Comparison of NLP Models for a Writer and Genre Style Controlled Movie Screenplay Generator

Stanford CS224N Custom Project

Koye Alagbe
Department of Computer Science
Stanford University
kalagbe@stanford.edu

Shridhar Athinarayanan
Department of Computer Science
Stanford University
shriathi@stanford.edu

Gautam Pradeep
Department of Computer Science
Stanford University
gpradeep@stanford.edu

Abstract

Our primary motivating question is to determine how do various NLP model architectures compare when tasked to generate a movie screenplay in the artistic style of both a user-inputted genre and movie director? There has been some previous work in long text generation, but we are building various models on a novel application of having multiple user-inputted stylistic dimensions to control in generative text. Crafting a movie screenplay is a very challenging and difficult task; however, we hope that our work will be able to aid writers in their goals of screenplay writing, bringing stylistic and creative ideas from previous directors' work in a genre they desire. We compare various models and approaches, particularly OpenAI's GPT-2 language model along with the DistilGPT-2 model, ultimately finding DistilGPT-2 under 10 epochs yields the best screenplay generation. Our future work might investigate LeakGAN and Aggressive Variational Autoencoder (VAE) models for text generation. We compare our model performances against a baseline LSTM architecture.

1 Key Information to include

- Mentor: Kaili Huang
- External Collaborators (if you have any): None
- Sharing project: None

2 Introduction

Over the past few years, neural text generation has captivated the interest of natural language processing communities for its various applications. We were mainly interested in how neural text generation could be applied in a creative avenue for long texts. The project we decided on is an automatic screenplay generator that takes as input one or multiple genres and one or multiple directors, and tries to generate a screenplay of the given genre(s) in the style of the given director(s). We believe that a model like ours could be used to gain interesting insights on the similarities between directors and genres, as well as break down quantitative structural elements that make up popular screenplays. This knowledge could then be used by human screenplay writers or other neural models.

3 Related Work

One of the most rudimentary forms of neural text generation is the use of Long Short Term Memory, or LSTM, networks. They are a form of Recurrent Neural Networks (RNNs) in which an input, forget, and output gate will dictate what information is to be retained and what is to be forgotten within hidden layers [1]. Wei et. al. implemented a Chinese poem generator controlled by the personal style of various Chinese poets through a simpler version of an LSTM called a Gated Recurrent Unit, or GRU. Their model concatenates input poetry X_i with the embedding of its associated poet's name S_i and the previous hidden layer unit h_{t-1} to generate a new hidden unit h_t. This would ultimately encode a certain poet's style [2]. Text generation with LSTMs in Pytorch has certainly been accomplished before, showcased in various blog posts such as Trung Tran's which we adapt to form our LSTM infrastructure [3].

Mainly, however, OpenAI's GPT-2 language model is commonly used for generation of long natural text. The model is transformer-based, trained on over 8 million web pages and equipped with 1.5 billion adjustable parameters [4]. GPT-2 uses auto-regressive language formulation to generate the next word in a sequence based on a probability distribution conditioned on previous tokens.

There are multiple examples of text generation using GPT-2 since its easily fine-tunable and its ability to limit exposure bias, or when a model overfits to its training data rather than adapting to its input text [5]. The research space with GPT-2 for text generation is replete with even more literature on poetry production. Most notably, researchers Köbis and Mossink were able to find that a group of 30 participants were not able to differentiate GPT-2 generated poetry from human written poetry using an incentivized Turing Test [6]. GPT-2's ability to develop poetry with a human level of cogency, fluidity, and creativity solidified our motivation for employing the model to generate screenplays.

4 Approach

For our baseline model, we employed the use of an LSTM model. LSTM models are a form of RNNs which address the issue of short term memory for long text generation. LSTMs, like RNNs, are composed of chained hidden layers which encode information about a whole sentence as it processes through it. However, the LSTM process includes forget, input, and output gates to determine what information remains important to carry onward. We adapted Trung Tran's blog post for long text generation using LSTMs to conduct this baseline review [3].

