DS-GS 1011 Fall 2018 HW1

zhe huang

October 10, 2018

https://github.com/hznyuds/nlp_1011/upload/master

1 Introduction and Data description

Totally 25000 training data, 25000 testing data from IMDB Movie review dataset.Split train dataset into 20000 train examples and 5000 validation examples. To save space, all figures and table will be appended in the last pages.

2 Tokenization Schemes

First, I experiment with 4 tokenization schemes: 1. only removing html tags 2. remove html tags, convert letter to lower case and remove punctuation 3. convert letter to lower case, remove punctuation and remove stop words 4. remove html tags, convert letter to lower case, remove punctuation, remove stop words and performing steming. From the result, tokenization scheme 3 is the best since the it achieves the best validation accuracy and validation accuarcy curve is almost always above the others. I will chose scheme 3 for the following experiments.

3 Varying n for n-gram

Under tokenization scheme3, I experimented with n equal 1, 2, 3, 4. The results show that for n equal 2, 4 the validation accuracy converges after 3000 steps under current setting. With n=4 slightly above n=2, considering that we need to experiment with different vocab size and embedding dimension, my intuition is n=4 will be more advantageous when we adjust other parameters. So I choose n=4 for the following experiments.

4 Varying vocabulary size

Under tokenization scheme3, n equal 4, I experimented with max vocabulary size of 1000, 10000, 100000, 1000000. Max vocab 1000000 gets slightly better

results than 100000, however, using max vocab size of 1000000 takes substantially longer time, while the gain is slim. So for the following experiment, I choose to use max vocabulary size of 100000.

5 Varying embedding size

Under tokenization scheme3, n equal 4, max vocab is 100000. I experimented with embedding size of 50, 100, 200, 300, 500. max validation accuracy achieved for embed100 is 89.28, max validation accuracy achieved for embed200 is 89.26. However we can see from the graph that after around 2000 iterations, embed200 is almost always performs better than embed100, given that we might need to increase epoch number later, and adjust other parameters, I will chose embed200.

6 Experiment with Optimizer

Using tokenization scheme3, n equal 4, max vocab 100000, embedding size 200. I experimented with Adam optimizer and SGD optimizer. We can see obviously that Adam reach a high validation accuracy much earlier in terms of epochs trained than SGD. But SGD takes 422.8 seconds to train 10 epochs while Adam takes 1318.2 seconds. So a fairer comparison would be to see which takes longer to reach the same level of accuracy, but for our purpose we don't need this since we will be specifying number of epochs to train.

7 Experiment with learning rate

Using tokenization scheme3, n equal 4, max vocab 100000, embedding size 200 and Adam optimizer, I experimented with learning rate 0.001, 0.01, 0.05, 0.1, as expected, learning rate of 0.001 start with very low validation accuracy, but as we train more steps, it catches up and get the best validation accuracy. I will choose learning rate to be 0.001. I also experimented with linear annealing of linear rate, setting the starting learning rate to be 0.1, and reduce the learning rate by 0.001 for every 300 iterations. Theoretically this scheme should reduce the time the algorithm takes to reach the same level of accuracy. However, since for the final model I planned to train for more epochs and save only the best model, I would choose the regular small learning rate not the linear annealing of linear rate in case the setting of reduction schedule is not good enough for the model.

8 Final Model

Configuration of the final model is learning rate 0.001, number of epochs is 50, batch size is 32, max vocab size is 100000, embedding dim is 200, max sentence length is 200. In the final model, I set the epochs number to be 50, and save

only the best model. The saved best model is then used to test on the test set. The test accuracy of the final model on the test set is 87.996. The training accuracy curve and validation accuracy curve are in Figure 8.

report 3 correct and 3 incorrect: Please refer to Jupyter notebook for full text in order to fit report into 4 pages. The examples are selected at random.

incorrect example1: true label is 0 predicted to be 1
since review 's film screening 's seen decade 's ago i 'd like add recent film open 's stock footage ¡unk¿ 's bombing germany film cut 's ¡unk¿ werner 's ¡unk¿ captain ¡unk¿ character aide running cover making way hitler 's ¡unk¿ bunker inside ¡unk¿ bunker staff personnel film cut 's conference scene 's ¡unk¿ ¡unk¿ giving decent impression adolf hitler ¡unk¿ officer 's ultimate victory...

