Proceedings of the 4th International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/48/20241154

Performance exploration of Generative Pre-trained Transformer-2 for lyrics generation

Yijia Hu

SWJTU-LEEDS Joint School, Southwest Jiaotong University, Chengdu, Sichuan, 611756, China

huyijia@my.swjtu.edu.cn

Abstract. In recent years, the field of Natural Language Processing (NLP) has undergone a revolution, with text generation playing a key role in this transformation. This shift is not limited to technological areas but has also seamlessly penetrated creative domains, with a prime example being the generation of song lyrics. To be truly effective, generative models, like Generative Pre-trained Transformer (GPT)-2, require fine-tuning as a crucial step. This paper, utilizing the robustness of the widely-referenced Kaggle dataset titled "Song Lyrics", carefully explores the impacts of modulating three key parameters: learning rate, batch size, and sequence length. The dataset presents a compelling narrative that highlights the learning rate as the most influential determinant, directly impacting the quality and coherence of the lyrics generated. While increasing the batch size and extending sequence lengths promise enhanced model performance, it is evident that there is a saturation point beyond which further benefits are limited. Through this exploration, the paper aims to demystify the complex world of model calibration and emphasize the importance of strategic parameter selection in pursuit of lyrical excellence.

Keywords: Natural Language Processing, Lyrics Generation, Large Language Model.

1. Introduction

Text generation has emerged as an imperative area of study in the broad landscape of Natural Language Processing (NLP). With the launch of Chat Generative Pre-trained Transformer (ChatGPT) by OpenAI, generative Artificial Intelligence (AI) applications have garnered significant attention from numerous enterprises. An increasing number of companies are beginning to roll out their generative models to boost their productivity. Text generation technology, as one of the core technologies of generative AI models, is also becoming increasingly significant for research. Text generation now serves as the backbone for a multitude of applications that span various sectors, from automated customer service solutions to more creative outlets like journalism and content creation [1].

Text generation algorithms can be used not only for generating general text forms but also for specialized sub-domain applications, one of which is lyric generation. Compared to general text, lyrics have higher requirements as they are a product of the combination of human speech and musicality. Lyrics not only need to maintain textual coherence but also need to adhere to specific stylistic norms, including rhythm and meter. In addition, lyrics must also be able to convey emotion and meaning.

In recent years, Transformer-based architectures have gained prominence for their ability to perform well in a variety of text-generation tasks. Specifically, the GPT series has achieved state-of-the-art

DOI: 10.54254/2755-2721/48/20241154

performance in multiple benchmarks [2]. Although these models have excellent capabilities in generating coherent and contextually relevant text, more research and exploration are still needed for specialized forms of application such as lyric generation.

This study aims to delve into this specific domain by investigating how various fine-tuning strategies can influence the performance of GPT-2 in lyric generation. The research of this paper focuses on manipulating three primary parameters: learning rate, batch size, and sequence length while keeping all other variables constant. This paper will study three parameters separately, keeping the other two constant in each study to isolate the impact of each parameter on the outcome. To measure the quality of the generated lyrics objectively, this paper employs Perplexity as the evaluation metric.

This paper seeks to answer the following three questions. Firstly, how does adjusting the learning rate affect the GPT-2 model's capability to produce high-quality lyrics? Secondly, how does the modification of the batch size influence the Perplexity of the lyrics generated? Thirdly, what effect does altering sequence lengths have on the overall quality of the generated lyrics?

Addressing these questions will not only deepen the understanding of the capabilities and limitations of GPT-2 in lyric generation but also pave the way for future research in this specialized area of text generation [3].

The insights gained from this research have significant implications for the broader field of text generation. Understanding the fine-tuning requirements for specialized tasks like lyric generation can greatly enhance the applicability and effectiveness of pre-trained models like GPT-2, paving the way for them to perform optimally in specialized subdomains.

