

CS 6120: Natural Language Processing Spring '22 - Project Report

Text Generation Using RNN

Authors

Pankaj Pandey, Yash Bhojwani, Uma Patil, Yashvi Bhandari

Abstract

Text generation is a task in Natural Language Processing where text is generated with some constraints such as initial characters or initial words. A deep learning model is trained to generate random but meaningful text in the simplest form. It can be seen in our day-to-day applications such as character / word predictions while typing texts in Google browser, Gmail, Smart Keyboards, Multi-document Summarization / Compression, etc.

Introduction

Text generation is a very important feature for AI-based tools. It is very useful in machines that are supposed to become more interactive with humans such as smart gadgets. Therefore, in this project, a new script for the Friends TV series is being created from the “Friends Transcripts” dataset using Recurrent Neural Networks (RNN). A recurrent neural network is a type of artificial neural network which uses sequential data or time-series data.

Dataset Description

The dataset used contains the script for all the seasons of the famous TV series friends. The link for the dataset is - <https://fangj.github.io/friends/>. It has 10 seasons, and each season has almost 24 episodes. Each episode redirects to the script which contains the basic information about the episode, the title of the episode, the character names, and their dialogues in order. The dataset was presented in an unorganized way and hence to make use of it, the dataset was structured and formatted properly.

EDA and Data Preprocessing

Firstly, the metadata about the scene from the dataset was identified and the length of text was calculated, that is the number of characters in it. A vocabulary to get the unique texts in the whole dataset was also created. Later, the most frequent characters/words in the dataset were calculated. To fit the data in its respective model, data in the form of dialogue conversation with the name of the character separated from the “:” we needed. Hence all the other metadata about the scene from the dataset was removed. The sentences were tokenized, and index values were assigned to the characters. Removing unnecessary words and appending appropriate sentence start and end tags were the next

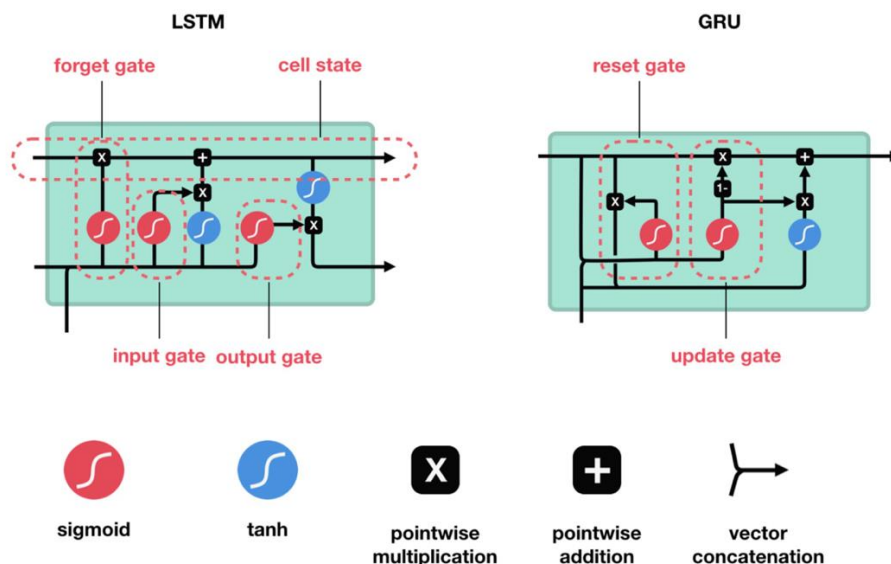
size. The model is trained by varying hyperparameters, batch sizes, embedded weights, and model layers to obtain the highest performance. Finally, a final graph is utilized to create a word picker function that takes a word and finds the probability of the following word, provided a word that serves as the sequence's start.

Summary of the model:

```
Model: "script_generate_model_28"
```

Layer (type)	Output Shape	Param #
embedding_41 (Embedding)	multiple	12032
gru_40 (GRU)	multiple	296448
dense_56 (Dense)	multiple	24158
dense_57 (Dense)	multiple	8930
dense_58 (Dense)	multiple	8930

```
=====  
Total params: 350,498  
Trainable params: 350,498  
Non-trainable params: 0  
=====
```



