Deep learning for visual understanding

Guo, Y.; Guo Y.

English Summary

It has long been the goal of computer vision researchers to develop an algorithm capable of understanding the visual information automatically and accurately. While this seems to be effortless to humans, there is no robust solution to date due to the well-known semantic gap between low-level features and the object or concept that is being modeled. In recent years, deep learning algorithms have been effective in closing this semantic gap due largely to the sophisticated visual representations they developed. This has resulted in major advances in diverse visual applications, such as image classification, object detection, image captioning and etc. The purpose of this thesis is to explore and design new deep learning algorithms for better visual understanding.

First, we present a comprehensive review of recent deep learning advances which targets general neural computing, computer vision and multimedia researchers who are interested in the state-of-the-art in deep learning in computer vision. Next, we establish our research on three visual applications: traditional image classification, hierarchical image classification and image captioning.

The traditional image classification task involves classifying an image into one pre-defined category, and has been widely studied in the computer vision community for decades. We proposed several new features, PPC and BoSP, to address this task. PPC is a straightforward scheme, which extracts and aggregates CNN features from different image regions, and utilizes PCA to reduce the feature dimension. BoSP regards the feature maps as surrogate parts, and proposes to assign the dense image regions to these surrogate parts by observing the activation

values. Both PPC and BoSP can be achieved without significantly increasing the computational cost.

Objects are often organised in a hierarchy. While the traditional image classification task only focuses on the leaf-level categories, we propose that providing an evolution of the image categories can better describe what the categories are. Accordingly, we introduce the hierarchical image classification task, which attempts to generate hierarchical coarse-to-fine labels rather than one leaf label, and develop the CNN-RNN framework to address this task. In this framework, the CNN is used to extract discriminative image features, and the RNN exploits the relationship between the hierarchical categories and generate sequential labels. In addition, we also investigate the effectiveness of utilizing this framework for the traditional image classification task and weakly supervised learning.

Another usage of the CNN-RNN framework is for image captioning, which is an important and challenging task in vision-to-language research. It aims to describe an image with meaningful and sensible sentence-level captions. We investigate the effects of different Convnets on image captioning, i.e. single-label Convnet, multi-label Convnet and multi-attribute Convnet. As these three Convnets focus on different visual contents in the image, we propose aggregating them together for a richer visual representation. Overall, we achieve competitive results with the state-of-the-art.