Report

Ramabhadra V
ramavyasa@gmail.com

1 Preprocessing

I am using 'Alice in Wonderland' (carroll-alice) dataset. I am converting all uppercase to lowercase and removing all non alpha-numeric characters.

2 Model for character level embedding

Input is a sequence of 20 characters and the corresponding output is the 21st character.

Training is on 108006 samples, testing is on 27002 samples (80:20).

The above input goes to a LSTM layer with 75 units and next to a densely connected output layer with softmax activation. The vocabulary size is 33 and the number of trainable parameters is 35,208.

I have run the above setup for 1000 epochs with batch size 10,000.
Figure 1 shows accuracy over epochs and figure 2 shows loss over epochs.
Some sample sentences:(seed is in quotes)
'when she got to' the garden and was to the hatter with the cat and she said the duchess went on it.
'she was in the' door and the march hare said the cat i m sure the hatter with it.

3 Model for word level embedding

Input is sequence of words in a sentence and corresponding output is the next word. Train on 20513 samples, validate on 5129 samples (75:25)

The above input goes embedding layer which learns vector representation of words,next to a LSTM layer with 128 units and next to a densely connected output layer with softmax activation. The vocabulary size is 2571 and the number of trainable parameters is 329,781. I have run the above setup for 2000 epochs with batch size 2,000.
Sample sentences:(seed is in quotes)
'I will be' a little girl as she had been to be a
'she came' on the same thing as she could not remember the

4 Observations

1. Character level:
a.The accuracy starts to decrease after 200 epochs on validation set. This is because model is overfitting. Accuracy is 54% on validation and 70% for training.
b. Perplexity: 4.52
c. However, if I use the weight's on which I get highest validation accuracy to generate sentences, the model repeatedly generates some random word.
d. Model with highest training accuracy is generating good sentences.
e. In both of the cases the words predicted are mostly spelled correctly.
2. word level:
a. The accuracy starts to decrease after 200 epochs on validation set. Training accuracy reaches 90% but validation accuracy is max 15%.
b. Perplexity: 424.11
c. model with highest validation accuracy generates better sentences.

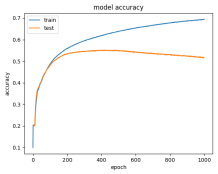

Figure 1: Accuracy vs Epochs.(character level)

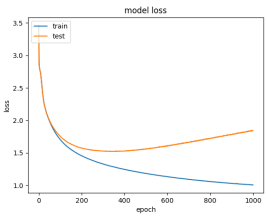

Figure 2: Loss vs Epochs.(character level)