NLP HW 1: Bag of N-Gram Document Classification

Hetian Bai | hb1500@nyu.edu

October 10, 2018

https://github.com/hb1500/NLP-Movie-Sentiment-Analysis

1 Problem Statement

This is a sentiment analysis of movie reviewers on IMDB dataset using Bag of N-Gram words model in PyTorch. In addition, I conducted an ablation study to find out the effects of hyper-parameters on model performance.

2 Ablation Experiments

In this section, I experimented over hyper-parameters to find out the effect of hyper-parameters on model by comparing the model performance. Following the fashion of ablation study, to study one hyper-parameters (i.e. n-gram), I control all other hyper-parameters fixed.
In this study I set control EPOCH = 3, using early stopping method, training model further leads to overfitting.

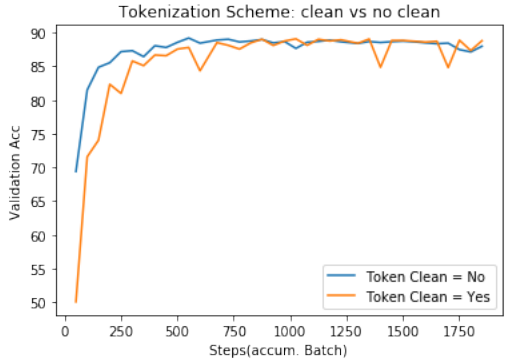


Figure 1: Tokenization scheme

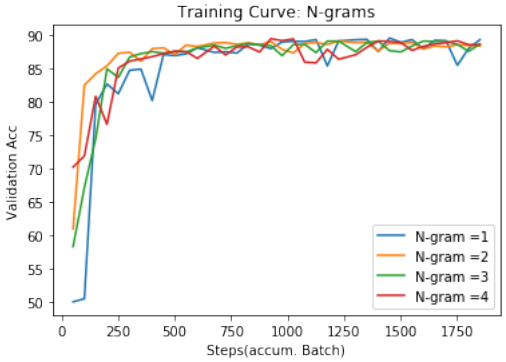


Figure 2: Compare N-grams

2.1 Tokenization Scheme

1. Controls: vocabulary size = 10000; learning rate: 0.1; Optimizer: Adam; word embedding dim = 100
   Compare tokenization with vs without [removing punctuations, change lower case, removing non-English words]
   See figure 1: with token-cleaning: the best validation accuracy is 89.08, better than non-cleaning scheme 89.02.

2.2 Model Hyper-parameters

1. N-Gram:
   Controls: vocabulary size = 10000; learning rate: 0.1; Optimizer: Adam; word embedding dim= = 100;
   Compare n-gram = 1, 2, 3, 4
   From graph 2, n-gram = 2 seems outperforms other models in general, but n-gram = 1 yields the highest validation accuracy (randiness plays role here). The highest validation accuracy rates are 89.62, 89.16, 89.16, 89.52.

2. Vocabulary size:
   Controls: n-gram = 1; learning rate: 0.1; optimizer: Adam; word embedding dim = 100;
   Compare vocabulary size =1000, 10000, 20000

See Figure 3 Vocabulary size is positively correlated with models' performance. Setting higher vocabulary size will increase model's accuracy rate.

3. Embedding size:
   Controls: n-gram = 1; learning rate: 0.1; Optimizer: Adam; vocabulary size = 10000;
   Compare word embedding dimension =50, 100, 200
   See figure 4 It seems that embedding dim does not plays a significant impact on model's perform.

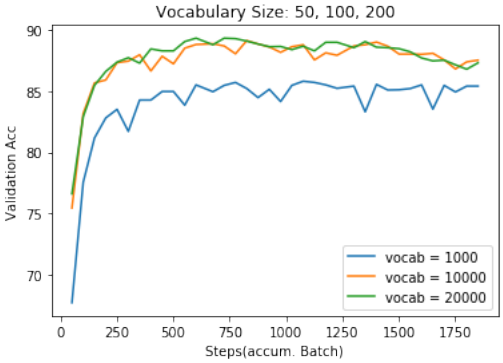


Figure 3: Vocabulary Size

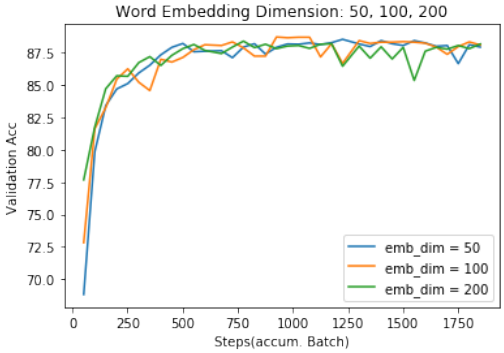
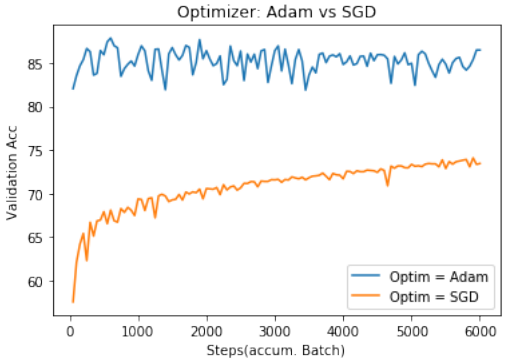


