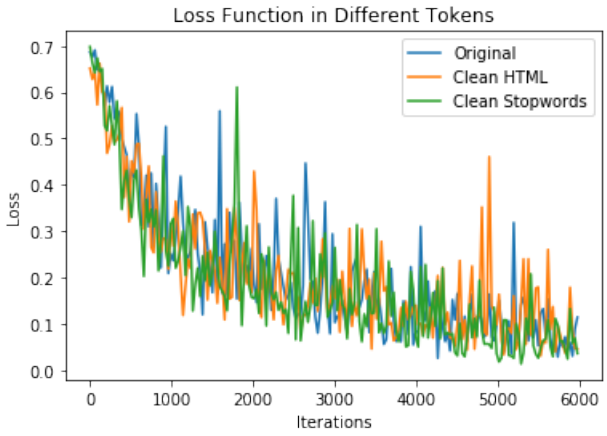
Natural Language Processing - Assignment 1

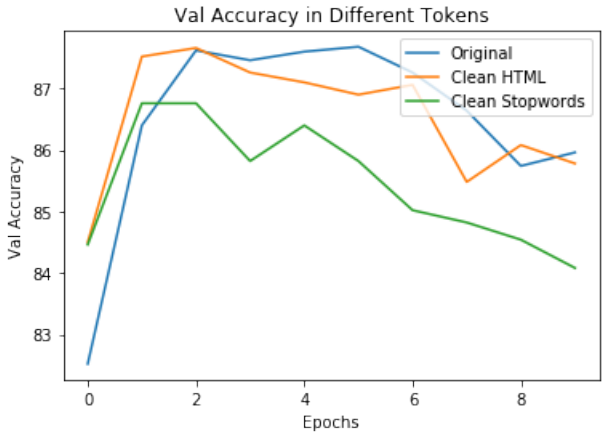
Weicheng Zhu, wz727

October 10, 2018

Source Code Repository: https://github.com/jackzhu727/NLP-1011.git

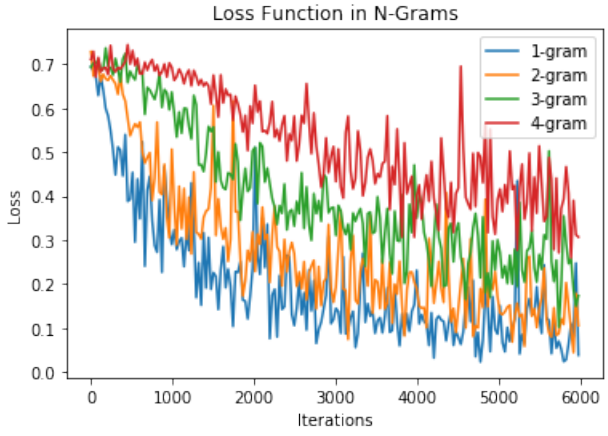
Tokenization schemes I compared three ways of tokens: first, I cleaned the punchuations and lowered case of letters in sentences. Then I observed that there are still a lot of HTML tags in paragraphs, which are meaningless to the sentiment analysis, so I applied BeatifulSoup to clean tags like breaks. After that, I also cleaned stopwords, like 'this', 'is', 'a' and etc., which are with no real meanings. The training curves show that the validation accuracy of first and second scheme are the highest. However, if we delete stopwords, the accuracy will be lower, probably because of more sparse feature vectors. Based on the comparsion, token that cleans html tags will be used in further experiments.

