Gary Ng
DS-GA 1011: NLP with Representation Learning

Assignment 1: Predicting Movie Review Sentiment

Please see accompanying code in Jupyter notebook at
https://github.com/nyjgary/nlp-sentiment-analysis/blob/master/nlp-sentiment-analysis.ipynb

I. Introduction

We aim to build a bag of n-grams model with embeddings for predicting the sentiment of the movie reviewers given their textual review of the movie. The dataset comprises 25,000 train and 25,000 test movie reviews scrapped from IMDB website, provided by Maas et. al. (2011)[1]. Each review is assigned a binary label: positive (label=1) if the review score is at least 7/10 or higher, or negative (label=0) if the review score is 4/10 or worse. In this assignment we build a binary classifier to predict whether the review is positive or negative.

II. Model Building

We begin our modeling process by first splitting the 25,000 train data into 20,000 train samples and 5,000 validation samples, using the latter to assess our model accuracy when selecting the optimal tokenization scheme, optimization hyperparameters and model hyperparameters. The actual test reviews are only used to evaluate the final model at the end.

A. Tokenization Scheme

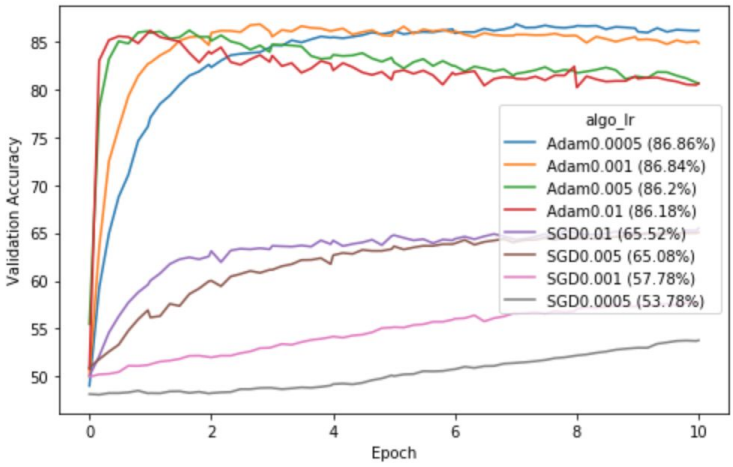
First we need to select a tokenization scheme to process the text data. The six schemes described below were attempted. The one that performed best on validation data involved converting the text to lowercase, removing punctuation, and then removing NLTK stopwords. Note that NLTK's was explored as an alternative to default SpaCy's stopwords because the latter contains a number of words that might be important for n-grams, e.g. 'not' and 'very'. In comparison, NLTK's stopword list appears more conservative, and indeed turns out to work slightly better than SpaCy. Additionally, lemmatization was also explored as another variation, but it did not appear to help, yielding worse accuracy on the validation dataset. We will select the winning tokenization scheme and use it for hyperparameter tuning.

Tokenization Scheme	Validation Accuracy
Lowercase only	84.52%
Lowercase + Removed Punctuation	85.02%
Lowercase + Removed Punctuation + Removed (SpaCy) Stopwords	86.06%
Lowercase + Removed Punctuation + Removed (NLTK) Stopwords	**86.20%**
Lowercase + Removed Punctuation + Removed (Spacy) Stopwords + Lemmatization	85.54%
Lowercase + Removed Punctuation + Removed (NLTK) Stopwords + Lemmatization	85.32%

[1] http://ai.stanford.edu/~amaas/data/sentiment/

B. **Optimization Hyperparameters**

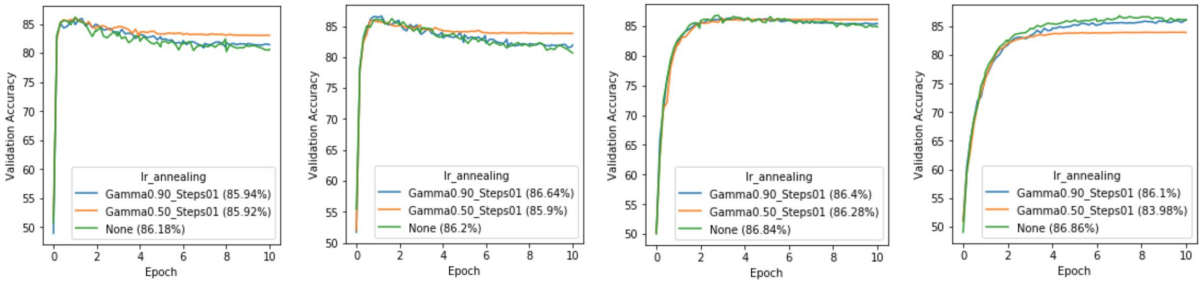
Next we explore optimization hyperparameters, namely the *optimization algorithm* itself, *learning rate*, and *learning rate annealing*. We are interested in finding a set of optimization hyperparameters that allows us to train effectively but quickly, so that we can search over combinations of model hyperparameters efficiently. We trained a model with each combination of optimization algorithms $\in$ [Adam, SGD] and learning rates $\in$ [.01, .005, .001, .0005], along with an arbitrary but constant set of model hyperparameters (namely: emb_dim=100, max_ngram=1, max_sentence_length=200, max_vocab_size=10000). Each model's learning curve is plotted below, along with its best validation accuracy recorded in the legend.



