CS670: Computer Vision - Fall 2017

Proposal: Image Captioning with Dilated CNNs and Attention Model

Members: Ao Liu, Weijie Shi, Zitao Wang **Date:** October 20, 2017

Problem: After deep learning introduced, more and more computer vision (CV) applications have done great jobs on different kinds of vision tasks, so as the natural language processing (NLP) applications. Image is the way that we learn from our world. We as human beings always see stuff in daily life and understand what we see with our brains, or even tell others about that since we were kids. Although some artificial intelligent (AI) and machine learning (ML) applications are already competitive with human performance, there is few that breaks the gap between CV and NLP. We aim to build a simple system that can describe a picture with a sentence of human readable natural language.

Methodology: Convolutional Neural Networks (CNNs) in general works well and fast on CV tasks. Dilated CNNs seems to be faster than traditional CNNs and removes redundant information. We aim to build our system based on this advantage. Also, some previous work show that dilated CNNs also works well on some NLP tasks, so that we may also want to try if it works well on text generation task.

Attention models are broadly used on many tasks. However, there are several alternatives of attention model. We also want to test which one works better for our task.

Dataset: The dataset we are going to use is Common Objects in COntext (COCO). COCO dataset is a large-scale object detection, segmentation, and captioning dataset. COCO has several features: Object segmentation, Recognition in context, Superpixel stuff segmentation, 330K images (>200K labeled), 1.5 million object instances, 80 object categories, 91 stuff categories and 5 captions per image.

Collaboration Plan:

Ao Liu will mainly work on adapting dilated CNNs on language generation of the task.

Weijie Shi will mainly work on choosing the attention model and evaluating the performance.

Zitao Wang will mainly work on improving the performance of vision part of the system.