COMS 4995-6 Final Project Proposal

Yiyang Li/Chenqin Xu/Danning Zheng

2018-03-22

1. Overview

Our project will be using deep learning to do text detecting in natural images. After implemented and trained the network, we may consider further application upon that network.

Our team is made of 3 people. Yiyang Li, Chenqin Xu and Danning Zheng.

2. Preliminaries

Text detection was formerly solved using traditional processing methods, like MSER[1] and SWT[2], which basically focus on the geometric and color features of pictures. They first extract candidate text regions through edge detection or clustering method and then eliminate non-text regions using heuristic rules. In this stage, detection precision was unsatisfactory and the speed was comparatively slow, because of all kinds of text-like structures which were not easy to be distinguished from real text regions by those traditional methods.

With the use of neural network, text detection performances have been greatly improved. R-CNN was applied into text detection in 2014, which firstly gets the candidate regions and then uses cnn to extract features and train classifiers. Later fast R-CNN became popular because of its single-stage training and higher detection quality. In recent years, YOLO[3] has also become popular. It is a unified and fast framework, although it may struggle with small objects that appear in groups.

3. Project Plan

Currently, our goal can be separated into 3 stages:

3.1 Preparing

There're many works have been done on this area. According to a survey at 2015[4], the complexity come from environment and text itself makes this a very tough mission.

However, in recent years, there're several amazing work using different deep learning methods have been deployed to this area and have achieved excellent result.

[1]Matas J, Chum O, Urban M, et al. Robust wide-baseline stereo from maximally stable extremal regions[J]. Image and Vision Computing, 2004, 22(10):761-767.

[2]Epshtein B, Ofek E, Wexler Y. Detecting text in natural scenes with stroke width transform[C] Computer Vision and Pattern Recognition. IEEE, 2010:2963-2970.

[3]Redmon J, Divvala S, Girshick R, et al. You Only Look Once: Unified, Real-Time Object Detection[J]. 2015:779-788.

[4]http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=6945320

We are considering implement our neural network based on one of former works that published on ECCV 2016 by Zhi Tian and Weilin Huang[5]. We will first re-implement their work since they have described their structure in detail witin their paper.

3.2 Implementing basic function–Detecting

As for training data, we are considering to use ICDAR dataset as Tian&Huang did. And since ICDAR dataset has grown, we can expecting a better result from them, or not. Also as a complementation, we can also use Google API to annotate some picture for our special training demands.

Their structure are based on a combination of VGG network and bi-LSTM, which is kind of fancinating and complicated.

As for the framework, we are going to use Tensorflow or Caffe, depends on which we found more efficient for our project.

3.3 Further Application

After that we will have to do some fine tuning work or transfer learning in order to fulfill our own goal. But we think it is more important to reach good enough accuracy first. So we will currently focusing on that.

There can be a lot of application based on a text detecting network. Like transform pictures of notes on the blackboard into some editable pdf files. Or extract text information from wild images. All these thought will depends on how the network can work.

[5]https://arxiv.org/pdf/1609.03605.pdf