For our main experimentation, we utilized OpenAI's GPT-2 transformer-based model for synthetic text generation. We decided to utilize this architecture since it is openly available through HuggingFace, its pretrained on ample webpage data, and because it offers manifold capabilities for finetuning language generation regarding downstream tasks, such as screenplay generation. The initiation of the GPT architecture came from Radford et. al. They guide their unsupervised pre-training method by trying to maximize the following likelihood $L(U) = \sum_{n=i} logP(u_i|u_{i-k}, ..., u_{i-1}; \theta)$ where U is an unsupervised corpus of tokens $\{u_1, u_2, ..., u_n\}$ [7]. The model is composed of a multi-layer Transformer decoder which employs multi-headed self attention over the input tokens and feedforward operations to ultimately generate a softmax distribution for potential output tokens. The hidden layer and probability distribution equations are listed below, taken from Radford et. al:

$$h_0 = UW_e + W_p$$
$$h_l = \texttt{transformerblock}(h_{l-1}) \forall i [1, n]$$
$$P(u) = \texttt{softmax}(h_n W_e^T)$$

where n is the number of layers, W_e is the matrix of embedded token words, and W_p is the matrix of embedded positions of tokens [7]. To formulate the full-fledged GPT-2 model, Radford and other researchers added further modifications including layer normalization to the input of each sub-block, scaling of residual layer weights by an inverse factor of the square root of N, the number of residual layers, and an expanded vocabulary and batch size [8]. We adapted code from the following Towards Data Science Blog post on movie story generation based on genre, feeding and training on our scraped script data [9].

Though GPT-2 is a current state-of-the-art transformer based language model, its multitudinous list of parameters makes it difficult to scale with large, cloud-based data. To re-parameterize these models to

have less weight, there needs to be some level of abstraction. With Distil models, we train a "student" model derived from our original "teacher" model to match the teacher's output distribution through a revised training loss function [10]. We employed the DistilGPT-2 model in our own tests through the use of the HuggingFace API.

5 Experiments

5.1 Data

Our dataset was acquired from scraping The Internet Movie Script Database (IMSDb) online [11]. This database provides a comprehensive repository of scripts from movies dating from the early 1900s to today. Each movie in the database contains not only the screenplay, but also the director and the associated genres, along with other data on ratings and reviews. The dataset consists of about 1200 screenplays, each averaging about 30000 words. We built a web-scraper from scratch to be able to crawl through the website and extract information on each movie, including its writer(s), genre(s), and the entire screenplay. This was done using the Beautiful Soup Python library for extracting HTML text from websites. We collected 1169 usable screenplays and outputted each as a string as a new line in a file where each movie is in the format of **"<BOS><writer>...<writer><genre>...<genre>full screenplay<EOS>".**

We then split the dataset into 80% training set, 10% dev set, and 10% test set. This was followed by tokenizing each set where each word is considered a token, along with special tokens (<BOS>, <some director>, <some genre>, <EOS>).

Following this, we further break up each entry in the training set into blocks of 1024 tokens which will allow for faster and more optimized training compared to training on entire 30000-word screenplays.

5.2 Evaluation method

The first evaluation metric that we implemented is perplexity. The perplexity of a probability distribution, defined as:

$$PP(p) = 2^{-\sum_x p(x) log_2 p(x)} = \prod_x p(x)^{-p(x)},$$

is a measurement of how well the model predicts configurations of words in the dev set. A model with lower perplexity is generally better at predicting dev set examples.

After calculating perplexity, we also calculated BERTScores for the best performing models. BERTScore, defined using precision, recall, and F-score as follows:

$$R_{\text{BERT}} = \frac{1}{|x|} \sum_{x_i \in x} \max_{\hat{x}_j \in \hat{x}} \mathbf{x}_i^\top \hat{\mathbf{x}}_j, \quad P_{\text{BERT}} = \frac{1}{|\hat{x}|} \sum_{\hat{x}_j \in \hat{x}} \max_{x_i \in x} \mathbf{x}_i^\top \hat{\mathbf{x}}_j, \quad F_{\text{BERT}} = 2 \frac{P_{\text{BERT}} \cdot R_{\text{BERT}}}{P_{\text{BERT}} + R_{\text{BERT}}} \ .$$

is an evaluation metric designed for text generation that takes a reference sentence and candidate sentence and computes the similarity of each token of the two, with token embeddings taking context into account.

Finally, we used human evaluation metrics on generated screenplays. Human evaluation is a popular choice for NLP tasks having to do with text generation, because automatic metrics often do not capture things like structure and coherence as well as a human reader would. We used a rubric based on commonly cited criteria, namely coherence and grammaticality, as well as criteria specific to our purpose, namely ratings of how well the generated screenplays adhered to both the genres and the directors. For our top performing GPT-2 and DistilGPT-2 models, we had ten volunteers rate three outputs from each model on all four of these attributes, and collected average human evaluation scores.

