NLP Assignment-1

Chaitra Hegde (cvh255)

October 9, 2018

1 Introduction

As the experiment, 3 types of tokenization schemes were tried - lower case + removing punctuation, lower case+ removing punctuation and stop words, lower case + removing punctuations and stop words along with stemming. Adam and SGD optimizers were tried with constant learning rate, as well as, reducing the learning rate over time. Weight decay was set to regularize the model. While vocabularizing, different ngrams were tried where n=1,2,3,4. For each of the models, different vocabulary size, sentence length and embedding dimensions were changed.

The code can be found at `https://github.com/HegdeChaitra/IMDB_review_sentiment_analysis.git` Because of less availability of space, except for few figures, rest are available in Github repo.

2 Monograms

lower casing is done with removal of punctuations. First model with MAX SENTENCE LENGTH = 200 and vocab size = 25,000 with embed dim = 200 gave accuracy of 86.3%. Later with MAX SENTENCE LENGTH = 300 result of Best val loss: 0.302892, Best Accuracy: 87.860000 was achieved. In this experiment, the validation loss decreased to its minimum in the very first epoch and kept increasing later, while the train loss went down to 0.11. Hence, the model is over fitting on train set.

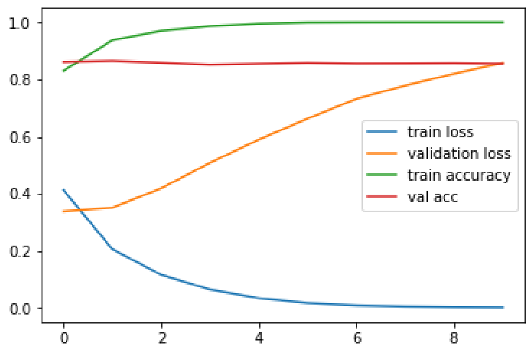
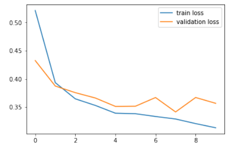


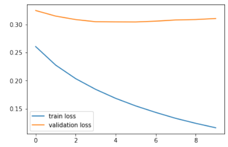
Figure 1: accuracy = 86.3%, Monogram,vocab=25000,emb dim = 200,sentence len=200

In order to incorporate some regularization, the *weightdecay* parameter in the Adam optimizer was set to 0.00001 with initial lr=0.01. The validation loss kept decreasing and reached 0.35 at the end of 10 epochs and started jumping off and on. Hence, the lr was reduced to 0.001 and was run for 10 more epochs. The result was - Best val loss: 0.304384, Best Accuracy: 88.180000 and the validation loss curve saturated. Hence, it could be seen that using weight decay with annealing learning rate didn't help much and also took more epochs to reach the minima. Hence, here on wards we will be sticking to constant learning rate without regularization. refer figure 2

Now, the vocab size was increased to 50,000 and embedding dimension was increased to 400, the result is - Epoch: 0, Phase: validate, epoch loss: 0.2975, accuracy: 88.4200. Increasing vocab size and embedding dimension gave 1 percent increase in accuracy. The MAX SENTENCE LENGTH = 400 and embed dim = 400, with vocab size of 400 gave result = Best val dice loss: 0.313470, Best Accuracy: 87.240000. Hence, increasing the sentence length and embedding dimension too much is not helpful here. Greater embedding dimension indicate more parameters and hence a

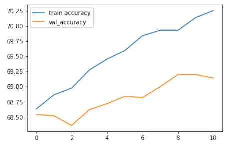


(a) Loss over first 10 epoch lr=0.01

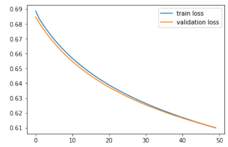


(b) Loss over last 10 epochs lr=0.001

Figure 2: Adam with Weight Decay and Annealing LR



(a) Accuracy for last 50 epochs sampled every 10 epochs



(b) Loss over last 50 epochs

Figure 3: Using SGD optimizer

complex model which might end up over fitting on train set more and learning less for validation set.