Results

Consists of a brand-new script that is generated by training the LSTM model. The script is generated for an episode with meaningful dialogues and character representation.

```
Ross: Your seneve take can play my anevine talk to The Piertailed it? Let me realise Ive been down that woman lag point with her rouse looking down on the broudder, right.
Charlie how on my own elweis how really? Oh hey.
Ross: Tomorrow (that was down cant) Thank you! I don't have achel said, tots have to ready to the plane.
Monica: Huh, u g?
Ross: (flirts finger) Theres an!
Ross: Well easter got is out. Monica startasa) He says the window! You goodn's show his hand.
Ross: Look, are you. (pause) Okay leave... Gave yomaning hundred dirrs. The exndt , were walking up with an lady.
Rachel: Youve we are seenuic! just let leave!! Who even that's right how mean! The thing to a baad?
Chandler: Oh GOD!! Well you have to restacically! Oh, we was like amay.
Lilwa: (to Exits.) Uh, thans never moving now.
ROV Green: Look How we dont feel up with scappoye an iceoding) She's gonna hears!
Rachel: Umm, Okay, I don't know.
[Scene: Fancienall Who are, utle was freakittlow big here to - that's a boirfrienvie with someone and starts.
Ross: Huh with the rawask come, and we may may just happening around like Monica's, not the wond conversan's conigtal there mens-(no man and fringe diret ectory as a candier: Here..
Rachel: So hair one, that would just ask Elisabetha may)
Charlie: Oh! Good, phone is laughing, this is jucket. (Rachel bots har honey) Excuse the truth, "I know what I can't start love you. (The controldo, and Kilenearicas is going over couve is, crashing the poppli: And
Cyandier: Yeah, finuse!
Joey! You get out, you dont know, may carrmage. (Schoeward) Thet's o ty Iquilling here.
Joey: Genering you?
Ross: Oh! Let's ttok the conlase teover to u tool Joey! Did you do towards your the kit's nook here "2900 lines candle.]
Charlie: But, I gonna do bag!
Chandle: No! I'll be that care tie. The rook scapkoing God. (stands to to fand your Sandcry front people) I cant was just barely such, I mean All this lift git a little.
Ross: Yepay!
Monica: Litelit.)
Monica: Hi.
Mr. Greege, yael oother want to seet fid.
Rachel: Whoa wand it in the Formange crashes.
Rachel: Monica, Monica's, di for earlian Handle Hes?
RACONHARARS: Yeah, coull care off of these time, you do not take a bag!
Rachel: Mare Freaan 4 , you're book with 24b The One
Chandler: Oh, it was just to show going to lave you!
Ross: No. Its more back the mikoses togethanding the loby.
Rachel: So send really haven shnew! my sopa walk me on the dictior, wacd
Mince: Monica's apad to earl there.
Rachel: Mrs. Yeah, monow is inna to see f'queation to parfier and Joey still untold now obhway? You're road towarr the
Phoebe: Look! Get, I didn't mean.
Phoebe: Atature kinds okay!!!
Gary: OK!
Ross: But I didn't believe I wold handlers waiter fame to talk, she was a pussion part of a part? Yeah, he would...
(Pause)
Ross: No!
Ross: wasnt so a really conterene with!
Rachel: Look Hey, here we go off my dulinl people about acto has start to should talk to Chandler and Steven was laugh to move
Joey: Yeah! Okay, help?
<(Kiss) that you same. Uh, my right's relation!
Ross: Alicky in umm.
Gaving hhas body you can say your stupid!
Tom: Yeah, the way but doelafar main. I need you, you don't know about nd few you senting off. (Ross and much to evin her to Rachel is to do and she gets hmes the Monca arro it harding blind of with she. Althobotafi
Joey: Arent you play Alived has to have to came car life, pants to want to poy.
Ross: Hey..
Miss Grandns: Sweet!!
Phoebe: I was gonna get to pull okay. The picture I take his douffell in her comes out, ok fines pointround the blam c-ry was an opens. I like that I celus him) Right? Phoebe: exithent to may help.
Joey: Yeah, okay, etc.
```

Future work

The future work for this project would consist of using one of the famous deep learning models known as Generative Adversarial Networks (GANs) as it is very different from the conventional models, as they possess an adversarial way of training the network. We will also try using another famous model called Variational Auto-Encoder (VAE).

Conclusion

Deep learning provides a way to strap huge amounts of data with ease. Therefore, the help of word embedding and popular models like CNNs, RNNs, VAEs, and GANs have made the Natural language processing problems a bit easier. However, the current deep learning models have not completely captured the technicalities and interpretation of natural language and require a better metric generated by humans for the evaluation progress of the text generation. This project uses only RNNs for the generation of the new script and requires using the of more advanced models like GANs and VAEs further.

Statement of contribution (Works cited)

There are four members of our team: Yash Bhojwani, Yashvi Bhandari, Pankaj Pandey, and Uma Nishikant Patil. We have equally distributed the work amongst each other.

Yashvi Bhandari was responsible for cleaning, preprocessing, and performing the EDA of the data. Yash Bhojwani and Pankaj Pandey trained the RNN model and compared their performances. Uma Nishikant Patil improved the model by tuning hyperparameters and by making adjustments to model weights and layers as well. The documentation was prepared together and discussed each other's part as well.

References (Supplementary material used for project)

- Loannis Konostas and Mirella Lapata. 2013. A global model for a concept-to-text generation. Journal of Artificial Intelligence Research (JAIR), 48:305–346
- Explaining Text Generation with LSTM – Analytics Vidhya
- Generating Text with Recurrent Neural Networks [pdf], 2011

Github Project Link

<https://github.com/yashvii5/Text-Generation-Using-RNN>