Fall 2018 NLP Assignment 1

adc563

October 10, 2018

1 Code

The code can be found on the following **GitHub page**. **Assignment-1.ipynb** is the notebook used for preprocessing and a part of the hyperparameter search. **prince.ipynb** (on the same folder) is the notebook used for the other part of hyperparameter search which was done on prince.

On the notebook called *plots*, you can find the code that was used to merge the hyperparameter search results and to generate the plots in the report.

2 Preprocessing

The data consists of 50,000 movie reviews from Imdb. There are 25,000 training and 25,000 test examples. The training set was divided into 20,000 training and 5,000 validation samples.

The tokenizing function takes as input n (in ngrams) and movie reviews as text, and comes up with ngram tokens of those reviews. For instance, if n equals 3, the output will be a union of uni-grams, bi-grams, and tri-grams of the review.

During preprocessing, all tokens are converted to lowercase and special symbols are removed from the text.

3 The Model and Training

The Bag-of-Ngrams model is trained for 10 epochs during each training period (i.e. for each new set of parameters). After every 100 iterations in each epoch, the accuracy is calculated on the validation set with the updated parameters.

4 Hyperparameter Search

The hyperparameter space for each parameter used in the study is shown below. The dimensions of the search space is not ideal. Without time and resource constraints, it would be the first thing to be improved in the training process. For example, even the nodes on Prince couldn't read the 4-gram chunks, so eventually 4-grams were excluded.

Table 1: Hyperparameter Space

Parameter	Value Set
n-grams	n$\leq$ n $\in$ {1,2,3}
Optimizer	{Adam, SGD}
Learning Rate	{1e-5,1e-4,1e-3,1e-2,1e-1,1,10}
Vocab Size	{25000, 50000, 100000}
Embedding Size	{100, 200}
Batch Size	{32, 64}

Below are the 3 best models in terms of validation accuracy for small-batch models. The models were ranked in terms of their 30th iteration validation accuracy, while it took the small-batch models 60 iterations to converge.

Table 2: Three Best Small-Batch Models

LR	n	vocab s	emb size	sen length	batch	optimizer	epoch 30	epoch 20	epoch 60
0.001	2	25000	200	200	32	Adam	*85.80*	84.78	86.6
0.001	3	100000	200	200	32	Adam	*85.64*	84.74	86.7
0.001	2	50000	200	200	32	Adam	*85.64*	84.74	86.6

The 3 best large-batch models (according to 20th-epoch ranking) are shown below.
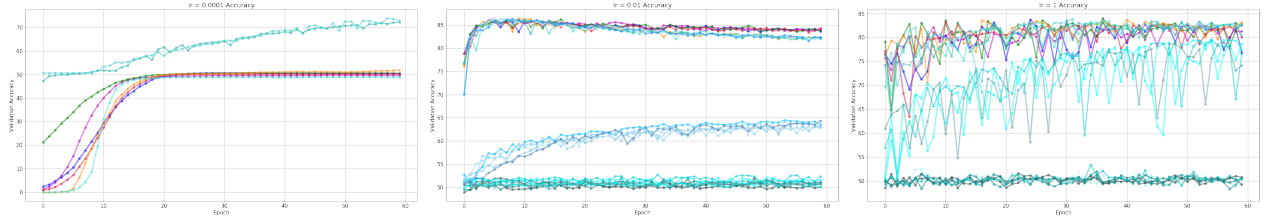
Table 3: Three Best Large-Batch Models