2. Method

2.1. Dataset

As is well-known, datasets play a crucial role in text generation, and their significance becomes even more pronounced in the specific area of lyric generation. After searching through numerous dataset files online and eliminating ones that were either too large or too small, the author eventually chose a song lyric dataset from Kaggle called "Song Lyrics" [4].

This file consists of lyrics from 49 different artists, with a total of 183K sentences and 1404K words. It contains 31,689 unique words. The size of this dataset makes it ideal for training with a pre-trained model like GPT-2. The average sentence in this dataset has 7.66 words, which greatly reduces the likelihood of generating lyrics that are either too long or too short.

Before using the data for training, data cleaning is performed by the author. This paper converted all uppercase letters to lowercase and replaced non-alphanumeric characters with spaces. Doing this maintains the consistency of the dataset format while simplifying the model's input, thereby reducing the likelihood of overfitting.

2.2. GPT-2 Model

This paper employs the GPT-2, a state-of-the-art language model developed by OpenAI. The primary function of this model is to perform natural language generation tasks. At the same time, because the size of this pre-trained model matches the size of its training set, it is the most suitable language model for this job.

2.2.1. Principles

GPT-2 is built on the Transformer architecture and consists of multiple layers with a total of 1.5 billion parameters. Unlike traditional RNNs and LSTMs, the Transformer architecture allows for parallelization during training, which significantly accelerates the learning process. This architecture is particularly effective in capturing long-range dependencies in text, a feature that is crucial for tasks like text generation [5].

DOI: 10.54254/2755-2721/48/20241154

2.2.2. Strengths and Weaknesses

GPT-2 has several advantages. One of the most significant advantages of GPT-2 is its ability to perform various tasks without requiring task-specific training data. This makes it a highly versatile tool for a range of natural language processing applications [6]. In additions, GPT-2 can be easily fine-tuned on a specific dataset, which allows for improved performance on specialized tasks. This is particularly useful in the study, where the focus is on lyric generation [7].

Despite its strengths, GPT-2 also has many weaknesses. It requires substantial computational resources for training. This could be a limiting factor for researchers with restricted access to high-performance computing facilities [8]. Moreover, GPT-2, like many deep learning models, suffers from a lack of interpretability. This makes it challenging to understand the reasoning behind its predictions and could be a potential drawback in certain applications [5].

2.2.3. Fine-tuning Strategy

To investigate the impact of different fine-tuning strategies on lyric generation, this paper employed the method of controlled variables to separately study the effects of learning rate, batch size, and sequence length on the generated outcomes. This article will conduct training over 5 epochs and record the perplexity after each epoch. First, this paper examines the learning rate, keeping the batch size at 32 and the sequence length at 50. This paper set the learning rate at 3e-2, 3e-3, and 3e-5, respectively, to explore its impact on the resulting perplexity. Next, this paper investigates batch size, keeping the learning rate at 3e-3 and the sequence length at 50. This paper set the batch size at 32, 64, and 128, respectively, to study its influence on the generated perplexity. Lastly, this paper focuses on the sequence length, maintaining the learning rate at 3e-3 and the batch size at 32. This paper set the sequence length at 50, 100, and 150, respectively, to examine its effect on the resulting perplexity.

3. Result

This paper will demonstrate the impact of three parameters on the generated results, starting with the learning rate, as demonstrated in Figure 1.

