NLU-Assignment2 Report

Akash Mishra
SR No- 14421
akashmishra@iisc.ac.in

Abstract

In Assignment-1, I had developed a tri-gram based language model and trained it over given Brown Corpus and Gutenberg Corpus. After that, I had developed the perplexity of that language model. In Assignment-2, we need to implement Neural Network, LSTM based model over Gutenberg corpus, generate sentances from that model and compare it's perplexity with previous N-gram model from assignment-1

Note: I have tested my model over only one file(austen-emma.txt) but if one has sufficient amount of resources(processor/GPU), s/he can train the same model over whole gutenberg corpus by just replacing the input data appropratly.

1 Introduction

This time, we have been given only one dataset, Gutenberg Dataset and we need to implement LSTM(long short term memory) based language model over some or all(whatever possible) files of this given dataset. Basically, we have 3 tasks to do-

1. Build the best token-level LSTM model over given dataset, find it's perplexity.
2. Build the best character-level LSTM model over given dataset, find it's perplexity.

Now, compare these perplexities with the N-gram based language model, build in Assignment-1.

3. Using models above mentioned, we need to build a sentence of 10 tokens

2 Language Model

We need to implement LSTM(Long-Short Term Memory) neural network based language model which is a variation of RNN(Recurrent Neural Network).

A recurrent neural network (RNN) is a class of artificial neural network where connections between units form a directed graph along a sequence. This allows it to exhibit dynamic temporal behavior for a time sequence. Unlike feedforward neural networks, RNNs can use their internal state (memory) to process sequences of inputs. This makes them very efficient for a language model generation.

RNN can generate text keeping in consideration n-previous inputs where ideally n can be a very large number but in practice this doesn't happen. Basic RNN fails when n is a very large number. But LSTM which is a variation of RNN itself overcomes this drawback significantly.

Long short-term memory (LSTM) units (or blocks) are a building unit for layers of a recurrent neural network (RNN). A RNN composed of LSTM units is often called an LSTM network. The expression long short-term refers to the fact that LSTM is a model for the short-term memory which can last for a long period of time.

We need to build our model with the help of LSTM over two different levels

1. Word/Token level
2. Character Level

A word level LSTM based language model is where we feed n words to our model and by using those words our model will generate new subsequent words. Whereas in character level LSTM based language model we need to feed in n characters insteade of n words and our model will predict the next character.

3 Preprocessing

3.1 Filtering Raw Data

Data given to us may have some arbitrary values some not useful information, punctuations etc

which may not help while training our model and sentence generation. So before start building our model, we need to filter our raw data.
for this process, I removed some punctuations, converted whole dataset to small letters, converted sentences into characters or words appropriately.

3.2 Generating Sequences

Since, in my assignment 1 I made a trigram based language model ,and since,we need to compare our current language model with previously generated language model, this time, for my word level language model, I feeded only 2 words as input ant generated next third word as output. So that, It will be more comparable with previously generated trigram model. But it should be noted that by just changing only two variables in the written code one can feed in n number of words where n is greater than 2.
For character level based language model, I feeded 50 characters as input and then predicted the consecutive words.

3.3 Split

As done before in assignment1, this time also I have divided my entire corpus into train and test where 80% of the corpus is used for training and remaining 20% is used for testing.

4 Results

It should be noted that due to lack of resources, I have trained my language model for only one file(”emma-austen) and for 50 epochs(for each char-lvl and token-lvl) but if one can train the model over a large dataset and with a higher number of epochs then high chances are that it may give a better accuracy and less perplexity. That model may generate more gramatically correct and meaningful sentences.
After training my model, I saved my model in a .h5 file and corrosponding tokenizer in a .pkl file so that I don’t have to train my model again until necessary. It is advisable that if model is trained again then it should be saved for further use because training this model may be a very time taking process.
I am using perplexity as a measure of comparison and since in assignment 1, I have used whole Gutenberg corpus for model training and testing, I trained that model again for only one file(”emma-austen”). I will attach that result also for better comparison.

Note: Below written losses, Accuracies and perplaxities are over test data not training data.

4.1 Word-Level Language Model Results:

loss = 8.411084093277172
Accuracy = 0.018786127167925038
perplexity = 340.3992606548815

4.1.1 Few Generated Sentences

1.) ”be paid” ==: to think of it and i am sure i had
2.) ”you you” ==: are not to be sure of succeeding emma was obliged
3.) ”with a” ==: smile to attach her and the very first sort of
4.) ”given her” ==: a little while i have been the first of the
5.) ”great deal” ==: of pleasure and manner were to be sure of succeeding

4.2 Character-Level Language Model Results:

loss = 7.5400957252515175
Accuracy = 0.06259373496685439
perplexity = 186.12083123866745

4.2.1 Few Generated Sentences

1.) Seed:” begin to exert ourselves oh mr weston i could not have believed it of you aye we men are s” ==: ure they were the sea walk of the sea walk
2.) Seed:” mr knightley success supposes endeavour your time has been properly and delicately spent if yo ” ==: u will be a great deal of her own mind i am
3.)Seed:” l and harriet not much otherwise emma would not interfere with she had an unhappy state of health ” ==: to her friends and the party of her being a grea
4.)Seed:” that pays woman a compliment he has the tenderest spirit of gallantry towards us all you must let ” ==: her to be a great deal of her own mind i am sur
5.)10.)Seed:” elf thoroughly if hungry that she would take something to eat that her own maid should sit up for ” ==: her fortunate and the wertons of her own conpanio

4.3 N-gram Language Model Results:

perplexity = 306.458604437

5 Conclusion

We can see that if training dataset is small or number of epochs are not sufficient then LSTM may not generate a relatively good model but if we can train our model over a large dataset and for a sufficient number of epochs, it has been seen that sentence generated using LSTM based model are more meaningful and contextual.

Github Link:
https://github.com/AkashMishra2k16/Assignment2_NLU