual regions and phrases. The attention mechanism has becoming an important technique in solving the task of weakly supervised phrase grounding, and can generally be divided into two types: the first type models the intra-modality compatibility that infers the latent correlations between different regions in an image or different words in a caption [151] based on self-attention mechanism. The other seeks to mine the cross-modal interactions between textual words and visual regions based on inter-modality compatibility [152]. That is, most of the previous methods only consider the correlations either in inter-modality or in intra-modality.

Another issue of weakly supervised phrase grounding is how to choose loss functions to obtain a better learning result. Recently, contrastive learning, e.g., InfoNCE [35], has shown promising results on a variety of applications. Gupta et al. [18] proposed a novel contrastive learning approach to the task of weakly supervised phrase grounding, which improved the performance by employing the InfoNCE loss defined on the positive and negative samples.

In this chapter, inspired by the advancements of contrastive learning [18] and phrase grounding [17], we introduce a new approach, called VRC-PG, to improve weakly supervised phrase grounding with visual representation contextualization (VRC). In our method, the inter- and intra-modality interactions are modeled for inferring the compatibility between phrases and visual regions. Here, we also call the phrase and visual region as the textual component and visual component, respectively. VRC-PG consists of three modules: object proposals pooling (OPP), visual self-attention (VSA) and visual-textual cross-modal attention (VTCA). In the visual representation contextualization, OPP is introduced to alleviate the suppression of object proposals (candidates) generated by object detectors. This benefits the visual representation contextualization in terms of trading off the richness of visual components and the computational efficiency. VSA aims to capture the correlation between visual object proposals for each image and generate the

representation of each candidate by incorporating the visual information of the other candidates. To measure the cross-modal compatibility at the level of topics, we subsequently introduce the VTCA module to distill the visual topic corresponding to each textual component, i.e., textual phrase, in a cross-modal common vector space, guided by the attention of a phrase to visual object proposals. In addition, we present a mixed contrastive loss function including two terms: one is to improve cross-modal compatibility in terms of topics of images and captions, and the other is to control the difference of the visual representations induced by the VSA module.

In summary, our contributions are three-fold: (1) we propose a novel approach to weakly supervised phrase grounding based on visual representation contextualization under the weak supervision of image-caption correspondences without region-phrase matching annotations. Moreover, a mixed contrastive loss is introduced to improve the performance of our model. (2) We present an architecture of visual representation contextualization that consists of object proposals pooling (OPP), visual self-attention (VSA) and visual-textual cross-modal attention (VTCA). (3) The proposed model is evaluated on Flickr30K Entities dataset and achieves the state-of-the-art performance, improving by 1.24% point and 3.88% point $Recall@1$ on the Flickr30K Entities test set when trained on the Flickr30K Entities training set and MS COCO, respectively.

5.1. RELATED WORK

5.1.1. PHRASE GROUNDING

The existing works are based on two different supervision processes, fully supervised learning and weakly supervised learning. Plummer et al. [23] proposed a global image-sentence canonical correlation analysis (CCA) model to analyze the region-phrase correspondence in the combined image-text embedding space, and achieved a state-of-the-art result for this task on the Flickr30K Entities dataset. Wang, and Sigal [153] used graphs to formulate more complex, non-sequential dependencies among object region proposals and phrase candidates. Most of these methods employ the annotations of region-phrase correspondences and are implemented under the supervised learning framework. Because manual labelling is expensive, also some other research has used the approach of weakly supervised learning. Plummer et al. [154] presented a weakly supervised learning method that modeled the appearance, object size and position of visual objects to localize phrases in images. Akbari et al. [36] proposed a multi-level multi-modal model to explicitly learn a non-linear mapping of the visual and textual modalities in a common semantic space, and do so at different granularities for each modality. Recently, the attention mechanism has been introduced to reconstruct the representation of vision and text guided by inter- or intra-modality. The result is a cross-modal attention mechanism with a fully supervised or weakly supervised learning framework. Chen et.al. [155] proposed a novel knowledge-aided network

which was optimized by reconstructing input information of queries and region proposals extracted by a region proposal network (RPN). These existing methods lack the ability to model image-caption paired supervision. This is essential for grounding phrases in the images based on weak supervision from caption-image pairs. In this chapter, we propose the VRC-PG approach and model the fine-grained interactions in the inter- and intra-modality by jointly considering the visual self-attention mechanism and cross-modal attention mechanism.

5.1.2. NON-MAXIMUM SUPPRESSION (NMS)

NMS [156] has been an important technique for computer vision tasks, such as object detection [30, 58] and edge extraction [157]. In object detection, NMS is a post-processing step adopted by a number of modern object detectors, which removes duplicate bounding boxes based on detection confidence. A major issue with NMS is that it sets the score for neighboring detection to zero. Thus, if an object is actually present in an overlap region with an IoU greater than the threshold it would be missed and this would lead to a drop in average precision.

To alleviate this problem, Bodla et al. [158] presented the Soft-NMS algorithm to decrease the confidence scores as an increasing function of overlap instead of setting the score to zero as in NMS. Softer-NMS [159] proposed a bounding box regression Kullback-Leibler loss for learning bounding box transformation and localization variance together. As a downstream task of object detection, language grounding methods have used NMS to align the language with the proposals. Chen et al. [160] used NMS to yield expression-aware region proposals to improve the performance language grounding. In our work, we use Soft-NMS to replace the NMS module in Faster R-CNN to keep more bounding box proposals, and introduce an extra object proposals pooling module with NMS to adaptively choose those proposals with high confidence scores and benefiting the weakly supervised phrase ground task.

5.1.3. CONTRASTIVE LEARNING IN CROSS-MODAL TASKS

Constrastive learning was first used as a powerful scheme for self-supervised representation learning [35, 161, 162, 163]. Until now, it has been explored to enforce consistency of different modal representations under different augmentations by contrasting positive pairs with negative ones. Zhang et al. [164] proposed a cross-modal model called XMC-GAN, which introduced an attentional self-modulation generator and a contrastive discriminator to maximize the cross-modal information between images and text. Dai and Lin [13] proposed a method that encouraged the distinctiveness of positive pairs, while maintaining the overall quality of the generated captions. Gupta et al. [18] built a weakly supervised phrase grounding model based on optimizing the lower bound of InfoNCE on Mutual Information (MI) with respect to parameters of a word-region attention model. Li et al. [165] proposed a framework combining a self-attention mechanism with contrastive feature construction so as to effectively summarize common information

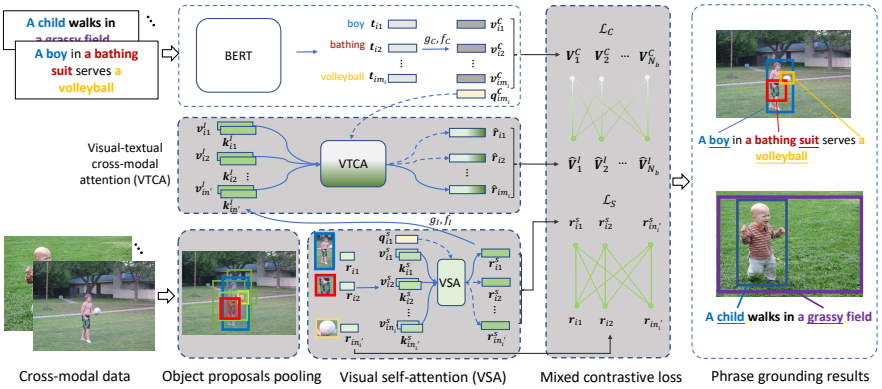


Figure 5.1: The framework of VRC-PG. The visual representation contextualization is comprised of three parts: 1) object proposals pooling, where thick bounding boxes (red, blue and yellow) are the output boxes and thin bounding boxes (green) are non-maximally suppressed, 2) visual self-attention, and 3) visual-textual cross-modal attention. The proposed model is trained with the contrastive learning paradigm by introducing our 4) mixed contrastive loss.

5

from each image group while capturing discriminative information between visual regions and phrases. CDMLMR [166] integrates the quadruplet ranking loss and semi-supervised constrastive loss for modeling cross-modal semantic similarity in a unified multi-task learning architecture. In our work, we learn our model with the contrastive learning paradigm and build a mixed contrastive loss function, which consists of two terms: one is control the difference of the visual representations induced by the VSA module, and the other is to improve the cross-modal compatibility in terms of the topics of images and captions.

5.2. METHODOLOGY

5.2.1. OVERVIEW

We are given a set of pairs, each consisting of an image and its caption. Formally, we have data $\mathscr{D}_i = \{(I_i, C_i)\}_{i=1}^N$, where I_i and C_i denote the i-th image and its corresponding caption, respectively. In general, the content of an image I_i can be described by a set of n_i visual object regions enclosed with bounding boxes $\mathscr{B}_i = \{b_{i1}, b_{i2}, \cdots, b_{in_i}\}$. The visual regions can be represented with the box location $\boldsymbol{B}_i = (\boldsymbol{b}_{i1}, \boldsymbol{b}_{i2}, \cdots, \boldsymbol{b}_{in_i})$, confidence score $\boldsymbol{S}_i = (s_{i1}, s_{i2}, \cdots, s_{in_i})$, visual features $\boldsymbol{R}_i = (\boldsymbol{r}_{i1}, \boldsymbol{r}_{i2}, \cdots, \boldsymbol{r}_{in_i})$, and category predictions $\boldsymbol{L}_i = (l_{i1}, l_{i2}, \cdots, l_{in_i})$. Regarding the textual modality, each caption C_i can be considered a sequence of m_i tokens $T_i = (t_{i1}, t_{i2}, \cdots, t_{im_i})$ and transformed to the token representation $\boldsymbol{T}_i = (\boldsymbol{t}_{i1}, \boldsymbol{t}_{i2}, \cdots, \boldsymbol{t}_{im_i})$ using the BERT-base model [28]. A phrase consists of one or multiple tokens of captions. In this manner, the training data can be described by $\mathscr{D}_i = \{(\boldsymbol{B}_i, \boldsymbol{S}_i, \boldsymbol{R}_i, \boldsymbol{L}_i), \boldsymbol{T}_i\}_{i=1}^N$.

In this chapter, we present a novel approach called VRC-PG to the task of

weakly supervised phrase grounding. As shown in Fig. 5.1, our VRC-PG approach includes four main parts: (1) object proposals pooling module, (2) visual self-attention module, (3) visual-textual cross-modal attention module and (4) mixed contrastive loss function. The proposed approach models visual representation contextualization by jointly considering the interactions in both the unimodal data and the cross-modal data, and trains the model with a contrastive learning paradigm under the weak supervision of the correspondence between images and text.

5.2.2. VISUAL REPRESENTATION CONTEXTUALIZATION MODEL

Feature extraction

The purpose of the visual representation contextualization model is to build the correspondence between the token representations $\boldsymbol{T}_i = (\boldsymbol{t}_{i1}, \boldsymbol{t}_{i2}, \cdots, \boldsymbol{t}_{im_i})$ and object candidate representations $\boldsymbol{R}_i = (\boldsymbol{r}_{i1}, \boldsymbol{r}_{i2}, \cdots, \boldsymbol{r}_{in_i})$ by measuring their attention.

We use the BERT-base model [28] to extract the text modal representation with caption as input.

$$\boldsymbol{t}_{ij} = BERT(C_i), \tag{5.1}$$

where $\boldsymbol{t}_{ij} \in \mathbb{R}^{d_t}$ is a dense vector representation.

