Deep learning for visual understanding

Guo, Y.; Guo Y.

Chapter 1

Introduction

1.1 Background

It is now the Age of Information. Each day, users produce enormous amounts of data by modern commonplace technologies. As a result, the amount of information is increasing exponentially, especially the multimedia information. Most of this information is stored digitally and available to the public. For example, it is reported that the Facebook users have uploaded over 250 billion photos, and are uploading 350 million new photos each day. Such a large amount of available data is a double-edged sword in our lives. On the one hand, if we can properly handle and analyze the data, we can have more alternatives for our queries. On the other hand, it is easy for us to get lost in the sea of data. Without computer-aided programs, it may take centuries for us to sift through the information and find what we want.

While search engines like Google and Yahoo can perform the textual analysis quite well, it is still challenging to fully exploit the visual content due to the well-known semantic gap. The key to bridging this gap is to develop or learn highly discriminative features to represent the images.

Generally, the development of image representation can be divided into three stages: In the first stage, images are described with low-level global features, such as color histograms, contour representations, shape descriptors, and texture features. These features represent the whole image with a single vector, and can capture the global image appearance well. However, they are sensitive to occlusion and clutter.

To incorporate more local context and obtain a more informative description, various local features are developed, such as SIFT [1], SURF [2], HoG [3], etc. The local features are descriptors of local image neighborhoods computed at multiple key points. Compared with global features, local features are more robust to image translation and occlusion. To aggregate spatial local features into a global image representation, these features are often encoded with Bag of Visual Words (BoW) [4], or its variants, such as VLAD [5] or Fisher Vector(FV) [6].

Both the traditional global features and local features are hand-crafted features, which often require expensive human labor and do not generalize well. Recent studies have shown that there are no universally best hand-crafted features for all datasets, and it would be more advantageous to learn features directly from the raw data [7]. Since 2006, deep learning has emerged as a new area of machine learning research [8], and it introduces the concept of end-to-end learning which means transformation from pixel level to real application. Deep learning algorithms typically attempt to distill high-level abstractions in data by utilizing hierarchical architectures. The output of each intermediate layer can be viewed as a representation of the original input data. Several deep representations have been repeatedly verified to be highly discriminative and achieved top tier performance on various benchmark datasets and international contests.

Due to advances in deep image representations, numerous breakthroughs have been made in diverse computer vision applications. For example, for the most intuitive and extensively studied task, image classification, many deep learning algorithms have reached comparable performance relative to the human performance on the large scale ImageNet dataset [9]. Aside from solely discovering the objects, there are some new emerging applications (e.g. image captioning, visual question answering) which aim to exploit more information (e.g. action, relation and etc) based on the deep image representation, and also achieved competitive results with the human performance [10].

1.2 Research Goals and Contributions

The main purpose of our work is to develop new algorithms which can improve the understanding of images. To fulfill this, we focus on two visual applications: image classification and image captioning.

Image classification aims to classify images into pre-defined categories, and helps people to know what objects the images contain. In the first part of this thesis, we propose new features which can improve the performance without significantly increasing the computational cost. Therefore, they may be utilized in many other applications.

The second part of the thesis proposes to address the hierarchical image classification task, which can generate multiple hierarchical labels in a coarse-to-fine pattern. By providing the evolution of the image categories, this task can better describe what the categories are, especially for the fine-grained categories.

For the third part, we investigate a more challenging and new emerging task, i.e. image captioning, which attempts to generate a sentence to describe the image. In contrast to image classification which only detects the existence of an object, the sentence generated by image captioning may also contain the action, relation and etc.

1.3 Thesis Overview

This thesis is based on articles where I have been a primary author that have been published or are currently under consideration at respected journals and conference proceedings. The following provides a brief description of each chapter.

