Project Proposal: Character level language models and sequence modelling

CSE 546: Machine Learning
Xiulong Liu

1 Datasets

a. Paul Graham's essays. Source: http://www.paulgraham.com/articles.html

b. Shakespeare works. Source: https://cs.stanford.edu/people/karpathy/char-rnn/shakespear.txt

c. Raw Wikipedia Source: https://cs.stanford.edu/people/karpathy/char-rnn/wiki.txt

2 Paper Readings

a. Generating Text with Recurrent Neural Networks, Ilya Sutskever, James Martens, Geoffrey Hinton.

b. Generating Sequences With Recurrent Neural Networks, Alex Graves.

3 Project Ideas

The motivation of this project comes from Karpathy's blog post on the Character level text generation using Recurrent Neural Network. It is interesting to discover how this powerful sequence model works on generating unstructrued texts and even structured texts. It is less magic for how RNN works for unstructured data after it's compared to a much more simpler and intuitive model, unsmoothed maximum-liklihood character level language models. However, for more structured data like Linux codes and XML, RNN really works magnificently in learning structure patterns itself, like nesting rules of braces and brackets.

Furthermore, I am really inpired by the visualization of the neuron firings which could account for what specific feature it learned from the structured text. And it somehow reveals a bit of underlying principles that RNN is by itself learning some high level features in order to generate similar patterns of structured text. But why it can learn itself is really worth studying and it's still an open question to answer.

4 Software

a. Python 3: Numpy, Scipy, Matplotlib packages.

b. Torch.

5 Will you have a teammate

No, I plan to work on my own.

6 Expected Experiment Results by Milestone

Complete training and evaluating unsmoothed maximum-likelihood character-model and LSTM with unstructured data and compare the results. On high order, two algorithms will show very similar performance in terms of accuracy rates of spelling, spacing and marks.