We utilize the Faster R-CNN [30] model trained on the Visual Genome dataset [22] to extract and represent the objects:

$$(\{\boldsymbol{b}_{ij}\}, \{s_{ij}\}, \{\boldsymbol{r}_{ij}\}, \{l_{ij}\}) = FasterRCNN(I_i), \tag{5.2}$$

where $\boldsymbol{b}_{ij} \in \mathbb{R}^4$ and $\boldsymbol{r}_{ij} \in \mathbb{R}^{d_r}$, s_{ij} is the maximum classification score among all categories. In this work, we do not employ the predicted category labels l_{ij} generated by Faster R-CNN for each object region in our task.

Object Proposals Pooling (OPP)

As weakly supervised phrase grounding is performed without phrase grounding annotations, its quality depends on the performance of object box proposals extracted with Faster R-CNN. In order to keep more effective object box proposals, we replace NMS used in Faster R-CNN by Soft-NMS [158]. The advantage of Soft-NMS is to keep more proposals for an object. However, it will cause the mapping accuracy to be lower if two objects overlap between each other. To alleviate this problem, we propose an object proposals pooling module based on NMS to further prune the detected objects and only keep boxes less than an IoU threshold θ in the training process. The OPP module can adaptively choose those proposals with high confidence scores $\{s_{ij}\}$ and benefit from the weakly supervised phrase ground task.

For an image I_i, the pruning starts with a bounding box b_{iz} with the highest confidence score $s_{iz} = \max_j(s_{ij})$. b_{iz} is kept as one of the bounding boxes produced the OPP module. Then, We update the confidence scores of all the bounding box b_{ij} by

$$s_{ij} = \begin{cases} s_{ij}, & IoU(\boldsymbol{b}_{ij}, \boldsymbol{b}_{iz}) < \theta, j \in 1, \ldots, n_i; \\ 0, & IoU(\boldsymbol{b}_{ij}, \boldsymbol{b}_{iz}) \geq \theta, j \in 1, \ldots, n_i. \end{cases} \tag{5.3}$$

Here, θ is a threshold to decide which object box should be directly excluded in each iteration of object proposals pooling. Based on the above process, we can choose more bounding boxes based on Eq. 5.3 until all the confidence scores are updated to zero. Finally, the OPP module produces n_i' object proposals. In this module, we do not employ the category predictions generated by Faster R-CNN.

Visual Self-attention (VSA)

In general, the visual components, i.e., the visual object region proposals comprised in an image, have spatial and semantic correlation with each other. We introduce a visual self-attention module to model the context of visual object regions and build their representations. The general attention mechanism can be formulated accordingly as follows:

$$\text{Attention}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \text{softmax}(\text{sim}(\boldsymbol{Q}, \boldsymbol{K})) \cdot \boldsymbol{V}, \tag{5.4}$$

where $\boldsymbol{Q}$, $\boldsymbol{K}$, $\boldsymbol{V}$ and $\text{Attention}(\cdot,\cdot,\cdot)$ refer to the query, key, value and output, respectively; and $\text{sim}(\cdot,\cdot)$ denotes a certain function to measure the corresponding of queries and keys. In this work, the query (key) and value are obtained by the projection functions $f_I^s(\cdot)$ and $g_I^s(\cdot)$, respectively, implemented with a fully-connected layer as follows:

$$\begin{cases} \boldsymbol{q}_{ij}^s, \boldsymbol{k}_{ij}^s = f_I^s(\boldsymbol{r}_{ij}), \ j = 1, \cdots, n_i'; \\ \boldsymbol{v}_{ij}^s = g_I^s(\boldsymbol{r}_{ij}), \ j = 1, \cdots, n_i'. \end{cases} \tag{5.5}$$

Where, $\boldsymbol{q}_{ij}^s, \boldsymbol{k}_{ij}^s$ and $\boldsymbol{v}_{r_{ij}}^s \in \mathbb{R}^{d_s}$ refer to the vector of query, key and value, respectively. The soft weight of self-attention from $\boldsymbol{r}_{ij}$ to $\boldsymbol{r}_{iu}$ can be measured by the corresponding between them defined as follows:

$$a_s(\boldsymbol{q}_{ij}^s, \boldsymbol{k}_{iu}^s) = \frac{e^{\boldsymbol{q}_{ij}^s \cdot \boldsymbol{k}_{iu}^s / \sqrt{d_s}}}{\sum_w e^{\boldsymbol{q}_{ij}^s \cdot \boldsymbol{k}_{iw}^s / \sqrt{d_s}}}. \tag{5.6}$$

Thus, the contextualized visual representation of an object region is obtained by considering the self-attention:

$$\boldsymbol{r}_{ij}^s = \sum_u a_s(\boldsymbol{q}_{ij}^s, \boldsymbol{k}_{iu}^s) \boldsymbol{v}_{iu}^s, \tag{5.7}$$

where $\boldsymbol{r}_{ij}^s$ denotes the contextualized visual representation for the object region $\boldsymbol{r}_{ij}$ that incorporates the global information of the i-th image.

Visual-textual Cross-modal Attention (VTCA)

To build an adaptive correspondence between the cross-modal components (i.e., object region proposals and tokens), we make a cross-modal alignment between the visual and textual components. Here, we introduce a visual-textual cross-modal attention module to find the semantically related components in the visual modality for a given textual component. First, we transform the representation of textual components generated by BERT and the contextualized visual representation into a common space of dimensionality d_c. In this module, we take the textual token as the query actor and measure the weight of attention to the visual components by computing the cross-modal correlation.

In the common space, the query and value for the token representation $\boldsymbol{t}_{ij}$ are generated by the functions $f_C(\cdot)$ and $g_C(\cdot)$, respectively, and the key and value for the visual region proposal o_{ij} are obtained by $f_I(\cdot)$ and $g_I(\cdot)$, respectively, as follows:

$$\begin{cases} \boldsymbol{q}_{ij}^C = f_C(\boldsymbol{t}_{ij}),\ j = 1, \cdots, m_i; \\ \boldsymbol{k}_{ij}^I = f_I(\boldsymbol{r}_{ij}^s),\ j = 1, \cdots, n_i'; \\ \boldsymbol{v}_{ij}^C = g_C(\boldsymbol{t}_{ij}),\ j = 1, \cdots, m_i; \\ \boldsymbol{v}_{ij}^I = g_I(\boldsymbol{r}_{ij}^s),\ j = 1, \cdots, n_i', \end{cases} \tag{5.8}$$

where $\boldsymbol{t}_{ij}$ refers to the representation of token t_{ij} generated by BERT, $\boldsymbol{r}_{ij}^s$ is the contextualized visual representation obtained with Eq. 5.7 and $\boldsymbol{q}_{ij}^C, \boldsymbol{k}_{ij}^I, \boldsymbol{v}_{ij}^C$ and $\boldsymbol{v}_{ij}^I \in \mathbb{R}^{d_c}$. In this work, $f_\cdot(\cdot)$ and $g_\cdot(\cdot)$ are implemented with fully-connected layers.

Given the representation of a token obtained from BERT as a query, i.e., $\boldsymbol{q}_{ij}^C$, based on the attention mechanism [167] , the cross-modal attention [18] is defined as follows:

$$a_c(\boldsymbol{q}_{ij}^C, \boldsymbol{k}_{iu}^I) = \frac{e^{\boldsymbol{q}_{ij}^C \cdot \boldsymbol{k}_{iu}^I / \sqrt{d_c}}}{\sum_{w=1}^{n_i'} e^{\boldsymbol{q}_{ij}^C \cdot \boldsymbol{k}_{iw}^I / \sqrt{d_c}}}, \tag{5.9}$$

$$\hat{\boldsymbol{r}}_{ij} = \sum_{u=1}^{n_i'} a_c(\boldsymbol{q}_{ij}^C, \boldsymbol{k}_{iu}^I) \boldsymbol{v}_{iu}^I, \tag{5.10}$$

where $\hat{\boldsymbol{r}}_{ij}$ represents a visual topic correlated to the semantics of the token t_{ij} by incorporating the textual token information with cross-modal attention.

5.2.3. MIXED CONTRASTIVE LOSS FUNCTION

For a mini-batch of size N_b in the learning process, we have N_b captions and images represented with $\boldsymbol{V}_i^C$ and $\hat{\boldsymbol{V}}_j^I$. Here, the textual representation $\boldsymbol{V}_i^C = [\boldsymbol{v}_{i1}^C, \boldsymbol{v}_{i2}^C, \cdots, \boldsymbol{v}_{im_i}^C]$ and visual representation $\hat{\boldsymbol{V}}_j^I = [\hat{\boldsymbol{r}}_{i1}, \hat{\boldsymbol{r}}_{i2}, \cdots, \hat{\boldsymbol{r}}_{im_i}]$ obtained from VTCA, we mea-

sure the similarity of two cross-modal samples as follows:

$$S(\boldsymbol{V}_i^C, \hat{\boldsymbol{V}}_j^I) = \frac{e^{\mathrm{tr}(\boldsymbol{V}_i^{C\mathsf{T}} \cdot \hat{\boldsymbol{V}}_j^I)}}{\sum_{k=1}^{N_b} e^{\mathrm{tr}(\boldsymbol{V}_i^{C\mathsf{T}} \cdot \hat{\boldsymbol{V}}_k^I)}}, \tag{5.11}$$

where $\mathrm{tr}(\cdot)$ and the superscript T denote the trace and transposition of a square matrix. Eq. 5.11 uses a softmax operator to normalize the similarity to sum 1.

For contrastive learning, in each mini-batch, an image and its matching caption are denoted a positive sample pair (i.e., $i = j$) and non-matching image-caption pairs are negative sample pairs (i.e., $i \neq j$). Based on the similarity measured by Eq. 5.11, we provide a contrastive loss function at the granularity of images and captions:

$$\mathscr{L}_C = -\frac{1}{N_b} \sum_{i=1}^{N_b} log(S(\boldsymbol{V}_i^C, \hat{\boldsymbol{V}}_i^I))/\mathscr{T}, \tag{5.12}$$

where $\mathscr{T}$ is a temperature hyper-parameter. The loss in Eq. 5.12 seems to only work on the positive pairs and do not involve the negative pairs. Actually, to maximize the similarity $S(\cdot,\cdot)$ in Eq. 5.12 for the positive pair will lead to the suppression of the similarity for the negative pairs due to the sum-to-one normalization in Eq. 5.11, which is just a manner of the contrastive learning.

In addition, we introduce a loss to force the outputs of the visual self-attention module to be close to its inputs. The visual self-attention loss is defined as follows:

$$\mathscr{L}_S = -\frac{1}{N_b} \sum_{i=1}^{N_b} \left(\frac{1}{n_i'} \sum_{j=1}^{n_i'} log \left(\frac{e^{(\boldsymbol{r}_{ij} \cdot \boldsymbol{r}_{ij}^s)}}{\sum_{u=1}^{n_i'} e^{(\boldsymbol{r}_{ij} \cdot \boldsymbol{r}_{iu}^s)}} \right) \right). \tag{5.13}$$

Clearly, the visual self-attention loss is also a contrastive loss.

Finally, we build a mixed contrastive loss function in the form of

$$\mathscr{L} = \alpha \mathscr{L}_C + \mathscr{L}_S, \tag{5.14}$$

where α is a hyper-parameter to control the balance of both terms.

5.3. EXPERIMENTAL RESULTS

In this section we first describe the datasets followed by the implementation details.