Figure 4: Word Embedding Size

2.3 Optimization Hyper-parameters

1. Optimizer:
   Controls: n-gram = 1; learning rate: 0.1; word embedding dim = 100; vocab. size = 10000;
   Compare fixed Optimizer = Adam vs SGD

Figure 5: Compare Optimizes



See figure 5 As compare from the validation accuracy curves, SGD requires more epochs to reach to a high accuracy rate on validation set. However, Adam performs better in terms of model validation accuracy. Adam is becoming overfit starts from epoch 2. In this setting with fixed learning rate, Adam works better than SGD. It is also interesting to see how SGD performs with adaptive learning rate.

2. Learning rate:
   Controls: n-gram = 1; optimizer: Adm; word embedding dim = 100; vocab. size = 10000;
   Compare fixed lr = 0.1, 0.01, 0.001
   See figure 6 When learning rate is 0.1, the training accuracy curve goes up in a very unstable way. While, when learning rate is 0.001, the learning process is very stable, but the model's over all accuracy rate is compromised. Hence, in this setting, I recommend set learning rate to 0.01 to achieve better results.

3. Learning rate style:
   Controls: n-gram = 1; optimizer: Adm; word embedding dim = 100; vocab. size = 10000;
   See figure 7 Compare fixed lr = 0.01 vs Cosine Annealing learning rate
   Annealing model is better with fewer steps, but both model are same in general. fixed: 89.34; Annealing: 89.58

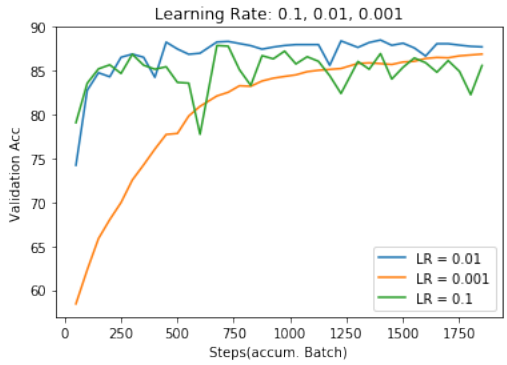


Figure 6: Learning Rate

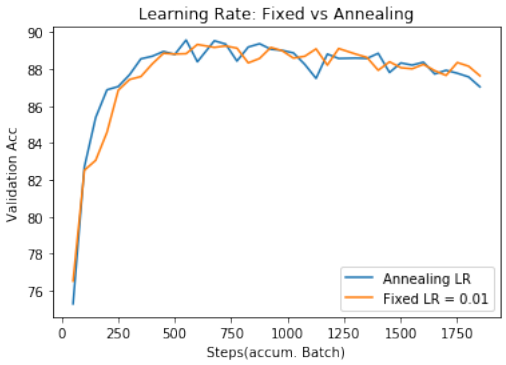


Figure 7: Learning Rate Style

2.4 Additional parameters

1. Batch size (see github readme), Max sentence length
2. Momentum in Adam optimization: by adjusting momentum, the optimizer can give faster convergence.

3 Result

3.1 Test Accuracy

Configuration: n-gram = 1; optimizer: Adm; word embedding dim = 100; vocab. size = 10000, learning rate = 0.01 –> My test accuracy rate is 88.76

3.2 Table Summary

	Hyper-parameters	Compare	Best Validation Acc.
0	Tokenization Scheme	Clean vs no Clean	89.08, 89.02
1	N-gram	1,2,3,4	89.62, 89.16, 89.16, 89.52
2	Vocabulary Size	5000, 10000, 20000	84.88, 89.02, 89.26
3	Embedding Size	50, 100, 200	88.54, 88.72, 88.4
4	Optimizer	Adam vs SGD	74.27, 87.93
5	Learning Rate	0.1, 0.01, 0.001	87.82, 88.44, 86.84
6	Learning Rate Style	Fixed vs Cosine Annealing	89.34, 89.58