We chose to move forward with Adam optimizer because it trains much faster than SGD. Next, we investigate the effects of learning rate annealing. For each learning rate (LR) $\in$ [.01, .005, .001, .0005], we test it with two LR schedules, one in which the LR is multiplied by gamma=0.9 and other by gamma=0.5 every epoch. The results are plotted below. Not surprisingly, LR annealing smoothes out the learning curves, reduced overfitting in cases where an aggressive LR (e.g. LR=[0.01, 0.005]) was chosen, but slowed down training in cases where a conservative LR (e.g. LR=0.0005).

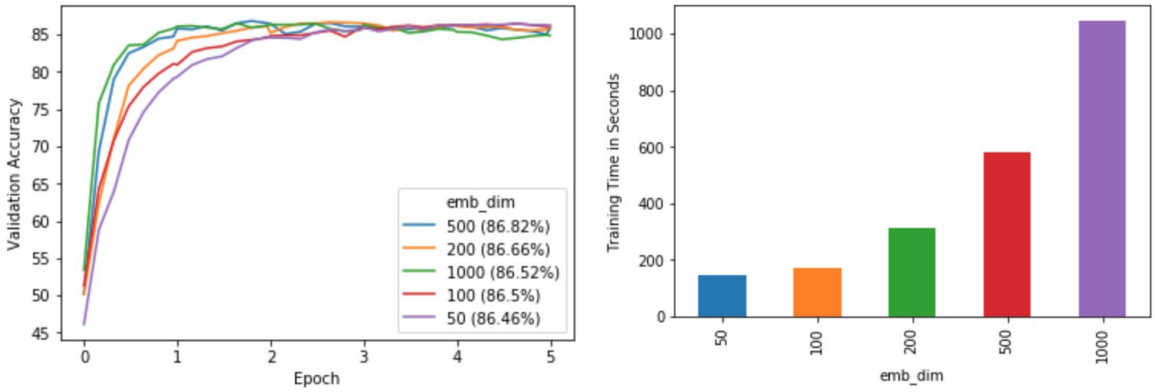
As far as tuning model hyperparameters go, we want to train quickly as well as base our decision on the best validation accuracy accorded. Since LR annealing slows down training, and yielded slightly lower best validation accuracy for each learning rate, we choose to move forward without LR annealing. We will however use LR annealing when we train our final model after the hyperparameters are chosen over much longer epochs to prevent overfitting.

Adam; LR = 0.01	Adam; LR = 0.005	Adam; LR = 0.001	Adam; LR = 0.0005



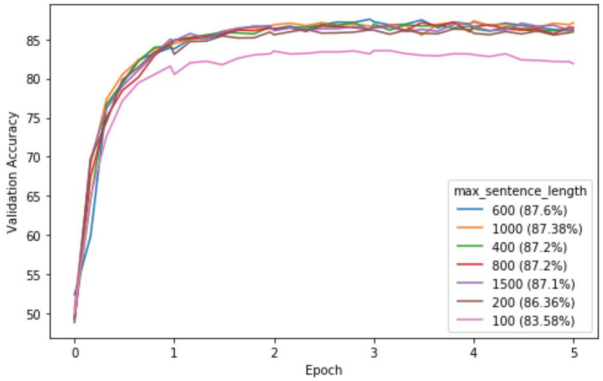
C. Model Hyperparameters

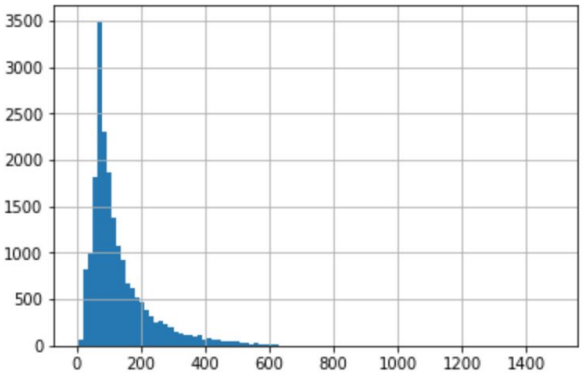
Finally, we explore the model hyperparameters. First we train models using different values of ***emb_dim*** ∈ [50, 100, 200, 500, 1000], with a moderately small LR of a moderately small LR of 0.001 with no annealing. Larger embeddings learned quickly, but the best validation accuracies obtained are actually remarkably similar, falling within a narrow range under 0.4% (86.46 - 86.82%). Since larger embeddings take longer to run, we use emb_dim=200 in subsequent hyperparameter tuning, but will use the best-performing emb_dim=500 in our final training model.