5.3.1. DATASETS AND METRICS

Datasets

The experiments are conducted on the Flickr30K Entities dataset and MS COCO 2014 dataset.

- Flickr30K Entities contains 31,873 images and 5 captions per image. Following Gupta et al. [18], we split the Flickr30K Entities in a training set with 29,783 images, a validation set with 1,000 images and a test set with 1,000 images. The Flickr30K Entites dataset provides the correspondence of phrases and visual object regions. Thus, the Flickr30K Entities validation set and test set are employed to validate the proposed model and test its performance, respectively, in this work.
- The MS COCO 2014 dataset contains 118,287 training images and 5,000 validation images, where each image is provided with 5 human-annotated captions. The MS COCO 2014 dataset does not contain the links between image regions and sentence phrases. We thus train our model on the MS COCO 2014 training set, validate and test on the Flickr30K Entities validation and test sets, respectively. In the training process, we randomly select one caption from 5 captions of each example as the textual segment.

Metrics

We use two standard metrics for this task:

- $Recall@K$ ($R@K$) for $K = 1,5$ and 10 measures the percentage of phrases for which $IoU > 0.5$ between the top K predicted bounding boxes and the ground truth boxes.
- Pt_Acc refers to pointing accuracy and is commonly used to evaluate weakly supervised phrase grounding models [18]. *Pt_Acc* is the proportion of phrases for which center point of the predicted bounding box falls in the ground truth box. Unlike $R@K$, pointing accuracy does not require identifying the IoU of the predicted object box. Generally, the center point of the selected bounding box is used as the prediction for each phrase for computing pointing accuracy.

5.3.2. IMPLEMENTATION DETAILS

Visual Feature Representation

We extract visual region proposals from each image using Faster R-CNN with a backbone ResNet-101 [30] based on the bottom-up attention method [11], which was trained on the Visual Genome dataset. The region proposals contain the bounding boxes, visual features and Faster R-CNN's confidence scores (after Soft-NMS thresholding). We choose 50 regions of interest (RoI) based on confidence scores and obtain 2048-dimensional visual representations (i.e., $d_r = 2048$). By the VSA module we will reduce the dimension of visual representations from 2048d to 768d (i.e., $d_s = 768$).

Textual Feature Representation

We follow the setting of the BERT model used in the work of Gupta et al. [18] for the generation of the textual representation, where a pre-trained BERT model [28]

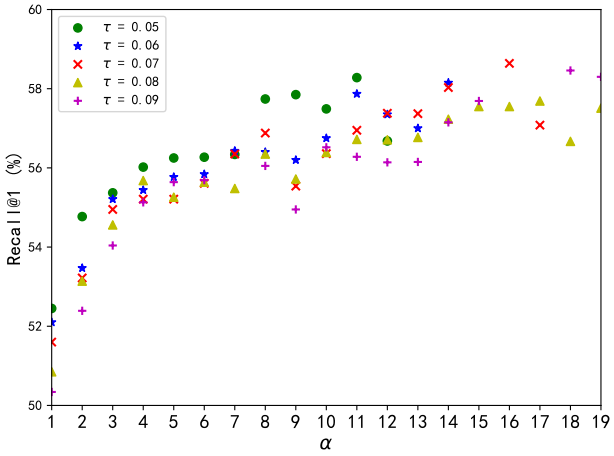


Figure 5.2: The hyper-parameters temperature τ and the loss function weight α optimized for Recall@1 on the validation set of Flickr30K.

is employed. A 768-dimensional token representation, i.e., $d_t = 768$, is generated for a word t_{ij} in captions with the BERT model. The dimension of the common space generated by the VTCA is set to is 384, i.e., $d_c = 384$.

Parameter Tuning

The hyper-parameters are determined with a grid searching on the Flickr30K Entities validation set. The threshold θ in Eq. 5.3 is set to 0.5, a same value as used in the evaluation of models in terms of the $R@K$ metrics. In our research, we perform grid search for determining the parameters. Fig. 5.2 shows the optimization result of the hyperparameters α from Eq. 5.14 and temperature $\mathscr{T}$ in Eq. 5.12. We train our model for 10 epochs with a batch size of 30 using an SGD optimizer with momentum 0.9 and a learning rate of 10^{-5}. We select the final checkpoints on the basis of the model's best performance in terms of $R@1$ on the Flickr30K Entities validation set. Based on the validation results, we set $\alpha = 16$ and $\mathscr{T} = 0.07$.

5.3.3. QUANTITATIVE RESULTS

Table 5.1 presents the experimental results of the compared methods on the Flickr30K Entities test set. From this table, we observe that our proposed approach outperforms the state-of-the-art work [15] by 1.24% point and 0.26% point in terms of $R@1$ and Pt_Acc, respectively, with the model trained on the Flickr30K Entities training set. For the models trained on MS COCO, our approach improves the performance by 3.88% point and 2.23% point in terms of $R@1$ and Pt_Acc, respectively, compared to the state-of-the-art work [18]. For the other cases, we observe that our approach is superior to the compared methods as a whole.

Table 5.1: The comparison of the results (%) of our approach with the state-of-the-art on the Flickr30K Entities test set. The models have been trained on Flickr30K Entities and MS COCO.

Methods	Training data	R@1	R@5	R@10	Pt_Acc
GroundeR [168]	Flickr30K	28.94	-	-	-
KAC Net [155]		38.71	-	-	-
InfoGround [18]		47.88	76.63	**82.91**	74.94
Wang et al. [14]		53.10	-	-	-
Liu et al. [15]		59.27	-	-	78.60
VRC-PG (ours)	Flickr30K	**60.51**	**78.77**	81.50	**78.86**
Fang et al. [65]	MS COCO	-	-	-	29.00
Akbari et al. [36]		-	-	-	69.19
Align2Ground [151]		-	-	-	71.00
InfoGround [18]		51.67	77.69	83.25	76.74
VRC-PG (ours)	MS COCO	**55.55**	**79.23**	**84.12**	**78.97**

In terms of R@10, our model obtains a lower performance (−1.41%) than InfoGround [18] when trained on the Flickr30K Entities training set. We analyzed this difference and found that our approach without the OPP module gets an R@10 of 83.86% which improves the performance of InfoGround by 0.95% point. The reason is that after the OPP module, we keep a smaller object proposals set as input to the next module than without the OPP module. The main contribution of the InfoGround model is that it uses the language model to generate a context-preserving negative caption set; the authors show that this improves the results in comparison to randomly sampling negatives from the training data. In our approach, we do not employ this negative caption set. In order to verify this, we re-train our model employing this negative caption set used in InfoGround [18]. Our proposed model with these negative captions results in 66.60% and 78.83% in terms of R@1 and Pt_Acc, respectively, with the model trained on the Flickr30K Entities training set. For the models trained on MS COCO, our approach with negative captions achieves 59.47% and 79.34% in terms of R@1 and Pt_Acc, respectively. Both of them demonstrate that our approach achieves much higher performances than InfoGround when employing the same settings of negative captions.

5.3.4. ABLATION STUDY

In Table 5.2, we report the quantitative performance of 8 different design choices, i.e., c1-c8, within our proposed model on Flickr30K Entities validation set. In this experiment, we take the design only consisting of the VTCA module as our baseline model, which is only supervised by image-caption pairs based on InfoNCE loss, similar to in the model by Gupta et al. [18]. The introduction of VSA improves Pt_Acc from 62.43% to 64.26%, but results in a drop of R@1 from 32.12% to 29.64% (c1 vs. baseline). Our OPP module, as shown in Table 5.2, brings a performance gain of 3.24% in terms of R@1, but a 1.46% lower Pt_Acc (c2 vs. baseline).

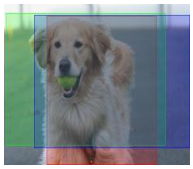
w/o VSA	VSA	w/o VSA	VSA
			
[0.15, 0.12, 0.11]	[0.19, 0.17, 0.09]	[0.73, 0.19, 0.05]	[0.82, 0.16, 0.02]
Two older men are sitting on opposite ends of a bench.		A blond man stands next to a cement mixer with mountains in the background.	
			
[0.67, 0.31, 0.01]	[0.79, 0.16, 0.04]	[0.15, 0.13, 0.12]	[0.21, 0.13, 0.11]
A man and a little girl happily posing in front of their cart in a supermarket.		Four girls in shorts on the beach throwing a football with the ocean behind them.	
			
[0.18, 0.18, 0.13]	[0.25, 0.2, 0.15]	[0.17, 0.12, 0.11]	[0.18, 0.14, 0.14]
A little white curly-haired dog runs across the pavement with a stick in its mouth.		A golden-colored dog , with his eyes alert , holds a brightly colored tennis ball in his mouth.	
			
[0.49, 0.3, 0.06]	[0.49, 0.45, 0.04]	[0.89, 0.03, 0.02]	[0.96, 0.02, 0.01]
A single man , riding his bike on the pier at sunset.		A young girl in a green shirt and shorts out riding her bike past a very nice apartment building.	

Figure 5.3: Attention scores achieved in Eq. 5.9 of region proposals on the Flickr30K Entities validation set for the setting without/with the visual self-attention module (i.e., w/o VSA and VSA). The visual regions surrounded by bounding boxes refer to the object proposals with top-3 cross-modal attention scores (colored by red, green and blue).

Table 5.2: Benefits of the different modules in our approach. All models are trained on the Flickr30K Entities training set and the results (%) are reported for the Flickr30K Entities validation set.

Methods	OPP	VSA	Loss	R@1	Pt_Acc
baseline	-	-	-	32.13	62.43
c1	-	✓	-	29.64	64.26
c2	✓	-	-	35.37	60.97
c3	✓	✓	-	39.21	63.61
c4	-	-	✓ w/o $\mathscr{L}_{\mathscr{S}}$	48.90	76.60
c5	-	✓	✓ w/o $\mathscr{L}_{\mathscr{S}}$	52.71	78.31
c6	-	✓	✓	53.20	78.27
c7	✓	-	✓ w/o $\mathscr{L}_{\mathscr{S}}$	55.64	77.58
c8	✓	✓	✓ w/o $\mathscr{L}_{\mathscr{S}}$	57.90	77.24
VRC-PG	✓	✓	✓	58.64	77.03

When we use these two modules together, the R@1 is improved from 32.13% to 39.21% and Pt_Acc from 62.43% to 63.61% (c3 vs. baseline). Thus, OPP is more positive for R@1 and VSA for Pt_Acc. If we want to simultaneously optimize both metrics, these two kind of modules can work in coordination with each other. We replace the InfoNCE loss in the baseline by our contrastive loss function (without $\mathscr{L}_S$), and achieve an improvement of 16.77% on R@1 and 14.17% on Pt_Acc (c4 vs. baseline). If we further add the visual self-attention loss $\mathscr{L}_S$, we can obtain a better result on R@1 and close result on Pt_Acc (c6 vs. c5 and VRC-PG vs. c8). This shows that our contrastive loss is very useful in the phrase grounding task.

In Fig. 5.3, we visualize a few examples of different model settings, i.e., with and without VSA, on the Flickr30K validation set. The figure indicates that the setting with VSA can lead to more attention being paid to the correct visual region corresponding to the phrase in the sentence than without VSA. For example, for the top-right example in the figure, we find that the setting with VSA gives a score (0.82) of attention to the bounding box (red) enclosing a man, while the setting without VSA generates a lower attention score (0.73) for the region (red) covering the man and a large area of background.

