Bag of Words Sentiment Analysis Report

Shubhankar Panse

October 10, 2018

Abstract

This is an ablation report of various hyperparameters used to tune the bag of words model for sentiment analysis of IMDb movie reviews dataset. The first section is varying n for n-grams. In this section, I have also experimented with different learning rates and optimizers to determine which works the best for each n-gram. Section 2 is experimenting with different number of embedding dimensions and section 3 is varying vocabulary size. Section 4 is varying the the Tokenization schemes. Section 5 discusses the effect of batch sizes and the last section reports the final hyperparameter configuration and reports 3 correct and incorrect results.

Github repo link

1 N-grams

I have varied the n-grams with n = 1 to n = 4. I checked the performance of each n-gram model with different learning rates.

The results are as follows:

1. Unigram using ADAM optimizer:

Learning Rate	Training Accuracy	Validation Accuracy
0.0001	74.90	72.12
0.001	96.05	86.66
0.01	99.505	82.84

2. Unigram using SGD Optimizer:

Learning Rate	Training Accuracy	Validation Accuracy
0.0001	54.90	53.12
0.001	57.87	56.36
0.01	65.01	64.06

We can see that the Learning Rate of 0.001 is the most appropriate. Learning Rate of 0.0001 is to small and hence the model never converges. A learning rate of 0.01 is too high while using the ADAM optimizer and this causes the optimizer to overshoot the minimum and diverge instead of converging. This can be seen in the Figure 1.

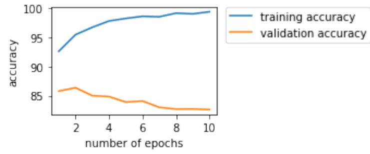

Figure 1: Validation and Training Accuracy (LR= 0.001, ADAM)

Thus we can see that Stochastic Gradient Descent performs very poorly even when the learning rate is very high. This can be attributed to the fact that ADAM is an Adaptive optimizer while SGD is not. That is why ADAM converges quicker than SGD.

However, there are certain problems with ADAM as well. It is not stable in the sense that the validation accuracy drops after reaching the peak value at the later epochs. This can be observed in the Figure 2.

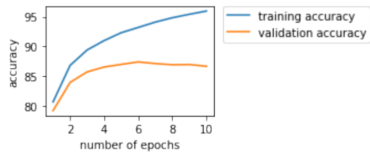

Figure 2: Validation and Training Accuracy (LR= 0.001, ADAM)

This is because the model converges quickly and then starts overshooting the minimum. Linear annealing of Learning Rate can help alleviating this problem. I halved the learning rate after every third epoch. Due to this, the learning rate dropped as the model reached closer to the minimum and hence a smooth curve can be seen in the Figure 3.

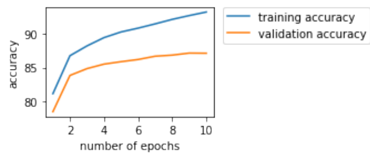

Figure 3: Validation and Training Accuracy (LR= 0.001, ADAM, Linear Annealing)

As we vary n-grams, the same configuration of LR=0.001 with Linear Annealing and using the ADAM optimizer seems to work the best. Following are the results and graphs for bi-gram, tri-gram and four-gram models.

n	Training Accuracy	Validation Accuracy
1	93.15	87.0
2	92.55	86.22
3	92.08	86.15
4	91.24	86.12

The model performance deteriorates with increase in 'n'. This could be because of the fact that adding higher order grams to the bag of words complicates the probability calculations and thus despite providing sequence information and negation information, it still underperforms as compared to the unigram model.

2 Embedding dimensions

From the previous section, the learning rate, linear annealing and unigram bag of words is fixed and embedding dimensions are varied. I tried 4 different values of embedding dimensions as follows:

Embed dims	Training Accuracy	Validation Accuracy
100	94.18	87.44
200	95.90	86.78
300	96.61	87.5
400	97.1	87.06

Thus we can see that embedding dimensions of 100 and 300 yield almost the same validation accuracy. Thus I chose to go with 100 embedding dimensions as it significantly reduces computation time.

3 Vocab size

Fixing all other hyperparameters, I changed the vocabulary size to 3 values as shown below:

Vocab size	Training Accuracy	Validation Accuracy
5000	90.85	85.8
10000	93.04	87.02
15000	94.32	86.92

We can observe that reducing the vocabulary size to 5000 leaves us with a high amount of unknowns and the model underperforms. The vocabulary size of 10000 and 15000 doesn't really make a difference so I chose to go with 10000.

4 Tokenization schemes

The first tokenization scheme used is without removing punctuations and without lowering all words. The results are poor as expected. This is because punctuations form a large part of the corpus and are largely non-informative. Also upper case versions of a word are considered different from their lower case versions which does not suit our task.

The next method I tried was removing punctuations and stemming the words. Stemming reduces the vocabulary by counting words and their different forms as the same token (eg. 'running' and 'run' are stemmed to 'run'). I have used the nltk PorterStemmer for this task. The results are much better and achieve a good accuracy on the validation dataset.

The last tokenization scheme was adding the removal of tags like tabs, newline characters,etc. This also bumped up the validation accuracy although only marginally.