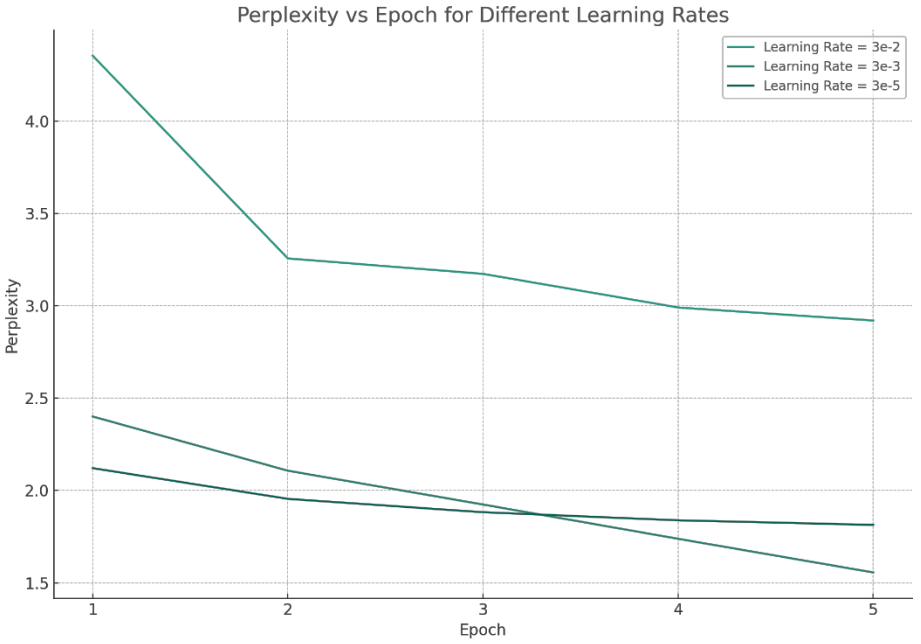

Figure 1. Perplexity at various epoch with different learning rates (Figure Credits: Original).

DOI: 10.54254/2755-2721/48/20241154

Next, Table 1 shows the generated lyrics.

Table 1. Representative generated lyrics at various learning rates.

	3e-2	3e-3	3e-5
we	we i you want want know	weve been trodding on the winepress much too long	we need a new way of life
here	here i want you be so just me you and a me	here we go my friend a homie lover friend	here we are now entertain us
why	why you what let know i cant	why do i hate her	why do you love me ask why do you love me

Results in Figure 2 are the perplexity at various batch sizes, and Table 2 displays the generated lyrics.

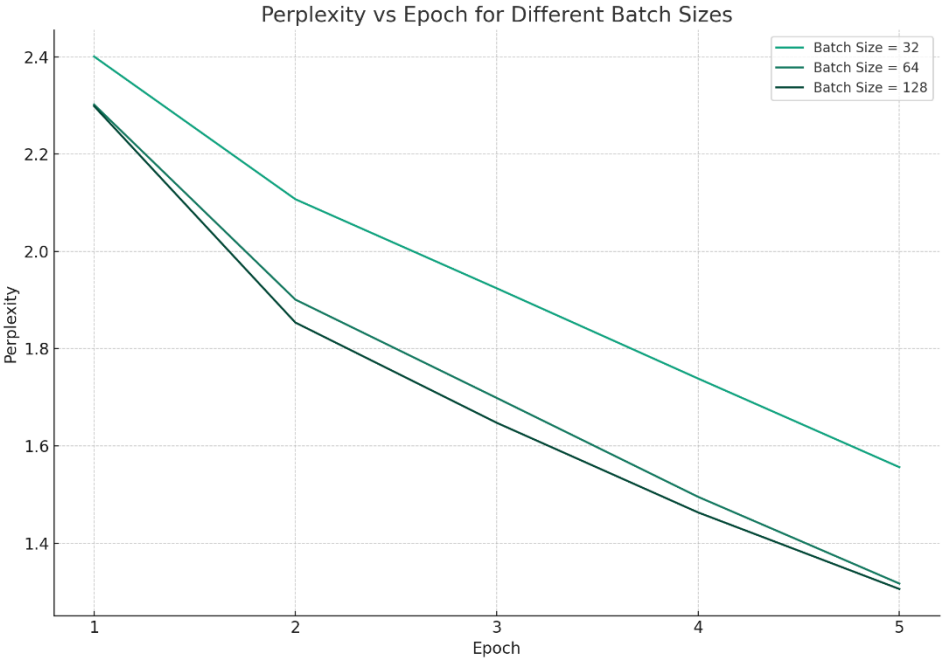

Figure 2. Perplexity at various epoch with different batch sizes (Figure Credits: Original).