5.3.5. QUALITATIVE RESULTS

In Fig. 5.4, we illustrate the qualitative results of visual grounding of phrases obtained by our approach on three image-caption pairs from the Flickr30K Entities test data. From this figure, it is evident that our model has the ability to localize phrases from the caption in the image. In Fig. 5.5, we show the attention scores obtained by Eq. 5.9 from the VTCA module in our model. For example, for the word 'old', our approach generates a high attention to visual region No. 17 (cf. Fig. 5.5(a)). It is visible in the image that this region contains a head with white hair and exhibits a kind of visual appearance of 'old'. Regions 29 and 3 are about the topic of scenes, and we can observe that the corresponding cells are high-

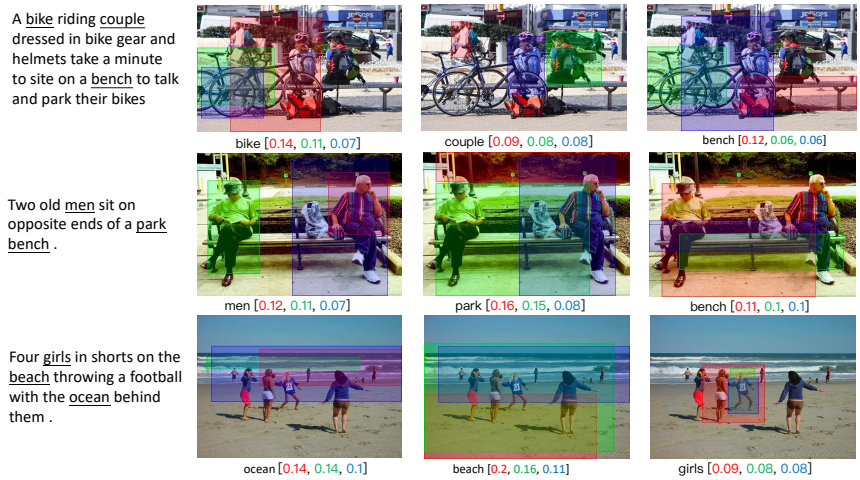
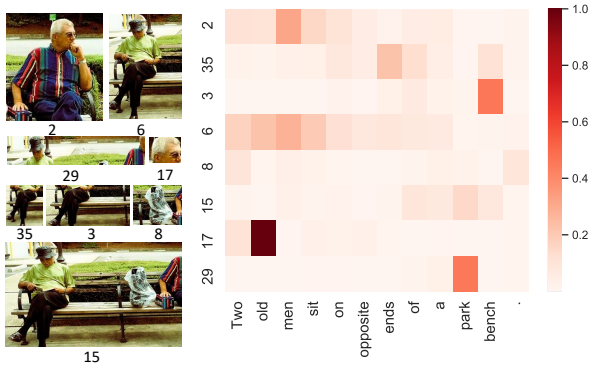


Figure 5.4: Visualization of weakly supervised phrase grounding. In each image, for a given word query, we show the visual regions in the form of bounding boxes with top-3 cross-modal attention scores (colored by red, green and blue) achieved in Eq. 5.9.



(a) Visual object proposals

(b) Attention weight map

Figure 5.5: Cross-modal attention scores achieved by Eq. 5.9 between visual object proposals and words. The darker cell color indicates that more attention is paid to the corresponding visual object proposals for a word query.

lighted in the attention weight map when the query of phrase is 'park' and 'bench'. Regions 2 and 6 both relate to 'men', and they are really paid much attention to for the query of phrase 'men' as shown in the attention weight map.

5.4. CONCLUSION

In this work, we have proposed a novel weakly supervised approach to phrase grounding under the supervision of the correspondence between images and captions. Our key contribution lies in systematically learning contextualized visual representations with a mixed contrastive loss function. In the visual representation contextualization, the three modules, OPP, VSA and VTCA, work in coordination with each other for representing local visual semantics by considering the unimodal and cross-modal contexts. In addition, we define a novel contrastive loss function on the intra- and inter-modal representations and clearly demonstrate that this leads to better results. Overall, we report the improvements of 3.88% point and 1.24% point on $R@1$, and 2.23% point and 0.26% point on Pt_Acc, with the models trained on the MS COCO and Flickr30K Entities training set, respectively, compared to the state-of-the-art methods. Our qualitative analysis using visualization of attention between words and image regions also illustrates the capability of our model to learn joint representations of image and text using the attention mechanism.

6

CONCLUSIONS AND DISCUSSION

6.1. MAIN CONTRIBUTIONS

The main contributions of the work presented in this thesis can be summarised by answering the six research questions as presented in Chapter 1 and as elaborated in different chapters as follows:

RQ1: To what extent is it possible to improve the representation of visual features detected by CNNs or the representation of textual features embedding and reduce the semantic gap between visual and textual information?

In Chapter 2, we proposed a novel visually supervised textual representation learning model (VS-Word2Vec), which learns the vector representation of relation words by jointly computing over visual modality and natural language. The framework of our model is inspired by the structure of CBOW of Word2Vec. In this method, we first compute the visual features based on deep networks over an image patch that reflects a relation word, and then achieve the visual similarity matrix for all relation words. The VS-Word2Vec model then resolves an optimization problem that consists of the terms related to the visual similarity and context in natural language. Our experiments demonstrate that our approach really changes the distribution of word representations and achieves more accurate similarity of words than the CBOW model and reduces the semantic gap between visual and textual information in a common embedding space.

RQ2: How to utilize an additional knowledge base to measure semantic matching?

We addressed this question in Chapter 3. First, we extract all noun words from the caption and analyze whether noun words and the coarse label belong to same synonym set in the knowledge base. We define the semantics between them to be similar if two words belong to same synonym set (i.e. we can use the noun words to replace the coarse label), otherwise they have different semantics (i.e. we cannot use the noun words to replace the coarse label). We employ the lexical database WordNet to analyze the connection between the coarse label and fine-grained label of object categories, then we build a semantic map to extend the coarse label to the fine-grained label. We propose a novel approach to the problem of fine-grained object label learning with the weak supervision of captions. Experimental results implemented on public datasets with fine-grained categories demonstrate the effectiveness of our approach. Through our semantic map, we can extend the 80 coarse categories to more than 160 fine-grained categories of the MS-COCO dataset and the number of fine-grained categories is decided by the words in caption. The new fine-grained object detection results show that our semantic map is helpful for improving the visual representation learning to measure semantic matching of cross-modal information.

RQ3: To what extent can curriculum learning measure the distribution of visual complexity and enhance weak supervision for semantic matching?

Data distribution can affect the accuracy of a learning model, and many datasets have a long-tail distribution problem. In Chapter 3, we build a semantic map to replace the object coarse categories with fine-grained categories, which is a weakly supervised learning process. The main challenge that is the new categories cause long-tail distribution of the visual labels. To address that problem, we build a learning process to address this long-tail distribution. Curriculum learning defines a learning process in which the samples are ranked from from easy samples to complex samples or from complex to easy. Based on the ranking of samples, the model can gradually learn the negative effects brought by noisy data in an early period of training. It can be also used for deciding the learning order of tasks. In this thesis, we introduce the label inference curriculum network with the consideration of the complexity of samples that describes the difficulty of fine-grained label learning. To evaluate the performance of fine-grained label learning, we construct multiple datasets based on widely-used public datasets. Experimental results demonstrate the effectiveness of our approach in the task of fine-grained label learning.

RQ4: How and with what quality can we model the semantic correlations between two different modalities?

In Chapter 4, we propose a new approach called kernel-based mixture mapping (KMM) to model the semantic correlations between web images and text. With this approach, we first construct latent high-dimensional feature spaces based on kernel theory to address the non-linearity of both the data distributions in the input spaces and the cross-model correlation. Second, we present a probabilistic neighborhood model to describe the spatial locality of semantics by assuming that proximate examples in feature spaces generally have the same semantics and a conditional model to describe cross-modal conditional dependency. Finally, we build a probabilistic mixture model to jointly model the spatial locality of semantics and the conditional dependency between different modalities. By combining nonlinear transformation and probabilistic models, KMM can address the nonlinearity of cross-modal correlation, the complexity of the semantic distributions at the global scale, and the continuity of semantic distributions at the local scale.

RQ5: What is the effect of the attention mechanism to eliminate the different modal representations produced in the common embedding space?

In recent years, transformer models have achieved state-of-the-art results in computer vision and NLP tasks. The transformer structure is based on the atten-

tion mechanism, which pays greater attention to certain factors when processing the data. In Chapter 5, we proposed an deep transformation net to embed visual features and textual features into a common vector space. Based on the paired image and caption, we optimize the parameters of the transformation net to achieve the best similarity score between the visual and textual representations that share the same semantics. Our deep transformation net for the visual contextualized representation is systematically learned in three stages: (1) object proposals pooling (OPP), (2) visual self-attention (VSA) and (3) visual-textual cross-modal attention (VTCA). OPP is utilized to alleviate the suppression of each object feature, which benefits the visual representation contextualization in terms of trading off the richness of visual components and computational efficiency. VSA aims to capture the correlation among object proposals of each image and generate the representation of each candidate incorporating the visual information of the other candidates. In order to measure the cross-modal compatibility in terms of topics, we subsequently introduce the VTCA module to represent the visual topic corresponding to each textual component (phrase) in the caption in a cross-modal common vector space, guided by the attention of a word to object proposals. Cross attention can discovers the latent alignment using both image regions and words in sentences as context via attention across modalities, which produces more accurate image–text similarity for matching.

RQ6: How to employ the correspondence between images and text as supervision instead of the matching annotations to address the limited data issue?

To address the issue of limited data, many prior methods are trained based on self-supervised and semi-supervised learning. The accuracy, however, is always lower than with fully supervised learning. In Chapter 5, we build models, i.e. VSA and VTCA, that are both based on a contrastive learning algorithm to improve the accuracy in image classification and image caption generation. A contrastive learning model or net can learn representations by maximizing agreement between differently augmented views of the same data example via a contrastive loss in the latent space. In this thesis, inspired by NCE loss, we have built a mixed contrastive loss function for a VTCA module including two terms: one is a contrastive loss function to improve cross-modal compatibility in terms of the topic of images and captions, and the other is to control the difference of the visual representations induced by the VSA module. Our model VTCA with a mixed contrastive loss function improves the phrase grounding accuracy both for the models trained on the MS COCO and Flickr30K Entities training set, compared to the state-of-the-art methods.

6.2. ACHIEVEMENTS OF RESEARCH PRESENTED IN THIS THESIS

Based on our analysis, we can see that the main challenge is to build a model that can serve as the bridge to connect visual representations and textual representations with the same semantics in the common semantic space. There are several types of models to build the common embedding space: 1) linear/nonlinear mapping, 2) probabilistic models, 3) knowledge-based correlation propagation methods, and 4) deep learning-based methods.

In this thesis, we used all types to build the common embedding space. First, we proposed the VS-Word2Vec model, which fuses the visual modality and natural language together to learn the relation words representation with visual supervision. The VS-Word2Vec model is a linear mapping with weights from visual features. Second, we propose a new approach called kernel-based mixture mapping (KMM) to model the semantic correlations between web images and text. KMM combines nonlinear transformation and probabilistic models, which can address the non-linearity of cross-modal correlation, the complexity of the semantic distributions at the global scale, and the continuity of semantic distributions at the local scale. Finally, we subsequently introduce the VTCA module to represent the visual topics corresponding to each textual component (phrase) in the caption in a cross-modal common vector space, guided by the attention of a word to object proposals.

