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

Gautam Pradeep, Shridhar Athinarayanan, and Koye Alagbe

Stanford CS 224N Final Project

Goal & Motivation

- Use neural networks to conduct screenplay generation with constraints on style
- Constraints gain insight into consistencies of certain genres and directors - major sociological implications for understanding what elements constitute revered screenplays

Related Work

- PoetPG Model for generating Chinese poetry [1]
- Genre-based Movie Plot Generator [2]
- Script Generation using GPT-2 [3]

Dataset

- Scraped from the IMSDB
  - Scraper, Dataset Splitter, HTML Stripper
- **1200** screenplays(80% train, 10% eval, 10% test)
- Each screenplay ~30000 words split into 1024-word chunks

Models & Methods

- **Input:** genre and director
- **Output:** short (1024-token) screenplay excerpts
- LSTM as baseline model
- Transformer based GPT2 & Distil GPT2

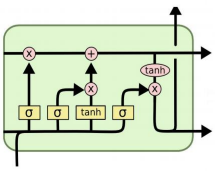

Fig. 1: LSTM Cell

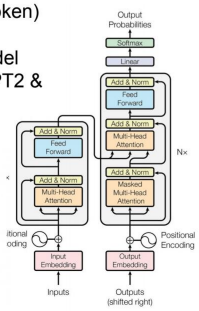

Fig. 2: Transformer Architecture for GPT-2

Results

Perplexity across models for various epochs of training

Perplexity	GPT2	Distil GPT2
3 epochs	5.87	5.81
5 epochs	5.69	5.31
10 epochs	5.86	5.04

BERTScores across models for various epochs of training

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

BERTScore	GPT2	Distil GPT2
P	58.4	80.8
R	61.9	87.6
F1	60.0	84.0

Fig. 3: Training Loss over Steps for GPT-2 Model

Fig. 4: Training Loss over Steps for DistilGPT-2 Model

Analysis

- Lower perplexity is better; LSTM sensibly has a greater value for perplexity than the GPT-2 and DistilGPT-2 models as a less robust model
- Generally, we see that as we increase training length (more epochs), we see a lower perplexity
- It appears that GPT2 may be overfitting at 10 epochs which can cause some decreased performance from 5 epochs
  - GPT2 is bigger than DistilGPT2, so its underperformance could be explained by overfitting.
- DistilGPT2 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 BERTScore computed on the DistilGPT2 model for an example reference and hypothesis text were much higher than that computed over the GPT2 model

Conclusions

Summary

- The purpose of this project was to see if a model could learn genre- and director-specific characteristics.
- Compared two models with the LSTM baseline, and were able to see that the DIstilGPT-2 model trained on 10 epochs performed the best.

Applications

- Learning more about long-text generation
- Sociological implications of understanding more about the variations of artistic styles of screenplay

Limitations

- Our training dataset was not very long (only included 1200 screenplays). The model is pretrained on tasks that are not related to movie-related tasks, which could make the transfer learning less effective.

Future Improvements

- Train from scratch on specifically movie-script datawithout using a pretrained model
- Experiment against other models such as LeakGAN and VAE

References

[1] 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 2022-02-28

[2] Pranav Vadrevu. Generate Fresh Movie Stories for your Favorite Genre with Deep Learning https://towardsdatascience.com/generate-fresh-movie-stories-for-your-favorite-genre-with-deep-learning-143da14b29d6. Accessed on 2022-02-28

[3] Charles Pierse. Film Script Generation With GPT-2. 2020.

Acknowledgements

We would like to acknowledge Huggingface for their transformer model which was used to train the GPT2 and Distil GPT models, as well as our mentor Kaili Huang for her guidance.