DSGA1011: Assignment #2

Haonan Tian
ht1151@nyu.edu

New York University — October 31, 2018

Overview

This is the report for DSGA1011 Natural Language Processing Assignment 2.

The project is designed to apply CNN and RNN models to classify the matches of two language sentences which are defined as premise and hypothesis. There are three types of matched of sentence which are treated as 'contradiction', 'entailment' and 'neutral'. All data used in this program comes from the Stanford Natural Language Inference which can be accessed by the following link: https://nlp.stanford.edu/projects/snli/. The definitions of the three types of matches of premise and hypothesis can be found on the website. In this project, these three types of match are treated as labels.

The data is first loaded from the original files and tokenized. In this project, no embedding layer is applied and jointly trained with the RNN or CNN model. Instead, a Fasttext pretrained embedding layer has been directly applied. The pretrained model can be found at https://fasttext.cc/docs/en/english-vectors.html. In this project, the model named 'wiki-news-300d-1M.vec' has been used.

Two different sets of datasets are downloaded and applied in this project. One is the data set which only contains single genre while the other one contains multi genres. The general logic of this project is that the training set of single-genre dataset is used to trained the model and then the validation set for single-genre dataset is used for scoring. Then consider the accuracy on the single-genre data set, the best performed CNN and RNN model will be selected to perform prediction on the multi-genre data set. Since there are multiple genres for the multi-genre data set. The dataset is initially divided to several subsets with each set contains the matched of sentence with the same genre then the prediction is carried out based on these subsets. In other words, we will perform a cross genre prediction for the multi-genre set.

Results

In order to test the performance of the two models, a set of hyper-parameters are modified to compare the performances of the models. For RNN, the hidden size and the interacting methods are modified. For hidden size, the program set the hidden size varies from 100 to 300 dimensions. For the interacting methods, the program has two approaches. The first one is the standard concatenation after the two hidden vectors are return from RNN and then feed the combined vector to the fully connected neural network. The results for RNN is shown in the following table:

For CNN, this program also modifies two hyper-parameters. The first hyper-parameter tuned in this program is the hidden size and the second hyper-parameter is the kernel size of the CNN. The results are list in the following table:

As can be seen from the above result tables, the best model for RNN trained by single genre training set occurs at the combination of 200 dimensional hidden layer with element-wise multiplication. The best

Hidden Size	Interacting method	Accuracy
100	concat	0.637
200	concat	0.652
1000	concat	0.616
100	multiplication	0.649
200	multiplication	0.667
1000	multiplication	0.654

Hidden Size	Kernel Size	Accuracy
100	3	0.677
200	3	0.674
1000	3	0.693
100	2	0.652
200	2	0.661
1000	2	0.673
100	5	0.622
200	5	0.684
1000	5	0.692

Genre	Model	Accuracy
fiction	CNN	0.443
fiction	RNN	0.412
telephone	CNN	0.421
telephone	RNN	0.465
slate	CNN	0.476
slate	RNN	0.452
government	CNN	0.432
government	RNN	0.465
travel	CNN	0.471
travel	RNN	0.421

model for CNN is the one with 1000 dimensional hidden layer and the kernel size is set to 3. Therefore, these two models are selected to perform the prediction on multi-genre validation set. The following table reveals the results from RNN and CNN.

Analysis and Conclusion

Firstly, at the training stage of this program, generally, the performance of these two methods are very close with CNN performs slightly better than RNN. The best prediction accuracies are around 0.7. For CNN, when the hidden size is relatively large, the learning is much faster compared with the case when hidden size is relatively smaller. The program constantly check the prediction accuracy by calling testmodel function. As can be seen when the hidden size is large, the accuracy jumps up very quickly and remain stable or slightly increasing at the rest of the epochs. When the hidden size id small for instance 300, then for the first epoch, there is a very apparent stage of continuously increasing accuracy. For RNN, despite the size of the hidden layer, there is an apparent stage of increasing accuracies. The kernel size seems have small influence on the CNN model. Since the length of the sentences varies from 1 to 80, this program choose to test kernel size 2, 3 and 5. basically for three choices, smaller kernel seems perform little bit better than large kernel.

Secondly, for the results on multi-genre data sets, the accuracies are much lower than the accuracies generated from single-genre validation sets. The probable reasons are that the sentences in certain genres have certain features and using Fasttext pretrained layers may constrain the model from adjusting the embedding layer to catch these features in the certain genres. Besides, since the model can be improved by giving more data. Since the model applies CNN, RNN with a 2 layers fully connected neural network. If more new data is given, the accuracy can go higher. Compared the accuracies across different genres, there is no obvious difference between these accuracies across genres. Compare with the performances across models, it is also hard to tell which model performs absolutely better.

Three correct and incorrect examples are provided as follows: Three correctly labeled examples: 1.The premice: Three cheerful ladies sitting at a table doing a yarn work in a room , at the background are similar groups of ladies doing similar work . The hypothesis: The ladies are discussing what they are going to do tonight . is correctly labeled as neutral

2.The premice: Nine women in blue and purple dresses and one man wearing a purple shirt and black pants , clap while a man dressed in black dances . The hypothesis: There are people clapping . is correctly labeled as entailment

3.The premice: Two men sitting on horses one wearing a cowboy hat , the other in a baseball cap , with a big tree behind them . The hypothesis: The men are practicing riding for a movie they are going to be in . is correctly labeled as neutral

Three Incorrect labeled examples: 1.The premice: A man in a brown jacket , white shirt , and dark is holding a book with his finger on the page while sitting on a wooden floor , and leaning against a yellow wall with a door on one side and cloths on on the other side . The hypothesis: A man sits on a wooden floor building a model ship . is incorrectly labeled as contradiction

2.The premice: An Asian man is standing on a rusty dock surrounded by ropes and tires and is sticking a fishing pole into the water in front of him . The hypothesis: A fisherman is working hard on the dock is incorrectly labeled as neutral

3.The premice: A young man in a green holds a young boy with a blue backpack and a yellow in front of a run-down building . The hypothesis: A man is holding a young boy . is incorrectly labeled as neutral

Probably the reason why the three samples above are incorrectly labeled is that there are relatively more less frequently occurred characters contained and makes the model hard to analyze.

For more details about the program, refer to the repository posted on the following Github:

https://github.com/haoNTT/dsga1011hw2.git