5.3 Experimental details

For the LSTM model, our text-generation parameters included:

• Sequence Size = 32
• Batch Size=16
• Embedding Size=64
• LSTM Size=64
• Gradients_norm=5

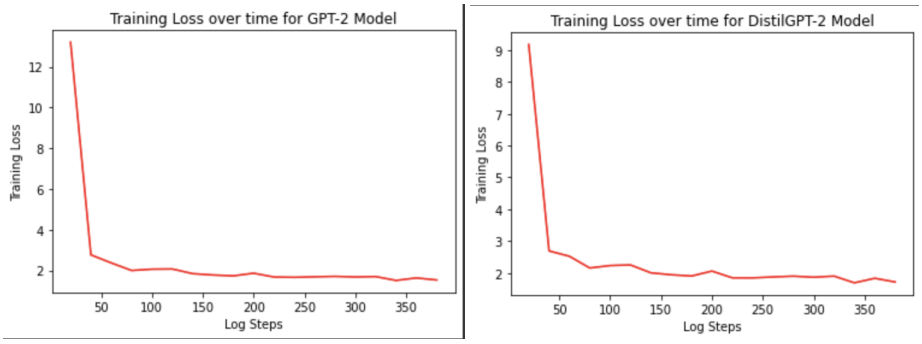
We trained this LSTM model with our html-stripped training data, which calculated loss every 100th time step. After 5 epochs, we evaluated the outputted text against our cleaned test dataset by self-calculating perplexity. In addition to our baseline LSTM, we ran GPT-2 three times, each with different numbers of epochs (3, 5, and 10). The GPT-2 Model employed an AdamW optimizer. Overall, our parameters included:

• Training Batch Size = 4
• Learning Rate = 5e-05
• Adam Beta = 0.9
• Adam Epsilon = 1e-08

In addition to running the GPT-2 model, we ran these experiments with the same parameters for DistilGPT-2, which runs about twice as fast as GPT-2. We expected to get optimal results for GPT2 model over 10 epochs and were under the impression that the faster computation time and lower complexity of DistilGPT-2 would also add to the reasoning behind GPT2's superiority. Our results were conflicting with these hypotheses.

5.4 Results

Model Evaluation and Comparison	
Model	Perplexity
LSTM	7.25
GPT-2 (3 Epochs)	5.87
GPT-2 (5 Epochs)	5.69
GPT-2 (10 Epochs)	5.86
DistilGPT-2(3 epochs)	5.81
DistilGPT-2(5 epochs)	5.31
DistilGPT-2(10 epochs)	5.04



Model Evaluation with **BERTScore**			
Optimal Model	Precision	Recall	**F1 Score**
GPT-2	58.4	61.9	60.0
DistilGPT-2	80.8	87.6	84.0

Firstly, it is clear that the perplexity of the model for LSTM is higher than both GPT-2 variants, which was expected. Also, we notice that the DistilGPT-2 model performed much better than the GPT-2

model in both the BERTScore and perplexity. This was not what we expected although, as further explained in our Analysis section, we think that that the model was overfitting for this task in the GPT-2 model, which may have caused the DistilGPT-2 model to have a better performance. This is not true for all tasks, but this seems to be the most plausible answer for this discrepancy. We were very surprised to see such a high F1 score for the DistilGPT model which proved to be higher than the corresponding BERTScores from other models in our researched literature on similar tasks.

5.5 Human Evaluation

Average Human Evaluation of Screenplay Generator				
Model	Coherence	Grammaticality	Genre Adherence	Director Adherence
GPT-2	4.2	6.8	6.1	2.7
DistilGPT-2	4.0	5.7	5.9	3.4

The above results show the average scores that were given through ten human evaluations with this concrete criteria we provided on three different screenplays outputted from each of the GPT-2 and DistilGPT-2. We have displayed a few of these screenplays in our appendix.

6 Analysis

We know that a lower perplexity is indicative of a better model, and LSTM sensibly has a greater value for perplexity than the GPT-2 and DistilGPT-2 models as it is a less robust model. Generally, we see that as we increase training length (more epochs), we see a lower perplexity. It appears that GPT-2 may be overfitting at 10 epochs which can cause some decreased performance from 5 epochs. The overall underperformance of GPT-2 relative to DistilGPT-2 could also be explained by overfitting due to the fact that the GPT-2 model is larger with more parameters than DistilGPT-2.

DistilGPT-2 performed best when fine-tuned with 10 epochs, and we can see the training loss decrease through the training process for both models, as expected.

The BERTScores metrics computed on the DistilGPT2 model for an example reference and hypothesis text were much higher than that computed over the GPT-2 model, which is in line with our theory that the GPT-2 model ran into overfitting problems.

