By: Pankaj Katkar
IIT GOA
pankaj.katkar.16001@iitgoa.ac.in

Text Generation using LSTM Recurrent Neural Network

Background: We were given assignment to use LSTM Recurrent Neural Network(RNN) for text generation .

Problem Statement: The problem statement was to generate text sequences by using LSTM RNN model trained on the book named as “Alice in Wonderland”. Text generated should follow the pattern as given in the book.

Data: The book was dowloaded through Gutenberg site. I downloade it from there and store it as a text file. The link for dataset with file name “alice.txt” in github is https://github.com/Pankaj01998/LSTM-RNN

Approach: I made .txt file of “Alice in wonderland”. Then convert whole text to lowercaase. Then for prediction of text, instead of predicting a word I used LSTM to predict character given sequence of character as input to LSTM. For that I trained LSTM model on the .txt file. Thus after training when we give some sentences as input to the model in the form of sequence of characters it outputs the next 1000 characters.

Solution: I first found all unique charactes used in the text. Then assign each character a unique number. Then I need to define training data for network. For that, I split the book text up into subsequences with a fixed length of 100 characters, an arbitrary length. Each training pattern of the network is comprised of 100 time steps of one character (X) followed by one character output (y). When creating these sequences, I slide this window along the whole book one character at a time, allowing each character a chance to be learned from the 100 characters that preceded it (except the first 100 characters of course). Then I transformed the each 100 character input pattern into corresponding unique number which I assigned earlier. Then I normalize the input pattern to rescale it to range 0-1. Also I created vector for output layer. Then I trained LSTM on this data for 2 hours with epoch 20. Then I choose the model with min loss and predicted the next 1000 characters with that model by giving some random input sentences from the book.

Link to code is – https://github.com/Pankaj01998/LSTM-RNN. In it filewith name “rnn.py” trains the LSTM model and file with name “trained_rnn.py” predicts the text.

Results: As I didn’t trained it for long I get the model with loss around 1.94. But the results were quite interesting.

1)The characters are separated into word-like groups and most groups are actual English words (e.g. “the”, “her” and “was”), but many do not (e.g. “tornd”, “lettle” and “hor”).

2)Some of the words in sequence make sense(e.g. “*and the turtle*“), but many do not (e.g. “*magch harerly*“).

```
pankaj@pankaj:~/rnn$ python trained_rnn.py
Using TensorFlow backend.
Total Characters:  144407
Total Vocab:  45
Total Patterns:  144307
2019-05-11 14:58:03.990085: I tensorflow/core/platform/cpu_feature_guard.cc:141] Your CPU supports instructions that this TensorFlow binary was not com
piled to use: AVX2 FMA

random Seed is:

"  now and then she had to stop and untwist it. after a while she
remembered that she still held the p "


################### predicting with model ########################
am oo the was oo tornd the was   and the west on tork inr teali of the sabd to the table  and the whrt hn dn anloeer tonee of the goure, and the war ao
ing to the tooe  and the was aoling to be in a lettle or too of the tabd to the table  and then the muckn wu tere the pucen  and the war oo taring to t
he tohee of the goushon th tee toae a futre to than the was aoi ali her foad to the tooe.
'the duehusd then ' said the muce turtle in a mor to cere the hoeae of the eareeriir.
''thal toe dare tiine would ' said the manch hare.

'iele you waid t ainine,' said the ming.
'no yhu whu the ' said the micg turtle an the cadt. ''than soese soued ' said the magch harerly.
ao the said the was so loce tored of the goure  and the wurdd her head oo hen tham wou doen the woide  aut oo hor to toe the somet '
'iot to the shre theng was ' said the manch hare.

'iele you waid t ainine,' said the ming.
'no yhu whu the ' said the micg turtle an the cadt. ''than soese soued ' said the magch harerly.
ao t
finish
```

We can see that generally there are fewer spelling mistakes and the text looks more realistic, but is still quite nonsensical. But if we trained it more it can give good result.