Project 4: Spring 2022

Austin Nguyen
Brian Tan
Ricky Segarra
Rosa Cho

Dataset description: The dataset was the script of the simpsons from seasons 10 to 31 provided by kaggle. Because there was a large amount of data (words, characters, etc.) used in the training, it took some time to test and preprocess the dataset. The dataset included things like character names and episode titles other than the script lines, but we did not use them.

Preprocess description:
First we created a list of all the characters used in our dataset, so that we could convert the text string to an integer array. We did not include characters from other languages in order to simplify training. We discarded any lines that contained characters outside of our alphabet. In addition, we had 3 special characters: ‘start token’, ‘end token’ ‘pad token’. We prepended a start token and appended an end token to each line. In order to have all of the lines be the same length, we added pad tokens to the end of each string.

It's likely that it would have been more interesting for our results had we included punctuation in our alphabet.

After we encode all the relevant text, we save the data in a numpy file.

Preprocessing is done in a separate script so that we don’t have to preprocess every time we train.

Model:
A simple multilayer GRU network inspired by Tenesorflow’s “Text generation with an RNN” tutorial. We used 1 layer as a baseline, then added more layers to increase the coherence of the generated text with not that much success unfortunately.

Training:
Our training setup is similar to the one found in the tutorial, however since we padded our sample texts, we used a masked loss function instead. Also, instead of using the built-in fit(), we did manual training using gradient tapes so that we could generate text after each epoch to track progress.

Training is performed in main.py

Results generated:

'sir you feed all right lisa if theres all night whens the plus in take it'
'oh yes dy shoe'

After more epochs
'no ive seen-up'
'marge a very man to think and driving for a cut it assolens were you right'
'about the saturd i were call of the last women of the virrs has broken of sushe this inspector it up'
'amour station'
'robes to my ride a man date junks your fever bring line or dog field boat homie but the family us'
'cincies temalious'
'oh i cant get in a whinn else'
'oh'
And more
haw has me from of that phones you put climbery criming this carvage man'
'why dollars prom buttons wake this polon'
'wait they will'
'americal stuck-are for the born alive im going and youre move are trying all'
'ow dumb its your cabbans'
'oh die'
'lisa everybody get the boses i know e stupid'
'its say the tabs kid whoa what will that ill look'
'homer whats your sister house bart'
'shuts i wear evening yard of the hardes in a higher the nerd mom help you holding any my cause you c'
'woo humiliate lisa eligist frindo'
'woo hoo of they do into you to our touch will be got you married that up there is this eall this'
'yeah imself l tired his will pramiss of whoop suit'
'these can ip even bart great danized tectok to plead'
'my vawent to gake thats noiken work was me'
'hmmm what was the cherie fifty-isabent the waited was for'
'ngind it about with me um hmading ill flighty one of anymoryee and of you'
'will in the shoilding a him barts advalking of them chepass night they well uh well the quiet'

We will provide a file contaning a full set of outputs for each epoch as well.