Exploring images with deep learning for classification, retrieval and synthesis

Liu, Y.

Chapter 9

Conclusions

In this thesis, we have devoted previous seven research chapters to address the eight research questions regarding three themes: classification, retrieval and synthesis. In this chapter, we derive main findings from our approaches and results. In addition, we discuss limitations of our approaches and possible solutions to address them. Lastly, we point out several directions for future work.

9.1 Main Findings

In each research chapter, we have proposed a new approach to answer the corresponding research question. In the next, we will conclude these approaches and present main findings inspired by experimental results and empirical analysis.

(1) We began the research part in **Chapter 2** by focusing on exploiting deep fusion networks for classification. We built a novel deep fusion architecture (*i.e.* CFN) on top of plain CNNs, and witnessed its effectiveness for diverse tasks ranging from image-level to pixel-level classification. In addition, it is promising to apply CFN to more applications such as object detection and visual tracking.

(2) In **Chapter 3** we further exploited CNNs to improve its robustness for edge detection. In contrast to using a general supervision, we proposed to develop relaxed deep supervision (RDS) to guide different intermediate layers. We observed that hierarchical supervisory signals with additional relaxed labels could be consistent with the diversities in different layers. We believe that it is feasible to adapt RDS to other pixel-level predictions, such as image segmentation and saliency detection.

(3) After investigating the classification theme, we then turned to address the questions about the retrieval theme in Chapters 4-7. In **Chapter 4**, we provided a good attempt to incorporate deep features into the inverted index scheme and exploited a novel DeepIndex framework for accurate and efficient image retrieval. In addition, we extended DeepIndex by integrating different deep features and built a 2-D DeepIndex structure that consists of two kinds of variants: intra-CNN and inter-CNN. We found that, Intra-CNN was simpler to build than Inter-CNN, but Inter-CNN could be viewed as a solution to bridge the gap between mid-level and high-level deep feature representations.

(4) Driven by the increasing popularity of large-scale multi-media data, we began to study the cross-modal retrieval task in **Chapter 5**. Specifically, we developed a deep matching network using recurrent residual fusion (RRF) as building blocks for improving visual-textual embeddings. Our work showed that RRF could recurrently improve feature embeddings while retaining the number of network parameters. In addition, the fusion module was efficient to integrate intermediate outputs during the recurrent stage. Potentially, RRF-Net would be seamlessly integrated into other multi-modal applications like image captioning and visual question answering.

(5) In **Chapter 6**, we proposed cycle-consistent embeddings in an image-text matching network, which could incorporate both inter-modal correlations and intra-modal consistency for learning robust visual and textual embeddings. During training, we integrated several ranking losses jointly to optimize the whole embedding learning. For a robust inference, we further leveraged two late-fusion approaches to integrate the matching scores of multiple embedding features. From the experimental results, we showed that cycle-consistency embeddings could effectively promote the cross-modal retrieval performance, compared to a single embedding.

(6) In an effort to accomplish both classification and retrieval, in **Chapter 7** we exploited a unified network for joint multi-modal matching and classification (MMC-Net). The experimental results demonstrated the robustness and effectiveness of the MMC-Net model, compared to the baseline models. On the one hand, the classification component was beneficial to alleviate the biased annotations, so that the model could learn more robust embedding features. On the other hand, the matching component was able to bridge the modality gap between vision and language, and thus combining visual and textual embedding features could produce a more discriminative multi-modal representation.

(7) After focusing on the classification and retrieval themes, our attention moved to the synthesis theme. In **Chapter 8**, we focused on addressing two research questions. The first one was what factors would affect the performance of generative models on the translation tasks. To answer this question, we extended the vanilla CycleGAN with new improvements and showed two extended models. First, we found that the long cycle could leverage more generators to further increase the generation abilities of the model and improve the quality of synthesized images. In addition, the additional inner cycles were able to directly connect the intermediate generators and provided more cycle-consistency losses to constrain the translation. The findings in this work could help in designing other cycle-consistent generative networks for solving image-to-image translation tasks.

(8) The second question we considered in **Chapter 8** was how we can exploit a generative model to transfer the fashion style between two person images. To this end, we interpreted the clothing swapping as a problem of pose-based person image generation and proposed a novel multi-stage generative framework (SwapGAN) to fulfill the clothing swapping from the condition person image to the reference one. The whole SwapGAN framework could be end-to-end trained with both adversarial loss and mask-consistency loss. Our work could be a benchmark study and help to drive future research on this task.

9.2 Limitations and Possible Solutions

Our methods in this thesis have addressed the eight research questions and achieved promising results in terms of the three research themes. However, they still have some limitations which can be discussed from the following three perspectives.

Algorithmic perspective

In Chapter 2, the proposed CFN uses a 1×1 kernel filter in the locally-connected fusion module. It can independently consider each spatial location over the feature maps, while may omit the relationships between different spatial locations. To solve it, a potential solution is to utilize larger kernel sizes such as 1×2 and 1×3, which can incorporate the contextual information in the feature maps. In addition, the adaptive weights learned in the fusion module are the same for all the images. An alternative is to learn dynamical weights conditioned on different input images. For example, Brabandere *et al.* [146] propose a Dynamic Filter Network (DFN), where filters are dynamically generated conditioned on an input image.

In Chapter 4, DeepIndex is designed for accurate and efficient retrieval, however, we can find its performance gap with recent state-of-the-art approaches [48]. It is straightforward to improve our results by using more powerful CNNs like ResNet-152. Besides, it is suggestive to extend multiple DeepIndex with three or more deep features, compared to the 2-D case.

