NLP Final Report

Gavri Kepets

May 2022

1 Introduction

For my final project in ECE-467 Natural Language Processing, I set out to build an RNN that can generate new movie scripts. The RNN is trained on the scripts of the original Star Wars movies, and the goal was to have the RNN generate a new Star Wars script. For this project, I used TensorFlow to build and train the RNN. I went through numerous iterations to arrive at a relatively successful result.

2 Design Process

2.1 Initial Attempt

For my first attempt, I fed the RNN the raw text from the movie script. The RNN trains on a word by word basis, so each word from the movie script was tokenized. Relatively generic hyperparameters were chosen for this first attempt. The RNN consists of an embedding layer, an LSTM layer, and a dense layer. The RNN had 4,614,403 parameters, an initial loss of 6.21, a final loss of 2.42 and an accuracy of 38.4%.

Table 1: Hyperparameters for first RNN

Batch Size	128
LSTM Size	128
Amount of LSTM Layers	1
Embedding Dimensions	200
Epochs	5
Learning Rate	0.01

This resulted in some pretty awful text. The RNN yielded mostly gibberish, and it barely resembled a movie script. One output from this RNN read as follows:

> pulls up a chain toward Luke, bloody and battered, hurry to his face
> and the robots turn and steam. Seeing an area At him closer. The

> droid and disintegrated the Wookiee, but Leia thinks he turns on two metal big No one shut with retreats before seems Luke's LUKE'S X-WING- COCKPIT- panics at the opening. There is Red Two. His tremendous explosion and the young senator, of the spacecraft in the side side of the troopers zooms over the surface to an of the bounty hunter's in terror. It stares on the Incom to his ignited on the lifter.

2.2 Preprocessing the Data

In order to improve upon the previous RNN, I decided that the data needed some preprocessing. This time, the punctuation and new lines were seperated from the words and tokenized seperately. Additionally, each line is designated as "Dialogue", "Exposition", or "Setting". That way, it is easy for the RNN to recognize what a setting, dialogue, or exposition in a script may look like. Settings are all capitalized and describe a location, dialogue is initialized by a character and is in first person, and exposition is in third person and typically describes what is happening in the scene. By specifying each line, the RNN can more easily recognize what type of text should be part of expositions, dialogues, and settings. The preprocessed data looks like this:

> |SETTING|: INT |.| MOS EISLEY SPACEPORT |-| DOCKING BAY 94 |NEWLINE|
>
> |EXPOSITION|: Luke |,| Ben |,| Threepio |,| and Artoo move toward the Millennium Falcon passing Solo |.| |NEWLINE|
>
> |DIALOGUE|: THREEPIO: Hello |,| sir |.| |NEWLINE|

This was done with a small algorithm that determines what each line is and writes it to a new file. More code was written convert this format back into the original move script format. The RNN had 3,722,743 parameters, an initial loss of 5.28, and a final loss of 2.82 and accuracy of 38.76%.

Table 2: Hyperparameters for the Improved RNN

Batch Size	256
LSTM Size	128
Amount of LSTM Layers	1
Embedding Dimensions	250
Epochs	10
Learning Rate	0.01

I trained it on more epochs and increased the embedding dimension and batch size. The results were much better- the RNN almost always successfully recognized when there was a setting, exposition, or dialogue, and followed through with that format. Unfortunately, it was still mostly gibberish, but it at least resembled a move script.

```
INT. STAR DESTROYER - TRACTOR

                                BIGGS:
                      Look, a new enemy - shrouded.

                              RED LEADER:
                              cross Wedge.

                                 BEN:
                    To you've picked one of Obi - Wan
                              Kenobi.....

124 EXT REBEL CRUISER - BRIDGE

The Dark Lord raises, clinging away, eager and attempts to increase,
louder and examines the hydraulic Nice. Come over, a large hypodermic
needle. Blood streams slackly to their breath.

EXT FOREST CLEARING - LEIA'S CRASH SITE

The procession moves unconscious.

EXT. SPACE - SNOWDRIFT - MAIN ICE offensive - LOWER

The AUNT fleet button, blocking Luke's radio transmissions disrupted
by a game Creature.

                                LUKE:
                    It is a thing! It's magnetically
                                sealed!

Biggs and goes off majestically by Vader appears to translate... and,
is separated appears - covered.
```

It even does a good job picking all capitalized words for the settings and characters who are speaking.

2.3 Tuning the RNN Hyperparameters

Finally, I tuned the hyperparameters of the RNN, and ran it for much longer. Most notably, I used three LSTM layers, and trained the model for 15 epochs (this took about nine hours on my PC!). The RNN had 3,963,543 parameters, an initial loss of 7.9464, and a final loss of 2.5825 and accuracy of 40.01%.

Table 3: Hyperparameters for the Final RNN

Batch Size	64
LSTM Size	128
Amount of LSTM Layers	3
Embedding Dimensions	250
Epochs	15
Learning Rate	0.01

This yielded better results.

```
INT. REBEL BASE- MAIN HANGER- SERVICE PANEL

Chewie barks a dirty sword system who stands followed by rolling
sideways.

                                YODA:
                    Mmm. Friends you can feel Vader we
                       built the dark times in the
                     security readout, as only hope.

                                THREEPIO:
                           We're a on for those Star
                            Destroyers something!

INT. LUKE'S X- WING- COCKPIT
                                RED LEADER:
                             Watch yourself !...

                                LUKE:
                                Ten...

Luke's voice turns off. His Imperial audience up to the intercom.

                                HAN:
                         It's a meteorite to Antilles as
                        they see, been lost to target...

Suddenly the Imperial probe Imperial TIE fighters run around
attending- Rebel speeders. Suddenly, something off ?

                                LUKE:
                                 Uh...
```

The sentences are much more coherent, but still not perfect.

3 Conclusion

The results of this project are mixed. While I could not successfully train an RNN to generate a coherent movie script, the RNN did successfully learn how to structure a move script. Dialogue (while generally incoherent) was structured as it should be, with first person perspective, a name before the speech, and more expressive punctuation. Settings typically are completely capitalized and describe the setting properly, including specifying whether a location is INT (interior) or EXT (exterior). Exposition is almost always in third person, and usually resembles what exposition would look like in a movie script.

One way to improve this would be to train multiple separate networks; one network for figuring out the layout and order of the script, and the others for specifically generating exposition, dialogue, and settings. Additionally, the model could train for even longer, as this project was built under a hardware and time restriction. Finally, I think contextual word embeddings may be extremely helpful for this type of RNN. With contextual word embedding, the RNN can differentiate between language used in exposition, settings, and dialogue. A character can be associated with an action (which would probably appear in exposition), a location (which would probably appear in a setting), or speech (which would appear as dialogue). Overall, these results are a great stepping stone towards making an even better RNN for generating move scripts.

4 References

- https://machinelearningmastery.com/choose-an-activation-function-for-deep-learning/
- https://shiva-verma.medium.com/generating-a-tv-script-using-recurrent-neural-networks-dd0a645e97e7
- https://machinelearningmastery.com/stacked-long-short-term-memory-networks/
- https://towardsdatascience.com/generating-tv-scripts-with-lstm-e94be65a179