Bidirectional LSTM for Shakespeare Text Generation

Rimon Dubnov
rdubnov@ucsd.edu

Abstract

This paper explores generating Shakespeare text using Bidirectional LSTMs. I trained multiple models with variable dropout rate, hidden size, and activation function. The results show that higher hidden size and lower dropout rate results in lower cross-entropy loss and more coherent text generation. Comparing RELU and Tanh activation functions shows that the RELU results in higher loss but more coherent text generation.

Introduction

Vanilla Recurrent Neural Network (RNN) architectures have shown great success in generating short text, however, are poor at generating longer text due to the vanishing gradient problem. The output of the current hidden state is affected by all of the previous hidden states, so the neural network has to backpropagate through all of the hidden states in order to update. As the RNN iterates through the previous states, the gradient slowly vanishes, which stops the network from updating and converging. An improved model should have a long term memory so that the current hidden states are heavily affected by the older hidden states. The Long Short Term Memory (LSTM) architecture fixes the vanishing gradient problem by replacing the RNN cells with LSTM cells that contain additional gates to provide the model with a short and long term memory. Using LSTM will improve the model's ability to produce long text that coherently connects the end of the text to the beginning.

Bidirectional RNNs are also known to increase a model's ability to generate long coherent sentences as the model gains more context from the sentence by reading it from beginning to end and reverse. Certain words in english sentences only make sense in the context of future words, thus bidirectional RNNs fix this issue by reading the text in reverse to give more context to the hidden state, then by just reading the text from beginning to end.

Architecture

The architecture used is from this source [1], a Bidirectional LSTM with a size of 256 and a RELU activation function, then a .6 dropout layer, and a softmax activation. The batch size is 32, the training sequence length is 30, and the model ran for 50 epochs with a learning rate of .001. The model uses early stopping so none of the models reach the 50th epoch. The data is a collection of Shakespeare plays and the training and validation dataset is a 90% and 10% split. I also trained models with a size of 128 and 512, dropout of .2 and .4, and a model with a tanh activation function.

Experiment

I ran the model with 6 different sets of hyperparameters with variable hidden size, activation function, and dropout. The model parameters of the experiments are (256, RELU, .2), (256, RELU, .4), (128, RELU, .6), (256, RELU, .6), (512, RELU, .6), and (256, Tanh, .6) where the arguments are hidden size, activation function, dropout rate. The results from the experiment are in the appendix.

Results

The final loss value of the 128, 256, and 512 hidden size models are 5.3362, 5.1236, and 4.5733 respectively. This shows a strong inverse correlation between hidden size and loss. Comparing the first 10 words from the generated text of each model shows that there is a weak correlation of text comprehension and model size

128: “on forfeit god life henry surely first , pains then were”

256: “shrinks come hear i mortal flowers have good come froth”

512: “the hath such, and than thee. you and I go“

The final loss value of .2, .4, and .6 dropout rates are 4.4860, 5.1418, 5.1236, showing that higher dropout increases loss. Comparing the first 10 words from the generated text of each model shows that there might be correlation of text comprehension and dropout rate.

.2: “the trust from ? with man and and to a 's a”

.4: “the pride of ' , to on the name : and that not”

.6: “shrinks come hear i mortal flowers have good come froth”

Tanh activation function has a final loss value of 4.8955 which is significantly smaller than RELU’s loss of 5.1236. However, looking at a sample text from the Tanh activation function shows that the model is very poor at generating text “a . , the , , with is ; , , , , , , , . is ; ; the to and”.

Conclusion

The Bidirectional LSTM architecture shows promising results that improve on traditional RNN architectures. After training the same architecture with different hyperparameters, I found that lower dropout rate, larger hidden size, and the RELU activation function increase the model’s performance. Future research could involve looking into different batch sizes, other loss functions, and variable learning rates.

References

1. https://github.com/campdav/Text-Generation-using-Bidirectional-LSTM-and-Doc2Vec-models/blob/master/1%20-%20generate%20sentence.ipynb
2. *The Unreasonable Effectiveness of Recurrent Neural Networks*, karpathy.github.io/2015/05/21/rnn-effectiveness/.
3. Campion, David. “Text Generation Using Bidirectional LSTM and Doc2Vec Models 1/3.” *Medium*, Medium, 14 Feb. 2018, medium.com/@david.campion/text-generation-using-bidirectional-lstm-and-doc2vec-models-1-3-8979eb65cb3a.

