NLP Homework1

Yiyan Chen

October 2018

1 Introduction

This homework is on one of the most classic NLP models, n-gram model. This homework used the IMDB movie review dataset provided by Stanford University to predict the sentiment of the movie review.

2 Details

2.1 Tokenization

The data was downloaded by the link attached by the assignment. In the train folder, there are pos(the positive reviews) and neg(the negative reviews) two sub folders containing multiple txt files. I assigned 1 value as the pos indicator and 0 as the neg indicator. Then I shuffled the training list and train_test_split to get the 20000 train values and 5000 validation values. Then in the prepossessing step, I omitted the punctuation, special characters, HTML tags and English stop word. And transform the words into their lemmenization form and lower case. I used the spacy.load('en_core_web_sm') as the tokenizer and tokenize train, test, val lists and dump them in pickle format.

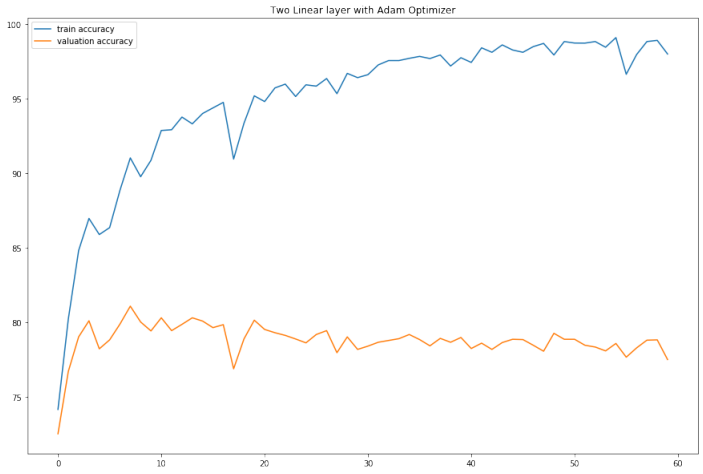
2.2 N-gram

I used the nltk.ngrams function to perform n-gram on train, test, val. I set the max_vocab_size = 10000 which means the vocabulary list has the top 10000 most common word. The words not in vocabulary list set as ¡unk¿. Then set the tokens in train, test, val with the corresponding number using token2id dictionary. And the index can also be transformed to the token. Then the max_sentence_length is 200. With token size that are longer than the max_sentence_length, the tokens list extract the max_sentence_length. With token size that are shorter than the max_sentence_length, the tokens list will be padded with zero. The default batch size is set to be as 32.

2.3 modelling

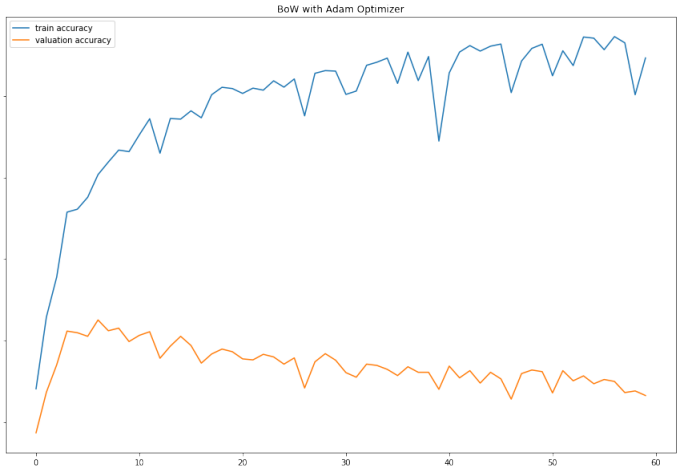
I used the BagOfWords() model with one linear layer and the NGramsModel() with two linear layers with a relu funcion in between. The loss function cross entropy loss. I have tried the two optimizers: Adam and SGD. The results look like the following:

```
After training for 10 epochs
Val Acc 78.86
Test Acc 77.092
```



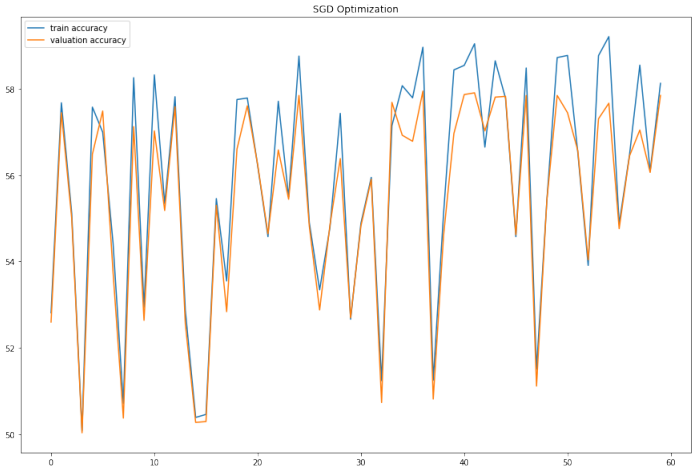
The NGramsModel with two linear lays and Adam Optimizer

```
After training for 10 epochs
Val Acc 76.88
Test Acc 75.156
```