Another solution is to improve the uni-modal representation so that it becomes closer to the other modality representation. In this thesis, we propose a novel approach called label inference curriculum network (LICN) to the problem of fine-grained object label learning with the weak supervision of captions. First, we construct a semantic map that builds a correspondence between the coarse category labels provided by public datasets and the fine-grained category labels extracted from captions based on the combination of embedding techniques and knowledge bases. Second, we present the label inference curriculum network with the consideration of the complexity of samples that describes the difficulty of fine-grained label learning.

6.3. FUTURE RESEARCH

Cross-modal semantic matching is a complex task. The different modality representations can be extracted based on different levels: visual representations are most based on image-level or region-level representations; textual representations are most based on word-level, phrase-level, expression-level[1]or sentence-level. The level will decide the quality of the model for representation of the common space. In this thesis, we focus on two different levels to improving the ac-

[1]The difference between phrase-level and expression-level is that phrase need to be extracted from the sentence first, while an expression is an independent unit.

curacy of semantic matching: (1) image-level and word-level; (2) region-level and word-level. We use the paired data on different levels to learn each single modality representation and build a model to transform a single modality representation to the common semantic space. However, we did not use the textual modality representation based on the phrase-level (expression-level) and not use this basic feature to understand the visual information more deeply, i.e, scene graph generation and image caption generation.

Based on the above analysis, we provide three recommendations for future research. The first recommendation is to build a dataset with more annotations for different tasks, e.g., image caption generation, visual grounding, object detection (coarse and fine-grained label), visual relationship detection and scene graph generation. Above tasks are all based on the connection between language and vision. As we know, MS COCO[21] was designed for object detection and image caption generation, and was annotated with object instances with category labels, each image paired with 5 captions; RefCOCO [169], RefCOCO+ [169] and RefCOCOg [170] were designed for expression grounding tasks. These datasets employ some images from the MS COCO dataset and replace the category label of the object instance by expressions; The Visual Genome [22] dataset [22] contains annotations for the densest representations and is the largest dataset of image descriptions, objects, attributes, relationships, and visual question answering. There are 76,631 shared images between Visaul Genome and MS COCO. In our research we exploited the overlapping part among these three datasets. Similar to the dataset sCOCO, which we construct in Chapter 3, we can construct a new dataset marked with different annotations of object instances for all above tasks.

Our second recommendation is to investigate how to improve the semantic matching based on textual representations on the expression-level (or sentence-level) and visual representations on the region-level in the common space. The main task for expression-level and region-level representation is Referring Expression Grounding (REG), and we can evaluate the model for this task on the RefClef, RefCOCO, RefCOCO+ and RefCOCOg. The framework takes two branches as input: one is for the textual representation on the expression-level, and the other is for the visual representation one the region-level. For each branch a model is built to embed the representations into a common embedding space. In the common embedding space, we can match the visual features and textual features based on their vectors. The main challenge is that for the expression-level representation it is difficult to learn to express the real semantic meaning, as the expression-level representation combines the word representations together and analyzes the contextualized connection between words in the expression.

One additional direction for future research comes from the argument that text is more ambiguous and diverse than vision, i.e., language can express the same concept with different words and different concepts with the same words, but vision labels are limited by the single selected label in the human annotation. Therefor, we should make use of the textual diversity of expressions to transmit or deliver the information between paired vision-language data and thereby solve

the limitation of visual data labels. For now, the Scene Graph Generation (SGG) task is more based on the image with the human scene graph annotation. However, there are little datasets for this task that can be used to train models based on full supervision. The image-caption (sentence) paired data is easy obtain from web or existing open source datasets. Based on the image-caption paired data, i.e., the different modal (vision and text) with the same semantics, we can obtain the shared fine-grained semantic matching, i.e., region-word connections or correlations, to generate the visual scene graph. This can be seen as a weakly supervised learning process for the SGG task, which is based on image-caption training data instead of not regions and scene graphs training data. The main challenge of weakly supervised SGG with image-caption paired data as training data is how to connect the regions in the image and the words in the caption. Our thesis can be seen as an solution of this challenge. We will evaluate in future work whether we can improve the accuracy of scene graph generation with our proposed model.

BIBLIOGRAPHY

[1] Sepp Hochreiter and Jürgen Schmidhuber. “Long short-term memory”. In: vol. 9. 8. MIT Press, 1997, pp. 1735–1780.

[2] Yann LeCun et al. “Gradient-based learning applied to document recognition”. In: *Proceedings of the IEEE* 86.11 (1998), pp. 2278–2324. DOI: 10.1109/5.726791.

[3] Ross Girshick. “Fast R-CNN”. In: *Proceedings of the IEEE International Conference on Computer Vision*. 2015, pp. 1440–1448.

[4] Karen Simonyan and Andrew Zisserman. “Very deep convolutional networks for large-scale image recognition”. In: *Proceedings of the International Conference on Learning Representations*. 2015.

[5] Kaiming He et al. “Deep residual learning for image recognition”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016, pp. 770–778.

[6] Kyunghyun Cho et al. “Learning phrase representations using RNN encoder-decoder for statistical machine translation”. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2014, pp. 1724–1734. DOI: 10.3115/v1/D14-1179.

[7] Jacob Devlin et al. “BERT: Pre-training of deep bidirectional transformers for language understanding”. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 2019, pp. 4171–4186. DOI: 10.18653/v1/N19-1423.

[8] Parminder Kaur, Husanbir Singh Pannu, and Avleen Kaur Malhi. “Comparative analysis on cross-modal information retrieval: a review”. In: *Computer Science Review* 39 (2021), p. 100336.

[9] Yunchao Gong et al. “A multi-view embedding space for modeling internet images, tags, and their semantics”. In: *International Journal of Computer Vision* 106.2 (2014), pp. 210–233.

[10] Viresh Ranjan, Nikhil Rasiwasia, and CV Jawahar. “Multi-label Cross-modal Retrieval”. In: *Proceedings of the IEEE International Conference on Computer Vision*. 2015, pp. 4094–4102.

[11] Peter Anderson et al. “Bottom-up and top-down attention for image captioning and visual question answering”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 6077–6086.

[12] Chenxi Liu et al. “Attention correctness in neural image captioning”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 31. 1. 2017.

[13] Bo Dai and Dahua Lin. “Contrastive learning for image captioning”. In: *arXiv Preprint arXiv: 1710.02534* (2017).

[14] Liwei Wang et al. “Improving weakly supervised visual grounding by contrastive knowledge Distillation”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 14090–14100.

[15] Yongfei Liu et al. “Relation-aware Instance Refinement for Weakly Supervised Visual Grounding”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 5612–5621.

[16] Damien Teney et al. “Tips and tricks for visual question answering: Learnings from the 2017 challenge”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 4223–4232.

[17] Tianyu Yu et al. “Cross-dodal omni interaction modeling for phrase grounding”. In: *Proceedings of the 28th ACM International Conference on Multimedia*. 2020, pp. 1725–1734.

[18] Tanmay Gupta et al. “Contrastive Learning for Weakly Supervised Phrase Grounding”. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2020.

[19] Long Chen et al. “Ref-NMS: Breaking Proposal Bottlenecks in Two-Stage Referring Expression Grounding”. In: *arXiv Preprint arXiv: 2009.01449* (2020).

[20] Sibei Yang, Guanbin Li, and Yizhou Yu. “Relationship-embedded representation learning for grounding referring expressions”. In: *arXiv Preprint arXiv: 1906.04464* (2019).

[21] Tsung-Yi Lin et al. “Microsoft coco: Common objects in context”. In: *European Conference on Computer Vision*. Springer. 2014, pp. 740–755.

[22] Ranjay Krishna et al. “Visual genome: Connecting language and vision using crowdsourced dense image annotations”. In: *International Journal of Computer Vision* 123.1 (2017), pp. 32–73.

[23] Bryan A Plummer et al. “Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models”. In: *Proceedings of the IEEE International Conference on Computer Vision*. 2015, pp. 2641–2649.

[24] Peter Young et al. “From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions”. In: *Transactions of the Association for Computational Linguistics* 2 (2014), pp. 67–78.

[25] Tomas Mikolov et al. “Distributed representations of words and phrases and their compositionality”. In: *arXiv Preprint arXiv: 1310.4546* (2013).

[26] Tomas Mikolov et al. “Efficient estimation of word representations in vector space”. In: *arXiv Preprint arXiv: 1301.3781* (2013).

[27] Jeffrey Pennington, Richard Socher, and Christopher D Manning. "Glove: Global vectors for word representation". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2014, pp. 1532–1543.

[28] Jacob Devlin et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. URL: https://www.aclweb.org/anthology/N19-1423.

[29] Xuejing Liu et al. "Adaptive reconstruction network for weakly supervised referring expression grounding". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019, pp. 2611–2620.

[30] Shaoqing Ren et al. "Faster R-CNN: Towards real-time object detection with region proposal networks". In: *Advances in Neural Information Processing Systems*. 2015, pp. 91–99.

[31] Kaiming He et al. "Mask r-cnn". In: *Proceedings of the IEEE International Conference on Computer Vision*. 2017, pp. 2961–2969.

[32] Ya Jing et al. "Learning Aligned Image-Text Representations Using Graph Attentive Relational Network". In: *IEEE Transactions on Image Processing* 30 (2021), pp. 1840–1852.

[33] Christiane Fellbaum. "WordNet". In: *Theory and Applications of Ontology: Computer Applications*. Springer, 2010, pp. 231–243.

[34] Robyn Speer, Joshua Chin, and Catherine Havasi. "Conceptnet 5.5: An open multilingual graph of general knowledge". In: *Thirty-first AAAI Conference on Artificial Intelligence*. 2017.

[35] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. "Representation learning with contrastive predictive coding". In: *arXiv Preprint arXiv: 1807.03748* (2018).

[36] Hassan Akbari et al. "Multi-level multimodal common semantic space for image-phrase grounding". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 12476–12486.

[37] David R Hardoon, Sandor Szedmak, and John Shawe-Taylor. "Canonical correlation analysis: An overview with application to learning methods". In: *Neural Computation* 16.12 (2004), pp. 2639–2664.

[38] David Grangier and Samy Bengio. "A Discriminative Kernel-Based Approach to Rank Images from Text Queries". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 30.8 (2008), pp. 1371–1384. DOI: 10.1109/TPAMI.2007.70791.

[39] Raman Arora and Karen Livescu. “Kernel CCA for multi-view learning of acoustic features using articulatory measurements”. In: *Symposium on machine learning in speech and language processing*. 2012.

[40] Douglas B Lenat. “CYC: A large-scale investment in knowledge infrastructure”. In: *Communications of the ACM* 38.11 (1995), pp. 33–38.

[41] Sheng Guo et al. “CurriculumNet: Weakly supervised learning from large-scale web images”. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018, pp. 135–150.

[42] Xin Huang and Yuxin Peng. “Deep cross-media knowledge transfer”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 8837–8846.

[43] Youtian Du and Kai Yang. “Learning semantic correlation of web images and text with mixture of local linear mappings”. In: *Proceedings of the 23rd ACM International Conference on Multimedia*. 2015, pp. 1259–1262.

[44] Xing Xu et al. “Cross-modal attention with semantic consistence for image–text matching”. In: *IEEE Transactions on Neural Networks and Learning Systems* 31.12 (2020), pp. 5412–5425.

