IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 12, Issue 6, June 2024

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.379

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 6, June 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1206084 |

Text Generation using Transformer

Kanishka Patil, Prof. Sonali Ajankar

Department of Master of Computer Applications, VJTI, Mumbai, India

ABSTRACT: We present a study with fine-tuning Transformer models for story generation. On their part, Transformer models are extremely powerful at capturing context through their inherent self-attention mechanism and have lately exhibited huge potentials in NLP tasks. However, the model's training on such diverse corpora often results in the output lacking two important elements: narrative coherence and creativity—important spices in compelling storytelling. In this approach of fine-tuning, the high-quality dataset shall be used to fine-tune a pre-trained transformer model. For instance, from this methodology of fine-tuning, techniques that are dynamic have been applied to ensure there is consistency in plot, the relevance of themes, and their context.

I. INTRODUCTION

One of the most significant innovations in the field of NLP is the capability of AI for creating coherent and interesting stories. The advancement is based on Transformer models as they have changed the approaches to understand and generate human language. The transformers were introduced by Vaswani et al in 2017 and are really beneficial for a number of NLP tasks as it employs self-attention for gathering context as well as interconnection in textual content. However, when applied directly to creative writing these approaches often result in a creation of a text that is not as tightly connected in its meanings as a good narrative and that does not possess the narrative charm of a good story.

The following are the limitations that were observed when using the Transformer models when applied ‘as is’ for story writing, which is the premise of this study. While they are able to write material that is contextually relevant and comprehensively professional in its writing, they can sometimes go weeks or months without creating characters or lacking a plot. These shortcomings are manifested more vividly in such work aspects as narration where it is imperative to create a particular emotional climate and a story. That speaks for itself and points to the fact that the further fine-tuning of pre-trained models is required in terms of the requirements for the generation of narratives to give the reader not only sensible but also engaging material

Some major advancements have been made in the different types of text generation, ranging from rule-based systems to the recent Recurrent Neural Network (RNN) and Long Short Term Memory (LSTM). However, these early models laid down the basic framework and were often hampered by their inability to navigate long-range dependencies well. Thus, using self-attention to handle context over long sequences, Transformer models reshaped this field and dispelled most of the challenges that their predecessors had. However, even today, the formation of coherent and engaging plots is still a challenging endeavor because storytelling requires the consideration of such factors as the character’s development and the plot’s continuous and meaningful themes.

The objective this paper seeks to narrow the space between the abilities present in pre-trained Transformer models and what is required for top-notch story generation. We plan to do this by fine-tuning a Transformer on a dataset that has been specifically curated for storytelling; our aim being making it more capable of producing coherent and engaging narratives. The approach we take will involve teaching the model how better understand and generate structural elements of stories such as plot arcs or character dynamics which are essential to creating interesting tales.

To achieve this objective, we employ a systematic approach that involves the following steps: First, we choose a big famous novel that fits the chosen style of narration as close as we can. This novel is pre-processed to construct the training data set that is then utilized to fine-tune the pre-trained Transformer model. The fine-tuning process involves tweaking the parameters of the model to match the style and structure of the novel, and to produce text that conforms to the stylistic and thematic characteristics of the original work. It also helps to transform the model so that it can create logical stories while also infusing the generated text with the tone of the source text.

The implications of this research can be observed in multiple ways for creative industries, education, and literature. Here, we examine how AI can mimic and enrich new forms of storytelling by fine-tuning models on specific works of literature. This is a good way to open up new opportunities for creating quality content that fits the specific style of the

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 6, June 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1206084 |

website. The presented approach can be useful for adaptive narrative environments, tools for creating style-specific content, as well as learning materials for reading comprehension and writing practice.

II. RELATED WORK

The generative pre-trained transformer language model, which was one of the first in this field, was introduced by Radford et al. in 2018. Trained on a large corpus of internet data, GPT showed impressive capabilities in producing human-like text across a variety of fields and also able to produce a similar style. However, while it was excellent at producing general text, it is not able to capture the distinctive stylistic nuances and narrative structures present in particular literary works. Acknowledging this limitation, scholars have investigated the optimization of transformer models using specific datasets to modify their language production capacities in specific fields.

