LSTM Document Classification

September 28, 2021

---

1 Introduction

I spent a little time looking at ways to classify documents. For most applications a bag-of-words model and a gradient boosting classifier works well despite ignoring contextual information. I am aware of cutting edge models like BERT which are supposed to capture contextual information, but in truth I am still catching up with my understanding of the state of the art of ML. However, in my reading on document classification approaches I found that LSTM models could be used for document classification [3]. Never having used a RNN or LSTM model and only having a conceptual understanding of how they worked I decided to try using a LSTM model for document classification.

2 Data Set

I chose try classifying Yelps reviews [1] using as "positive" or "negative", defined as greater than a 3 star review and less than or equal to a 3 star review respectively. The data set consists of $8,635,403$ samples. 20% of the data I reserved for testing and the remainder I use for training as I did no hyper-parameter tuning and hence did not require a validation set.

I pre-processed the reviews to remove punctuation and numbers. I also replace all words that occur less than 1000 times in the training set with the token UNKW00. The resulting vocabulary consists of 17,524 words. All words in the test set not in this vocabulary were also assigned the token UNKW00.

3 LSTM Model and Training

I use the LSTM model implemented in Pytorch [2]. Instead of the LSTM predicting the next word in a text, I simply take the log softmax of the mean output value of the output associated with each word in the text. This is output is compared to the class labels; 1 for a good review, 0 for a bad review. The error in the models output relative to each training batches labels is used to train the LSTM model.

Pytorch expects the words in a text to be mapped to integers, the number of integers the words are mapped to is called the embedding dimension. But this if not a vector embedding such as word2vec, it is a mapping of words to integers. The embedding dimension of this model is 1000 meaning if words are mapped uniformly to integer id's one would expect 17.524 words to be mapped to each integer id. The hidden dimension is the number of hidden features in the model, I set it to 100. In training and testing I used a batch size of 64, I trained the model for 10 epochs.

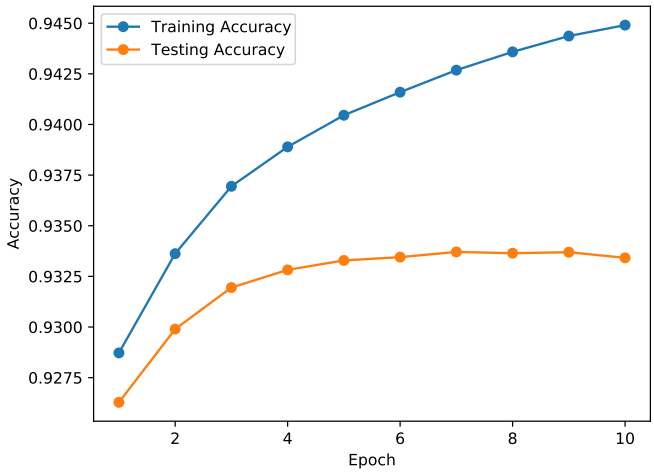

Figure 1: Epoch Accuracy

As mentioned above I have not tuned any of the models hyper-parameters. One would expect that increasing the embedding dimension would increase the models accuracy to a point. I am less clear how the hidden unit dimension or batch size might effect the models performance. However since it took 2 days to train this model with an i5 and $60 GPU, I was constrained in my hyper parameter exploration due to computational limitations. Therefore I simply choose an embedding dimension and hidden dimension that seemed reasonable and increased the batch size until I started running out of GPU memory.

4 Results

The per epoch accuracy of the model on the training and testing set are shown in Figure 1. Unsurprisingly the models accuracy on the training set outperforms that on the test set. The fact that while the accuracy of the test set plateaus without decreasing significantly suggests that the model lacks the capacity to over fit and the embedding dimension or the hidden dimension could be increased to improve accuracy in the training set if not the test set.

The fraction of class 0 and class 1 in the training and testing sets are nearly identical. In the training set 33.6% are class 0 and 66.4% are class 1. In the testing set 33.5% are class 0 and 66.5% are class 1. Numerical results for the model with the highest accuracy on the test set are given for the training and testing set in Section 4.

Dataset	TP	TN	FP	FN	TPR	TNR	Accuracy
Training	4432033	2080322	241265	154704	0.9663	0.8961	94.27%
Testing	1101484	511100	68147	46348	0.9596	0.8824	93.37%

Overall I was surprised how well this unoptimized model performed on such a large and diverse data set. I give a few examples of preprocessed Yelp reviews classified as true positives, true negatives, false positives, and false negatives below. I'll add a link the code associated with this project in the near future.

True Positives:

- I called UNKW00 on the recommendation of a couple of friends who had used them in the past and thought they did a nice job Im a fan now too UNKW00 Evan and Cody showed up right on time for

my move this past weekend They were friendly and energetic working quickly but carefully to get all my things moved out of the old place and into the new one in less than UNKW00 hours All of my heavy furniture arrived in perfect condition and they took extra care not to scratch the wood floors in the process UNKW00 I hope not to move again anytime soon but next time I do Ill be calling UNKW00

- I work in the UNKW00 and this is the most affordable and tasty place in the food court UNKW00 deals where a meal is $$ and the chicken pesto is really good UNKW00 UNKW00 I am not a UNKW00 person but all there soups I have had are pretty damn good UNKW00 Broccoli chicken is my favorite UNKW00 Also probably the most personable Food court staff I have ever had the pleasure of ordering from

