# DS-GS 1011 Bag of N-Gram Document Classification

Yulin Shen

**Assignment 1**

## 1 Introduction

In the assignment, I use IMDB Movie review dataset to split 25000 train samples into 20000 train samples and 5000 validation samples to train the model with hyperparameter tuning. Then, I can test the models by the other 25000 test samples to find a best one. In order to find the best model, I design my model in these ways including tokenization schemes, vary for n-grams, vocabulary sizes, and embedding dimensions, also optimization hyperparameters.

## 2 Model Designing

### 2.1 Tokenization Schemes

I use 2 simple tokenization ways in the assignment. One is keeping all punctuations, and another is ignoring all punctuations as same as lab 3.

### 2.2 Model Hyperparameter

I try to set n numbers in n-grams as 1, 2, 3, 4 to get 4 different combinations. With 2 tokenization ways above, I have 8 combinations tokens sets now. In the lab 3, we use vocabulary size as 20000 and embedding dimension as 200, so I add a 30000 vocabulary size and a 300 embedding dimension to see the differnces.

### 2.3 Optimization Hyperparameter

I both try SGD and Adam optimizer to train the models. I set a small learning rate 0.005, and I find most of them are diverged and overfitted in 3 to 5 epochs in this small learning rate. So, I set each training process in 4 epochs with all 0.005 learning rate without using a linear annealing of learning rate.

# 3 Results

Because I have 2 different tokenization schemes, 4 different n-grams, 2 different vocabulary sizes, 2 different embedding dimensions, and 2 different optimizers, I could have 64 different results. I incorporate all results into 2 tables in the last page.

In the table, we can see the best model with highest test accuracy is to keep punctuations with 2 n-grams with 30000 vocabulary size and 300 embedding dimension. Its test accuracy is 83.784.

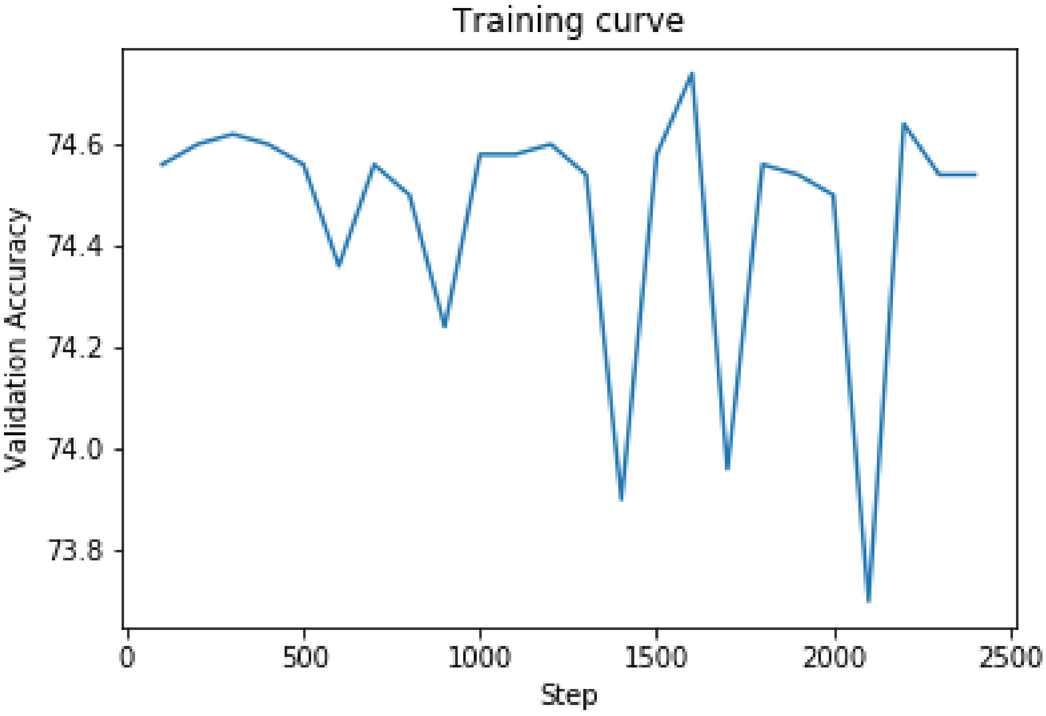
Also, I plot all training curves, and I pick two examples here in the next page. They could show the validation accuracy changing in the training processing.

