Kenneth Hua

Module 6 Final Report

Background:

Content Creation on Youtube is a large and growing business, which allows individual creators to earn revenue from videos hosted on the youtube platform. Creators are always seeking out ways to make more revenue from their videos by increasing viewership.

Data:

The Data was obtained using Youtube Data API. For our training, the data collected was limited to Travel Videos. The acquisition was done in a few steps:

1. Videos were searched through the Search Query by keyword and ordered by popularity. Keywords were “[City Name] + Travel”. The City Names are ordered from a list of high population cities in the world
2. Then from the video_IDs, the title and viewcount of each video can be extracted.
3. Extract information from the json, clean the title strings, and store in a pandas dataframe

	Unnamed: 0	video_id	video_title	view_count
0	0	3r4x4tztZ7E	11 Things To Do in Yangon, Myanmar (Are You Re...	949267.0
1	0	QFsmfF77GEI	Vlog - 3 | Yangon, Myanmar Diaries | Travel vl...	635046.0
2	0	z1D1P9-_oNA	MYANMAR STREET FOOD TOUR in Yangon | Delicious...	540365.0
3	0	UivJLq-anyk	A Taste of Yangon, Burma (Myanmar) - Burmese S...	440729.0
4	0	8Ai6AZfUcSM	On Board With President Obama - Rangoon, Burma	304705.0
...	...	...	...	...
16610	0	e6UMHe5HW_c	AKCAKOCA | ZONGULDAK | ANKARA | CANAKKALE | TE...	115.0
16611	0	mBZqS3a85Ws	Usa ko new city travel TEKIRDAG to KOCAELI🚃🚃🚃🚃...	17.0
16612	0	fky2PiOz-Rs	10º Viaje. Tallin (Estonia) - Tekirdag (Turquí...	16.0
16613	0	78MYia_NaHk	Travel TEKIRDAG TO SANLIURFA BUS SIMULATOR UNL...	8.0
16614	0	hd5AeN8dQ_U	Tekirdağ'ın turkuaz plajları her mevsim gü...	22.0

16615 rows × 4 columns

The video_title data was cleaned to remove non alphabetical characters, make all characters lowercase, and remove spaces. Spaces were an issue originally, so spaces were removed.

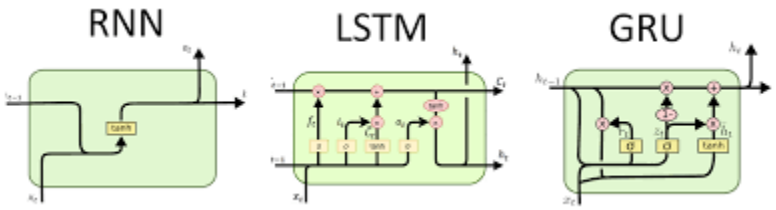
From the cleaned text, we generated sequences of fixed length. This was done with an offsetting method, so that each sentence can generate multiple sequences with different offsets.

The sequences were one hot encoded, and split into X and y, where y is the last character of the sequence, and X is the rest of the sequence. X and y were then split into train and test sets for further work.

Model:

For text generation, the data is sequential so we need sequential neural networks to be able to tackle the problem.

There are a few models that we can explore with this type of problem: RNN, GRU, and LSTM. After doing some experimentation with all three models, I chose LSTM, because it is capable of handling the memory of longer texts and does not fall into loops as frequently.



My model consists of

1. LSTM with 4 layers
2. Linear Layer

The input is a batch of the one-hot encoded sequences and the output is a one-hot encoding of the most likely next character.

Results:

Our training loss decreased over epochs trained, so the model is learning to our data quite effectively. The final accuracy was 60.53%

By looking at a few examples we can see:

Seed Word	**Generated Phrase**
“Walk around”	“Walk around ianamia travel vlog”
“Europe advice”	“Europe advice of air ane around ianamia travel”
“Japan places”	“Japan places to visiting and angalore”

“Shanghai food”	“Shanghai food ange road trips olone road trip”
“Street Food”	“Street Food tourist an travel vlog”

As you can see from some of the examples, there is some coherence, but certainly a lot of room for improvement. Text Generation is quite a challenging problem, so a lot of improvements/new model architectures need to be experimented with to further improve our results.

Conclusion and Next Steps:

Conclusion:

- Our text generator is able to generate text from a seed word, and can pick up some fairly common patterns
- However, in its current form, the generator could be higher quality, and can get stuck in loops with certain seed words.

Next Steps:

- There are a few approaches that can be done to improve the model
- Give more training data or adjust the parameters of our current LSTM model to minimize the probability of pattern convergence.
- Instead of character predictions, we can build a model to perform word predictions. Because the amount of words in the corpus will be very large, this will need to be done by creating embeddings of the words.
- There is also the option of creating a BiDirectional-LSTM model, which will have the capability to

Main Tools:

Google Client - To connect with Youtube API more effectively
Numpy - For data manipulation
Pytorch - For deep learning