E1 246 Assignment-2

Rahul Chittimalla
Sr no: 06-02-01-10-51-17-1-14585
rahulc@iisc.ac.in

For the purpose of language modeling, the gutenberg corpus is scaled down to 80% of the total sentence for the purpose of computation/running the neural model. This dataset is further split into 90%, 5% & 5%, train, valid and test sets respectively.

Task 1

For the token level LSTM-based model, first the data is preprocessed to feed into the LSTM. One-hot vector representation of the words is made and then that is fed into the model to generate dense vector embeddings of size 500. There are two LSTM hidden layers. The first LSTM layer takes the input of 500 for each batch of inputs being fed.

The model data i.e., weights are stored after each epoch so as to not loose the task done so far.

Task 2

In the character level LSTM-based model, first the vocabulary of unique characters is built to get the one-hot encoding of each character in the corpus. The input to the LSTM is a 100 length characters each encoded in one-hot representation. It is like a sliding window of 100 characters moving one step everytime. The model is trained and saved after each epoch. The trained model is then used to generate the sentences by loading the pre-trained model using the training set.

To generate the sentence, a seed of the window size 100 is randomly picked and fed into the LSTM which has already learnt the weights

Task 3

The generation of the text involves randomly picking a seed and then generating the sentences according to the next probable word. The code for generating the senetence is in the file `char_lstm_generate.py`. First the model is saved from Task1 and Task2. The best model among the two is loaded into this file to generate the sentences. Some of the sentences generated from the previous language model used in assignment-1 are:

- as the junior mates were hurrying to execute the warrants
- upon it in truth in judgment and in lovingkindness and
- of course who keep it alive and preserve it so
- the box would break open a hamper and produce filets
- miles a day under sub freezing temperature conditions attendants inactivation
- always wished to be a christian means to say yes

The sentences generated using LSTM model are:

- as you can conveniently leave town and we must put
- they will bring him to the king of babylon and
- armies rush to battle in thine own people for ever

- his own robust rapidity asserted itself unconsciously and he walked
- thus shall ye deal with them ye shall not eat

It is obvious that the LSTM model generates more meaningful sentences than the previous model. The code is available at github repository.