Deep learning for visual understanding

Guo, Y.; Guo Y.

Chapter 7

Conclusions

7.1 Conclusions

In this thesis, we explored and designed deep learning algorithms for better image understanding. The topic of image understanding has long been an active research field, and it aims to visualize and understand the image content in a way that is consistent with human perception. To this end, there are many related tasks, such as image classification, object detection, image retrieval and image captioning to name a few. While all these tasks may seem disjoint, developing a good image representation is essential for all of them.

In Chapter 2, we presented a comprehensive review of the developments of various deep learning algorithms. This chapter is intended to be useful for general neural computing, computer vision and multimedia researchers who are interested in the state-of-the-art in deep learning in computer vision. Generally, the deep learning methods can be divided into four categories according to the basic method they are derived from: Convolutional Neural Networks (CNN), Restricted Boltzmann Machines (RBM), Autoencoder and Sparse Coding. Among these four categories, CNN is the most commonly used for the computer vision area and also the basis of the work in this thesis.

Chapter 3 presents an effective scheme to achieve image features (referred to

as PPC). PPC is derived from the fully-connected CNN activations (referred to as CNN). It aims to learn more information from the image and proposes to extract CNN features from multiple spatial sub-regions, and aggregates the multiple CNN together. Without increasing the complexity during the test phase, the feature is further reduced to the same dimension with CNN (i.e. 4096-D) using PCA. Although it is straightforward to achieve, PPC consistantly delivers better performance than the commonly-used CNN, and has the potential to be useful for many tasks.

Initially, researchers have focused on employing CNN activations based on the fully-connected layers. However, current research studies are giving increased attention to the convolutional layers, since they can preserve the spatial information and contain rich semantic information. A common usage of the convolutional activations is to encode them with the Bag-of-Words (BoW) variants, such as VLAD and Fisher Vector. This pipeline not only preserves high discrimination of the CNN activations, but also incorporates the 'bag' conception to improve the invariance property to scale changes, location changes and occlusions. Motivated by this pipeline, Chapter 4 proposes a novel method to incorporate the CNN feature with the BoW framework. In contrast to the common practice, we do not explicitly generate the codebook, and extensively assign the features to the generated visual words according to the similarity. Instead, we take the feature maps as the 'surrogate' pars, and take the activation values as the assignment strengths for these surrogate parts. As a consequence, our novel feature, i.e. BoSP, is much easier to compute, and has a significantly lower dimension than the common usage of 'CNN+BoW'.

Aside from the traditional image classification task, Chapter 5 suggests addressing the hierarchical image classification task, by incorporating the Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN). This task is intended to generate multiple image labels in a coarse-to-fine pattern, and thus can provide a better understanding of the categories, especially the fine-grained categories. In addition to addressing the hierarchical image classification task, the CNN-RNN paradigm also has the following potential advantages: (1) It can

improve the traditional leaf-level classification performance by exploiting the relationship between hierarchical labels; (2) It can be built on top of any CNN architecture which is primarily designed for leaf-level classification. Accordingly, we built a high-performance baseline network, i.e. wider-ResNet, based on which the CNN-RNN paradigm achieves remarkably better performance than the state-of-the-art on CIFAR-100. (3) It can enhance the image classification performance when part of the training data is only annotated with coarse labels. This provides a promising direction for weakly supervised learning.

The application of CNN-RNN paradigm is not limited to the image classification task. A more common usage is for the image captioning task. As is similar with the image classification, the captioning performance is also closely dependent on the discriminative capacity of CNNs. In Chapter 6, we investigate the effects of different Convnets on image captioning, i.e. single-label Convnets, multi-label Convnets and multi-attribute Convnets. Since the three Convnets focus on different visual contents in one image, we propose aggregating them together to generate a richer visual representation.

7.2 Research Limitations

Although our research has reached its aims, we cannot neglect its limitations and weaknesses.

First, from the general point of view, deep learning is often considered as a black box. It can generate relevant results for the given input, but it is not clear what the final learned network means - under what conditions will it work correctly. In addition, the designing and training processes of the deep neural networks may be sensitive: some small architectural and optimization differences may lead to substantial variance in the final result.

Second, it is not clear how well the algorithms and architectures generalize to images from other domains, such as biological images and medical images, since they are established based on the off-the-shelf models which are pretrained on the ImageNet dataset, and ImageNet consists primarily of accurately annotated

natural images. Moreover, we only evaluated our algorithms on the relatively clean benchmarked datasets, and therefore it is difficult to predict how well the methods will work on imagery containing more complexity and noise.

Third, the designing of some algorithms in this thesis has limited theoretic foundation. Take the proposed BoSP feature as an example, it was inspired by the intuition of attempting to use bag-of-words approach on the feature maps from the learned network because bag-of-words had significant success in image understanding based on salient features. However, there was no guarantee that it would work well, nor is there strong theory which would predict the weaknesses of the learned network.

7.3 Future Work

In the future, we will extend our work in the following directions:

Fusing hand-crafted and deep learned features for image representation: The hand-crafted feature can be seen as a particular form that a human designer thinks can represent the images well. Before the surge of CNN, hand-crafted feature has long been a key component in the competition-winning systems for visual understanding. In the future, we would like to employ the idea of hand-craft features to design the deep networks, in order to make the networks focus more on the important areas.

Image captioning with grammar supervision: Image captioning is a new emerging research area which can describe the image with more informative contents, including the objects, actions, relations and etc. In order to generate novel sentences for unseen scenes, most of the current works employ the generative approaches, such as Baby Talk [301] and LRCN [266]. These approaches generate the words one-by-one, and as a consequence, the whole generated sentences may be oddly organised. To obtain sentences that are more consistent with our language, in our future work, we propose to provide grammar supervision during the training of the network.

Designing more comprehensive CNN models: CNN models have achieved significant success in various computer vision tasks, including image classification, object detection, image retrieval, image captioning, to name a few. These seemingly disjoint tasks do have some fundamental similarities. For example, Liu et al. [318] proposed to utilize the segmentation annotations to help the edge detection. Oquab et al. [37] took advantage of the object location to improve the image classification performance. While most of these works end up with task-specific CNN models, we assume that the 'real' artificial intelligence should be capable of tackling a broad set of computer vision problems. Therefore, in our future work, we want to exploit the synergy between different visual tasks, and design a universal network that can solve multiple tasks,