- Ordered the original tonkotsu base ramen and a char siu don for my UNKW00 year old UNKW00 Loved the soup base but especially loved the size of the UNKW00 most ramen place the size of the bowl reflected the quantity of the content in it UNKW00 Totally looking forward to going back so I can customize my next bowl of UNKW00 select thicker noodle for sure and perhaps add more ramen Fabulous service and loved that they had a sign advising ppl to come in write down name then wait outside so not to clog the UNKW00 UNKW00 Great ramen place but wont go with a group larger than UNKW00 ppl UNKW00 Heads up there was only one high chair

True Negatives:

- The setting is perfectly adequate and the food comes close UNKW00 UNKW00 The dining chains like Chilis and Victoria Station do barbecue better UNKW00 Its no surprise you can always pick up coupons for UNKW00 at restaurantcom

- I love ale house Im here all the time The UNKW00 star isnt for the place or for the service The UNKW00 star is for the management You give absolutely zero respect to the game of soccer UNKW00 Ive been coming here since you opened and Ive seen numerous times where you UNKW00 for finals of UNKW00 where years past have shown you will be at capacity UNKW00 UNKW00 Its a shame you have a great establishment and great employees If only management could sort it out this place would be UNKW00 stars

- Bummer I was very disappointed UNKW00 So I finally stopped in for take out UNKW00 I got the fried haddock sandwich with fries UNKW00 The fries were frozen and greasy UNKW00 The haddock sandwich was not good UNKW00 Tasted UNKW00 very fishy and frozen UNKW00 I had such high hopes their take out would be good

False Positives:

- By no means is me giving UNKW00 UNKW00 stars a bad thing This is my go to spot for Mexican food when my family comes to visit me from NY Its cheap its quick and its surprisingly good The establishment offers a wide selection of Mexican dinners and combination specials common in other traditional Mexican restaurants If youre a fan of Mexican food especially if youre from out of town just go here and eat Parking isnt bad your seated immediately and there is always a table available Unfortunately there can sometimes be minor communication errors with your waitresses and the staff so make sure to be very clear with what you want when placing your UNKW00 The food is good nothing amazing but it definitely hits the spot

- Visiting Atlanta for a conference and I needed a bit of respite so I searched for my favorite comfort food pho I was happy to find this place less than a mile from the UNKW00 UNKW00 I ordered the house pho took my number and headed downstairs Set in five points this place is busy but they have plenty of seating The pho was delivered shortly after i found a seat The broth at first taste is a bit

sweet and doesnt exactly have a beefy pho taste I will say that once lime basil and a bit of chili oil were added the flavor developed and became more savory UNKW00 The flank was well done when it got to the table I would suggest meatball or the brisket before the flank While this was not my favorite pho it still hit the spot and provided a nice getaway from the hustle and bustle of attending a conference

- I shall give UNKW00 stars The dining environment is great and you can tell how food is cooked by sitting on the bar seats The food is nice but a little salty Other is good

False Negatives:

- I HAD ZERO UNKW00 OMG I could only fit in a UNKW00 minute deep tissue UNKW00 I walked in crooked and walked out UNKW00 as an UNKW00 UNKW00 The tiny little thing jumped on the table and used her knees to UNKW00 my lower back and hips UNKW00 GET OUT OF UNKW00 UNKW00 I was skeptical but walked out a believer

- I wish Menchies had more flavors There are roughly a dozen choices versus UNKW00 at local ice cream shops Their pineapple is my favorite but they just UNKW00 UNKW00 which is also excellent The variety of toppings is their strength but their product is a little expensive For two cups of yogurt we never escape for under $ When John UNKW00 UNKW00 about the cost of a shake in UNKW00 UNKW00 he was astonished to pay five dollars for an Ice cream drink If he came here he might tear off his hair piece in frustration So unless you are a UNKW00 UNKW00 you will likely enjoy Menchies

- I love Forest Park I live in St Johns and go there at least UNKW00 times a week with my dog to wear her out I spend most of my time on the Newton Road UNKW00 and have a few loops around there that I love to watch change with the seasons UNKW00 UNKW00 UNKW00 Forest Park has a strict leash requirement which I consistently break with my dog I know this is bad and naughty I do keep my dog under control and as a UNKW00 good citizen and wildlife UNKW00 dog she is very well trained UNKW00 UNKW00 That said I have a big concern The last time I was in the park a park employee or volunteer was out spraying UNKW00 on the blackberries to get rid of them They told me my dog should be on leash because UNKW00 UNKW00 the rules and UNKW00 the UNKW00 can cause serious eye damage to animals that come in contact with it As long as we stay on the trail it should be fine they told me This concerned me greatly because of course I dont want my dog near any kind of UNKW00 that can make it go blind and more importantly no one gave that UNKW00 to the wild birds UNKW00 and UNKW00 that live in the park Such a bummer Id happily thrown in extra time to help UNKW00 blackberries manually or contribute to a fund that would employ workers to tear out the blackberries in a way that doesnt involve dangerous poison UNKW00 UNKW00 Anyway Yelp always filters my reviews so this will probably never see the light of day but thought park fans should know that there are dangerous UNKW00 being sprayed throughout the park and to be careful not to let your dog come into contact with them

References

[1] *Yelp dataset*, `https://www.yelp.com/dataset`, Accessed: 2021-09-19.

[2] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala, *Pytorch: An imperative style, high-performance deep learning library*, Advances in Neural Information Processing Systems 32 (H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, eds.), 2019, pp. 8024–8035.

[3] Xiang Zhang, Junbo Zhao, and Yann LeCun, *Character-level convolutional networks for text classification*, 2016.