Style-imitative Text Generation using LSTM-RNN

Justina (Hyunjii) Cho, Madhav Datt, Andrew Zhou
hjcho@g.harvard, madhav_datt@college.harvard, andrewzhou@college.harvard
CS182: Artificial Intelligence, Fall 2016

Introduction

We describe an application of a recurrent neural network with Long Short-Term Memory (RNN-LSTM) units to generate text in the style of Sherlock Holmes. We construct such a neural network, along with a generator based on a Markov model, and investigate how choice of system and choice of parameters affect the training of our character-level language model and consequently the output of our the model. In particular, we evaluate how choosing system and parameters influences the quality, defined by both general coherence and potential overfitting, of the generated text, and we discuss the tradeoffs and relationship between these two criteria.

The objective of our investigation was to:

1)Design, construct, and train a system capable of producing text in a specific style.
2)Employ an evaluation metric to assess the quality of generated text.
3)Adjust the parameters and underlying methodologies of the system(s) and compare the quality of their results.

To this end, we implement a recurrent neural network with Long Short-Term Memory units (LSTM-RNN), as well as a second system based on a Markov model for comparison.

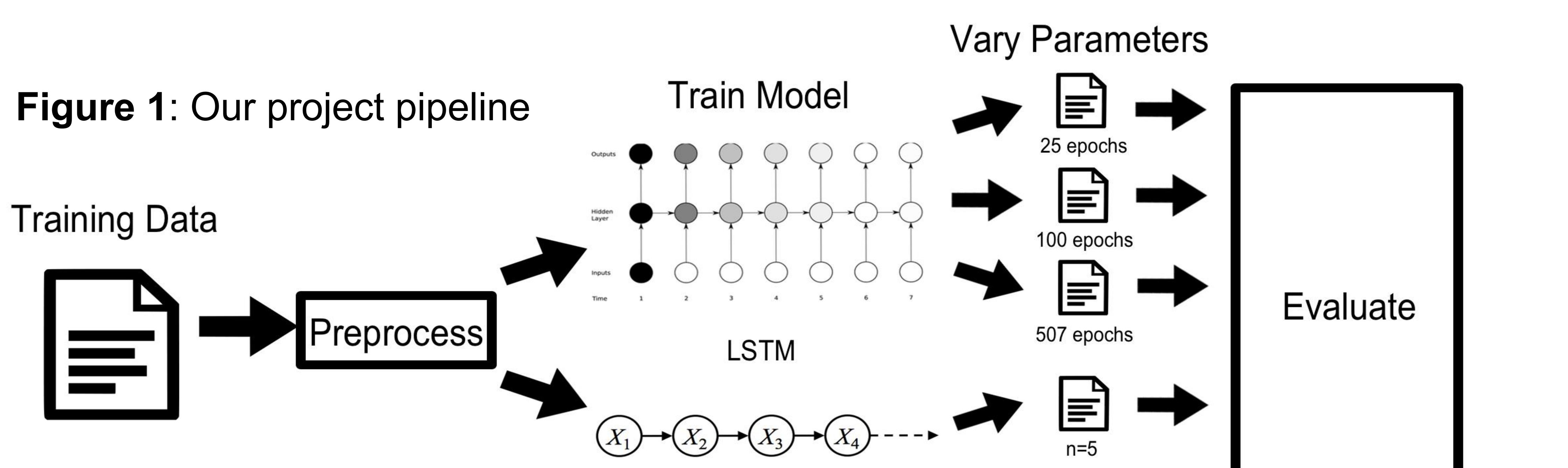
Approach

Our LSTM-RNN and Markov Model systems both use a character-level language model, train on a training set, and use an initial seed to generate text. For our LSTM-RNN, we use the Keras Python libraries with Theano as backend, and the Adam algorithm (Kingma, D. and Ba, J., 2014) as our optimizer. We train with GPUs on Harvard's Odyssey computing cluster due to the intensive computational power required.

We choose to vary the following parameters: the number of layers and units and the training epochs for our LSTM-RNN, and the number of characters in each Markov state. We train a separate language model for each set of parameters, and then evaluate the text generated by different models for its coherence and potential overfitting to the training set using the BLEU metric.

Figure 1 is an overview of our process.