incorrect example2: true label is 0 predicted to be 1
linda ¡unk¿ victim sadistic woman hater chuck ¡unk¿ i n't understand having sex dog animal abuse found entertaining funny linda ¡unk¿ virtual prisoner ¡unk¿ making films i know people criticize comment i feel strongly types films fuel fire hatred misogynistic feelings women this society...

incorrect example2: true label is 0 predicted to be 1
when employees theater find old reel film decide midnight screening night living dead assuming 's old preview reel unfortunately 's actually old nazi mind control experiment turns audience ¡unk¿ mindless...

correct example1: true label is 0 predicted to be 0
this movie absolute worst movie seen my sister boyfriend went rent zodiac 2007 got accident thought joke actual movie terrible waiting scary movie actual facts real zodiac killer the filmmakers clearly n't bother research killings ...

correct example2: true label is 0 predicted to be 0
jane russell proved delightful musical comedy performer similarly titled gentlemen prefer blondes sadly film ¡unk¿ skills there budget nice paris photography film n't work ms...

correct example3: true label is 0 predicted to be 0
so thats i called bad bad film ... poor acting poor directing terrible writing i cant stop laughing scenes story meaningless dont waste time watching film ...

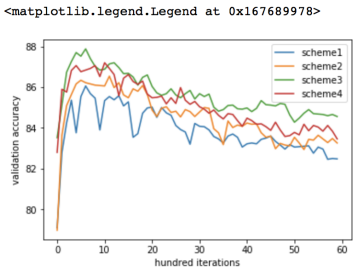

Figure 1: tokenization scheme

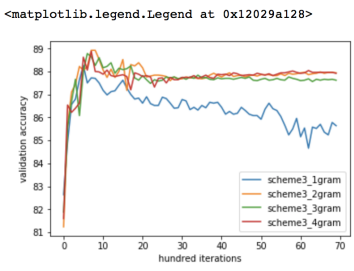

Figure 2: n gram

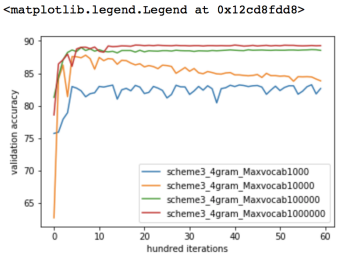

Figure 3: Vocab size

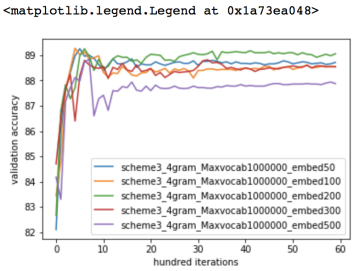

Figure 4: embed size

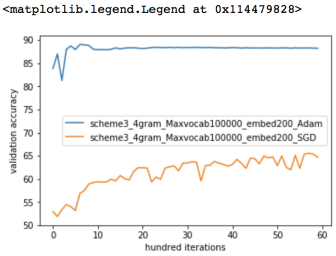

Figure 5: optimizer

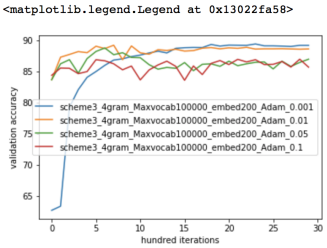

Figure 6: learning rate

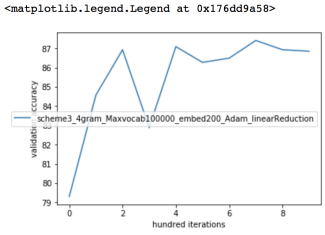

Figure 7: annealing linear of lr

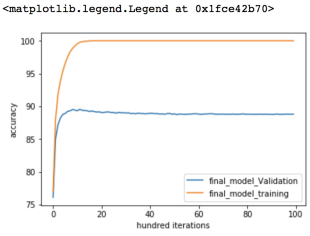

Figure 8: FinalModel Training Validation Curve