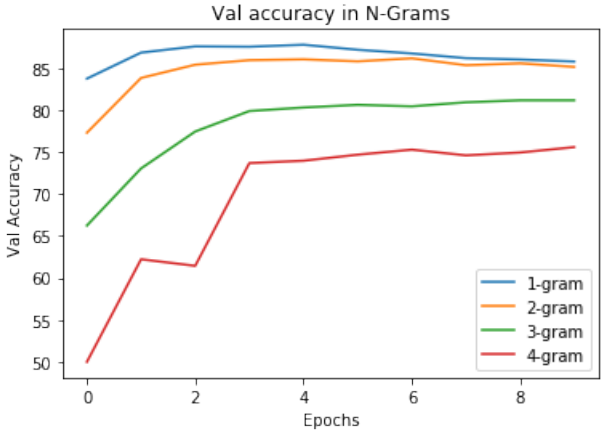




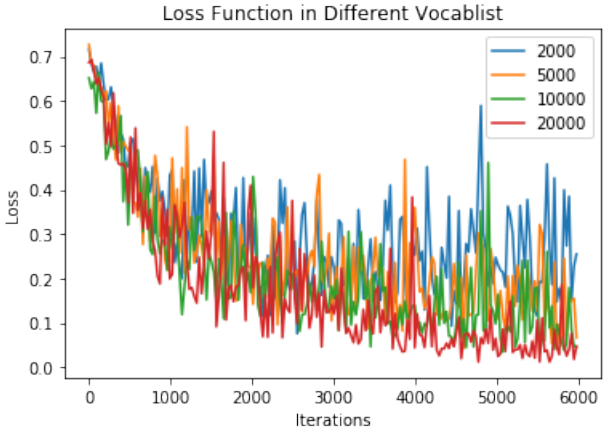
Model hyperparameters

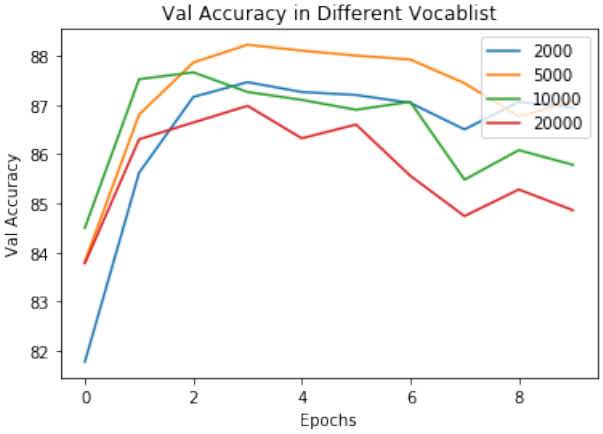
n-gram I splited tokens with 1,2,3 and 4 grams and experimented the model with this dictionaries. From these figures, we can learn that 1-gram and 2-gram have better performance, and 1-gram obtains highest validation accuracy. This probably because 2 or more grams may connected unrelated words and leads to more sparse feature vectors. Thus, I used 1-gram in future experiments.



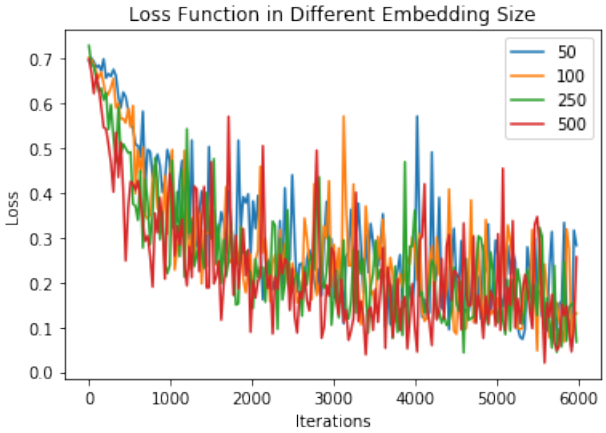


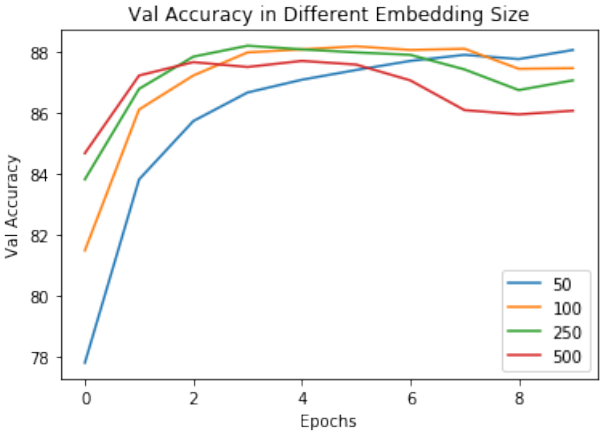
Vocabulary size From these figures, the maximum vocabulary size of 5000 has best performance, because small vocab list holds less information, while large vocab list holds too much distracting information. Size of 5000 holds the proper amount of information required to fit the model.





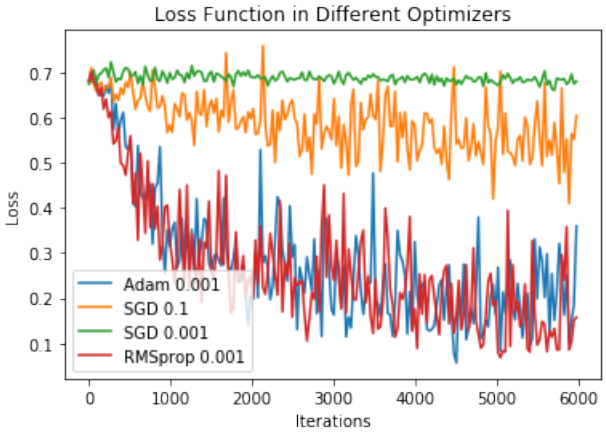
Embedding size I experienced with 4 embedding sizes, and figures imply that large embedding size make the validation accuracy increases fast in first several epoches but easily overfitting, because a larger embedding matrix can carry more information of words at same time. To balance between convergence speed and over-fitting, I chose a reasonable embedding size of 100.

