Deep Learning Final Project Proposal

Project: Text Categorization

Members: Cong Gao, Zhaohao Fu, Liujiang Yan

1) A definition of your problem

The task of this project is to design a deep neural network to classify topics from text. The input of the model includes news headlines, news articles, tweet comments and other public news or social networking related text files, which can be of different length, but limit to English language. These texts will be labeled to finite set of classes. Our goal is to compare the classification performances under several datasets.

2) some details of how you propose to solve your chosen problem including a breakdown into approaches/components.

There are two general approaches for this problem. One is combining Wording Embedding and Convolutional Neural Network (CNN). Another is applying Recurrent Convolutional Neural Network (RCNN). We propose to empirically explore the performance of the two different approaches on different datasets. We will follow and reimplement some benchmark architectures from previous studies using Pytorch.

3) the datasets you’ve identified

There are several public datasets on text categorization that we could utilize for this project.

A data set of Trump and Clinton comments

* The dataset has binary category: from Trump / from Clinton

A data set of Trump and Clinton tweets

* The dataset has binary category: from Trump / from Clinton

A data set of news headlines from 4 categories

* the category of the news item:
  + -- *b* : business -- *t* : science and technology -- *e* : entertainment -- *m* : health

A data set of complete news articles with many categories:

* 90 classes, 7769 training documents and 3019 testing documents.

Additional dataset will be added during our exploration.

Team member assignments

All the team members Cong Gao, Zhaohao Fu, Liujiang Yan will contribute to the model implementation. We will set up the Pytorch framework together. Cong will focus on the implementation of Wording Embedding, Zhaohao will focus on designing CNN and Liujiang will contribute to the implementation of RCNN. We will make full use of our interdisciplinary background and knowledge set and team members will review and test other’s work.