Hunter Camfield, Vishnu Kartha, George Amarhanov

NLP Project

Concepts Used/Approach:

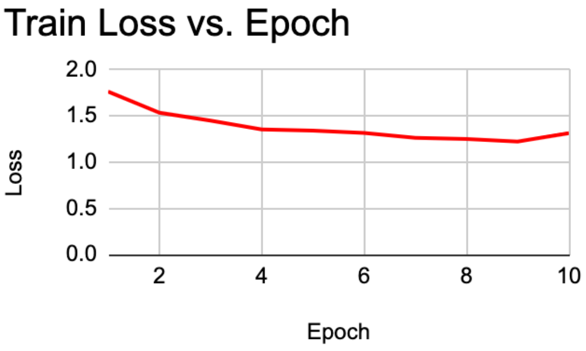
The first approach we used was to make sure that we understood the MyProgram framework and get the docker system working. Thus our first model was simple and returned an arbitrary 3 characters.

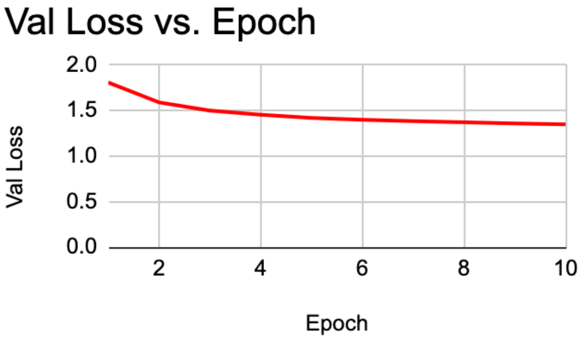
The second approach we used was a character level trigram model where we turned the input dataset into trigrams for training. During the prediction stage we looked at the last two characters in an input line and selected the 3 highest probability characters for the next unknown character.

The Final approach we used was training a character level RNN specifically an LSTM(to perform better with longer input lines) to predict the most probable 3 subsequent characters given an input. We used Adam to optimize the neural network.

Data Collected.

Based on the data we collected per epoch we can see the training loss and the val loss are approximately similar.





Dataset Used

The datasets we used were the torchText WikiText2 library. This library has over 100 million tokens and over 33,000 words in its vocabulary. We used this data set both in the RNN model and the trigram model. In the trigram model we only needed to add start tokens and in the RNN model we embedded each word and used it in testing.

Libraries and Packages

Pytorch: an open source machine learning framework that is flexible and uses a Python interface that is easy to use. Pytorch allows for high level tensor computation with strong acceleration via graphics processing units.

Torchtext: a companion package to Pytorch that includes popular datasets for natural language processing as well as data processing utilities. Torchtext allowed for easy to use portable datasets.

Work Cited

-Google. (n.d.). *Google colaboratory*. Google Colab. Retrieved March 14, 2022, from https://colab.research.google.com/github/abhaysrivastav/ComputerVision/blob/master/Chararacter_Level_RNN.ipynb#scrollTo=cLtLn8jnu8wS

Karpathy, A. (n.d.). The unreasonable effectiveness of recurrent neural networks. Retrieved March 14, 2022, from http://karpathy.github.io/2015/05/21/rnn-effectiveness/

• The report must be no more than one page (references don’t count against the page, figures/tables do), letter size, 1-inch margins, 11-point Times font, submitted as a pdf.

•Describe your approach, making use of concepts and methods learned in class. -Vishnu

• Describe the data you collected, existing datasets you used, -Hunter
and existing code libraries or packages you used.-George