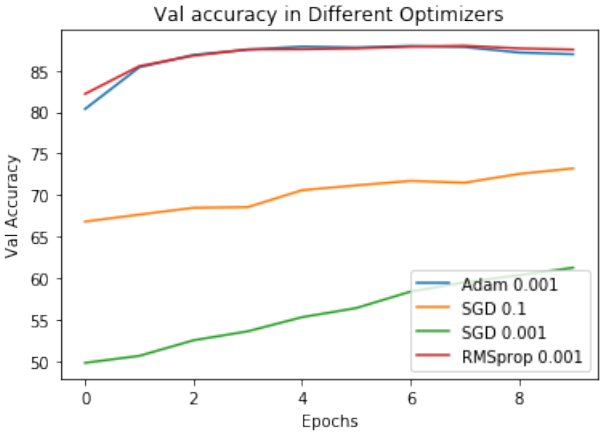




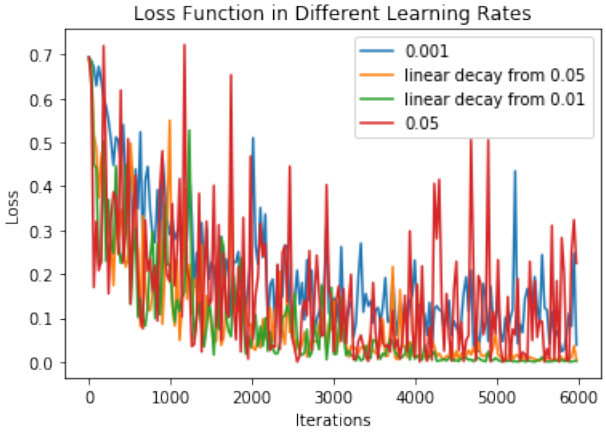
Optimization hyperparameters

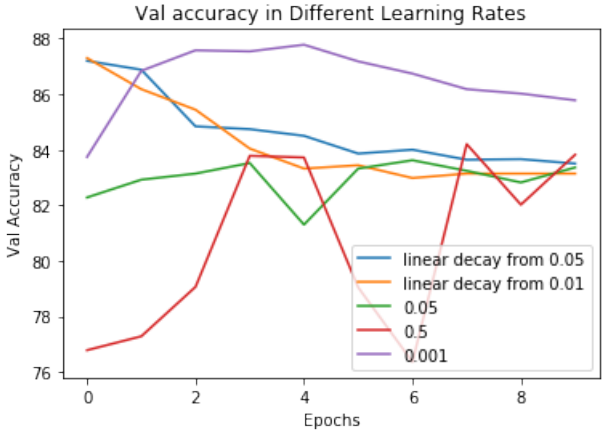
Optimizer I compared values of empirical lose function and validation accuracy with different optimizers. RMSprop and Adam have good performance, but SGD has much lower convergences speed. Even though I used a larger learning rate, SGD still converges much slower than Adam and RMSprop.





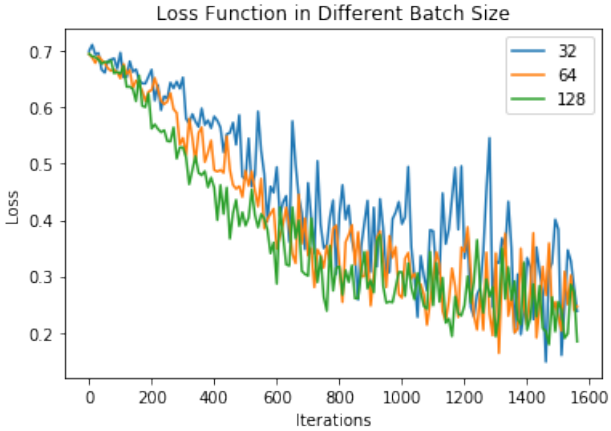
Learning Rate I compared training loss using different learning rates: $\alpha = 0.001, 0.05, 0.5$, linear decay from 0.01, linear decay from 0.05. From the loss curve, we learnt that with learning rate 0.5, the accuracy is low and loss function was volatile; with learning rate 0.001, the loss converges slow, but we have the highest validation accuracy; with linear decay learning rates, the loss function has both fast convergence and less long-term volatility, and validation accuracy is high. From the validation accuacy curve, the accuracy decreases after the maximum, it implies the overfitting of the model. Thus, 0.001 leads to best validation accuracy; linear decaying learning rate lead to highest efficiency.