The BagOfWord model with one linear lay and Adam Optimizer

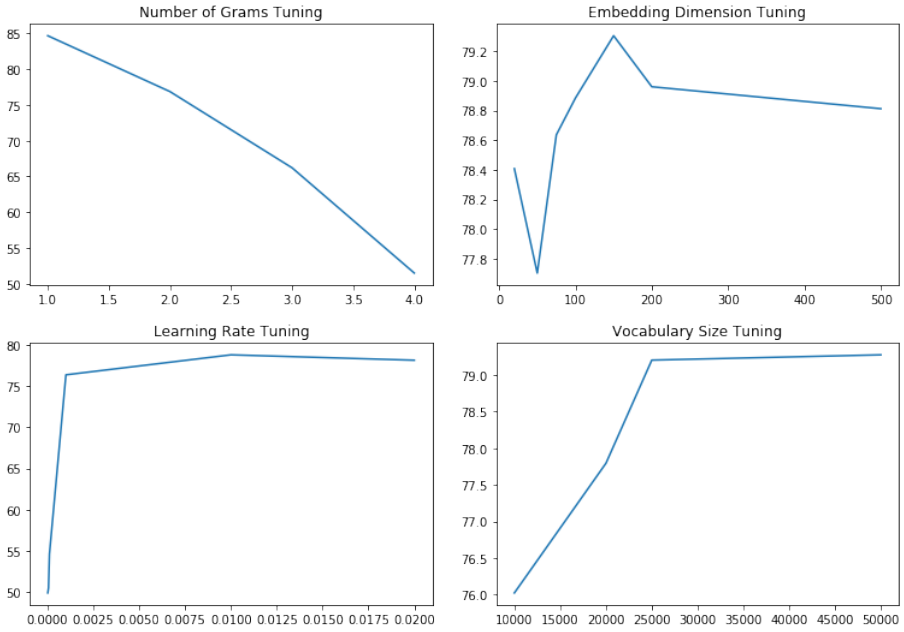
```
After training for 10 epochs
Val Acc 57.46
Test Acc 57.292
```



The BagOfWord model with one linear lay and Adam Optimizer

2.4 Tuning Hyperparameters

I tuned the number of grams, the embedding dimensions, the learning rate and the vocabulary size. The results look like the following:



3 Right/Wrong Prediction

I created the function to catch the 3 right predictions and 3 wrong predictions.

The right predictions:

```
In [87]: find_original_sentence(r[0])
```

```
Out[87]: 'Although others have commented that this video is just an edited version of the two shows: "Fire in Space" and "Livi
ng Legend", if you watch the original shows you\'ll find that dialogue from this video edition was edited out. I foun
d this video version much better because scenes and lines were added to it. I would say if you want to see the show i
n its original version, see the video versions on VHS. They have more to offer the fan than the original episodes bei
ng offered on DVD now. Another good video is Conquest for the Earth, which had more scenes from Galactica 1980 than d
id the actual broadcasts themselves. Overall, I rate this as a 10 because it gives you more to enjoy than what the ne
tworks wanted to show in the time slot they gave the producers.'
```

```
In [94]: find_original_sentence(r[3])
```

```
Out[94]: 'This movie is the perfect illustration of how NOT to make a sci fi movie. The worst tendency in sci-fi is to make yo
ur theme an awful, sophomoric, pseudo-Orwellian/Huxleyan/whateverian "vision" of "the human future."<br /><br />Scien
ce fiction filmmakers (and authors), as geeks, take themselves very seriously given the high crap-to-good-stuff ratio
of their genre. I think other genres with a high CTGSR (yes, I just made it up, relax), like horror or action or even
romantic comedy, seem to have a little better grasp of the fact that they are not changing the world with some profou
nd "message."<br /><br />Sci fi can certainly be successful on a serious level, as numerous great filmmakers have pro
ven. But there is an immense downside to the whole concept, which is represented by "Robot Jox," with its low-rent co
nstruction of "the future" (lone good design element: the bizarre, slick-looking billboard ads all over the place tha
t encourage women to have more babies) and its painfully heavy-handed "Iliad" parallels (He\'s NAMED ACHILLES FOR GOD
\'S SAKE! I actually didn\'t pick up on this until I saw the film for like the tenth time, but I went to public schoo
l, so the filmmakers are not exonerated.)<br /><br />Of course, if you\'re a crazy movie freak like me, this downside
has a great upside. I absolutely LOVE movies like this, because bad movies are quite often more fun and sometimes eve
n more interesting than good ones. It\'s kind of a Lester Bangs approach to movie viewing, I guess.<br /><br />Note:
The lead in this movie (Gary Graham? Is that his name? I refuse to go check.) is really not that bad. He makes a go o
f it. He\'s kind of cool, especially when he\'s drunk/hung over.'
```

```
In [89]: find_original_sentence(r[2])
```

```
Out[89]: 'In my opinion, this movie\'s title should be changed from "Only the Brave" to "All About Lane". I went to a screenin
g of this film a few months ago and was quite disappointed with the outcome. Although, I appreciate that the director
made a movie about the men of 442nd - a subject matter that long deserved addressing in the film industry - the actin
g in some parts of film was quite stale. The performances of Marc Dacascos, Tamlyn Tomita, and Jason Scott Lee were a
ll great. However, the director should have NEVER put himself as the main character in the movie. Sorry Lane, you are
just not a film actor. Stick to what you\'re good at - theater acting. Gina Hiraizumi\'s performance in this film was
also horrible. She should never have been given a speaking role and her looks were unfit to play the part of a Miss N
isei queen. There were other young actresses in the film who were naturally beautiful and whose performances were won
derful... Why weren\'t they cast for that role? Another major problem with this film were its action sequences. The J
apanese-American soldiers don\'t look like they were fighting German soldiers... let alone anyone. Granted this was a
low budget feature, but since this was a war-based film, isn\'t it important to show some actually fighting? This fil
m was a worthy attempt, but definitely not worth a major distribution.'
```

