Bag of N-Gram Document Classification

Jyotirmoy Mohapatra

10/10/2018

GitHub Repo

1 Comparing Performance by tuning N-Gram size, Optimizer and Learning Rate

In this section, an experiment was conducted with several combinations of N-gram size, Optimizer and learning rate in order to find the combination that gave the best validation accuracy. The tokenization scheme used here removes the punctuation and converts every token to lowercase. Other hyperparameters that were kept constant were : Batch Size= 32, Vocabulary Size = 10000, Embedding Size = 100.

Hyperparameter Comparison				
N-Grams	Optimizer	Learning Rate	Train Acc.	Validation Acc.
1	Adam	0.01	99.445	78.62
1	Adam	0.001	96.155	83.78
1	SGD	0.01	64.785	65.06
1	SGD	0.1	71.095	70.34
1	SGD	0.1($\gamma = 0.03$/epoch)	66.77	67.54
1	Adam	0.001($\gamma = 0.03$/epoch)	90.135	83.76
1	Adam	0.001($\gamma = 0.17$/epoch)	92.75	85.18
2	Adam	0.001	94.225	85.02
2	Adam	0.001($\gamma = 0.03$/epoch)	88.435	84.6
2	Adam	0.001($\gamma = 0.17$/epoch)	91.145	85.66
3	Adam	0.001	93.29	85.12
3	Adam	0.001($\gamma = 0.03$/epoch)	87.995	83.96
3	Adam	0.001($\gamma = 0.17$/epoch)	92.56	85.24
4	Adam	0.001	92.225	84.76
4	Adam	0.001($\gamma = 0.03$/epoch)	87.543	84.54

From the table above, it was observed that Adam optimizer performs considerably better than the Stochastic Gradient Descent (SGD). SGD with a lower learning rate tends to overshoot and does not converge well. Hence, a high learning rate is better for SGD optimizer in this case. Adam optimizer performs better with a lower learning rate. Thus, in order to further improve the performance with Adam, learning rate annealing was performed with different decay rates. It was observed that a decay rate of 0.5 every 3 epochs performs better than a decay rate of 0.1 every 3 epochs. Secondly, different n-gram sizes were experimented upon. In this experiment, a unigram model consisted of single word tokens, a bi-gram model consisted of both single word tokens and 2 word tokens and a similar approach was used for tri-gram and 4-gram models. The unigram, bi-gram and tri-gram models gave comparable accuracy but there was a slight dip in accuracy for the 4-gram model. The Bi-gram model with adam optimizer and learning rate of 0.001 with decay rate of 0.5 every 3 epochs was used for conducting further experiments.

2 Exploring different Tokenization Schemes

To further improve the performance of the model, three tokenization schemes were explored. The first tokenization scheme involved no pre-processing of tokens. The second tokenization scheme removed the punctuations and converted every token to lowercase. The third tokenization scheme replaced each token with its lemma and also removed few HTML tags.

Tokenization Scheme Comparison			
Tokenization Scheme	N-Grams	Train Acc.	Validation Acc.
No Pre-Processing	1	92.16	83.84
No Pre-Processing	2	90.005	84
No Pre-Processing	3	89.5	83.84
Lower Case and no punctuation	1	92.75	85.18
Lower Case and no punctuation	2	91.145	85.66
Lower Case and no punctuation	3	92.56	85.24
Stemming	1	91.41	83.4
Stemming	2	89.675	83.74
Stemming	3	88.445	83.02

The tokenization scheme with punctuation removal gave the best results. The stemming tokenization scheme reduced the dictionary size and gave more meaningful vocabulary, however, it did not perform as well as expected. One possible reason for this might be because the punctuation were not removed in this scheme which might have diluted the quality of the vocabulary.

3 Exploring Other Hyperparameters

Further experiments were performed using the Bi-gram model with Adam optimizer, a learning rate of 0.001 and decay factor of 0.5 every 3 epochs along with tokenization scheme which removed punctuation and converted every token to lowercase.

3.1 Vocabulary Size and Embedding Size

Vocabulary and Embedding size Comparison			
Vocab Size	Embedding Size	Train Acc.	Validation Acc.
10000	100	91.145	85.66
10000	300	92.87	85.28
20000	100	94.32	85.14
20000	300	95.595	85.38
40000	100	95.42	85.18
40000	300	95.445	85.41

Varying the embedding size and the vocabulary size did not really have a huge impact on the model's performance. Increasing the vocabulary size and embedding size increased the number of computation and thus increased the overall training time.

3.2 Batch Size

Batch size Comparison		
Batch Size	Train Acc.	Validation Acc.
16	92.48	85.06
32	91.145	85.66
64	89.155	84.72

Batch sizes of 16 and 32 had comparable performance. However, increasing the batch size to 64 resulted in dip in accuracy.

4 Comparison Graphs

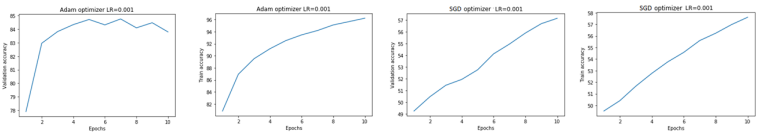

Figure 1: Adam vs SGD Optimizer.

