CS571 Nature Language Processing
Homework 2 Report: Word Segmentation

Zhexiong Liu

1 Problem Description

This project aims to implement a CNN to classify the sentiment of a given sentence. The corpus includes training, validation, and testing dataset, which could be embedded using FastText.

2 Modeling

2.1 Data Preprocessing

'fasttext-50-180614.bin' is used to embed each words in the sentences into 50-dimensional vectors. The longest sentence in the training set includes 61 words, therefore, the longest dimension of the vectors is 61. Regrading this, all the other sentences are automatically augmented a 61-dimensional vectors in training, validation, and testing dataset.

2.2 CNN model 1

The structure of the CNN 1 includes a simple convolutional layer, a max pooling layer, a fully connected layer. convolutional layer includes 128 filters with sieze (4×50). The corresponding max pooling size is (58×1). Fully connected layer has 128 notes.

2.3 CNN model 2

The structure of the CNN includes two simple convolutional layer, two max pooling layer, a fully connected layer. convolutional layer includes 10 filters with sieze (3×25). The corresponding max pooling size is (2×2). Next convolutional layer includes 100 filters with size (3×25), and the corresponding pooling size is (56×1). Fully connected layer has 100 notes.

3 Experiments

In the experiment part, two models are implemented in Pytorch framework. For the first model, the batch size of each epoch is 64, learning reate is 0.01. In each batch, the samples

are scuffled. As can be seen in Figure 1, model 1 arcieves 43.7% accuracy on the validation dataset, and 94% accuracy on the training dataset in 50 epoches. As can be seen in Figure 2, model 2 arcieves 42.6% accuracy on the validation dataset, and 98% accuracy on the training dataset in 50 epoches.

In both Figure 12, the models performs overfitting after 25 epoch. The training accuracy increases while the models fit greatly on the training set. The models achieve similar performance even though they have different structures and parameters.

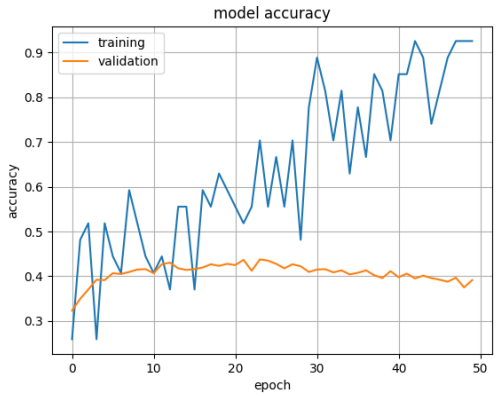

Figure 1: Model 1 accuracy

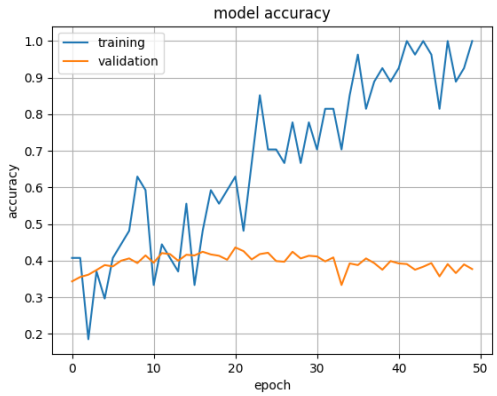

Figure 2: Model 2 accuracy

4 Problems

The sentences are embedded into 61 dimensions, in which each word is 50-dimensional vector. In this case, multiple zero vectors are introduced into the training, validation, and test data, which decrease the model accuracy and speed. Another observation we can obtain in this experiments is that sentiment analysis is difficult to predict.