Table 2. Representative generated lyrics at various batch sizes.

	32	64	128
we	weve been trodding on the winepress much too long	we gonna light up the night like shooting stars	we will stand tall
here	here we go my friend a homie lover friend	here we go my friend a homie lover friend	here comes the sun and i say
why	why do i hate her	why you wanna trip on me	why did you take him away

Figure 3 and Table 3 demonstrate the results and generated lyrics at different sequence lengths.

DOI: 10.54254/2755-2721/48/20241154

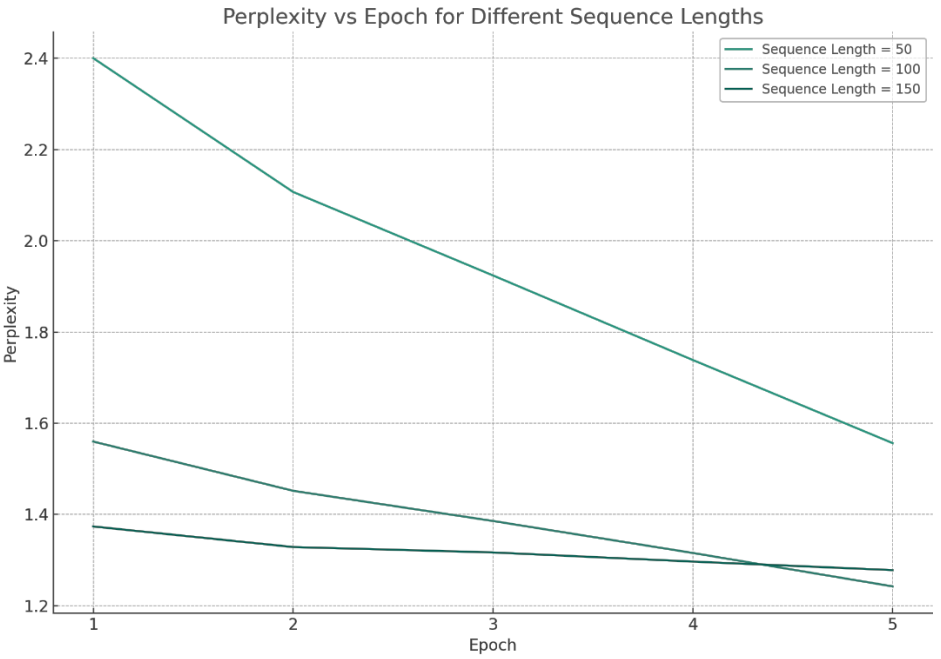

Figure 3. Perplexity at various epoch with different sequence lengths (Figure Credits: Original).

Table 3. Representative generated lyrics at various batch sizes.

	50	100	150
we	weve been trodding on the winepress much too long	we were sad of getting old	Weve been everywhere but we got the whole time
here	here we go my friend a homie lover friend	here i am baby	here we need a friend who
why	why do i hate her	why dont you scream and shout it	why theres nothing that i got to do

The author will first analyze and research based on the results of the learning rate. Based on the results from the table, the higher the learning rate, the higher the initial perplexity, and the faster the perplexity decreases from the first epoch to the second epoch. In this dataset, within the range of learning rates from 3e-2 to 3e-3, higher learning rates are more helpful in reducing the model's perplexity. However, when the learning rate reaches 3e-5, the model's generalization ability is worse than with lower learning rates.

Overall, at high learning rates, although the perplexity decreases quickly, the quality of the generated text is relatively low. At low learning rates, the initial perplexity is the lowest, but the rate of decrease in perplexity is smaller. The quality of the generated text is also relatively better. At medium learning rates, a good compromise is offered, showing relatively balanced performance in terms of both decreasing perplexity and generating text of good quality.