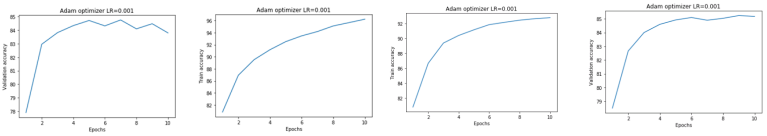

Figure 2: Learning Rate annealing

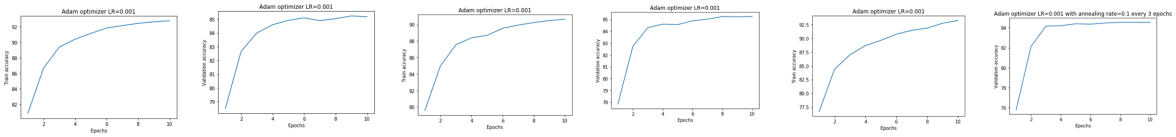

Figure 3: n-gram sizes

5 Correct and Incorrect Review

5.1 Correct

vincent ¡unk¿ plays the part of paul an ex con ¡unk¿ to an office job where he meets ¡unk¿ a secretary who is ¡unk¿ ¡unk¿ when she has her hearing ¡unk¿ in ¡unk¿ ¡unk¿ when not played by ¡unk¿ ¡unk¿ together they help each other to develop as ¡unk¿ /¿¡br /¿what was particularly interesting about this film was the ¡unk¿ of the characters not fitting into obvious stereotypes paul appears ¡unk¿ in the office environment is it that he 's just not cut out for work this belief is ¡unk¿ when he gets a job in a bar and ¡unk¿ /¿¡br /¿the film has a certain ¡unk¿ which i find refreshing and showed how easy is to act ¡unk¿ even if we think it is ¡unk¿ or ¡unk¿ /¿¡br /¿finally it is a film full of great ¡unk¿ both touching and humorous one is when ¡unk¿ is ¡unk¿ and is trying to ¡unk¿ a screaming baby she continues to ¡unk¿ it but takes her hearing ¡unk¿ out for her own ¡unk¿ ¡unk¿ ¡unk¿ plays the the part part of ¡unk¿ ¡unk¿ ¡unk¿ ¡unk¿ ¡unk¿ ¡unk¿ to an ¡unk¿ ¡unk¿ ¡unk¿ where he he meets ¡unk¿ ¡unk¿ ¡unk¿ ¡unk¿ who is ¡unk¿ ¡unk¿ ¡unk¿ when she she has -positive

i can agree with other comments that there was n't an enormous amount of history ¡unk¿ in the movie but it was n't a documentary it was meant to entertain and i think it did a very good job at it.¡br /¿¡br /¿i agree with the black family the scenes with them seemed out of place like all of a sudden it would be thrown in but i did catch on to the story and the connection between the families later on and found it pretty good.¡br /¿¡br /¿despite it was n't a re ¡unk¿ of the 60s it did bring into the light very big and important ¡unk¿ ¡unk¿ of the decade i found it very entertaining and worth my while to watch i can ¡unk¿ agree with ¡unk¿ ¡unk¿ ¡unk¿ that there there was was n't ¡unk¿ ¡unk¿ ¡unk¿ amount of ¡unk¿ ¡unk¿ ¡unk¿ in the the movie movie but but it it was was n't n't a a documentary ¡unk¿ it was was meant meant to ¡unk¿ ¡unk¿ and i i think think it it did did a a very very good good job ¡unk¿ ¡unk¿ it.¡br /¿¡br /¿¡br /¿i ¡unk¿ agree with with the the black ¡unk¿ ¡unk¿ the scenes scenes with with them ¡unk¿ ¡unk¿ out of of place ¡unk¿ like all all of of a ¡unk¿ ¡unk¿ it would would be ¡unk¿ thrown in ¡unk¿ but i i did ¡unk¿ ¡unk¿ on to to the the story story and and the ¡unk¿ -positive