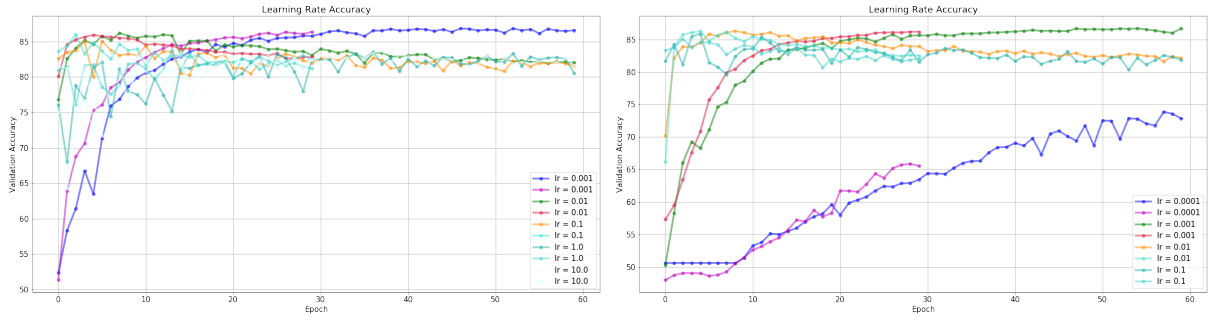| LR	n	vocab s	emb size	sen length	batch	optimizer	epoch 30	epoch 20	epoch60
0.001	2	50000	200	200	64	Adam	86.50	*85.74*	-
0.001	2	50000	100	200	64	Adam	86.44	*85.68*	-
0.001	2	25000	200	200	64	Adam	85.38	*85.62*	-

5 Ablation Study

5.1 Learning Rate

Below, the validation accuracy of each model in the parameter space was plotted against 3 different learning rates: $(10^{-4}, 10^{-2}, 1)$
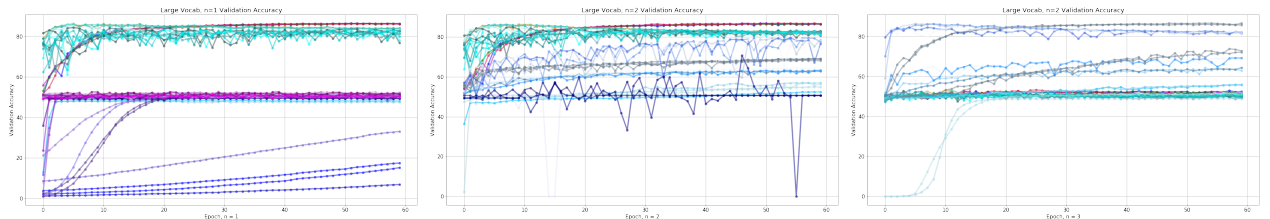

To generate the plot on the (below) left, the parameter space was first set to ($n = 2, vocab\_size = 25000, , emb\_size = 200, sentence\_length = 200, batch = 32, optim = "Adam"$) and the validation accuracy was plotted for all learning rates, keeping the other parameters fixed. Similarly, for the one on the right, the validation accuracy was plotted for all learning rates, keeping the parameters fixed at: ($n = 3, vocab\_size = 100000, , emb\_size = 100, sentence\_length = 200, batch = 32, optim = "Adam"$).
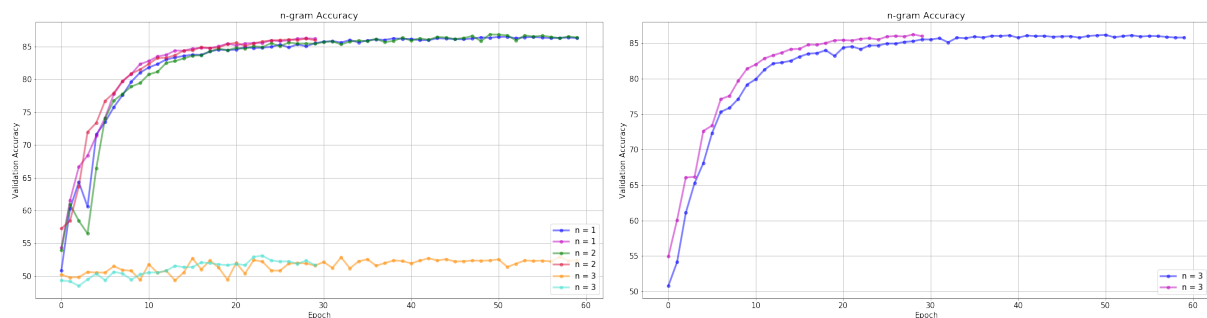