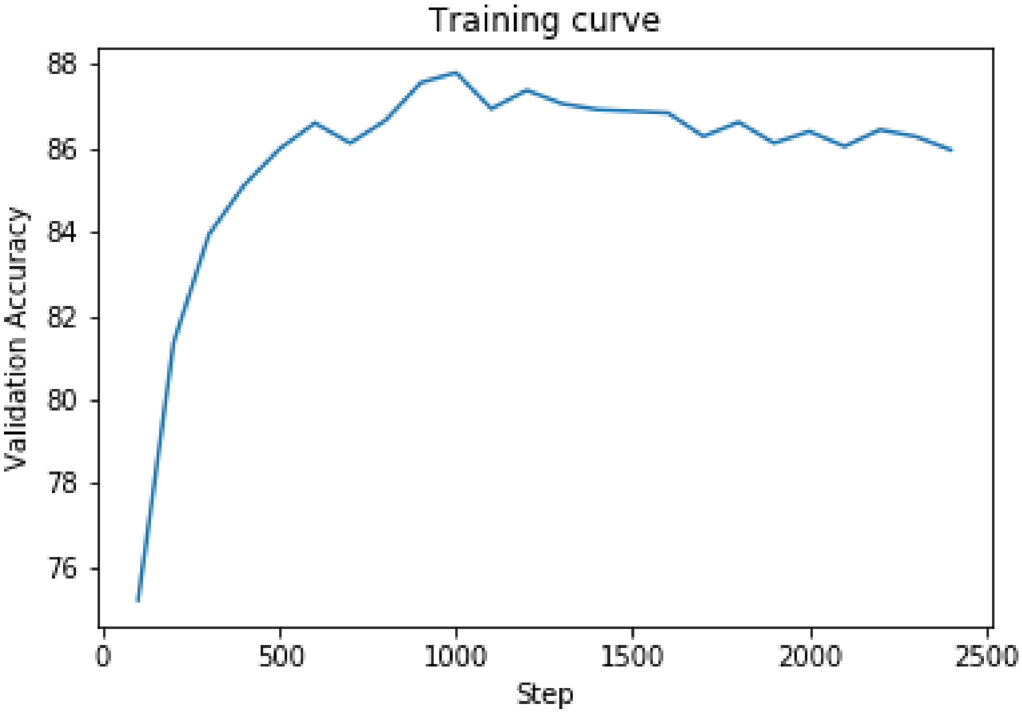
# 4 Conclusion

In my processing, I find n-grams number is the most important variable in the model. 1 and 2 grams perform much better than 3 and 4. Keeping or not keeping punctuations does not affact a lot in the result, and bigger vocabulary and embedding sizes seems better. Adam and SGD optimizer give very similar results.

# 5 Code Repository

I push my code into a github repository. The link is https://github.com/ys2542/NLP_Assignment_1. There are two python notebook files. Preprocess.ipynb is to tokenize the datasets with different n-grams into pickle files. After getting all pickle files, we only need to run train.ipynb.


Training curve
Validation Accuracy
74.6
74.4
74.2
74.0
73.8
Step
0
500
1000
1500
2000
2500


Training curve
Validation Accuracy
88
86
84
82
80
78
76
Step
0
500
1000
1500
2000
2500