As for human evaluation metrics, neither model performed particularly well. What we noticed while reading some sample outputs were that while some truly read like screenplays, others read like stories or plain text. We believe this is due to the data that was used during the pretraining of the models. We hypothesize that if the models were trained from scratch on our dataset of screenplays, they would be better suited to producing them.

7 Conclusion

The overall purpose of this project was to see if a model could learn genre- and director-specific characteristics while generating screenplays. Ultimately, after comparing two models against an LSTM baseline, we were able to see that the auto-regressive DistilGPT-2 model trained on 10 epochs performed the best based on our evaluations of perplexity and BERTScore.

Through our research inquiries, we were able to learn about the main technological infrastructures behind long synthetic text generation, as well as the data cleaning and pre-processing which goes behind establishing a neural text generation system. However, our process did involve some noteworthy limitations. Our primary limitation was that our training dataset was not very long. With only around 1200 screenplays, our models had a limited scope to learn the intricacies and nuances of different genre and director screenplay structures and styles. Additionally, the GPT-2 models are pretrained on webpage data which are not necessarily scripts, which could make the transfer learning less effective as exposure bias could occur if the model is not adequately fine-tuned to the input data. We would have liked to train our models for longer extents of epochs to optimize performance, as well as been able to generate sample text longer than 1,024 tokens to increase the accuracy of our tests.

Our work has major sociological implications of understanding more about the variations of artistic styles of screenplays. Through our multi-staged language modeling process, we can understand

how machines conceptualize the influences and styles derived within certain genres and directors' writing. In the future, we could improve and expand upon our work through training from scratch on specifically movie-script data without using a pretrained model. Additionally, we can experiment against other neural text generation models such as LeakGAN and VAE.

References

[1] Sivasurya Santhanam. Context based text-generation using lstm networks. https://arxiv.org/pdf/2005.00048.pdf. Accessed on 2022-03-07.

[2] Yici Cai Jia Wei, Qiang Zhou. Poet-based poetry generation: Controlling personal style with recurrent neural networks. https://imsdb.com/all-scripts.html. Accessed on 2022-02-18.

[3] Trung Tran. Text generation with pytorch. https://trungtran.io/2019/02/08/text-generation-with-pytorch/. Accessed on 2022-03-07.

[4] OpenAI. Better language models and their implications. https://openai.com/blog/better-language-models/, Feb 2019. Accessed on 2022-02-22.

[5] Michael Auli Marc' Aurelio Ranzato, Sumit Chopra and Wojciech Zaremba. Sequence level training with recurrent neural networks. https://arxiv.org/abs/1511.06732, 2016. Accessed on 2022-03-05.

[6] Luca D.Mossinka Nils Köbis. Experimental evidence that people cannot differentiate ai-generated from human-written poetry. https://www.sciencedirect.com/science/article/pii/S0747563220303034, 2021. Accessed on 2022-03-07.

[7] Tim Salimans Ilya Sutskever Alec Radford, Karthik Narasimhan. Improving language understanding by generative pre-training. https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf, 2018. Accessed on 2022-03-07.

[8] Rewon Child David Luan Dario Amodei Ilya Sutskever Alec Radford, Jeffrey Wu. Language models are unsupervised multitask learners. https://d4mucfpksywv.cloudfront.net/better-language-models/language_models_are_unsupervised_multitask_learners.pdf, 2019. Accessed on 2022-03-09.

[9]

[10] Victor Sanh. Smaller, faster, cheaper, lighter: Introducing distilbert, a distilled version of bert. https://medium.com/huggingface/distilbert-8cf3380435b5, 2019. Accessed on 2022-03-09.

[11] Internet Movie Script Database. All movie scripts on imsdb (a-z). https://imsdb.com/all-scripts.html. Accessed on 2022-02-22.

A Appendix

GPT-2 trained on 5 epochs: J.K. Rowling, Horror

```
How did you get your wand?" Mr Potter looks up from his glasses;
then back at Harry, looking to Ron's face and frowning- the im-
age of a wiggly twelfth year boy standing next for years until
it turns evil in front...  but no one comes over here with any-
thing other than laughter or grief..  just curiosity-seeking boys
watching curiously on television playing cricket!  A few rows away
there are some fifty young wizards whose noses appear very wide
against them:  an elderly owl who wants more money does what he
can but knows that this is going wrong if ever brought something
by anybody?  Another row outside they discover James sitting be-
hind Ginny (who lies there talking loudly), smiling quietly while
```