Next, the author will first analyze and research based on the results of batch size. Based on the results shown in the table, the larger the batch size, the stronger the model's generalization ability and the lower the perplexity. At the same time, the larger the batch size, the faster the perplexity decreases from the first epoch to the second epoch. Comparing the lines for batch size=64 and batch size=128 in the table, it can be seen that these two lines almost overlap. This indicates that once the batch size reaches a certain level, further increases offer little benefit in terms of enhancing the model's generalization ability. During the experiment, training with a batch size of 128 took approximately twice as long as with a batch size of 64, yet the results were quite similar. Therefore, setting an excessively high batch size for

DOI: 10.54254/2755-2721/48/20241154

this dataset would be a waste of computational resources. A moderate batch size allows the model to achieve good generalization without consuming excessive computational resources.

Lastly, the author will first analyze and research based on the results of sequence length. Based on the results from the table, the impact of sequence length on the generated results is very similar to that of the learning rate. Upon comparison, it can be observed that the larger the sequence length, the greater the decrease in each epoch. Based on this information, this paper found that when the sequence length is 150, the model's generalization ability is worse than when the sequence length is 100.

In the experiments, this paper found that, unlike the learning rate, adjusting the sequence length significantly increases the training time for the model. In this trial, training with a sequence length of 150 took 1.7 times longer than training with a sequence length of 100. Consuming more computational resources for worse generalization ability is not appropriate.

Now, the author will analyze the impact of three parameters on the generated results simultaneously. Based on the generated lyrics, the learning rate is the most critical factor for producing coherent and meaningful lyrics. When trained with a low learning rate, the lyrics generated by the model are incoherent and unclear. The influence of other parameter adjustments on the coherence of the generated results is not as significant as that of the learning rate. Overall, among the three parameters studied, the batch size has the least impact on the model's generalization ability. In the study of both the learning rate and sequence length, it was observed that moderate values yielded the best results

4. Discussion

In recent years, with the rapid advancement of generative models and their increasing applications, lyric generation as a specialized sub-domain of generation tasks has gradually become a research hotspot. Therefore, fine-tuning GPT-2 for lyric generation has become increasingly important for research. The findings of this study offer valuable insights into the impact of various fine-tuning strategies on the quality of generated lyrics.

One of the primary observations from this study is the significant impact of the learning rate on the generated outcomes. As highlighted, higher learning rates lead to a rapid decrease in perplexity but compromise the quality of the generated text. This aligns with the findings of Zhang et al., who noted that while higher learning rates can accelerate convergence, they can also lead to overfitting and reduced generalization in certain tasks [9]. On the other hand, lower learning rates, although starting with the lowest initial perplexity, have a slower rate of decrease in perplexity. This is consistent with the observations of Liu et al., who emphasized the importance of carefully selecting the learning rate to balance convergence speed and model generalization [10].

The study's findings on batch size are particularly intriguing. While larger batch sizes enhance the model's generalization ability and reduce perplexity, there seems to be a threshold beyond which further increases in batch size offer diminishing returns. This observation is in line with the research by Kim et al., which suggests that while larger batch sizes can provide more stable gradient estimates, they can also lead to over-smoothed optimization landscapes, making it harder for the model to escape local minima [11].

The impact of sequence length on the generated results mirrors that of the learning rate. However, it's worth noting that increasing the sequence length significantly prolongs the training time without necessarily improving the model's generalization ability. This resonates with the findings of Chen et al., who pointed out that while longer sequences can capture more contextual information, they also introduce more noise and can lead to computational inefficiencies [12].

In conclusion, when analyzing the combined effects of these three parameters, it is evident that the learning rate is the most crucial factor for generating coherent and meaningful lyrics. In contrast, the batch size has the least impact on the generated results. This highlights the disparities in how different parameters affect the outcomes and underscores the importance of parameter selection in fine-tuning. Prioritizing the fine-tuning of parameters with a significant impact on the results is paramount for generating high-quality lyrics. Although this study provides a comprehensive analysis of the effects of

DOI: 10.54254/2755-2721/48/20241154

various fine-tuning strategies, future research could delve deeper into the interactions between these parameters and explore other factors that might influence the quality of generated lyrics.