In 2019, Radford et al. explored way of optimizing GPT-2 for use on particular domains like news stories and webpages, The results showcased the potential of fine-tuning a model has to enhance the model's coherence in the targeted domain

Building upon this idea, several studies have focused on fine-tuning transformer models on literary datasets to capture distinct writing styles and narrative conventions. Guan et al. (2020) explored fine-tuning GPT-2 on a corpus of science fiction stories, aiming to generate coherent and stylistically consistent narratives within the genre. Similarly, Khalifa et al. (2021) investigated the use of transformer models for generating poetry by fine-tuning GPT-2 on a diverse collection of poetic works, enabling the model to capture rhythmic and stylistic elements of poetry.

III. PROPOSED METHODOLOGY AND DISCUSSION

The proposed methodology for fine-tuning a transformer model to capture the storytelling style of a single novel involves the following key steps
Data Preparation-Its dataset is just the text from one massive novel that's used for training: In order for the model to understand the text, it first must be preprocessed so that it satisfies the model requirement as follow:

Clean The Text: The novel is cleaned to remove all of the non-story cluttered headers, footnotes, and other information any real book should not have.

Tokenization: The text is tokenized using the tokenizer of a pre-trained transformer model (like GPT-2 tokenizer) to obtain a token-by-token sequence Tokenizer: A tokenizer is a text processing facility that splits the text into sub word units and keeps the meaning and punctuation which helps in vocabulary management and model efficiency improvements.

Chunking: To deal with a limited context size (usually 1024 tokens for GPT-2), the text is chunked into chunks that fall within this range.

Encoding: Each chunk is encoded into input tensors that the model can process.

Model Selection and Pretraining

Model: Select a pretrained transformer model i.e. GPT-2.An improvement on the GPT language model is GPT-2 (Generative Pre-trained Transformer 2), released by OpenAI in February 2019, is a massive step forward in natural language processing (NLP). GPT-2 expands on the foundation set by GPT, with a three-fold larger capacity and application in analytically diversified domains; at the time, GPT-2 established a benchmark for text production and language comprehension.

The architecture of GPT-2 is built on the basis of the transformer model with the use of a decoder-only model. This design makes use of self-attention mechanism in order to allow the model to process and generate text adequately. In this way, using self-attention, GPT-2 can decide in which degree each word in the sequence is important for the model and thus concentrate only on the important information while generating the text.

This capability allows GPT-2 generate text that is coherent, in the context, and natural as opposed to the previous models used. This opportunity to interact with a broad class of web data further increases the ability of the model to

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 6, June 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1206084 |

imitate various language patterns and contexts making the model useful in a multitude of NLP applications including content creation, text summarization, and dialogues generation.

GPT-2 was prepared with a data set called WebText consisting of approximately 8 million documents, and the total size of the training data is over 40 gigabytes of text. There a large amount of text that is extracted from the web and this means that GPT 2 has training data on different languages and context. The training procedure was quite different from the supervised learning as the GPT-2 system made an attempt to guess the following word in the sentence using the input text data without the labels. This form of training is what actually enabled GPT-2 to learn a lot about various language tasks.

Although, GPT-2 is dependent on several critical aspects of the transformer architecture. Thus, through the multiple-headed self-attention mechanism, GPT-2 is capable of attending to how important certain words of a sequence are and pay utmost attention to critical input when generating text. The third concept is positional encoding that is very important as it informs the model of the position of the words

The significance of GPT-2 is not limited by its functions and potential in developing the language model. Its capability of generating text that is as good as that written by human beings brought it into the limelight due to the possibilities it holds. Due to its highly desirable text generation ability, GPT-2 was applied to any undertakings where original content writing is required such as in writing articles, creating poetic and fictional text, and for use in simple-to-advanced dialog systems of chatbots. Besides those uses, GPT-2 has been helpful in programming the field for code snippets

Initialization: Load the pre-trained weights of the model. The model is then set to use the tokenized dataset of the book.

Training: The novel dataset is used to train the model with the following specifics:
- Learning Rate - learning rate is the small learning rate (1e-5) to tune the model won't wipe out the pre-trained knowledge.
- Batch Size: 2 (a small batch size because the number of parameters of a model are huge and the GFX memory are limited).
- Epochs: Many epochs must be used (e.g., 100) to give the model time to learn the storytelling attitude.
- Optimization**:** The AdamW optimizer is employed to minimize the loss function efficiently.