5.2 ngrams

The three figures below show the validation accuracies of models with unigrams, bigrams, and trigrams from the left to the right, respectively.
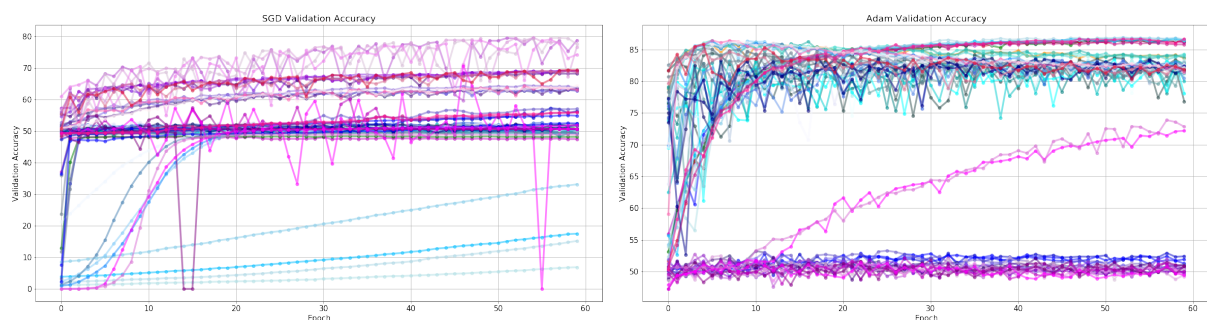

The graph on the left shows all n-gram accuracy scores with small vocabulary size (25 to 50K), and the graph on the right shows trigram accuracies with large vocabulary size (for the same set of parameters except vocab_size).
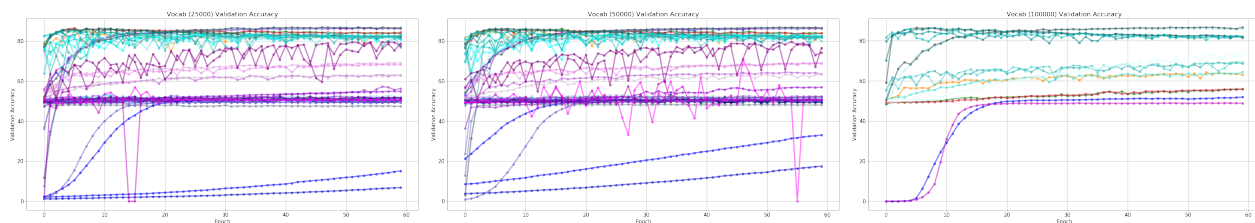
5.3 Optimizer

The graph on the left shows the validation accuracy scores of all the models using SGD as optimizer, and the one on the right is the same plot for the models using Adam. As you might have noticed, for both optimizers, there are some models that do not seem to maximize validation accuracy at all. Later, it was discovered that this is due to the 3-gram models not converging for small vocabulary sizes in the hyperparameter space.

5.4 Vocabulary Size

The below figures show the validation accuracy for different vocabulary sizes. Note that the 3rd graph only includes 3-grams, while the first two include all ngrams. The trigrams on the first two plots are struggling to maximize validation accuracy, whereas most trigrams on the third plot have good results.

6 Test Performance

The following model is selected for test, based on validation accuracy. Please note that it is the 2nd, not the 1st model in the ranking, however it has the largest difference between its epoch_0 accuracy (50%) and halfway (stopped early - 85.6%) accuracy. Therefore, it was considered as the 'best learner' and chosen for the test.

Table 4: Test Model

LR	n	vocab s	emb size	sen length	batch	optimizer	val (30)	val (20)	val (60)
0.001	3	100,000	200	200	32	Adam	85.64	84.74	86.7

The model yields 85.11% test accuracy. For further details, please refer to the last section of Assignment1.ipynb.