Appendix

Model:
rnn_size = 256, RELU activation, .6 dropout

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
bidirectional_1 (Bidirection (None, 512)               25794560
_________________________________________________________________
dropout_1 (Dropout)          (None, 512)               0
_________________________________________________________________
dense_1 (Dense)              (None, 12338)             6329394
_________________________________________________________________
activation_1 (Activation)    (None, 12338)             0
=================================================================
Total params: 32,123,954
Trainable params: 32,123,954
Non-trainable params: 0
_________________________________________________________________
```

Error Values:

```
Train on 25241 samples, validate on 255 samples
Epoch 1/50
25241/25241 [==============================] - 126s 5ms/step - loss: 7.0484 - categorical_accuracy: 0.0534 - val_loss: 6.3338 - val_categorical_accuracy: 0.0667
Epoch 2/50
25241/25241 [==============================] - 125s 5ms/step - loss: 6.0867 - categorical_accuracy: 0.0826 - val_loss: 6.3197 - val_categorical_accuracy: 0.0667
Epoch 3/50
25241/25241 [==============================] - 127s 5ms/step - loss: 5.8142 - categorical_accuracy: 0.0981 - val_loss: 6.3442 - val_categorical_accuracy: 0.0784
Epoch 4/50
25241/25241 [==============================] - 126s 5ms/step - loss: 5.5986 - categorical_accuracy: 0.1019 - val_loss: 6.4520 - val_categorical_accuracy: 0.0784
Epoch 5/50
25241/25241 [==============================] - 126s 5ms/step - loss: 5.3912 - categorical_accuracy: 0.1075 - val_loss: 6.5142 - val_categorical_accuracy: 0.0824
Epoch 6/50
25241/25241 [==============================] - 125s 5ms/step - loss: 5.1236 - categorical_accuracy: 0.1227 - val_loss: 6.7283 - val_categorical_accuracy: 0.0824
```

Example Output:

"shrinks come hear i mortal flowers have good come froth cunning the be redress my blister'd seeming said afterward hands they your loath or for love that betide god grace , honour ; than o policy the now thoughts hast 'll that porch what the broken escalus their very or stormy ill give man good foes care , times little george fortune 'll thy sworest o'erwhelm with come king were with grumio this virtues but , my julietta are for happy honourable doit and she ? lowest richard liking you nor proclaim'd is sun him brother i expedition the thou"

Model:
rnn_size = 128, RELU activation, .6 dropout

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
bidirectional_2 (Bidirection (None, 256)               12766208
_________________________________________________________________
dropout_2 (Dropout)          (None, 256)               0
_________________________________________________________________
dense_2 (Dense)              (None, 12338)             3170866
_________________________________________________________________
activation_2 (Activation)    (None, 12338)             0
=================================================================
Total params: 15,937,074
Trainable params: 15,937,074
Non-trainable params: 0
_________________________________________________________________
```

Error Values:

```
Train on 25241 samples, validate on 255 samples
Epoch 1/50
25241/25241 [==============================] - 112s 4ms/step - loss: 7.0334 - categorical_accuracy: 0.0677 - val_loss: 6.3794 - val_categorical_accuracy: 0.0667
Epoch 2/50
25241/25241 [==============================] - 111s 4ms/step - loss: 6.1264 - categorical_accuracy: 0.0781 - val_loss: 6.3369 - val_categorical_accuracy: 0.0667
Epoch 3/50
25241/25241 [==============================] - 112s 4ms/step - loss: 5.8790 - categorical_accuracy: 0.0889 - val_loss: 6.4091 - val_categorical_accuracy: 0.0706
Epoch 4/50
25241/25241 [==============================] - 110s 4ms/step - loss: 5.6903 - categorical_accuracy: 0.0998 - val_loss: 6.4763 - val_categorical_accuracy: 0.0706
Epoch 5/50
25241/25241 [==============================] - 111s 4ms/step - loss: 5.5043 - categorical_accuracy: 0.1067 - val_loss: 6.6198 - val_categorical_accuracy: 0.0824
Epoch 6/50
25241/25241 [==============================] - 110s 4ms/step - loss: 5.3362 - categorical_accuracy: 0.1138 - val_loss: 6.7898 - val_categorical_accuracy: 0.0745
```