3.3 Correct and incorrect samples

Correct:
1. [("there are enough sad stories about women and their oppression by religious political and societal means not to diminish the films and stories about genital mutilation and reproductive rights as well as wage inequality and marginalization in society all in the name of allah or god or some other ridiculous justification but sometimes it is helpful to just take another approach and shed some light on the subject the setting is the 2006 match between iran and bahrain to qualify for the world cup passions are high and several women try to disguise themselves as men to get into the match the women who were caught played by sima mobarakshahi shayesteh irani ayda sadeqi golnaz farmani and mahnaz zabihi and detained for prosecution provided a funny and illuminating glimpse into the customs of this country and most likely all muslim countries their interaction with the iranian soldiers who were guarding and transporting them both city and villagers and the father who was looking for his daughter provided some hilarious moments as we thought about why they have such unwritten rules it is mainly about a paternalistic society that feels it has to save it's women from the crude behavior of it's men

rather than educating the male population they deny privilege and rights to the women seeing the changes in the soldiers responsible and the reflection of iranian society it is nos surprise this film will not get any play in iran but jafar panahi has a winner on his hands for those able to see it", tensor([1]),1)
2. ("moonstruck is a lovely little film directed by superb story teller norman jewison in the heat of the night fiddler on the roof the hurricane the film is great on many levels it shows a good slice of italian culture has a touching romance and best of all is a hilarious comedy one thing i liked most about the film was the relative unconventional looks of the actors nicolas cage looks positively odd for most of the film and cher well cher always looks a little odd overall it's a fun film and easy to recommend 74 out of 10", tensor([1]),1)
3. ("i think i truly love this film prix de beaute was originally a silent film but later dubbed into french in 1930 despite having someone else's voice dubbed over hers this remains a stunning tour de force for louis brooks the fact that her singing voice is dubbed by the legendary edith piaf helps to mollify us purists about the dubbing deception this is the story of lulu and we first see her at a resort with her macho boyfriend andre georges charlia and their friend antonin augusto bandini lulu enters the frame as a pair of legs we see her inside the car changing into her bathing costume lulu is very free with showing off her body and this does not sit well with the irksome andre when lulu considers applying for the title of 'miss europe' we know that a happy ending is not going to be sitting at the end of easy street the film seems to focus a lot on men ogling beautiful women we see plenty of bathing beauties and the reactions of the men staring at them but at the center of it all is the magnificent louise brooks if you don't mind watching films from the bygone eras then consider checking out this one louise brooks is not a name that most average movie buffs may readily know but as soon as you see her you will be mesmerised and you'll want to know more also check her out in 'pandora's box' if you can find it be wary of the us kino dvd release i don't know if their projection speed is correct a lot of the scenes appear to be shown at too fast a speed this may have been the way they were shot i don't know but since it's the only way to see this film it's worth swallowing that one minor bitter pill", tensor([1]), 1)]

Incorrect:
1. [('a series of shorts spoofing dumb tv shows groove tube hits and misses a lot overall i do really like this movie unfortunately a couple of the segments are totally boring a few really great clips make up for this a predecessor to such classics like kentucky fried movie', tensor([0]), 1)
2. ("i love ghost stories in general but i particularly love chilly atmospheric and elegantly creepy british periodstyle ghost stories this one qualifies on all counts a naive young lawyer solicitor in britspeak is sent to a small village near the seaside to settle an elderly deceased woman's estate it's the 1920s a time when many middleclass brits go to the seaside on vacation for their health well guess what there's nothing healthy about the village of crythin gifford the creepy site of the elderly woman's hulking brooding victorian estate which is located on the fringes of a fogswathed salt marsh when the lawyer saves the life of a small girl none of the locals will help the endangered tot you find out why later on in the film he inadvertently incurs the wrath of a malevolent spirit the woman in black she is no filmy gauzy wraith but a solid black silhouette of malice and evil the viewer only sees her a few times but you feel her malevolent presence in every frame as the camera creeps up on the lawyer while he's reading through legal papers you expect to see the woman in black at any moment when the lawyer goes out to the generator shed to turn on the electricity for the creepy old house the camera snakes in on him and you think she'll pop up there too waiting for the woman in black to show up is nailbitingly suspenseful we've seen many elements of this story beforethe locked room that no one enters the fog the naive outsider who ignores the locals' warnings but the director somehow manages to combine them all into a completely newseeming and compelling ghost story watch it with a buddy so you can have someone warm to grab onto while waiting for the woman in black ",tensor([0]),1)
3. ('sunny a cocktail waitress in the dc area is a bit dim to put in mildly she drives an old clunker and rents a tiny room from a gay male couple however she saves the life of a prominent arab by taking a bullet in the behind that was meant for the official she charms the national press with her zany remarks and her sweet looks sniffing an opportunity presidential aides get her installed in the protocol department for the us government even then she messes things up at times but she tries hard and learns a lot she even grabs the romantic attention of a state department official but is there another sinister plot in the making involving an arab man who wishes to take another wife a blonde one lol lol lol this movie features goldie as pretty as a picture and as dumb as a fox as they say sunny learns her way around the jungle of the us government very very well she even has important things to say about honesty and the lack of it in her protocol surroundings perhaps the arab community would be less than thrilled with this work but for those who like to laugh rent this today', tensor([0]),1)]