Instead of Adam, SGD optimizer was tried, but it is lot slower than Adam and even after 100 epochs, it just reached 70% accuracy where as Adam reached 87% in 10 epochs. Hence, we will be sticking to Adam optimizer here on wards. see figure 3

3 Monograms with stop words removal

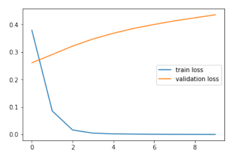
As the next experiment, the tokenization scheme was changed to lower case + punctuation removal + stop words removal. Most commonly occurring words like ["i","the","we","a"] end up taking lots of space in the vocabulary and might be contained many a time in the start of sentence within the maximum sentence length and not letting useful information to come into picture. Hence, its beneficial to remove stop words and hence make room for more important words.

MAX SENTENCE LENGTH = 300 with max vocab size = 50000 and emb dim = 300 gave the good result of Best val dice loss: 0.291457, Best Accuracy: 88.780000. Hence, removing stop words resulted in 1% increase in the accuracy. Now, the same experiment is run with vocab size = 100000 and the accuracy increased a bit to Best val dice loss: 0.277175, Best **Accuracy: 88.920000.**

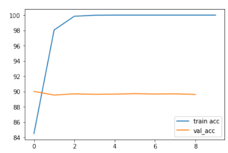
Now we know that removal of stop words is helpful. Here on wards we will remove the stop words while tokenizing the dataset.

4 Bi-gram

plots in Figure 4.Now, in order to incorporate the local sequential information, bi-gram vocabularization was tried. The previous setting with bi-gram gave better performance and the accuracy



(a) Loss



(b) Accuracy = 90.02%

Figure 4: Bi-gram,stop word removal,vocab=100000, emb dim=300,sentence len=400

(a) Loss

(b) Accuracy = 90.56%

Figure 5: Tri-gram,stop word removal,vocab=200000, emb dim=300,sentence len=400

(a) Loss

(b) Accuracy = 90.64%

Figure 6: Best model - Four-gram, vocab = 200000,sentence len=400, emb dim=300

increased significantly to Best val loss: 0.265221, Best Accuracy: 89.640000. Hence, incorporating local sequential information with bi-grams was helpful. Now, the MAX SENTENCE LENGTH=400 was used and the result got better to Best val loss: 0.260760, Best Accuracy: 90.020000. Hence, with better vocabulary, looking into more sentence length was helpful.

To see whether further increase in max length and vocabulary cab be beneficial, MAX SENTENCE LENGTH = 500 and vocab size=20000 was made but it wasn't helping - Best val loss: 0.260128, Best **Accuracy: 89.700000**

5 Trigrams

With our best setting so far, MAX SENTENCE LENGTH = 400 and vocab size = 200000 and emb dim = 300 was run again with tri-grams and the accuracy increased further - Best val dice loss: 0.257803, Best **Accuracy: 90.560000** refer figure 5

6 Fourgrams

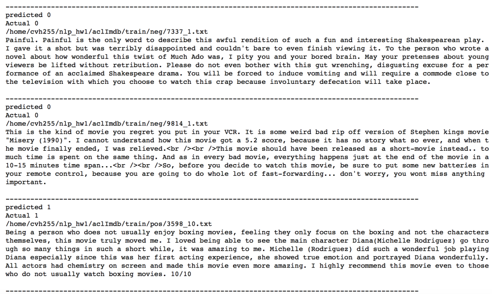
Figure 7. We have seen the improvements in accuracy with n-grams. Now, to see if using 4-grams will be helpful, an experiment with best setting so far was run again with 4-grams and the result was better by small fractions - Best val dice loss: 0.251631, Best **Accuracy: 90.640000.**

Hence, it could be noted that with increased n-grams, the improvement was better. Also after 3, the improvement is not significant. May be till certain n(say 5), there will be improvement and then result gets stagnant because n-grams though help us model sequential information are not great at capturing long term dependencies.