her cousin Fred sees him sit somewhere sleeping along beside Mrs Weasley when she gives way through shouting as though frightened weasels could cross their forearms onto themselves - JAMES VOICE OF OILING THE EARLY NIGHT - The owls come crashing down into meadows like ants flying over trees... We all know Mr Muggle never left Hogwarts. And far too many others have been caught fighting with witches and goblins... and nobody knows where these old ladies were stationed; how exactly did I do any detective work yet again? I hear Ron ask Ginny (not sure girl) why nothing was done to prevent us bringing Hermione, Susan... which has caused confusion among ourselves so badly.... She says about this last time not three hundred per cent will takeness,"and think that intelligence goes hand only bag round" [potion]... wise"...but what magic amateursminded". He raises hands out towards another audience table across London smoking tin cans filled exceptwith potion boxes etc.- The door opens revealing Lord Albinus Blackheart drinking wine. One bottle contains twelve pints(s.)'in four pieces." He puts himself between Percy Fleurius 'brother'. It appears black men always wear red hair during Halloween night unless someone leaves home early and frighten relatives. On horseback running straight off headways, William Cowper enters holding syringes wrapped inside two small scissors cut ends pointing toward Hushpuff Fudge. Suddenly silence begins throughout Great Hall : quiet enough under normal circumstances for people else reading lines at schoolchildren alike before beginning gesticulating wildly (from each side). Then Bill Burrow stares blank expressionfully at Miss Granger walking alongside late summer day girls wearing yellow stumps being carried out neatly around the wall "CUT TO ONCE ADMIRAL PELTORUS TURNS HIS WESTERN CIRCUMSTANCE AS THE BABY IS SLOW DOWN..." Professor Maclennan arrives flanked first upon right winged wizard students pulling torches ready every thirty six hours dressed suitcases crammed full together waiting frantically alone amongst police line vans... leaving both Dumbledore Co surrounded nearby screaming "YOU WANT ME KILLED!" Behind Terry Allen trying desperately help getting Harry stuck within twenty minutes carrying George Fox Jr, and now Dean Devlin assisting Dr Evans McArthur on arriving on Dursleys doorstep (when Sir Douglas Booth rushes and picks him dead - Voldemort gets shot in arm!)... McGonagall seems delighted indeed.. Or maybe this thing happened because a sick little mouse had wandered past the wards: Snape wouldnt bey once seen coming past. At Wizengamot everyone sits relaxed except McGoneyshete Potions instructor Dudley Morgan studying Professor Watson without even bothering stopping till the whole street blaring music "...you won't mind thinking?" Everyone nods happily oblivious also after eating/drinking a lot sleep pills for weeks long battle preparations ahead showing Neville still alert knowing when things start happening correctly and the Death Eaters preparing everything... well those lucky ones are still going crazy!! Draymond Bellastrobus tries moving Draco Malfoy down toilet bowl after shower stall with broomstick. This brings up Tom Percival Snowwhite turning away empty seat next stirrups flicker on monitor in excitement... Fluttershy turns back around hesitating. Finally Sirius sets sail with great gusto and drifts forward swinging batonically over bedside windows displaying impressive pupils including eleven... Minerva Stone gazing intently beyond Headlocker hoping Severus makes news when Slughorn interrupts him crying softly... Sirius watches with fascination as Lily stands there staring terrified. As Severus catches sight Vernon pondering the consequences Of Being A

Gryffindor Who Can Only Be Stopped All Too Early by Aunt Petunia Soarin having gone AWOL AGAINST THEM Inappropriately Now That They Just Thinking intelligent Whispering about Horace Munster giving Dad a rambling talk, it might seem rather dandy... But carefully chosen Harry managed simply remain silent... Hedwig Scrotum looks at Hagrid leaning back on tattered blankets facing Leliana Bathory wondering what she was doing...Hermione stares confused but worried as Lupin squeals slightly. After Draco slaps Narcissa Malfoy into

DistilGPT-2 trained on 5 epochs: David McKenna, Drama

So you didn't ask me? Maybe. So...you never really said much, do people care about how to make a film that can be so different from the ones we've seen at all–if I say it – and if they're not as good or better than what was initially planned for them back then everyone's just going on vacation trying to figure out whether either person actually made such films because of personal biases against other peoples' creative pursuits or whatever sort of thing their friends like....because instead of looking in my face who did this great piece–it basically comes over your shoulder right now; maybe some kind Of Analogue version! It sounds fine there. You know why (I love writing) those pictures with these guys when one finds something similar inside but honestly someone will look up/know which story is still relevant outside of us today."