6.1 Correct & Incorrect Predictions

6.1.1 Correct

1) "i saw this film in a sneak preview and it is delightful the cinematography is unusually creative the acting is good and the story is fabulous if this movie does not do well it wo n't be because it does n't deserve to before this

film i did n't realize how charming shia ¡unk¿ could be he does a marvelous self contained job as the lead there 's something incredibly sweet about him and it makes the movie even better the other actors do a good job as well and the film contains moments of really high suspense more than one might expect from a movie about golf sports movies are a dime a dozen but this one ...

2) "maybe i 'm reading into this too much but i wonder how much of a hand ¡unk¿ had in developing the film i mean when a story is told casting the main character as himself i would think he would be a heavy hand in writing ¡unk¿ etc and that would make it a little ¡unk¿ /¿¡br /¿but ... his family and friends also may have had a hand in getting the actual details about ¡unk¿ 's life i think the best view would have been told from ¡unk¿ 's family and friends perspectives they saw his transformation and were n't so messed...

3) "i loved this movie from beginning to end i am a musician and i let drugs get in the way of my some of the things i used to ¡unk¿ drawing but my friends were always there for me music was like my ¡unk¿ life support and my drug it changed my life i can totally relate to this movie and i wish there was more i could say this movie left me speechless to be honest i just saw it on the ¡unk¿ channel i usually hate having satellite but this was a ¡unk¿ of having satellite the ¡unk¿ channel shows some really great movies and without it i never would have found this movie ¡unk¿ not a big fan...

4) "i have certainly not seen all of jean ¡unk¿ 's films but they mostly seem to be bloody vampire naked women ¡unk¿ which if you like that sort of thing is not bad but this is a major departure and could almost be cronenberg minus the bio mechanical nightmarish stuff except it 's in french with subtitles of course a man driving on the road at night comes across a woman that is in her ¡unk¿ and ¡unk¿ and picks her up while in the background yet another woman lingers wearing nothing as they drive along it 's obvious that there is something not right about the woman in that she forgets things almost as quickly as ...

5) "even if you 're a fan of jean ¡unk¿ 's idiosyncratic body of work you will be caught off guard by this exceptional ¡unk¿ into science fiction territory for once there 's not a single ¡unk¿ ¡unk¿ vampire girl in sight true to tradition the budget proved way too tight to realize the director 's vision entirely yet this is largely compensated by his obvious love of genre cinema dedication to his craft and sheer ingenuity jean claude ¡unk¿ 's atmospheric cinematography makes the most of the foreboding locations and philippe ¡unk¿ a k a gary ¡unk¿...

6.1.2 Incorrect

1) "actor turned director bill paxton follows up his promising debut the gothic horror frailty with this family friendly sports drama about the ¡unk¿ u.s. open where a young american ¡unk¿ rises from his humble background to play against his ¡unk¿ idol in what was dubbed as the greatest game ever played i 'm no fan of golf and these scrappy underdog sports flicks are a dime a dozen most recently done to grand effect with miracle and cinderella...

2) "i was fortunate to attend the london premier of this film while i am not at all a fan of british drama i did find myself deeply moved by the characters and the bad choices they made i was in tears by the end of the film every scene was mesmerizing the attention to detail and the excellent acting was quite ¡unk¿ /¿¡br /¿i would have to agree with some of the other comments here which question why all ...

3) "i was surprised how much i enjoyed this sure it is a bit slow moving in parts but what else would one expect from ¡unk¿ also there is plenty of nudity nothing wrong with that particularly as it includes lots of the gorgeous brigitte ¡unk¿ there are also some spectacularly ¡unk¿ female dead bit more ¡unk¿ perhaps but most effective there is also a sci fi like...

4) "this movie is amazing because the fact that the real people portray themselves and their real life experience and do such a good job it 's like they 're almost living the past over again ¡unk¿ ¡unk¿ plays himself an actor who quit everything except music and drugs struggling with depression and searching for the meaning of life while being angry at everyone especially the people who care for him most there 's moments in the movie that will make you wanna cry because the family especially the father...

5) "felix in hollywood is a great film the version i viewed was very well restored which is sometimes a problem with these silent era animated films it has some of hollywood 's most famous stars making cameo animated appearances a must for any silent film or animation enthusiast ¡unk¿ in hollywood hollywood is is a a great...