Tokenization schemes	Training Accuracy	Validation Accuracy
No pre-processing	92.67	85.8
Punctuation removal and stemming	92.59	87.02
Stemming and removing punctuation and other tags	92.55	87.10

5 Batch Size

Varying batch-size also has an effect on the performance. This can be seen in the table below:

Batch size	Training Accuracy	Validation Accuracy
16	94.22	87.24
32	92.44	86.94
64	90.615	86.16

Thus we can see that a decrease in batch-size results in a marginal increase in the validation accuracy which comes at the cost of more training time.

6 Final hyperparameter values

Thus, as per above experiments, I have arrived at the following hyperparameter values:
1. n-grams: unigrams
2. Vocabulary size: 10000
3. Embedding dimensions: 100
4. Learning rate: 0.001 with linear annealing 0.5x after every 3 epochs.
5. Optimizer: ADAM
6. Tokenization scheme: Removing punctuations and special tags and lowering tokens followed by stemming.
7. Batch size=16

The final test accuracy using the above hyperparameter configuration is 84.86.

Examples: (removed padding from all examples)

Correctly Classified no.1

Review: 1st watch ¡unk¿ 4 out of ¡unk¿ jim ¡unk¿ brian smith drab and un spectacular suppos sequel to the origin classic anim 101 dalmatian ye the movi continu where it end in the first one but the problem is that it play out much like the origin one of the great thing about the origin wa the pace of the stori which thi one doe n't have the anim is also veri un spectacular for disney and all we get is the same charact go thru the same kind of stori all over again when is disney go to stop bore us with sequel and re do 's etc .. etc probabl when we stop rent or buy thi mediocr fare that they have put out
Actual label: neg
Predicted label: neg

Correctly Classified no.2 Review: after read the comment to thi movi and see the mix review i decid that i would add my ten cent worth to say i thought the film wa excel not onli in the visual beauti the write music score act and direct but in put across the stori of joseph smith and the road he travel through life of hardship and persecut for believ in god the way he felt and knew to be hi path i am veri pleas inde to have had a small part in tell the stori of thi remark man i recommend everyon to see thi when the opportun present itself no matter what religi path he or she may be walk thi onli ¡unk¿ one with more determin to live the life that we should with true valu of love and forgiv as the savior taught us to do
Actual label: pos
Predicted label: pos

Correctly Classified no.3 Review: i have to say that the event of 9/11 did n't hit me until i saw thi documentari it took me a year to come to grip with the devast i wa the one who wa chang the station on the radio and channel on tv if there wa ani talk about the tower i wa sick of hear about it when thi wa air on tv a year and a day later i wa ¡unk¿ my eye out it wa the first time i had cri sinc the attack i highli recommend thi documentari i am watch it now on tv 5 year later and i am still cri over the tragedi the fact that thi contain one of the onli video shot of the first plane hit the tower is amaz it wa an accid and look where it got them these two brother make me want to have been there to help
Actual label: pos
Predicted label: pos

Incorrectly classified no.1 Review: i thought the origin of thi film wa quaint and charm as well as have me sit on the edg of my seat tri to figur it out.¡br /¿sinc i had alreadi seen the origin when i saw thi on sci fi ¡unk¿ i do n't know if thi remak wa deliber made for sci fi i knew what it wa within the first few minut sinc i like richard ¡unk¿ as a charact actor i want to see how he would pull it off.¡br /¿the writer produc etc modern the film a bit by tri to explain the plight of the alien they could no longer reproduc their own kind and need help use the same pseudo scienc that ha been cram in our ear in the 90 's mayb it ad a bit of polish to the film or not.¡br /¿thi film film thi product take on a more sinist edg than the origin ¡unk¿ the origin end with a confront between the young woman and the alien and an understand of sort took place although no resolut of the alien 's ¡unk¿ /¿i sort of rememb that in thi remak the woman becam rather hostil toward the
Actual label: pos
Predicted label: neg

Incorrectly classified no.2 Review: it 's terrif when a funni movi doe n't make smile you what a piti thi film is veri bore and so long it 's simpli ¡unk¿ the stori is stagger without goal and no fun.¡br /¿you feel better when it 's finish
Actual label: neg
Predicted label: pos

Incorrectly classified no.3 Review: everi onc in a while someon out of the blue look at me a littl sideway and ask what 's with ¡unk¿ i know immedi they have a case of bare hidden amus horror you see i wa the cinematograph on the film.¡br /¿let me clarifi some point regard thi interest life ¡unk¿ ¡unk¿ ¡unk¿ wa call one hard hit i met jame cahil in juli of 1999 a day after i wrap triangl squar a great littl 35 mm featur that like so mani indi featur of the era never got distribut despit festiv accolad ... it fell etern victim to the fine print of ¡unk¿ 's notori experiment featur contract but i digress ... /¿i though i wa on a roll and when jame ask me to shoot hi littl gangster flick in 16 mm with a shoot budget of about ¡unk¿ not want to break pace i took it after all clerk el ¡unk¿ ... i too believ the myth back ¡unk¿ /¿let 's just chalk it up as film school for mani involv myself includ ¡unk¿ wa shot over two week in august 1999 in ¡unk¿ ¡unk¿ and santa ¡unk¿ ca cahil taught drama at a
Actual label: neg
Predicted label: pos