The wrong predictions:

```
find_original_sentence(w[0])
```

```
"Freddy's Dead: The Final Nightmare (1991) was the last film to feature Freddy Krueger as a solo act (not as an entit
y or a co-star). The years of killing have taken a toll upon the town of Springwood. It has gotten to the point that
the little city has become a virtual ghost town. The parents who killed Freddy Krueger so many years ago have all pai
d the ultimate price. Only the mad inhabit the town and the survivors are scattered everywhere. But that doesn't stop
Freddy from seeking out his final revenge. No matter how they try to stop him, he always comes back for more. But thi
s time he finds out a little more about his old life. Can the kids finally stop Freddy for good? What is this secret
that is buried in Freddy's twisted mind? to find out you'll have to watch Freddy's Dead. the end was originally filme
d in 3-D.<br /><br />A fitting way to end the franchise. Freddy learns something about himself and his perverted life
and he gets to go out in a bang! Lisa Zane, Yaphet Kotto and Freddy Krueger star in this final installment. Rosanne,
Tom Arnold and Johnny Depp make special appearances. A whole lot better than the last one but it's filled with a few
dated jokes. If you enjoy the series then you don't want to miss out on this one.<br /><br />I have to recommend this
movie for Freddy fans."
```

```
find_original_sentence(w[1])
```

```
'This movie was portrayed in the trailer as a comedy. It is an extreme tragedy. It left me sick to my stomach. I hate
d it. I think if they want to make a movie like this than they should be man enough to reflect the true intentions of
the movie in the trailer. I would not have seen this movie if I would have known. I think the trailer should reflect
the theme and intentions of a movie. I am tired of it. I really wanted to have a fun comedy and I am extremely disapp
ointed. It has been several days now and I still have a bad taste in my mouth from this movie. I have never been more
disappointed in a movie, nor have I ever written a comment on a bad movie. I really think that true deception was inv
olved in this trailer because if they showed the true intention of the movie, no one would have seen it.'
```

```
find_original_sentence(w[2])
```

```
"If you're amused by straight-faced goings-on that are logical within a given illogical situation, you'll enjoy this
whimsical 8-minute Spanish film.<br /><br />A woman enters a small café. The scene looks ordinary, but the counterman
, customers, and two musicians seem somehow oddly subdued.<br /><br />Suddenly, the musicians play and one man begins
to sing the title song , dancing across table tops with musical-comedy gestures. The customers, at first immobile, at
intervals chime in (badly but gamely) with phrases from the song, read from slips of paper in their palms. On and off
they jump up and dance (awkwardly but earnestly) in choreographed motions, like backup singers.<br /><br />But why???
the woman wonders. The answer is revealed as the soloist's jacket opens and she sees what's strapped across his chest
-- just before the explosive climax...<br /><br />Even if you don't catch the song's (probably ironical) lyrics, the
situation-perfect performances should give you a grin and a chuckle... I'd love to see it again!"
```