[45] Michael Gutmann and Aapo Hyvärinen. “Noise-contrastive estimation: A new estimation principle for unnormalized statistical models”. In: *Proceedings of the thirteenth international Conference on Artificial Intelligence and statistics*. JMLR Workshop and Conference Proceedings. 2010, pp. 297–304.

[46] John Lafferty and Chengxiang Zhai. “Document language models, query models, and risk minimization for information retrieval”. In: *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2001, pp. 111–119.

[47] Angeliki Lazaridou et al. “Compositionally derived representations of morphologically complex words in distributional semantics”. In: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2013, pp. 1517–1526.

[48] Kevin Lund and Curt Burgess. “Producing high-dimensional semantic spaces from lexical co-occurrence”. In: *Behavior Research Methods, Instruments, & Computers* 28.2 (1996), pp. 203–208.

[49] Marco Baroni and Alessandro Lenci. “Distributional memory: A general framework for corpus-based semantics”. In: *Computational Linguistics* 36.4 (2010), pp. 673–721.

[50] Ronan Collobert and Jason Weston. “A unified architecture for natural language processing: Deep neural networks with multitask learning”. In: *Proceedings of the 25th International Conference on Machine Learning*. 2008, pp. 160–167.

[51] Piotr Bojanowski et al. “Enriching word vectors with subword information”. In: *Transactions of the Association for Computational Linguistics* 5 (2017), pp. 135–146.

[52] Satwik Kottur et al. “Visual word2vec (Vis- W2V): Learning visually grounded word embeddings using abstract scenes”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016, pp. 4985–4994.

[53] Cewu Lu et al. “Visual relationship detection with language priors”. In: *European Conference on Computer Vision*. Springer. 2016, pp. 852–869.

[54] Radityo Eko Prasojo, Mouna Kacimi, and Werner Nutt. “Modeling and summarizing news events using semantic triples”. In: *European Semantic Web Conference*. Springer. 2018, pp. 512–527.

[55] Qiuhao Lu and Youtian Du. “Wikipedia-based Entity Semantifying in Open Information Extraction”. In: *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*. Vol. 1. IEEE. 2017, pp. 765–770.

[56] Daniela Gerz et al. “Simverb-3500: A large-scale evaluation set of verb similarity”. In: *arXiv Preprint arXiv: 1608.00869* (2016).

[57] Ali Diba et al. “Weakly supervised cascaded convolutional networks”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 914–922. ISBN: 9781538604571. DOI: 10.1109/CVPR.2017.545.

[58] Joseph Redmon et al. “You only look once: Unified, real-time object detection”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016, pp. 779–788.

[59] Augustus Buonviri et al. “Survey of Challenges in Labeled Random Finite Set Distributed Multi-Sensor Multi-Object Tracking”. In: *2019 IEEE Aerospace Conference*. IEEE. 2019, pp. 1–12.

[60] Wei Du, Ronald Phlypo, and Tülay Adalı. “Adaptive Feature Selection and Feature Fusion for Semi-supervised Classification”. In: *Journal of Signal Processing Systems* 91.5 (2019), pp. 521–537.

[61] Hakan Bilen and Andrea Vedaldi. “Weakly supervised deep detection networks”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016, pp. 2846–2854.

[62] Fang Wan et al. “Min-entropy latent model for weakly supervised object detection”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 1297–1306.

[63] Keren Ye et al. “Cap2Det: Learning to amplify weak caption supervision for object detection”. In: *Proceedings of the IEEE International Conference on Computer Vision*. 2019, pp. 9686–9695.

[64] Christopher Thomas and Adriana Kovashka. “Predicting the politics of an image using webly supervised data”. In: *Advances in Neural Information Processing Systems*. 2019, pp. 3630–3642.

[65] Hao Fang et al. “From captions to visual concepts and back”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015, pp. 1473–1482.

[66] Ishan Misra et al. “Seeing through the human reporting bias: Visual classifiers from noisy human-centric labels”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016, pp. 2930–2939.

[67] Mingda Zhang, Rebecca Hwa, and Adriana Kovashka. “Equal but not the same: Understanding the implicit relationship between persuasive images and text”. In: *arXiv Preprint arXiv: 1807.08205* (2018).

[68] Weifeng Ge, Sibei Yang, and Yizhou Yu. “Multi-evidence filtering and fusion for multi-label classification, object detection and semantic segmentation based on weakly supervised learning”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 1277–1286.

[69] Peng Tang et al. “PCL: Proposal cluster learning for weakly supervised object detection”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 42.1 (2018), pp. 176–191.

[70] Xiaolin Zhang et al. “Adversarial complementary learning for weakly supervised object localization”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 1325–1334.

[71] Yu Zhang et al. “Weakly supervised fine-grained categorization with part-based image representation”. In: *IEEE Transactions on Image Processing* 25.4 (2016), pp. 1713–1725.

[72] Youtian Du et al. “Fundamental visual concept learning from correlated images and text”. In: *IEEE Transactions on Image Processing* 28.7 (2019), pp. 3598–3612.

[73] Yale Song and Mohammad Soleymani. “Polysemous visual-semantic embedding for cross-modal retrieval”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2019, pp. 1979–1988.

[74] Jiu-le TIAN and Wei Zhao. “Words similarity algorithm based on Tongyici Cilin in semantic web adaptive learning system”. In: *Journal of Jilin University (Information Science Edition)* 6.010 (2010).

[75] Zhendong Dong, Qiang Dong, and Changling Hao. “HowNet and the computation of meaning”. In: (2006).

[76] Qun Liu. “Word similarity computing based on HowNet”. In: *Computational Linguistics and Chinese Language Processing* 7.2 (2002), pp. 59–76.

[77] Jonathan Krause et al. “A hierarchical approach for generating descriptive image paragraphs”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 317–325.

[78] Changliang Li et al. “Measuring word semantic similarity based on transferred vectors”. In: *International Conference on Neural Information Processing*. Springer. 2017, pp. 326–335.

[79] Maxime Oquab et al. “Is object localization for free?-Weakly-supervised learning with convolutional neural networks”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015, pp. 685–694.

[80] Bolei Zhou et al. “Learning deep features for discriminative localization”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016, pp. 2921–2929.

[81] Vadim Kantorov et al. “ContextLocNet: Context-aware deep network models for weakly supervised localization”. In: *European Conference on Computer Vision*. Springer. 2016, pp. 350–365.

[82] Peng Tang et al. “Multiple instance detection network with online instance classifier refinement”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 2843–2851. ISBN: 9781538604571.

[83] Yunchao Wei et al. “TS2C: Tight box mining with surrounding segmentation context for weakly supervised object detection”. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018, pp. 434–450.

[84] Yoshua Bengio et al. “Curriculum learning”. In: *Proceedings of the 26th Annual International Conference on Machine Learning*. 2009, pp. 41–48.

[85] Dingwen Zhang et al. “Leveraging prior-knowledge for weakly supervised object detection under a collaborative self-paced curriculum learning framework”. In: *International Journal of Computer Vision* 127.4 (2019), pp. 363–380.

[86] Miaojing Shi and Vittorio Ferrari. “Weakly supervised object localization using size estimates”. In: *European Conference on Computer Vision*. Springer. 2016, pp. 105–121. ISBN: 9783319464534. DOI: 10.1007/978-3-319-46454-1_7.

[87] Jiasi Wang, Xinggang Wang, and Wenyu Liu. “Weakly- and semi-supervised Faster R-CNN with curriculum learning”. In: *2018 24th International Conference on Pattern Recognition (ICPR)*. IEEE. 2018, pp. 2416–2421.

[88] Guy Hacohen and Daphna Weinshall. “On the power of curriculum learning in training deep networks”. In: *arXiv Preprint arXiv: 1904.03626* (2019).

[89] Christopher D Manning et al. “The Stanford CoreNLP natural language processing toolkit”. In: *Proceedings of 52nd Annual Meeting of the Association for Computational linguistics: system demonstrations*. 2014, pp. 55–60.

[90] Nikhil Rasiwasia et al. “A new approach to cross-modal multimedia retrieval”. In: *Proceedings of the 18th ACM International Conference on Multimedia*. 2010, pp. 251–260.

[91] Xixuan Wu et al. “Cross matching of music and image”. In: *Proceedings of the 20th ACM International Conference on Multimedia*. 2012, pp. 837–840.

[92] Yue-Ting Zhuang, Yi Yang, and Fei Wu. “Mining semantic correlation of heterogeneous multimedia data for cross-media retrieval”. In: *IEEE Transactions on Multimedia* 10.2 (2008), pp. 221–229.

[93] Yansong Feng and Mirella Lapata. “Automatic caption generation for news images”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35.4 (2012), pp. 797–812.

[94] Lei Wu, Rong Jin, and Anil K Jain. “Tag completion for image retrieval”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35.3 (2012), pp. 716–727.

[95] Meng Wang et al. “Assistive tagging: A survey of multimedia tagging with human-computer joint exploration”. In: *ACM Computing Surveys (CSUR)* 44.4 (2012), pp. 1–24.

[96] Hai-Feng Guo et al. “Deep multi-instance multi-label learning for image annotation”. In: *International Journal of Pattern Recognition and Artificial Intelligence* 32.03 (2018), p. 1859005.

[97] Jinhui Tang et al. “Cross-space affinity learning with its application to movie recommendation”. In: *IEEE Transactions on Knowledge and Data Engineering* 25.7 (2012), pp. 1510–1519.

[98] Jose Costa Pereira et al. “On the role of correlation and abstraction in cross-modal multimedia retrieval”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36.3 (2013), pp. 521–535.

[99] Tao Jiang and Ah-Hwee Tan. “Learning image-text associations”. In: *IEEE Transactions on Knowledge and Data Engineering* 21.2 (2008), pp. 161–177.

[100] Florent Monay and Daniel Gatica-Perez. “Modeling semantic aspects for cross-media image indexing”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 29.10 (2007), pp. 1802–1817.

[101] Xiaohua Zhai, Yuxin Peng, and Jianguo Xiao. “Effective heterogeneous similarity measure with nearest neighbors for cross-media retrieval”. In: *International Conference on Multimedia Modeling*. Springer. 2012, pp. 312–322.

[102] Aviv Eisenschtat and Lior Wolf. “Linking image and text with 2-way nets”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Rcognition*. 2017, pp. 4601–4611.

[103] Liwei Wang et al. “Learning two-branch neural networks for image-text matching tasks”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41.2 (2018), pp. 394–407.

[104] David Grangier and Samy Bengio. "A discriminative kernel-based model to rank images from text queries". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 30.8 (2008), pp. 1371–1384.

[105] Tao Jiang and Ah-Hwee Tan. "Learning image-text associations". In: *IEEE Transactions on Knowledge and Data Engineering* 21.2 (2009), pp. 161–177.

[106] Pereira J Costa et al. "On the Role of Correlation and Abstraction in Cross-Modal Multimedia Retrieval." In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36.3 (2014), pp. 521–535.

[107] Hong Zhang, Yueting Zhuang, and Fei Wu. "Cross-modal correlation learning for clustering on image-audio dataset". In: *Proceedings of the 15th ACM International Conference on Multimedia*. 2007, pp. 273–276.

[108] Hong Liu et al. "Supervised matrix factorization for cross-modality hashing". In: *arXiv Preprint arXiv: 1603.05572* (2016).

[109] Ting Zhang and Jingdong Wang. "Collaborative quantization for cross-modal similarity search". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016, pp. 2036–2045.