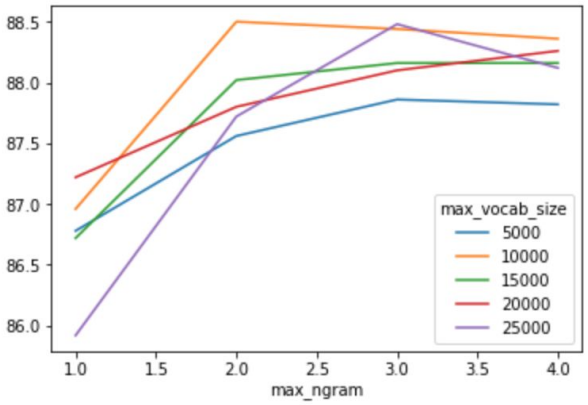
Next we address ***max_sentence_length***, which is a bit of a misnomer. It is in fact the maximum number of tokens that were taken into consideration for each data point, such that the post-padding training data is of the dimension (n, max_sentence_length). While this was not suggested as a hyperparameter to tune, it in fact has a huge impact because in the sample code, max_sentence_length was hard-coded as 200 by default, but the distribution of the tokenized (1-gram) dataset suggests that a nontrivial number of data points are longer than 200-tokens, which meant valuable information might be effectively left out. We experimented with a range of max_sentence_length ∈ [100, 200, 400, 600, 800, 1000, 1500], and 600 (just enough to cover the full text for vast majority of reviews) yielded the best performance, nearly 1.4ppt better than the baseline performance for length 200.

We will use max_sentence_length=600 for our baseline 1-gram model, but will expand to max_sentence_length=600*max_ngram for larger n-gram models. This is necessary because our dataset, which is a concatenated list of [1-grams; … ; n-grams], scales with max_ngram. Since we derived the insight from previous experiments that it is optimal to choose high enough max_sentence_length to cover the full text for vast majority of reviews, we shall adhere to this principle by scaling max_sentence_length with max_ngram.





Finally, we tune ***max_ngram*** and ***max_vocab_size*** together, since they are naturally related. Here we try 20 combinations of max_vocab_size ∈ [5K, 10K, 15K, 20K, 25K] and max_ngrams ∈ [1, 2, 3, 4]. The chart below summarizes the best validation performance for each combination. We see that expanding beyond 1-grams greatly improved our model accuracy. max_ngram=3 slightly edged out max_ngram=2 and max_ngram=4 in average performanced aggregated over the different values of max_vocab_sizes attempted. However, the top combination was (max_ngram=2, max_vocab_size=10K). While we probably could have tried even larger values of max_vocab_size for max_ngram=3 (its performance peaked at max_vocab_size=25K, which is the largest value attempted) to rival max_ngram=2, we are content with the simple (max_ngram=2, max_vocab_size=10K) model that performs exceedingly well. We proceed to use this in our final model.



III. Model Evaluation

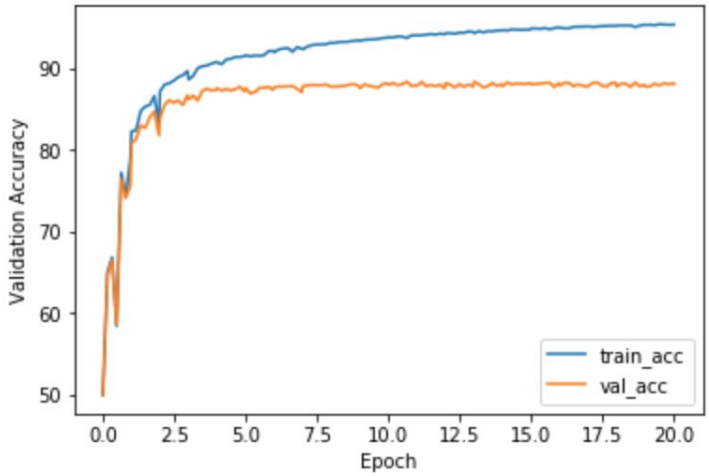
A. Final Model

We trained our final model for 20 epochs using the following set of hyperparameters:

- Tokenization: Lowercase + Removed Punctuation + Removed (NLTK) Stopwords
- Max N-gram = 2 (i.e. unigrams + bigrams)
- Max Sentence Length = 1200 (i.e. 600 * Max N-gram)
- Max Vocab Size = 10,000
- Embedding Dimension = 500
- Optimization Algorithm = Adam
- Learning Rate = 0.0005
- Learning Rate Annealing = Scale LR by 0.9 every epoch

B. Learning Curve

With learning rate annealing implemented in the final model, our validation accuracy plateaus off at ~88% rather than dip previously due to severe overfitting. As such, we can use the final trained model (after 20 epochs) without resorting to early stopping.



C. Validation Examples

- *(Positive: Correctly Predicted) "This movie is so great. Its set back in probably the 40's and Meg Ryan's character struggles to be known as 'smart.' Plus Tim Robbins is so cute in this movie. And everything about it has a magical feeling towards it. Everytime I watch it I feel happy. It's definitely a girl movie, and I'm a girl, so I like it. I also love the music. The violin is awesome. but besides that I think it's a cute story and everyone should watch it."*
- *(Negative: Correctly Predicted) "A dedicated fan to the TLK movies, with the first one being a milestone and the second probably the best sequel Disney has produced, along comes this film... Now I'm not arguing with animation, voice work, music, but this is no more than a Timon/Pumbaa screwloose in the TLK atmosphere. Although it isn't bad, it doesn't add anything. Basically this movie is one big joke... and that's about all that saves it. Make a real TLK3, Disney! The potential is there.

4/10"*
- *(Negative: Correctly Predicted) "This is, quite literally, the worst movie I have ever watched in my life. It may be the worst movie possible. Some movies are so bad that they're good; this movie is so bad that it goes past enjoyable camp and simply becomes unwatchably awful. It is the anti-enantiodromia. We bought it with the intent to heckle, and all of my family gathered around for a fun evening of clever remarks; instead, we sat in stunned silence, pitying poor Peter Sellers.

This is worse than the animated Lord of the Rings. It is worse than the Matrix sequels. It is worse than Krull. It is worse than any Batman movie.

Do not, under any circumstances, let this movie approach within ten feet of your television."*
- *(Positive: Incorrectly Predicted as Negative) "Even without speaking a word, Billy Connely is wonderful as a zombie... Carrie Ann Moss as "Mom"?, even better. Zombie girlfriends?

"...My father thied to eat me... I never tried to eat Timmy."

And I thought Dawn of the Dead was good. It\'s kinda like Airplane meets (meats?) Night of the Living Dead, sponsored by Zomcom..

And don\'t forget my head coffin

And Fido in an Aloha shirt is just way cool!

And yes, the social comment is just too much to even begin to comment on.

Sufice it to say, it all really works!"*
- *(Positive: Incorrectly Predicted as Negative) "so halfway through the season, i got so caught up in school and my activities that i didn't realize that the show had been canceled halfway through, which is crap.

i think the followers of this show should write fox and ask them to at least finish filming so that a the season can be released on DVD later. maybe then they'll see how many people were disappointed that the show didn't survive its first season.

i loved the show and looked forward to it every thursday after the OC. can you imagine my disappointment when i came back to try and watch the show only to discover that it had disappeared? needless to say, i'm not very happy with fox right now. even more so after discovering that NO ENDING WAS FILMED. i mean, if you're going to work on a project, at least finish it to see what happens. a half filmed show is like a half made car, it's pretty much useless. fox, film the damn ending and give some of the show's fans some peace."*
- *(Negative: Incorrectly Predicted as Positive) "The first episode set the bar quite high i thought. It starred William Hurt as a hit-man who is contracted to kill a toymaker. We are given very little information on his character or who is paying him to kill, indeed the episode is notable for having no dialogue at all. Returning to his modernist penthouse he is delivered a package containing toy soldiers, this gives him a smile but he dismisses it and goes about his business. But he is in for a night of hell, the soldiers are alive and are about to wage war, driving jeeps, shooting machine guns and bazookas and even flying helicopters!. The special effects are good for a TV show and it becomes quite tense as he dodges around the apartment using his wits to survive, sometimes getting the upper hand and other times not. I wont spoil the ending but suffice to say it was a clever little twist. This gave me hope for the rest of the series but i was in for a disappointment, the other episodes were all rubbish and i lost interest by the fourth one. Stephen King adaptations are always a mixed bag and these are no exception"*

D. Test Performance

Finally, our model yielded 88.58% accuracy on the unseen test set of 25K examples.