Multi modal representation learning and cross-modal semantic matching

Wang, X.

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.