Text generation -
After fine-tuning, the model generates text based on prompts related to the novel. The process involves:
- Prompting: The model is given a prompt that aligns with the novel's themes or starting sentences. This prompt sets the context for the generated text.
- Decoding: The model generates text using strategies such as top-k sampling or nucleus sampling (top-p sampling) to produce coherent and contextually relevant text.

Challenges and Limitations:
However, in the course of the project, a few hurdles had to be faced:

Overfitting: The data available for training is small (a single novel); the model may be encouraged to generate text that is essentially a copy of the lines it has seen in training, rather than generalizing to new contexts.

Context length: The model had a context length of 1024 tokens, which limited the length of the output to the smallest of the blockchain action, the intention (1014-1021 tokens depending on the model).

IV. RESULTS

The findings from fine-tuning the Transformer model on a dataset derived from a single large novel. The results are evaluated based on their conformity to narrative coherence g, style, and overall quality. The evaluation consists of a quantitative metric to provide a comprehensive view of the model's performance.

Quantitative Evaluation:
To access the performance of the model after fine tuning, we have many options in evaluation metrics which are commonly used. One of which is perplexity. The fine-tuned model achieved a perplexity of 140. This suggests that the

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 6, June 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1206084 |

fine-tuned model generates text that is on an average decent in aligning with the patterns observed in the novel, considering the limited dataset and the computation available at the time.

Metric	Score
perplexity	140

table 1: .This table shows the perplexity metric.

V. CONCLUSION

In this study we explored the fine tuning of the transformer specifically to generate text that mimics the style and narrative flow of a novel or text it is fine-tuned on. By fine tuning the model on a dataset we wanted to enhance the text generation style and consistency in generating coherent text which is similar to that of the dataset.

The Finding indicates that fine tuning significantly improves the ability of the model to generate text that is similar to the style and flow of the original work. The generated story exhibited similar words, nuances and thematic plot of the novel thus suggesting the effectiveness of fine tuning a model to adapt to generate a text which pertains to a specific style.

Quantitative metrics also support this observation. The average Perplexity score achieved across the various generated segments demonstrated a decent degree of similarity to the reference text, highlighted the model's capacity to predict subsequent tokens correctly to a good extent, reflecting an improvement in the coherence and fluency of the generated text.

However, the evaluation also revealed some limitations. Despite the improved performance, the model sometimes struggles with maintaining long-term coherence in the generated text, especially over extended length of passages. This suggests a potential area for further improvement such as integrating techniques for better context management over longer text spans.

The model's performance is dependent on the quality and specificity of the fine-tuning dataset. Given that the dataset was derived from a single novel, the quality and the flow of the generated text will have hallucinations and the coherence of the generated text will sometimes will not be in style of the original text. Future work could address this by selecting a dataset high quality and larger size.

fine-tuning Transformer models on specific literary work presents a promising approach for generating text that closely mimics the style and structure of the source material. While this method enhances the model's ability to produce text with style and narratives similar to source material, this area still in continued research and development to address challenges related to long-term coherence of the text.

REFERENCES

1. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.
2. Howard, J., & Ruder, S. (2018). Universal language model fine-tuning for text classification. arXiv preprint arXiv:1801.06146.
3. Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training.
4. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI blog, 1(8), 9.
5. Keskar, N. S., McCann, B., Varshney, L. R., Xiong, C., & Socher, R. (2019). Ctrl: A conditional transformer language model for controllable generation. arXiv preprint arXiv:1909.05858.
6. Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Askell, A. (2020). Language Models are Few-Shot Learners Advances in Neural Information Processing Systems 33.
7. Dong, L., Yang, N., Wang, W., Wei, F., Liu, X., Wang, Y., ... & Hon, H. W. (2019). Unified language model pre-training for natural language understanding and generation. Advances in neural information processing systems, 32.
8. Peng, P., & Wang, J. (2020). How to fine-tune deep neural networks in few-shot learning?. arXiv preprint arXiv:2012.00204.
9. Li, J., Tang, T., Zhao, W. X., Nie, J. Y., & Wen, J. R. (2022). Pretrained language models for text generation: A survey. arXiv preprint arXiv:2201.05273.
10. Chen, Y. C., Gan, Z., Cheng, Y., Liu, J., & Liu, J. (2019). Distilling knowledge learned in BERT for text generation. arXiv preprint arXiv:1911.03829.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi crossref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details