Multi modal representation learning and cross-modal semantic matching

Wang, X.

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.