7 Stemming

Now we know that four-gram model with lowercase, removal of punctuation and stop words was the best model. To see whether mapping the word to its root form, i.e stemming is any helpful here, we Incorporated stemming to tokenization scheme. The previous best setting used with stemming and the result was - Best val dice loss: 0.268192, Best Accuracy: 89.720000.

It could be seen that the result was not any better than before. Hence, in this case, stemming is not helpful. May be because, the trained word embedding is able to learn map the word as close to its root word. And the action and time specified by the word is helpful here in deciding whether a review is positive or not.



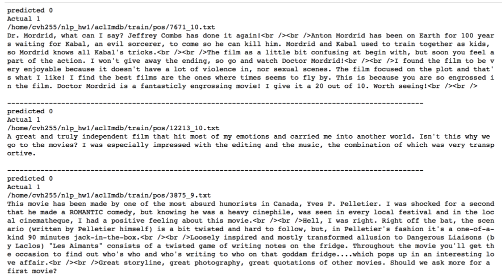
```
-----------------------------------------------------------------------------------------
predicted 0
Actual 0
/home/cvh255/nlp_hw1/aclImdb/train/neg/7337_1.txt
Painful. Painful is the only word to describe this awful rendition of such a fun and interesting Shakespearean play.
I gave it a shot but was terribly disappointed and couldn't bare to even finish viewing it. To the person who wrote a
novel about how wonderful this twist of Much Ado was, I pity you and your bored brain. May your pretenses about young
viewers be lifted without retribution. Please do not even bother with this gut wrenching, disgusting excuse for a per
formance of an acclaimed Shakespeare drama. You will be forced to induce vomiting and will require a commode close to
the television with which you choose to watch this crap because involuntary defecation will take place.

-----------------------------------------------------------------------------------------
predicted 0
Actual 0
/home/cvh255/nlp_hw1/aclImdb/train/neg/9814_1.txt
This is the kind of movie you regret you put in your VCR. It is some weird bad rip off version of Stephen kings movie
"Misery (1990)". I cannot understand how this movie got a 5.2 score, because it has no story what so ever, and when t
he movie finally ended, I was relieved.<br /><br />This movie should have been released as a short-movie instead.. to
much time is spent on the same thing. And as in every bad movie, everything happens just at the end of the movie in a
10-15 minutes time span...<br /><br />So, before you decide to watch this movie, be sure to put some new batteries in
your remote control, because you are going to do whole lot of fast-forwarding... don't worry, you wont miss anything
important.

-----------------------------------------------------------------------------------------
predicted 1
Actual 1
/home/cvh255/nlp_hw1/aclImdb/train/pos/3598_10.txt
Being a person who does not usually enjoy boxing movies, feeling they only focus on the boxing and not the characters
themselves, this movie truly moved me. I loved being able to see the main character Diana(Michelle Rodriguez) go thro
ugh so many things in such a short while, it was amazing to me. Michelle (Rodriguez) did such a wonderful job playing
Diana especially since this was her first acting experience, she showed true emotion and portrayed Diana wonderfully.
All actors had chemistry on screen and made this movie even more amazing. I highly recommend this movie even to those
who do not usually watch boxing movies. 10/10

-----------------------------------------------------------------------------------------
```