Figure 1: Our project pipeline



Images from Icons DB, Berkeley Slides, and http://doi.ieeecomputersociety.org/cms/Computer.org/dl/trans/tp/2009/05/figures/ttp200905085510.gif

Data and Results

Text generated by our Markov model text generator, training on the first 100,000 characters of *The Sign of the Four.* The model calculates the probability of the next character given n previous characters.

n = 5, seeded with "dy di":

vidualities and the said her an

n = 10, seeded with "lem which ":

you might swarm up, wooden legs. on one occasional glimmering eyes upon us. "you will excuse me, miss," he said, "because you do not feel some uneasiness at the side of the water-pipe near. roof quite out of reach. yet a man has mounted the steps again. "facts are better than you think that the hiding-place was on his very late, and we came to heated words. morstan and i have made no use of it as any other man--?" asked holmes, "whatever you may say will be in vain. your

Text generated by our LSTM-RNN, training on the same set. We used 2 hidden layers and 512 units, with a training window of 100 characters and a batch size of 128. We did not decay so as to observe the training process from under to overfitting.

25 epochs, seeded with " who had addressed us mounted to the box, while we took our places inside. we had hardly done so befo":

re the sroang of the coor which i had the forr of the conroats should be a conticerable sroange of the coorer and plont pf the conroats and plont pf the conroats and plont pf the conroats and plont pf the conroats and plont pf the conroats and plont

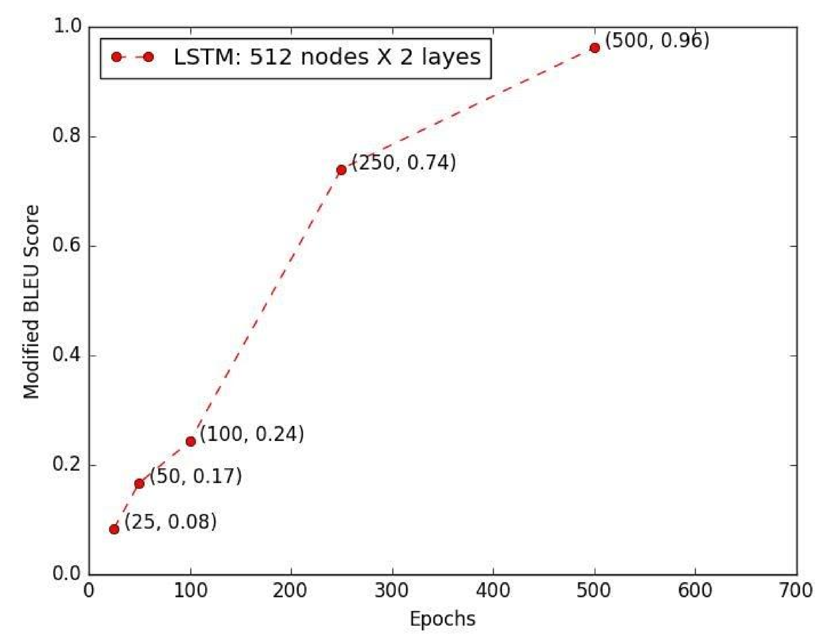
100 epochs, seeded with "earance of captain morstan. you will see from what i say that i can hardly be blamed in the matter. ":

a lost comice of the lanpern. the old coack shad it which h had thised his hands and helpless are all one out that i had a thirered langer, miss morstan, aut i am the only one of the fnure, i should be black, i have a shrire of his head, and had been

507 epochs, seeded with "final s. they are undoubtedly by the same person. i should not like to suggest false hopes, miss mor":

stan, but is there any resemblance between this hand and that of your father?" "nothing could be more unlike." "i expected to hear you say so. we shall look out for you, then, at six. pray allow me to keep the papers. i may look into the matter befor

Figure 2: BLEU score vs epochs (sans overfitting penalty)



Conclusions

We see that the Markov generator is repetitious for n=5 and runs into overfitting problems at n=10, where our generated text mimics the original precisely. This behavior is reasonable given each state has 26^n configurations, each of which will appear very sporadically in text.

Our LSTM-RNN demonstrates more diverse behavior. Though repetitious at 25 epochs, at 100 epochs we see novel constructions, structures, and words. At 507 epochs, however, the model has overfitted and replicates the original text verbatim. Figure 2 demonstrates the expected increase of BLEU score with epochs, though we have not yet penalized for potential overfitting.

References

1) Kingma, D. and Ba, J., 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.
2) Papineni, K., Roukos, S., Ward, T. and Zhu, W.J., 2002, July. BLEU: a method for automatic evaluation of machine translation. In Proceedings of the 40th annual meeting on association for computational linguistics (pp. 311-318). Association for Computational Linguistics.
3) Sutskever, I., Martens, J. and Hinton, G.E., 2011. Generating text with recurrent neural networks. In Proceedings of the 28th International Conference on Machine Learning (ICML-11) (pp. 1017-1024).

John A. Paulson School of Engineering and Applied Sciences