Example Output:

"on forfeit god life henry surely first , pains then were night should nurse well inforced ! coriolanus richard yourself ? well john limed she noise dead day follow'd the a glorious make him - memory lies bed long have pair i while name thy crown dare we hope must how my sweet shalt you frame shall face be good gone enfoldings blessed with know court put current bed lend provost that i long the his bloody holy name long a servants thee long morrow the inforced reason wife there senator one nine day blasphemy thanks this which feasting learned"

Model:
rnn_size = 512, RELU activation, .6 dropout

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
bidirectional_3 (Bidirection (None, 1024)              52637696
_________________________________________________________________
dropout_3 (Dropout)          (None, 1024)              0
_________________________________________________________________
dense_3 (Dense)              (None, 12338)             12646450
_________________________________________________________________
activation_3 (Activation)    (None, 12338)             0
=================================================================
Total params: 65,284,146
Trainable params: 65,284,146
Non-trainable params: 0
```

Error Values:

```
_________________________________________________________________
Train on 25241 samples, validate on 255 samples
Epoch 1/50
25241/25241 [==============================] - 157s 6ms/step - loss: 6.8248 - categorical_accuracy: 0.0739 - val_loss: 6.2513 - val_categorical_accuracy: 0.0667
Epoch 2/50
25241/25241 [==============================] - 154s 6ms/step - loss: 5.9686 - categorical_accuracy: 0.0943 - val_loss: 6.1998 - val_categorical_accuracy: 0.0784
Epoch 3/50
25241/25241 [==============================] - 160s 6ms/step - loss: 6.3813 - categorical_accuracy: 0.1049 - val_loss: 6.1286 - val_categorical_accuracy: 0.0863
Epoch 4/50
25241/25241 [==============================] - 157s 6ms/step - loss: 5.4501 - categorical_accuracy: 0.1193 - val_loss: 6.2030 - val_categorical_accuracy: 0.1098
Epoch 5/50
25241/25241 [==============================] - 160s 6ms/step - loss: 5.1778 - categorical_accuracy: 0.1313 - val_loss: 6.2785 - val_categorical_accuracy: 0.1020
Epoch 6/50
25241/25241 [==============================] - 157s 6ms/step - loss: 4.9190 - categorical_accuracy: 0.1430 - val_loss: 6.4267 - val_categorical_accuracy: 0.1020
Epoch 7/50
25241/25241 [==============================] - 159s 6ms/step - loss: 4.5733 - categorical_accuracy: 0.1655 - val_loss: 6.5168 - val_categorical_accuracy: 0.1137
```

Example Output:

"the hath such, and than thee. you and I go . : then , the tricks : hastings : you , to you and you . do : then you , you , you you you , for ? , : you , you you , you have you you you . i : you , i it , you it it . me : you , you ? is : you you you , you you you you , it it it it it is it is . it is . then you you you you"

Model:

rnn_size = 256, RELU activation, .4 dropout

Error Values:

```
Train on 25241 samples, validate on 255 samples
Epoch 1/50
25241/25241 [==============================] - 259s 10ms/step - loss: 6.6443 - categorical_accuracy: 0.0702 - val_loss: 6.3150 - val_categorical_accuracy: 0.0667
Epoch 2/50
25241/25241 [==============================] - 257s 10ms/step - loss: 6.0555 - categorical_accuracy: 0.0782 - val_loss: 6.3392 - val_categorical_accuracy: 0.0667
Epoch 3/50
25241/25241 [==============================] - 259s 10ms/step - loss: 5.8083 - categorical_accuracy: 0.0935 - val_loss: 6.3411 - val_categorical_accuracy: 0.0902
Epoch 4/50
25241/25241 [==============================] - 258s 10ms/step - loss: 5.4849 - categorical_accuracy: 0.1032 - val_loss: 6.4657 - val_categorical_accuracy: 0.0980
Epoch 5/50
25241/25241 [==============================] - 261s 10ms/step - loss: 5.1418 - categorical_accuracy: 0.1208 - val_loss: 6.5093 - val_categorical_accuracy: 0.1098
```

Example Output:

"the pride of ' , to on the name : and that not of most . we : : i , and you , and , to to my , my my , on , and to by king as . : richard , , tell , , you me . gloucester : thou 's stand , , and , and time which that . no . : : my , a there , i as ? love : so 's he , i your , go than , i , for him and to day , it to the true"

Model:

rnn_size = 256, RELU activation, .2 dropout

Error Values:

Train on 25241 samples, validate on 255 samples
Epoch 1/50
25241/25241 [==============================] - 261s
10ms/step - loss: 7.5039 - categorical_accuracy: 0.0531 - val_loss: 6.3358 - val_categorical_accuracy: 0.0667
Epoch 2/50
25241/25241 [==============================] - 257s
10ms/step - loss: 6.0489 - categorical_accuracy: 0.0778 - val_loss: 6.3177 - val_categorical_accuracy: 0.0667
Epoch 3/50
25241/25241 [==============================] - 258s
10ms/step - loss: 5.7801 - categorical_accuracy: 0.0806 - val_loss: 6.3718 - val_categorical_accuracy: 0.0706
Epoch 4/50
25241/25241 [==============================] - 256s
10ms/step - loss: 5.4316 - categorical_accuracy: 0.0949 - val_loss: 6.4877 - val_categorical_accuracy: 0.0824
Epoch 5/50
25241/25241 [==============================] - 259s
10ms/step - loss: 4.9879 - categorical_accuracy: 0.1214 - val_loss: 6.5580 - val_categorical_accuracy: 0.0863
Epoch 6/50
25241/25241 [==============================] - 258s
10ms/step - loss: 4.4860 - categorical_accuracy: 0.1580 - val_loss: 7.0239 - val_categorical_accuracy: 0.0941

Example Output:

“the trust from ? with man and and to a 's a of a ; of the into of the , the of ; the a that , you , the that of not of and , the , . the : , a , the the , a and the with of and , and , , and me , and , you , you , in , and the that , . : he is in to . ; , for the - of , , the you , , you you , and you of , , and”

Model:

rnn_size = 256, Tanh activation, .6 dropout

Error Values:

Train on 25241 samples, validate on 255 samples
Epoch 1/50
25241/25241 [==============================] - 124s
5ms/step - loss: 6.5579 - categorical_accuracy: 0.0765 - val_loss: 6.3471 - val_categorical_accuracy: 0.0667
Epoch 2/50
25241/25241 [==============================] - 122s
5ms/step - loss: 6.1628 - categorical_accuracy: 0.0782 - val_loss: 6.3199 - val_categorical_accuracy: 0.0667
Epoch 3/50
25241/25241 [==============================] - 126s
5ms/step - loss: 5.9554 - categorical_accuracy: 0.0784 - val_loss: 6.3524 - val_categorical_accuracy: 0.0667
Epoch 4/50
25241/25241 [==============================] - 126s
5ms/step - loss: 5.6815 - categorical_accuracy: 0.0930 - val_loss: 6.3361 - val_categorical_accuracy: 0.0902
Epoch 5/50
25241/25241 [==============================] - 128s
5ms/step - loss: 5.3288 - categorical_accuracy: 0.1115 - val_loss: 6.4966 - val_categorical_accuracy: 0.0980
Epoch 6/50
25241/25241 [==============================] - 125s
5ms/step - loss: 4.8955 - categorical_accuracy: 0.1461 - val_loss: 6.7565 - val_categorical_accuracy: 0.1059

Example Output:

“a . , the , , with is ; , , , , , , , . is ; ; the to and , ; . on the is ; , . . ; , ; . , ; , . . . , ; life ; ; , , ; . ; . , , , ; is , ! , ; is , , ; , ; . , , ; , , . , ; , ; . , ; ! , , , , the , of the we , , , , ; , ; as ,”