[110] Ran He et al. "Cross-modal subspace learning via pairwise constraints". In: *IEEE Transactions on Image Processing* 24.12 (2015), pp. 5543–5556.

[111] Jinhui Tang et al. "Cross-space affinity learning with its application to movie recommendation". In: *IEEE Transactions on Knowledge and Data Engineering* 25.7 (2013), pp. 1510–1519.

[112] Antoine Deleforge, Florence Forbes, and Radu Horaud. "High-dimensional regression with gaussian mixtures and partially-latent response variables". In: *Statistics and Computing* 25.5 (2015), pp. 893–911.

[113] Lauren A Hannah, David M Blei, and Warren B Powell. "Dirichlet process mixtures of generalized linear models." In: *Journal of Machine Learning Research* 12.6 (2011).

[114] Yan Hua et al. "Cross-modal correlation learning by adaptive hierarchical semantic aggregation". In: *IEEE Transactions on Multimedia* 18.6 (2016), pp. 1201–1216.

[115] Liang Zhang et al. "Cross-modal retrieval using multiordered discriminative structured subspace learning". In: *IEEE Transactions on Multimedia* 19.6 (2017), pp. 1220–1233.

[116] Xing Xu et al. "Learning discriminative binary codes for large-scale cross-modal retrieval". In: *IEEE Transactions on Image Processing* 26.5 (2017), pp. 2494–2507.

[117] Jiwoon Jeon, Victor Lavrenko, and Raghavan Manmatha. "Automatic image annotation and retrieval using cross-media relevance models". In: *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*. 2003, pp. 119–126.

[118] Yansong Feng and Mirella Lapata. "Automatic caption generation for news images". In: *IEEE transactions on Pattern Analysis and Machine Intelligence* 35.4 (2013), pp. 797–812.

[119] Ruofei Zhang et al. "A probabilistic semantic model for image annotation and multimodal image retrieval". In: *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*. Vol. 1. IEEE. 2005, pp. 846–851.

[120] Ying Wu, Qi Tian, and Thomas S Huang. "Discriminant-EM algorithm with application to image retrieval". In: *Proceedings IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2000 (Cat. No. PR00662)*. Vol. 1. IEEE. 2000, pp. 222–227.

[121] Yangqing Jia, Mathieu Salzmann, and Trevor Darrell. "Learning cross-modality similarity for multinomial data". In: *2011 International Conference on Computer Vision*. IEEE. 2011, pp. 2407–2414.

[122] Anh Pham et al. "Multi-instance multi-label learning in the presence of novel class instances". In: *International Conference on Machine Learning*. PMLR. 2015, pp. 2427–2435.

[123] Wanxia Lin, Tong Lu, and Feng Su. "A novel multi-modal integration and propagation model for cross-media information retrieval". In: *International Conference on Multimedia Modeling*. Springer. 2012, pp. 740–749.

[124] Michalis Lazaridis et al. "Multimedia search and retrieval using multimodal annotation propagation and indexing techniques". In: *Signal Processing: Image Communication* 28.4 (2013), pp. 351–367.

[125] Jiao Xue, Youtian Du, and Hanbing Shui. "Semantic correlation mining between images and texts with global semantics and local mapping". In: *International Conference on Multimedia Modeling*. Springer. 2015, pp. 427–435.

[126] Dong Liu et al. "Image retagging using collaborative tag propagation". In: *IEEE Transactions on Multimedia* 13.4 (2011), pp. 702–712.

[127] Lei Zhang et al. "Full-space local topology extraction for cross-modal retrieval". In: *IEEE Transactions on Image Processing* 24.7 (2015), pp. 2212–2224.

[128] Fei Yan and Krystian Mikolajczyk. "Deep correlation for matching images and text". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015, pp. 3441–3450.

[129] Liwei Wang, Yin Li, and Svetlana Lazebnik. "Learning deep structure-preserving image-text embeddings". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016, pp. 5005–5013.

[130] Yuxin Peng et al. "CCL: Cross-modal correlation learning with multi-grained fusion by hierarchical network". In: *IEEE Transactions on Multimedia* 20.2 (2018), pp. 405–420.

[131] Richang Hong et al. “Coherent semantic-visual indexing for large-scale image retrieval in the cloud”. In: *IEEE Transactions on Image Processing* 26.9 (2017), pp. 4128–4138.

[132] Bo Wang et al. “Movie question answering: Remembering the textual cues for layered visual contents”. In: *Thirty-Second AAAI Conference on Artificial Intelligence*. 2018.

[133] Matt P Wand and M Chris Jones. *Kernel smoothing*. CRC Press, 1994.

[134] Christopher M. Bishop. “Pattern recognition and machine learning”. In: Springer, 2006.

[135] Jingdong Wang, Jianguo Lee, and Changshui Zhang. “Kernel trick embedded Gaussian mixture model”. In: *International Conference on Algorithmic Learning Theory*. Springer. 2003, pp. 159–174.

[136] Jeff A Bilmes et al. “A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models”. In: *International Computer Science Institute* 4.510 (1998), p. 126.

[137] Baback Moghaddam and Alex Pentland. “Probabilistic visual learning for object representation”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 19.7 (1997), pp. 696–710.

[138] Micah Hodosh, Peter Young, and Julia Hockenmaier. “Framing image description as a ranking task: Data, models and evaluation metrics”. In: *Journal of Artificial Intelligence Research* 47 (2013), pp. 853–899.

[139] Benjamin Klein et al. “Associating neural word embeddings with deep image representations using fisher vectors”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015, pp. 4437–4446.

[140] Tat-Seng Chua et al. “Nus-wide: A real-world web image database from national university of singapore”. In: *Proceedings of the ACM International Conference on Image and Video Retrieval*. 2009, pp. 1–9.

[141] Quoc Le and Tomas Mikolov. “Distributed representations of sentences and documents”. In: *International Conference on Machine Learning*. PMLR. 2014, pp. 1188–1196.

[142] Martin Engilberge et al. “Finding beans in burgers: Deep semantic-visual embedding with localization”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 3984–3993.

[143] A. Karpathy and L. Fei-Fei. “Deep visual-semantic alignments for generating image descriptions”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39.4 (2017), pp. 664–676.

[144] Ivan Vendrov et al. “Order-embeddings of images and language”. In: *arXiv Preprint arXiv: 1511.06361* (2015).

[145] Quanzeng You, Zhengyou Zhang, and Jiebo Luo. “End-to-end convolutional semantic embeddings”. In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2018.

[146] Yu Liu et al. “Learning a recurrent residual fusion network for multimodal matching”. In: *Proceedings of the IEEE International Conference on Computer Vision*. 2017, pp. 4107–4116.

[147] Xinlei Chen et al. “Microsoft COCO captions: Data collection and evaluation server”. In: *arXiv Preprint arXiv: 1504.00325* (2015).

[148] Stanislaw Antol et al. “VQA: Visual question answering”. In: *Proceedings of the IEEE International Conference on Computer Vision*. 2015, pp. 2425–2433.

[149] Alane Suhr et al. “A corpus for reasoning about natural language grounded in photographs”. In: *arXiv Preprint arXiv: 1811.00491* (2018).

[150] Rowan Zellers et al. “From recognition to cognition: Visual commonsense reasoning”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 6720–6731.

[151] Samyak Datta et al. “Align2ground: Weakly supervised phrase grounding guided by image-caption alignment”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019, pp. 2601–2610.

[152] Farley Lai et al. “Contextual Grounding of Natural Language Entities in Images”. In: *arXiv Preprint arXiv: 1911.02133* (2019).

[153] Mohit Bajaj, Lanjun Wang, and Leonid Sigal. “G3raphground: Graph-based language grounding”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019, pp. 4281–4290.

[154] Bryan A Plummer et al. “Phrase localization and visual relationship detection with comprehensive image-language cues”. In: *Proceedings of the IEEE International Conference on Computer Vision*. 2017, pp. 1928–1937.

[155] Kan Chen, Jiyang Gao, and Ram Nevatia. “Knowledge aided consistency for weakly supervised phrase grounding”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 4042–4050.

[156] A. Neubeck and L. Van Gool. “Efficient Non-Maximum Suppression”. In: *18th International Conference on Pattern Recognition (ICPR’06)*. Vol. 3. 2006, pp. 850–855. DOI: 10.1109/ICPR.2006.479.

[157] C Lawrence Zitnick and Piotr Dollár. “Edge boxes: Locating object proposals from edges”. In: *European Conference on Computer Vision*. Springer. 2014, pp. 391–405.

[158] Navaneeth Bodla et al. “Soft-NMS – Improving Object Detection With One Line of Code”. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. Oct. 2017, pp. 5561–5569.

[159] Yihui He et al. “Softer-NMS: Rethinking bounding box regression for accurate object detection”. In: *arXiv Preprint arXiv: 1809.08545* 2.3 (2018).

[160] Long Chen et al. “Ref-NMS: Breaking proposal bottlenecks in two-stage referring expression grounding”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 35. 2. 2021, pp. 1036–1044.

[161] Kaiming He et al. “Momentum contrast for unsupervised visual representation learning”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 9729–9738.

[162] Ting Chen et al. “A simple framework for contrastive learning of visual representations”. In: *International Conference on Machine Learning*. PMLR. 2020, pp. 1597–1607.

[163] Zhirong Wu et al. “Unsupervised feature learning via non-parametric instance discrimination”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 3733–3742.

[164] Han Zhang et al. “Cross-modal contrastive learning for text-to-image generation”. In: *arXiv Preprint arXiv: 2101.04702* (2021).

[165] Zhuowan Li et al. “Context-aware group captioning via self-attention and contrastive features”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 3440–3450.

[166] Xin Huang and Yuxin Peng. “Cross-modal deep metric learning with multitask regularization”. In: *2017 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE. 2017, pp. 943–948.

[167] Ashish Vaswani et al. “Attention is all you need”. In: *arXiv Preprint arXiv: 1706.03762* (2017).

[168] Anna Rohrbach et al. “Grounding of textual phrases in images by reconstruction”. In: *European Conference on Computer Vision*. Springer. 2016, pp. 817–834. ISBN: 9783319464473. DOI: 10.1007/978-3-319-46448-0_49.

[169] Licheng Yu et al. “Modeling context in referring expressions”. In: *European Conference on Computer Vision*. Springer. 2016, pp. 69–85.

[170] Junhua Mao et al. “Generation and comprehension of unambiguous object descriptions”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016, pp. 11–20.

SUMMARY

Humans perceive the real world through their sensory organs: vision, taste, hearing, smell, and touch. In terms of information we consider this different modes also referred to as different channels of information or modals. Considering multiple channels of information at the same time, is referred to as multimodal and the input as multimedia. By their very nature, multimedia data are complex and often involve intertwined instances of different kinds of information. We can leverage this multimodal perspective to extract meaning and understanding of the world. This is comparable to how our brain processes these multiple channels, we learn how to combine and extract meaningfull information from it. In this thesis the learning is done by computer programs and smart algorithms. This is referred to as artificial intelligence. To that end, in this thesis, we have studied multimedia information, with a focus on vision and language information representation for semantic mapping. The aims of the semantic mapping learning in this thesis are: (1) visually supervised word embedding learning; (2) fine-grained label learning for vision representation; (3) kernel-based transformation for image and text association; (4) visual representation learning via a cross-modal contrastive learning framework.