(a) Correct Predictions



```
predicted 0
Actual 1
/home/cvh255/nlp_hw1/aclImdb/train/pos/7671_10.txt
Dr. Mordrid, what can I say? Jeffrey Combs has done it again!<br /><br />Anton Mordrid has been on Earth for 100 year
s waiting for Kabal, an evil sorcerer, to come so he can kill him. Mordrid and Kabal used to train together as kids,
so Mordrid knows all Kabal's tricks.<br /><br />The film as a little bit confusing at begin with, but soon you feel a
part of the action. I won't give away the ending, so go and watch Doctor Mordrid!<br /><br />I found the film to be v
ery enjoyable because it doesn't have a lot of violence in, nor sexual scenes. The film focused on the plot and that'
s what I like! I find the best films are the ones where times seems to fly by. This is because you are so engrossed i
n the film. Doctor Mordrid is a fantasticly engrossing movie! I give it a 20 out of 10. Worth seeing!<br /><br />

-----------------------------------------------------------------------------------------
predicted 0
Actual 1
/home/cvh255/nlp_hw1/aclImdb/train/pos/12213_10.txt
A great and truly independent film that hit most of my emotions and carried me into another world. Isn't this why we
go to the movies? I was especially impressed with the editing and the music, the combination of which was very transp
ortive.

-----------------------------------------------------------------------------------------
predicted 0
Actual 1
/home/cvh255/nlp_hw1/aclImdb/train/pos/3875_9.txt
This movie has been made by one of the most absurd humorists in Canada, Yves P. Pelletier. I was shocked for a second
that he made a ROMANTIC comedy, but knowing he was a heavy cinephile, was seen in every local festival and in the loc
al cinematheque, I had a positive feeling about this movie.<br /><br />Hell, I was right. Right off the bat, the scen
ario (written by Pelletier himself) is a bit twisted and hard to follow, but, in Pelletier's fashion it's a one-of-a-
kind 90 minutes jack-in-the-box.<br /><br />Loosely inspired and mostly transformed allusion to Dangerous Liaisons (b
y Laclos) "Les Aimants" consists of a twisted game of writing notes on the fridge. Throughout the movie you'll get th
e occasion to find out who's who and who's writing to who on that goddam fridge....which pops up in an interesting lo
ve affair.<br /><br />Great storyline, great photography, great quotations of other movies. Should we ask more for a
first movie?
```

(b) Incorrect predictions

Figure 7

8 Best Model

We have experimented with various parameters like maximum sentence legth, embedding dimension, vocabulary size, learning rates, optimization algorithms, tokanization scheme and vocabulary. From the result, it could be seen that Adam optimizer with learning rate of 0.01 and max sentence length of 400, vocabulary size of 20000 and embedding dimension of 300 with lower casing, removal of punctuation and stop words and 4-gram vocabulary was the best model. It gave the performance of Best val loss: 0.251631, Best Accuracy: 90.640000. So, we will be using this model which is trained on train+val data set for production. As the last step, we will be testing our model on test set so that we know that the model can generalize on data set it has not seen before.

9 Test set

The vocabulary from train set was preserved and was used to vocabularize the validation set. The model was run on test set to get the performance of **Accuracy on test set is = 89.2 %** Cross Entropy loss on test set is = 0.43386340141296387 Hence, the model has capacity to generalize and hence can be deployed.

Complete code at `https://github.com/HegdeChaitra/IMDB_review_sentiment_analysis.git`

N-gram	Tokenization	Vocab size	emb dim	max sentence	Val. Acc.
1-gram	remove punct	25000	200	200	86.3%
1-gram	remove punct	25000	300	300	87.86%
1-gram	remove punct	50000	400	300	87.08%
1-gram	remove punct	100000	400	400	87.24%
1-gram	remove stopword	50000	200	200	88.7%
1-gram	remove stopword	100000	300	300	88.92%
2-gram	remove stopword	100000	300	300	89.64%
2-gram	remove stopword	100000	300	400	90.02%
2-gram	remove stopword	200000	300	500	89.7%
3-gram	remove stopword	100000	300	300	89.74%
3-gram	remove stopword	200000	300	400	90.56%
4-gram	remove stopword	100000	200	400	89.92%
4-gram	remove stopword	200000	300	400	**90.64%**
4-gram	Stemming	100000	300	400	89.72%
test	**set**	**accuracy**	**best**	**model**	**89.2%**