Chapter 2 presents a survey which reviews about 200 papers published between 2010 and 2016 in the area of deep learning for visual understanding. The survey provides a comprehensive background for this research area, including the development of the relevant methods, the applications, and the directions that the field is moving towards. This survey has been published by :

- Neurocomputing (journal)

Chapter 3 introduces an effective and straightforward feature, called Principal Pyramidal Convolution (PPC). This feature is derived from the commonly-used CNN feature (i.e. the fully-connected activation), and demonstrates superior performance than the baseline for different datasets and different dimensions. This work has been published in the conference proceeding:

- 16th Pacific-Rim Conference on Multimedia (PCM2015) in Gwangju, Korea.

Chapter 4 presents a new feature called Bag of Surrogate Parts (BoSP). This feature is motivated by the well-known Bag of Words (BoW) scheme, and aims to integrate the advantages of CNN and BoW (i.e. high discrimination for CNN, and scale/position/occlusion invariance for BoW). Together with the feature, several enhancements are also proposed, including spatial aggregation, scale pooling and global-part prediction. An early version of this work was presented at:

- 27th British Machine Vision Conference (BMVC2016) in York, UK.

Chapter 5 aims to give better understanding of the objects by tracing how the semantic categories evolve, and utilizes the CNN-RNN framework to fulfill the hierarchical image classification task. This framework can not only generate hierarchical labels for images, but also improve the traditional leaf-level classification performance by incorporating the relationship between hierarchical labels. In addition, we also investigate how we can utilize the framework to benefit the classification when a fraction of the training data is coarse-labeled. This work has been submitted to:

- Multimedia Tools and Applications (journal)

Chapter 6 focuses on a new emerging research area, i.e. image captioning, and investigates the effects of different Convnets. To obtain a richer visual representation, we propose aggregating their activations and achieve promising performance. This work has been published in the conference proceeding:

- 23rd International Conference on Multimedia Modeling (MMM2017) in Reykjavik, Iceland.

Chapter 7 concludes the thesis and reflects on our future work.

These are the publications which are related to the contents of this thesis:

- **Guo Y.**, Bai L., Lao S., Wu S., and Lew M.S., "A Comparison between Artificial Neural Network and Cascade-Correlation Neural Network in Concept Classification." 15th Pacific Rim Conference on Multimedia, 2014.
- **Guo Y.**, Lao S., Liu Y., Bai L., Liu S., and Lew M.S., "Convolutional Neural Networks Features: Principal Pyramidal Convolution." 16th Pacific Rim Conference on Multimedia, 2015.
- **Guo Y.**, and Lew M.S., "Bag of Surrogate Parts: one inherent feature of deep CNNs." 27th British Machine Vision Conference, 2016.
- Liu Y.*, **Guo Y.***, and Lew M.S., "What Convnets Make for Image Captioning?" 23rd International Conference on Multimedia Modeling, 2017 (* means equal contribution).
- **Guo Y.**, Liu Y., Oerlemans A, Lao S., Wu S., and Lew M.S., "Deep learning for visual understanding: A review." Neurocomputing, vol 187, 2016.
- **Guo Y.**, Liu Y., Lao S., Bakker E.M., Bai L., and Lew M.S., "Bag of Surrogate Parts for Visual Recognition." IEEE Transactions on Multimedia (submitted).
- **Guo Y.**, Liu Y., Bakker E.M., Guo Y., and Lew M.S., "CNN-RNN: A Large-scale Hierarchical Image Classification Framework." Multimedia Tools and Applications (submitted).
- Liu Y., **Guo Y.**, Wu S., and Lew M.S., "DeepIndex for Accurate and Efficient Image Retrieval." Proceedings of the 5th ACM on International Conference on Multimedia Retrieval, 2015.
- Liu Y., **Guo Y.**, and Lew M.S., "On the Exploration of Convolutional Fusion Networks for Visual Recognition." 23rd International Conference on

Multimedia Modeling, 2017 (Best Paper).

- Liu Y., **Guo Y.**, Bakker E.M., and Lew M.S., "Learning a Recurrent Residual Fusion Network for Multimodal Matching." International Conference on Computer Vision, 2017.