for those of you that do n't that reference ¡unk¿ was 4 ¡unk¿ hitting one body ... ¡br /¿¡br ¡unk¿ onto the ¡unk¿ /¿¡br /¿i miss ¡unk¿ saturday night some of my favorite wrestling moments took place on this stage i remember watching stunning steve ¡unk¿ rick ¡unk¿ brian ¡unk¿ ¡unk¿ jack ¡unk¿ ¡unk¿ johnny b. ¡unk¿ ¡unk¿ in his ¡unk¿ days lord steven ¡unk¿ ¡unk¿ heat ¡unk¿ ¡unk¿ ¡unk¿ ... ¡unk¿ be here a while ¡unk¿ everyone point is ¡unk¿ had an awesome ¡unk¿ in the pre ¡unk¿ days and they were producing entertaining television ¡unk¿ ¡unk¿ on commentary in it 's later years gave me a whole new reason to watch when i started smoking ¡unk¿ as a teenager ... i really wish ¡unk¿ would put him on the ¡unk¿ for a show or two maybe at the next great american ¡unk¿ they ¡unk¿ here comes ¡unk¿ ¡unk¿ he was ¡unk¿ /¿¡br ¡unk¿ for those those of of you you that that do do n't n't that ¡unk¿ ¡unk¿ ¡unk¿ ¡unk¿ ¡unk¿ ¡unk¿ ¡unk¿ ¡unk¿ ¡unk¿ ... ¡br ¡br /¿¡br ¡unk¿ ¡unk¿ onto the ¡unk¿ ¡unk¿ /¿¡br /¿i ¡unk¿ ¡unk¿ ¡unk¿ saturday night ¡unk¿ some of of my my favorite ¡unk¿ ¡unk¿ ¡unk¿ ¡unk¿ ¡unk¿ on this ¡unk¿ ¡unk¿ i remember ¡unk¿ ¡unk¿ ¡unk¿ ¡unk¿ -positive

5.2 Incorrect

sitting down to watch the ¡unk¿ season of the ¡unk¿ on the ¡unk¿ of love i knew i would be in for an interesting time i had watched some of the previous seasons of the ¡unk¿ in passing watching an episode or two and missing the next three or so i find that the ¡unk¿ is often appealing and intriguing though its quality and morality are often ¡unk¿ /¿¡br ¡unk¿ the ¡unk¿ of love details the journey taken by jake a ¡unk¿ year old commercial pilot from ¡unk¿ texas to find true love as true a love as one can find in a season long reality drama ¡unk¿ show jake meets 25 beautiful girls from all over the country he begins to get to know them a bit but it is mostly superficial how well can you get to know someone in a few 5 minute conversations jake tries to make his true intentions known from the very beginning at least to the audience he noted that he does n't just want love or a good time but he wants a ¡unk¿ or wife we can only assume that he has made this clear to the women in the competition -negative

anyone who has a remote interest in science fiction should start at the ¡unk¿ everyone says star wars and star trek are the best science fiction films to begin at which is fine but the truth is the ¡unk¿ and this movie ¡unk¿ green are far better choices than those series ¡unk¿ is probably science fiction 's best kept secret it remains one of the biggest yet most forgotten films but the impact of its setting is becoming more a reality with each passing day ¡unk¿ heston ¡unk¿ his role yet it works edward ¡unk¿ ¡unk¿ in his final role makes the most out of it in ¡unk¿ green more than anyone else and his final scenes are ¡unk¿ /¿¡br /¿it is manhattan in ¡unk¿ the world is ¡unk¿ and food is an unbelievable fortune a small ¡unk¿ of ¡unk¿ ¡unk¿ costs ¡unk¿ a big executive for the ¡unk¿ company is murdered and police detective ¡unk¿ is on the ¡unk¿ /¿¡br /¿the secret of ¡unk¿ green is not a mystery if you do research on the movie ¡unk¿ is enjoyable to watch but the whole screenplay is a joke it is just as cheap as the entire production the screenplay -negative

an ¡unk¿ although repetitive and rather ¡unk¿ ¡unk¿ of exploitation and ¡unk¿ in a situation where there is no way forward or up where the attempts to make yourself feel better by ¡unk¿ and putting down whoever is below you seems to be the only ¡unk¿ but even here in this ¡unk¿ ¡unk¿ of lost dreams and no future that does not work and ¡unk¿ out to something or someone to ¡unk¿ and share with a simple act of ¡unk¿ gives some ¡unk¿ even if it just makes the present ¡unk¿ by ¡unk¿ memories of the ¡unk¿ /¿¡br /¿although there is little actual on screen violence this is a harsh and brutal film about the small ¡unk¿ of ¡unk¿ ¡unk¿ and personally that does not make for easy entertainment clearly based on a play with a small cast a ¡unk¿ more ¡unk¿ ¡unk¿ to the general social and political environment would possibly have helped the film to reach a ¡unk¿ audience ¡unk¿ ¡unk¿ ¡unk¿ ¡unk¿ ¡unk¿ ¡unk¿ ¡unk¿ ¡unk¿ ¡unk¿ ¡unk¿ ¡unk¿ ¡unk¿ ¡unk¿ in a ¡unk¿ ¡unk¿ ¡unk¿ there is is no no way ¡unk¿ ¡unk¿ ¡unk¿ ¡unk¿ where the ¡unk¿ attempts to to make ¡unk¿ ¡unk¿ ¡unk¿ ¡unk¿ ¡unk¿ ¡unk¿ ¡unk¿ ¡unk¿ ¡unk¿ ¡unk¿ ¡unk¿ ¡unk¿ -positive

6 Conclusions

Multiple experiments were conducted to find out a combination of hyper-parameters that gave the best accuracy on the validation set. The following hyper-parameter combination gave the best result: n-grams:2, optimizer: adam, learning rate: 0.001 with 0.5 decay every 3 epochs, vocab size: 10000, embedding size: 100, batch size: 32, tokenization scehme: no punctuation. The above model achieved an accuracy of 85.18 on the Test dataset.