Shakespearean Text Generation using LSTM

Mateusz Kulinski

Introduction

- NLG is a subset of NLP focusing on generating natural language from structured data.
- Possibilities for text generated vary wildly; from IM messages, to report summaries, to technical documents.
- A now popular method of achieving this task is through the use of LSTM, an RNN architecture.

Problem Statement

- Problem: Is it possible to generate text in Shakespearean style, such that it is indistinguishable from genuine text?
- Approach: LSTM RNN operated at character level to generate a Shakespearean string.
- Evaluation: Two-step evaluation; quantitative—maximise accuracy, followed by qualitative—human opinion.

Dataset

- 4.4MB text file containing **all** of Shakespeare's plays.
- Each line in a play has its own line in the file.
- Indication of character is removed, only the speech is needed.
- Over 100,000 lines in total.

```
"So shaken as we are, so wan with care,"
"Find we a time for frighted peace to pant,"
"And breathe short-winded accents of new broils"
"To be commenced in strands afar remote."
"No more the thirsty entrance of this soil"
"Shall daub her lips with her own children's blood,"
"Nor more shall trenching war channel her fields,"
"Nor bruise her flowerets with the armed hoofs"
"Of hostile paces: those opposed eyes,"
"Which, like the meteors of a troubled heaven,"
"All of one nature, of one substance bred,"
"Did lately meet in the intestine shock"
```

Figure 1. Dataset Sample

Method

- Pre-process the dataset
  - First by removing unnecessary characters
  - Second by creating a corpus
- Create a sliding window model to get samples from the dataset
- Convert all of the samples into a one-hot representation
- Train the model with the samples
- Manually evaluate generated text

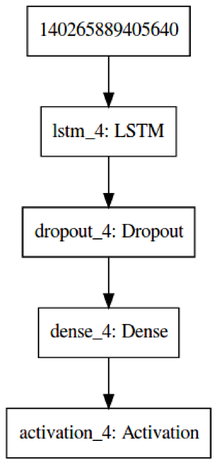

Figure 2. Model

Model

- Sequential model, with LSTM layer providing memory to remember previous characters. 200 units with 50.48 input
- Dropout following LSTM to help with overfitting
- Dense layer with 48 units, one for each character in the corpus
- Softmax activation

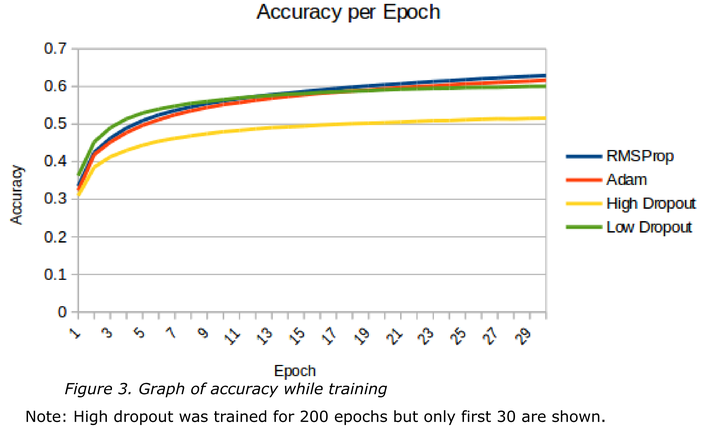

Figure 3. Graph of accuracy while training

Note: High dropout was trained for 200 epochs but only first 30 are shown.

Results

RMSProp:
"strook, and as the duke of york and the third will be relent to the crown and the counterfeit on my son to the crown, and therefore i shall then in the towns and the duke of suffolk, and the commons have i shall then, and therefore shall he that i"
Adam:
"sirghom of the realm of warwick, and there is my soul with thee to the courtesy will be a colour of the world, and therefore i will see them all the death with the duke of york, and therefore i will be a commonwealth of the death. the third of al"
High Dropout:
"and so the son to the stare and stand to the stare and stand to the stare and stand to the stare..."
Low Dropout:
"the sons, and then the state of the fields of the state. the sun in the field of the field. what say you to the protector of the see. the lord of were the commonwealth of the commons to the part of the seas of france, and then the state of the field"

In the future

- There are many directions to take the work presented here, here are some considerations

1. Higher level model (paragraph level?)
2. Stacked LSTM architecture
3. Even more samples, find a way to get around memory limitation.
4. Further hyperparameter tuning; minimise training time to achieve good results
5. Evaluation methods; accurate quantitative measures for performance.

References

[1] B. Sherman, and Z. Hammoudeh. "Make Deep Learning Great Again: Character-Level RNN Speech Generation in the Style of Donald Trump". Dept. of Comp. Sci., Cali. Univ.

[2] S. Xie, and R. Rastogi. "Deep Poetry: Word-Level and Character-Level Language Models for Shakespearean Sonnet Generation". Dept. of Comp. Sci., Stanford Univ.

[3] I. Sutskever, J. Martens, and G. Hinton. "Generating Text with Recurrent Neural Networks". Toronto Univ.