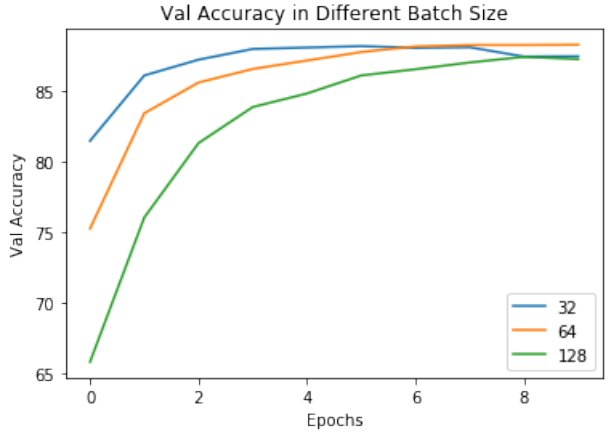




Other Hyperparameters(BONUS)

Batch Size I experimented with several different batch sizes. Large batch sizes cost less time each epoch, but they need more epochs to obtain minimums; small batch sizes are in the contrary. From figures, batch size of 64 have the fastest convergence speed considering number of epochs and time cost per epoch.





I also tried to build the neural network by adding non-linear layers, like ReLu, Tanh and Sigmoid, and I found the Tanh has best performance that improves the accuracy of a single linear layer.

Overall Performance I used token that cleans HTML tags with 1-gram, ADAM optimizer with learning rate 0.001, neural network with a Tanh layer and linear layer, 64 batch size and bag of word model with vocabulary list of 5000, embedding matrix of 100.
After training for 10 epochs, the model reached following accuracy:
Val Acc 88.22; Test Acc 87.968

Correct Examples Texts:
1. Predict: Neg, Label: Neg
a christmas story is one of many people 's all time most beloved films acs was able to take the viewer to a time and a place in such a way that very few films ever have it had a sweetness and goodwill to it that is rare so i awaited and awaited its sequel it runs in the family the film was almost released a couple of times

only to be pulled at the last minute when it finally came out iritf was and is i guess a total failure the sets and cinematography were just fine but the directing totally completely missed the mark the film was nothing more than a cash flow formula of lazy casting lazy writing and disconnected acting the narrator jean shepard who was one of america 's great humorists and story tellers forced upon us a false reprise of the warm wit he used in acs he over emoted and why he did that i 'll never know he somehow managed to become an annoying overwrought parody of himself the writing and acting in iritf is inauthentic and forced the actors may have seen acs but whatever wit and nuance that was in acs must n't have registered at all on any of them the acting was embarrassingly slapstick and bereft of any of shepard 's dry humor acs will always be a real treasure but to call iritf a sequel is to insult all of the fans of jean shepard and acs
2. Predict: Pos, Label: Pos
the acting in this movie was superb but mixed with the truth about the condition of many africans in south africa made it heart wrenching it was good that the writer isolated boesman and lena from others run from their homes so we could share fully in their triumphs and defeats the conflicts they shared as they grew together and apart worth seeing when you put the movie in it 's proper context
3. Predict: Pos, Label: Pos
an adult realistic cruel dark story like a second part of les roseaux savages the wild reeds plenty of beauty and sadness ellipsis and silences shadows and little sparks of hope a man searching for a warm companion a better life a sincere attitude

Wrong Examples Texts
1. Predict: Pos, Label: Neg
this is another north east florida production filmed mainly in and near by to fernandina beach and the kingsley plantation i was rather surprised the company was able to take over the main street of fernandina beach as long as was necessary to achieve the street scenes the film is pretty and pretty bad tami erin is cute but overacts eileen brennan overacts even more good for small kids or for those who like fluff in large doses a 4 from the miller movies formula
2. Predict: Pos, Label: Neg
i very much looked forward to this movie its a good family movie however if michael landon jr. 's editing team did a better job of editing the movie would be much better too many scenes out of context i do hope there is another movie from the series they 're all very good but if another one is made i beg them to take better care at editing this story was all over the place and did n't seem to have a center which is unfortunate because the other movies of the series were great i enjoy the story of willie and missy they 're both great role models plus the romantic side of the viewers always enjoy a good love story
3. Predict: Neg, Label: Pos
gung ho tries to express many ideas and entertain us with a wiseguy comedy at the same time the result is uneven but generally entertaining keaton balances all three aspects of his lead character quite well wantabedde is even better one warning george wendt is very poor in his supporting role otherwise this is quite enjoyable time capsule

1-10 Score Classification(BONUS) I also ran the classification on 1-4 and 7-10 scores. After training for 20 epochs, val Acc 42.02, test Acc 42.58.
After analyzing the prediction result of the valiation set, I found most of the prediction are correct in positive or negative attitudes. Most of the descrapencies are like predicting a 1-score comment as 0. This problem is hard to solve, because even a human can hardly tell whether a comment is 0-score or 1-score. Thus, the test accuracy of 42.58 is acceptable.