Looking forward, as the field of NLP continues to evolve, it will be essential to revisit and refine these fine-tuning strategies. The emergence of newer models and architectures may introduce additional parameters that could further influence the quality of lyric generation. Moreover, as computational resources become more accessible, experimenting with even larger batch sizes or longer sequence lengths might yield different insights. Collaboration and joint efforts between NLP researchers, lyricists, and machine learning practitioners can pave the way for more refined and meaningful lyric generation techniques, further pushing the boundaries of possibilities in this intersection of musical art and NLP technology.

5. Conclusion

Embarking on the journey of text generation, especially when dealing with stalwarts like GPT-2, one is met with a delicate balance between the artistic nuances and scientific precision. The methodological approach of this paper, dissecting and analyzing the nuanced roles of learning rate, batch size, and sequence length, offers a panoramic view into the intricacies involved in customizing GPT-2 for niche endeavors like lyric crafting. The empirical evidence is compelling, catapulting the learning rate to a preeminent position, shaping the very essence and finesse of the lyrics brought to life. Although batch size and sequence length wield considerable influence, their utility has discernible boundaries, reinforcing the philosophy of moderation and equilibrium. The insights gleaned from this study underscore the quintessence of astute parameter choices, ensuring that generative models are primed to deliver peak performance. As the field of NLP continues to evolve, it is paramount to continually refine methodologies and strategies. With the impending arrival of next-gen models and the democratization of computational resources, the synergy between the soulful artistry of music and the analytical rigor of NLP signals an era filled with boundless innovation and seminal discoveries.

References

[1] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. Advances in neural information processing systems, 30, 1-11.

[2] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI blog, 1(8), 9.

[3] Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., et al. (2020). Transformers: State-of-the-art natural language processing. In Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations, 38-45.

[4] Song Lyrics, URL: https://www.kaggle.com/datasets/paultimothymooney/poetry?select=nursery_rhymes.txt. Last Accessed: 2023/09/12

[5] Wallace, E., Feng, S., Kandpal, N., Gardner, M., & Singh, S. (2019). Universal adversarial triggers for attacking and analyzing NLP. arXiv preprint arXiv:1908.07125.

[6] Budzianowski, P., & Vulić, I. (2019). Hello, it's GPT-2--how can I help you? towards the use of pretrained language models for task-oriented dialogue systems. arXiv preprint arXiv:1907.05774.

[7] Xie, W., Que, M., Yang, R., Liu, C., & Yu, D. (2019). BLCU_NLP at SemEval-2019 Task 8: A Contextual Knowledge-enhanced GPT Model for Fact Checking. In Proceedings of the 13th International Workshop on Semantic Evaluation, 1132-1137.

[8] Wu, Q., Zhang, Y., Li, Y., & Yu, Z. (2019). Alternating recurrent dialog model with large-scale pre-trained language models. arXiv preprint arXiv:1910.03756.

[9] Mishra, P., & Sarawadekar, K. (2019). Polynomial learning rate policy with warm restart for deep neural network. In IEEE Region 10 Conference, 2087-2092.

[10] Yang, J., & Wang, F. (2020). Auto-ensemble: An adaptive learning rate scheduling based deep learning model ensembling. IEEE Access, 8, 217499-217509.

DOI: 10.54254/2755-2721/48/20241154

[11] Kochura, Y., Gordienko, Y., Taran, V., Gordienko, N., Rokovyi, A., Alienin, O., & Stirenko, S. (2020). Batch size influence on performance of graphic and tensor processing units during training and inference phases. In Advances in Computer Science for Engineering and Education II, 658-668.

[12] Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. Advances in neural information processing systems, 27, 1-9.