Project Update: Science Fiction Text Generation with GPT-2

Yu Xia
University of Michigan
xiayuu@umich.edu

1 Introduction

The field of natural language processing (NLP) has made significant strides in recent years, with the development of advanced language models such as GPT-2. These models have the ability to generate coherent and fluent text, opening up new possibilities for creative writing and content creation. One area of particular interest is the generation of science fiction stories. Science fiction is a genre that relies heavily on imagination and creativity, making it a challenging task for automated systems.

In this project, we aim to fine-tune the GPT-2 model on a large corpus of science fiction text to produce imaginative and coherent stories. Our objective is to create science fiction stories that are indistinguishable from those written by human authors. Solving this problem would have important implications for the NLP community and practical applications for creative writing, content creation, and technology industries.

The results of this project will be of interest to NLP researchers, creative writers, and technology companies. By demonstrating the ability of advanced language models to generate high-quality science fiction stories, we hope to inspire further research and development in this area.

Our proposed method for solving this problem involves fine-tuning the GPT-2 model on a large corpus of science fiction text. We will adopt some evaluation metrics to assess the quality and coherence of the generated stories.

2 Data

For this project, we will use the science fiction stories text corpus available on Kaggle . This dataset was created by Jannes Klaas and contains a collection of over 100 science fiction stories from the 1950s to the 1960s. The stories were obtained from Project Gutenberg and have been preprocessed to remove headers and footers.

The dataset is provided as a single text file. We will preprocess the data by splitting the text into sentences and form 10 sentences as a paragraph. We will then split the data into train and test sets for fine-tuning and evaluating our model.

Here are some rough statistics about the dataset:

- Number of words: 1,072,830
- Number of sentences: 2,212,421
- Number of paragraphs: 2,212,42

Example story from the dataset:
Title: The Variable Man Author: Philip K. Dick

He fixed things—clocks, refrigerators, vidsenders and destinies. But he had no business in the future, where the calculators could not handle him. He was Earth's only hope—and its sure failure!

Security Commissioner Reinhart rapidly climbed the front steps and entered the Council building. Council guards stepped quickly aside and he entered the familiar place of great whirring machines. His thin face rapt, eyes alight with emotion, Reinhart gazed intently up at the central SRB computer, studying its reading.

3 Related Work

There has been significant work in the field of text generation using deep learning models, including various variations of recurrent neural networks (RNNs) and transformers. OpenAI's GPT-3 (Brown et al., 2020) is a large-scale transformer-based language model trained on a massive corpus of text data. Generative Adversarial Networks (GANs) (Nie et al.; Yang et al., 2020) have been used to generate realistic and diverse text samples. In this project, we will use OpenAI's open-sourced GPT-2 (Radford et al., 2019) as the base model for our text generation task.

Several studies have explored text generation in various domains. For example, (Keskar et al., 2019) introduced a method for controlling the style and content of generated text by conditioning the model on control codes. (Dathathri et al., 2019) proposed a plug-and-play method for fine-tuning GPT-2 on specific tasks without the need for task-specific training data.

Our approach differs from previous work in that we will fine-tune GPT-2 specifically on a large corpus of science fiction text to generate imaginative and coherent stories. By focusing on a specific genre and providing the model with a large amount of relevant training data, we hope to improve the quality and coherence of the generated stories. We believe that our approach has the potential to produce science fiction stories that are indistinguishable from those written by human authors.

4 Methodology

In this project, we will fine-tune the GPT-2 model on a large corpus of science fiction text to generate imaginative and coherent stories. Our methodology involves the following steps:

Data preprocessing: We will use the science fiction stories text corpus available on Kaggle as our training data. This dataset contains a collection of over 100 science fiction stories from the 1950s to the 1960s. We will preprocess the data by splitting the text into sentences and then form paragraphs. We will then split the data into train and test sets for fine-tuning and evaluating our model.

Model fine-tuning: We will use OpenAI's open-sourced GPT-2 model as the base model for our text generation task. We will fine-tune the model on our training data using the transformers library.

5 Evaluation and Results

In this project, we will evaluate the science fiction text generated by finetuned GPT-2 model in terms of Perplexity against base GPT-2 model without any finetuning.

Perplexity: This measures the likelihood of the generated text given the input. A low perplexity score indicates that the generated text is similar to the validation data and is therefore of high quality.

The result for base GPT-2 model is shown as follows:

- Cross-Entropy Loss: 93.63
- Validation Perplexity: INF

6 Work Plan

1. Data preparation: Clean and preprocess the sci-fi text data. **[Done]**
2. Fine-tuning GPT-2: Fine-tune the GPT-2 model on the sci-fi text data. **[Doing]**
3. Model evaluation: Evaluate the model on the metrics of coherence and novelty. **[TBD]**
4. Results analysis: Analyze the results of the evaluation and discuss the performance of the model in generating new and coherent sci-fi text. **[TBD]**

References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2019. Plug and play language models: A simple approach to controlled text generation. *arXiv preprint arXiv:1912.02164*.

Nitish Shirish Keskar, Bryan McCann, Lav R Varshney, Caiming Xiong, and Richard Socher. 2019. Ctrl: A conditional transformer language model for controllable generation. *arXiv preprint arXiv:1909.05858*.

Weili Nie, Nina Narodytska, and Ankit Patel. Relgan: Relational generative adversarial networks for text generation. In *International conference on learning representations*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Yang Yang, Xiaodong Dan, Xuesong Qiu, and Zhipeng Gao. 2020. Fggan: Feature-guiding generative adversarial networks for text generation. *IEEE Access*, 8:105217–105225.