We first address the task of improving the representation learning for the textual representation of relation words based on visual supervision. In our work, we propose the VS-Word2Vec model to learn the vector representation of relation words by jointly compute over the visual modality and natural language. This can reduce the semantic gap between vision and text. Our method can compute the visual features based on deep networks over an image patch that reflects a relation word, and then achieve the visual similarity matrix for all relation words. Based on our embedding method, the user can achieve the relation word embedding which has a more accurate similarity of words than the original CBOW model.

In an image, the object can be recognized and labeled as category label, or as subcategory of that label; if we consider the subcategory that is referred to as fine-grained label learning. For vision representation it aims to answer the question of how to learn the fine-grained object labels in object detection with the help of auxiliary information attached to images. In this thesis, we propose a novel approach called label inference curriculum network (LICN) to the problem of fine-grained object label learning with a weak supervision of captions. This method can build a mapping based on the correspondences between the coarse category labels provided by public datasets and the fine-grained category labels extracted from captions based on the combination of embedding techniques and knowledge bases. By this semantic map, the user can mark the object with fine-grained labels and learn a detector about the objects. Experimental results obtained with

public datasets as well as our constructed datasets demonstrate the effectiveness of our approach and show that it is helpful to structure the training process by ranking from easy to hard samples. This approach is known as the framework of curriculum learning.

Another approach that we have probed is the kernel-based transformation for image and text associations. This aims to build a probabilistic mixture model, called KMM, for modeling the semantic correlation between web images and text. A KMM was built based on the assumption that the relationship between different modalities follows multiple basic transformations, each working over a local region described by a neighborhood model in the input space. Our model can address the nonlinearity of the data distribution and cross-modal mapping via kernel-based theory. As a solution for the nonlinearity transformation for cross-modal semantic mapping, our model addresses the complexity of the semantic distribution over the global input space, and its continuity at the local scale.

Finally, in this thesis, we probed the visual representation learning via the cross-modal contrastive learning framework. This approach aims to find a method for cross-modal mapping based on weak supervision. Weakly supervised phrase grounding intends to map the phrases in an image caption to the objects appearing in the image under the supervision of image-caption correspondences. We have proposed a novel weakly supervised approach to phrase grounding under the supervision of the correspondence between images and captions. Our key contribution lies in systematically learning contextualized visual representations with a mixed contrastive loss function. Overall, our model achieves state-of-the-art accuracy on the MS COCO and Flickr30K Entities test set.

SAMENVATTING

De mens neemt de wereld waar met zijn zintuigen, te weten: zicht, smaak, gehoor, geur en aanraking. In termen van informatie overdracht beschouwen we deze zintuiglijke waarnemingen als verschillende informatiekanalen of *modalities*. Wanneer we meerdere informatiekanalen tegelijk beschouwen dan spreken we van multi-modale overdracht en de invoer is bekend als multimedia. Multimedia-data zijn van nature complex en bevatten verschillende soorten informatie die met elkaar verweven zijn. We kunnen dit multimodale perspectief gebruiken om betekenis en begrip toe te kennen aan data. Dit is vergelijkbaar met de verwerking van informatie in het menselijke brein, we leren hoe we waarnemingen kunnen combineren en daar zinnige informatie uit halen. In dit proefschrift wordt het leren gedaan met computers en slimme algoritmen. Dit wordt aangeduid met *kunstmatige intelligentie*. Vanuit dat perspectief hebben we, in dit proefschrift, multimedia informatie bestudeerd, met een focus op beeld en tekst voor semantische *mapping*. De doelstellingen van het leren van deze semantische mappings zijn: (1) het leren van word embeddings via afbeeldingen van objecten en relaties; (2) het fijnmazig labellen van objecten in afbeeldingen; (3) *kernel-based data transformation* voor beeld- en tekstassociatie; (4) het leren van beelrepresentaties via een cross-modaal *contrastive learning* framework.

Het eerste doel was het verbeteren van het leren voor de tekstuele representaties van relatiewoorden op basis van visuele supervisie van afbeeldingen. In ons werk stellen we het VS-Word2Vec-model voor om de vectorrepresentatie van relatiewoorden te leren door tegelijkertijd de visuele modaliteit en natuurlijke taal te analysieren. Dit kan de semantische kloof tussen beeld en tekst verkleinen in het vinden van multi-modale informatie. Onze methode kan de visuele *features* van objecten berekenen op basis van analyse van de delen van afbeeldingen die relatiewoorden representeren, en vervolgens de visuele overeenkomstmatrix voor alle relatiewoorden berekenen. Op basis van onze *embedding*-methode krijgen de embeddings van relatiewoorden een betere representatie dan met het originele CBOW-model.

In een beeld kunnen objecten worden herkend en gelabeld als een categorie label, of in een subcategorie van dat label. Als we de subcategorie beschouwen dan spreken we van het aanleren van fijnmazige labels. Het tweede doel was om de vraag te beantwoorden hoe fijnmazige objectlabels bij objectdetectie kunnen worden aangeleerd met behulp van aanvullende informatie gekoppeld aan afbeeldingen. In dit proefschrift stellen we een nieuwe methode voor, genaamd *label inference curriculum network* (LICN), voor het probleem van het fijnmazig leren van objectlabels met *weak supervision* van bijschriften bij afbeeldingen. Met deze methode kan een mapping worden gemaakt op basis van de overeenkomst

tussen de grove categorielabels uit openbare data-sets en de fijnmazige categorielabels die worden geëxtraheerd uit bijschriften. We gebruiken hiervoor een combinatie van embeddingstechnieken en databases. Door deze semantische mapping kunnen objecten beter geïdentificeerd worden met fijnmazige labels en via die labels kunnen betere detectors voor die objecten ontwikkeld worden. Experimentele resultaten op openbare datasets en onze geconstrueerde datasets demonstreren de effectiviteit van onze aanpak en laten zien dat het nuttig is om het trainingsproces te structureren in de volgorde van eenvoudige voorbeelden naar moeilijke voorbeelden. Deze aanpak is bekend als tet zogenaamde curriculum learning framework.

Een volgende aanpak die we hebben bestudeerd is de *kernel-based data transformation* voor beeld- en tekstassociatie. Dit heeft tot doel een probabilistisch *mixture model*, het KMM, te tranen voor het modelleren van de semantische correlatie tussen web-afbeeldingen en tekst. Het KMM leert op basis van de veronderstelling dat de relatie tussen verschillende modaliteiten meerdere basistransformaties volgt, die op een bepaald deel van de afbeelding van toepassing zijn, te weten de regio. Die regio wordt gerepresenteerd door een *neighborhood model* in de vector ruimte. Ons model geeft een oplossing voor de nonlineariteit van de datadistributie en cross-modale mapping via een op kernels gebaseerde theorie. Als een oplossing voor de niet-lineariteitstransformatie voor cross-modale semantische mapping richt ons model zich op de complexiteit van de semantische distributie over de input ruimte, en daarbij de continuïteit ervan op lokale schaal.

Tenslotte hebben we in dit proefschrift het leren van visuele representaties via het cross-modale *contrastive learning* framework bestudeerd. Deze aanpak heeft tot doel een methode te vinden voor cross-modal mapping op basis van *weak supervision* van de bijschriften van afbeeldingen. Het doel is om frases uit afbeeldingsbijschriften te koppelen aan de objecten in de afbeelding. In dit proefschrift hebben we een nieuwe benadering van deze *weakly supervised phrase grounding* voorgesteld op basis van de correspondentie tussen afbeeldingen en bijschriften. Onze belangrijkste bijdrage ligt in het systematisch leren van gecontextualiseerde visuele representaties met een *mixed contrastive loss function.* Met ons model zijn we in staat de best mogelijke nauwkeurigheid te realiseren op de MS COCO- en Flickr30K Entities-testsets.

CURRICULUM VITAE

Xue Wang was born in Qiqihar, Heilongjiang, China in 1989. In 2008, she graduated from High School and started her bachelor study in Information and Computing Sciences at Qiqihar University, Qiqihar, Heilongjiang, China. She finished the four-year bachelor courses and got her bachelor degree of science in 2012. In 2013, she began her master study of statistics at Xinjiang University of Finance and Economics. In 2016, she obtained her master degree of science statistics. In the same year she started her PhD study at Xian Jiaotong University. In 2018, she obtained a grant to participate in the joined PhD trajectory of Leiden University and Xian Jiaotong University. In this she was supported by Graduated School of Xi'an Jiaotong University Scholarship for 6 months to visit Leiden University in 2018. In 2019, she was supported by the Chinese Scholarship Council for a 2-year PhD study in Leiden University in the Netherlands. During the PhD, she investigated cross-modal information extraction and matching based on the same semantic. In Leiden she was embedded in the Computational Bio-imaging group as well as in the data science group supervised by Prof. Fons Verbeek and Dr. Suzan Verberne respectively. From Xi'an Jiaotong University supervision was given by Dr. Youtian Du. The techniques she has used for her research include vision feature extraction, object detection, word embedding, algorithm design, statistical modelling and data visualization. She likes thinking ahead and providing predictive and analytical models for projects in different fields.

ACKNOWLEDGEMENTS

Three years ago, the journey of being a Ph.D. candidate and study at Leiden University started. It was full of excitement and passion. For me, the first challenge was getting comfortable with the language and how to arrange life in this new country. Friends and supervisors give me a lot of encouragement to fit in with the new life. I would, therefore, like to express my sincere gratitude to my supervisors, colleagues, friends and families.

I would like to thank my daily supervisor Suzan Verberne for her guidance, discussions and comments in the research process. She is like a best friend who can always make me feel free to explore, and at the same time encourage and support me in overcoming the difficulties when in trouble. I would like to thank my promoter Fons Verbeek for his guidance, support and comments in vision analysis. He is like my parents who can care about me and support me pick up challenges. I would also like to thank my supervisor Youtian Du from Xi'an Jiaotong University for supporting and guiding me to step into the research in cross-modal information analysis and learning. As my supervisor at the onset of PhD, he continuously inspired me to start my research with big efforts. Furthermore, I would like to thank Marloes and Marcello for their help and care about my studies at Leiden University.

Many thanks to my office colleagues: Anne, Antonio, Daniela, Gerrit-Jan and Yuchen. These two years with COVID19 were a special time with working mostly from home. Nevertheless, at each encounter you always gave me a warm feeling, even although we spend little time working together in our mutual office. I would like to thank all my colleagues in each group. For data science group, Prajit, Anne, Daniela and Hugo you give me a lot of help when I first arrived in Leiden. For bio-image group: Chen, Yi, Feibo, Jia, Danyi, Xiaoqin, Erick, Lu, Sacha, Solomiia, Shima, Mehrdad, Katy, Mariam, Leon, Irene and Rohola; thanks for being around with me. In the XJTU research group: Minhua ZhangYujie Xie and Hang Wang, for sharing so many academic ideas and technical methods in our discussions.

In addition, I would like to thank all the colleagues and friends for your companionship including Wenjing, Yuchen, Zhuoyi, Zhengyang, Jia, Chen, Ruochen, Yu, Jiadong, Anne, Daniela, Erick, Haiyang Wu, Hongfeng Niu, and Qiaolan Fan. The daily life with you brought memorable moments to my life. Special thanks go to my friend for sharing time with me and encouraging me when I felt lonely. With pleasure I also would like to thank my long-term friends, Lianyin Jiang, Haiwan,

Hui, Yue and Jinlong, you give me a lot of support in solving problems in China.

The great gratitude goes to thank my father Yanbin Wang and my mother Jing Jia. Thank you for your endless support, encouragements and unconditional love. Every time when I think of you, I always gain new strength to find joy in life.

Xue Wang
May 2022