Multi modal representation learning and cross-modal semantic matching

Wang, X.

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-

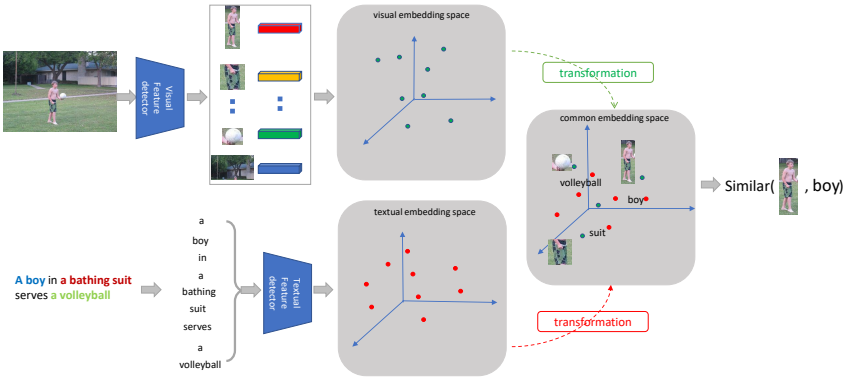

Figure 1.1: Cross-modal semantic matching is the task of matching between a textual description and an image. (a). Some methods learn different modality representations in embedding space respectively. The semantic matching performance is capped by the quality of the different modality representations. (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.

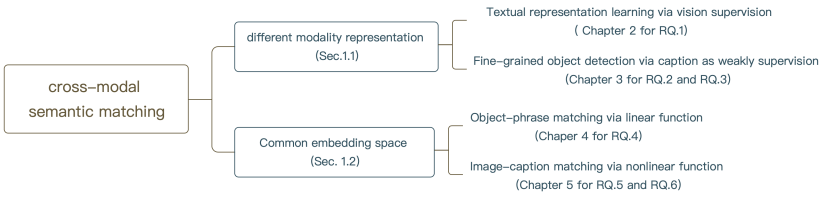

Figure 1.2: The structure of this thesis.

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.