In Chapter 8, the extended CycleGAN models, *i.e.* Long CycleGAN and Nest CycleGAN, can improve the generated quality, however, they will increase the training cost due to using more generators. One promising solution is to introduce a weight-sharing mechanism to avoid increasing the cost. In terms of the proposed SwapGAN for person-to-person clothing swapping, it is hard to preserve rich color and texture information in the clothes. This problem may be caused by the limited capability of the original adversarial loss. To overcome it, we can make use of additional losses (*e.g.* perception loss [80]) to help enhance the synthesis process. However, they will increase the memory cost and training time.

Theoretical perspective

In Chapter 3, we have discussed our motivation for exploiting relaxed deep supervision for robust edge detection. Nevertheless, we should still realize that it still lacks of theoretical insights into interpreting the benefit of diverse supervision for training deep neural networks. Recent works [253, 254] propose theoretical approaches to interpreting deep visual representations learned in CNNs. It is encouraged to use these approaches to achieve deeper insights regarding the utility of diverse supervision.

In Chapter 5, we develop a building block based on recurrent residual fusion (RRF) to advance the visual-textual embedding features. We notice that, using more recurrent steps may decrease the performance. One reason is attributed to the potential over-fitting issue while training the model, however, it is hard to prove it in theory. This issue limits further performance improvements. One alternative is to impose the RRF block on more layers, since RRF is a general structure that can potentially be applied to many existing layers in a deep network.

Practical perspective

In Chapter 6, we apply the proposed CycleMatch to solve the task of cross-modal retrieval between images and texts. Although we witness its promising performance for this task, it is encouraged to transfer our method to other challenging tasks, like visual grounding, visual relationship detection and visual reasoning. In addition to the global image-text matching, we should take into account local similarities between visual regions and phases.

In Chapter 7, the proposed MMC-Net, which can jointly accomplish multi-modal matching and classification, requires ground-truth class labels in addition to the paired information. However, some multi-modal datasets (*i.e.* Flickr30K) do not provide the class labels. Therefore, it is infeasible to train the full MMC-Net model. One potential alternative is to automatically construct a dictionary by parsing all the textual descriptions. Then we can label each image with its key words derived from the dictionary. In this way, it is still feasible to accomplish the classification task based on the word-level labels instead of unavailable class labels.

9.3 Future Research Directions

In the previous seven chapters, we have presented many methods to address the research questions regarding the three research themes. A wide variety of future research is also encouraged to advance these themes. In this section, we briefly discuss future research directions regarding each theme.

Zero-shot classification

Zero-shot classification (ZSC) [255] aims to solve the task where not all the classes are represented in the training set. In ZSC, the training and test class sets are disjoint. It needs to learn a visual classifier based on the seen images and their semantic categories, and then transfers the classify to recognize images of unseen classes. Existing approaches can be summarized in three groups. (1) Direct mapping: learning a mapping function from visual features to semantic representations.

(2) Common space learning: constructing a common embedding space where visual features and semantic representations can be correlated. (3) Model parameter transfer: exploiting the inter-class relationship between seen and unseen classes and then transferring the model parameters of seen classes to the unseen ones.

In recent years, deep neural networks have been widely used for solving the ZSC task due to their powerful representation capabilities [60, 75, 256]. Nevertheless, this task remains challenging in discovering the relations between visual features and semantic knowledge, as well as generalizing the relations to unseen classes. Since ZSC relies on discovering the sematic relations between visual and textual features, it is encouraged to incorporate a visual-textual matching component into a ZSC system. Our research on classification and retrieval is related to this future direction.

Generation for cross-modal retrieval

Recall that cross-modal retrieval needs to overcome the semantic gap between two different modalities like vision and language. To achieve it, one common approach is to project visual and textual features into the same embedding space where we need to compare their correlations. However, in most existing datasets, each matched image-text pair has limited samples, for example one image is labeled with one or five descriptions. This issue will hinder the learning capabilities of deep neural networks. Recently, Zheng *et al.* [257] propose to use generation networks to produce more image samples to extend the datasets. Driven by this idea, it is feasible to use GANs to alleviate the lack of image-text samples for cross-modal retrieval. For example, we can generate more realistic-looking images based on the text description, and also create additional descriptions for each image. In addition to cross-modal relations, we can add intra-modal constraints between the real and generated samples. Integrating both cross-modal and intra-modal matching could be beneficial to learn better embedding features. Our research on retrieval and synthesis can be adopted to this future direction.

Unified image synthesis

Recent studies on image-to-image translation have achieved encouraging results for a range of different domain-specific image sets. However, most of existing approaches are inefficient for jointly modeling multi-domain image translation tasks, because they need to train individual generative networks for every two domains, *i.e.* , in order to learn all mappings among N domains, $N \times (N-1)$ generators need to be learned. To address this problem, StarGAN [258] recently proposes a unified generative adversarial network, which allows to translate a range of image domains by using a single generative network. The key point in StarGAN is that it uses

a label (*e.g.* binary or one-hot vector) to represent the domain information. In addition, StarGAN can incorporate different labels from multiple datasets using a simple mask vector to indicate the dataset information. However, one potential issue may make StarGAN fail when different datasets have some overlapped labels. One potential solution is that, we can extend the mask vector to be consistent with the number of domains, rather than with the number of datasets. In this way, StarGAN can discard overlapped domain labels in different datasets. We believe that exploring a unified image generative network is still a promising future work.