| tokenization | n-grams | vocabulary | embedding | optimizer | val accu | test accu |
|---|---|---|---|---|---|---|
| wp | 1 | 20000 | 200 | Adam | 86.44 | 82.404 |
| wp | 1 | 20000 | 200 | SGD | 86.34 | 82.588 |
| wp | 1 | 20000 | 300 | Adam | 85.82 | 81.748 |
| wp | 1 | 20000 | 300 | SGD | 85.6 | 82.364 |
| wp | 1 | 30000 | 200 | Adam | 85.86 | 82.532 |
| wp | 1 | 30000 | 200 | SGD | 86.1 | 82.24 |
| wp | 1 | 30000 | 300 | Adam | 85.76 | 81.676 |
| wp | 1 | 30000 | 300 | SGD | 85.7 | 82.092 |
| wp | 2 | 20000 | 200 | Adam | 83.04 | 82.448 |
| wp | 2 | 20000 | 200 | SGD | 83.5 | 82.964 |
| wp | 2 | 20000 | 300 | Adam | 83.2 | 82.46 |
| wp | 2 | 20000 | 300 | SGD | 83.26 | 82.52 |
| wp | 2 | 30000 | 200 | Adam | 84.06 | 83.624 |
| wp | 2 | 30000 | 200 | SGD | 84.2 | 83.628 |
| wp | 2 | 30000 | 300 | Adam | 83.84 | 83.784 |
| wp | 2 | 30000 | 300 | SGD | 83.88 | 83.728 |
| wp | 3 | 20000 | 200 | Adam | 79.32 | 78.644 |
| wp | 3 | 20000 | 200 | SGD | 79.12 | 78.604 |
| wp | 3 | 20000 | 300 | Adam | 78.24 | 77.836 |
| wp | 3 | 20000 | 300 | SGD | 78.64 | 78.12 |
| wp | 3 | 30000 | 200 | Adam | 79.68 | 79.092 |
| wp | 3 | 30000 | 200 | SGD | 79.84 | 79.296 |
| wp | 3 | 30000 | 300 | Adam | 79.62 | 79.1 |
| wp | 3 | 30000 | 300 | SGD | 79.42 | 79.2 |
| wp | 4 | 20000 | 200 | Adam | 71.16 | 70.748 |
| wp | 4 | 20000 | 200 | SGD | 73.5 | 73.316 |
| wp | 4 | 20000 | 300 | Adam | 74.06 | 74.512 |
| wp | 4 | 20000 | 300 | SGD | 74.02 | 74.516 |
| wp | 4 | 30000 | 200 | Adam | 75.12 | 75.216 |
| wp | 4 | 30000 | 200 | SGD | 74.94 | 75.384 |
| wp | 4 | 30000 | 300 | Adam | 74.62 | 74.828 |
| wp | 4 | 30000 | 300 | SGD | 74.62 | 74.784 |

| tokenization | n-grams | vocabulary | embedding | optimizer | val accu | test accu |
|---|---|---|---|---|---|---|
| np | 1 | 20000 | 200 | Adam | 85.82 | 82.54 |
| np | 1 | 20000 | 200 | SGD | 85.98 | 82.852 |
| np | 1 | 20000 | 300 | Adam | 85.72 | 82.384 |
| np | 1 | 20000 | 300 | SGD | 85.68 | 82.468 |
| np | 1 | 30000 | 200 | Adam | 86.04 | 82.86 |
| np | 1 | 30000 | 200 | SGD | 86.04 | 82.872 |
| np | 1 | 30000 | 300 | Adam | 85.76 | 82.516 |
| np | 1 | 30000 | 300 | SGD | 86.06 | 82.288 |
| np | 2 | 20000 | 200 | Adam | 83.56 | 82.584 |
| np | 2 | 20000 | 200 | SGD | 83.74 | 82.812 |
| np | 2 | 20000 | 300 | Adam | 83.16 | 82.712 |
| np | 2 | 20000 | 300 | SGD | 83.3 | 82.728 |
| np | 2 | 30000 | 200 | Adam | 84.64 | 82.876 |
| np | 2 | 30000 | 200 | SGD | 84.2 | 83.368 |
| np | 2 | 30000 | 300 | Adam | 84.28 | 83.448 |
| np | 2 | 30000 | 300 | SGD | 84.32 | 83.312 |
| np | 3 | 20000 | 200 | Adam | 78.72 | 78.552 |
| np | 3 | 20000 | 200 | SGD | 79.6 | 78.76 |
| np | 3 | 20000 | 300 | Adam | 78.6 | 78.608 |
| np | 3 | 20000 | 300 | SGD | 79.06 | 78.552 |
| np | 3 | 30000 | 200 | Adam | 80.2 | 79.316 |
| np | 3 | 30000 | 200 | SGD | 80.48 | 79.664 |
| np | 3 | 30000 | 300 | Adam | 79.28 | 79.048 |
| np | 3 | 30000 | 300 | SGD | 79.3 | 79.08 |
| np | 4 | 20000 | 200 | Adam | 72.1 | 71.224 |
| np | 4 | 20000 | 200 | SGD | 74.28 | 73.636 |
| np | 4 | 20000 | 300 | Adam | 72.42 | 72.552 |
| np | 4 | 20000 | 300 | SGD | 73.38 | 73.063 |
| np | 4 | 30000 | 200 | Adam | 74.32 | 73.504 |
| np | 4 | 30000 | 200 | SGD | 74.32 | 73.468 |
| np | 4 | 30000 | 300 | Adam | 73.94 | 73.564 |
| np | 4 | 30000 | 300 | SGD | 74.56 | 74.176 |