Sampling Methods on Language Models

Arda Orcun

Yıldız Teknik Üniversitesi

23023608

Abstract

This study investigates the effects of various sampling techniques, with a particular focus on top-k sampling, on the performance and sustainability of language models. By fine-tuning models on both sampled and non-sampled datasets, we compare metrics such as training time, loss, GPU power consumption, and overall efficiency. Our findings reveal that top-k sampling significantly enhances model performance and reduces computational costs, demonstrating its potential for sustainable AI development. The results suggest that employing effective sampling methods can lead to more efficient and environmentally friendly training processes for language models.

Key Words: language models, sampling techniques, top-k sampling, sustainability, performance, GPU power consumption

Introduction

The rapid advancement of language models has significantly improved natural language processing (NLP) tasks, enabling applications such as text generation, translation[1], and sentiment analysis. However, training and fine-tuning these models often require substantial computational resources, raising concerns about their environmental impact and sustainability. Efficient training methods that reduce resource consumption without compromising performance are therefore essential.

Sampling techniques play a critical role in optimizing the training process of language models. By selecting a subset of data that accurately represents the entire dataset, sampling can reduce the computational load and enhance model efficiency. Among various sampling methods, top-k sampling has emerged as a particularly effective approach. Top-k sampling involves selecting the k most probable tokens at each step of the training process, ensuring that only the most relevant data points are used for model updates. All code is available at https://github.com/Stealeristaken/Sampling2Project

This paper focuses on the impact of top-k sampling on the performance and sustainability of language models. We utilize the "Open Platypus"[2] dataset, a diverse collection of text datasets, to evaluate the effects of top-k sampling compared to non-sampled training. Our study involves fine-tuning two separate models with identical parameters for a fixed number of epochs, one using the sampled dataset and the other using the non-sampled dataset. The models were trained using 2xT100 GPUs, and we monitored key metrics such as training time, loss, GPU hour, and GPU power consumption to assess the efficiency and effectiveness of the sampling approach.

Our findings reveal that top-k sampling not only improves model performance by reducing mean loss but also significantly decreases training time and power consumption. These results underscore the potential of top-k sampling as a sustainable and efficient training method for language models. By minimizing the computational resources required, top-k sampling contributes to lower energy usage and reduced carbon footprints, aligning with the goals of sustainable AI development.

Methodology

Sampling Techniques Overview

Sampling techniques are essential for optimizing the performance and efficiency of language models by selecting a representative subset of data for training. In this study, we focus on several common sampling techniques, providing a foundation for understanding the specific impact of top-k sampling.

Greedy Search

Greedy search is a straightforward decoding method where the model selects the highest probability token at each step. While it ensures that each token chosen is the most

likely given the previous tokens, greedy search can lead to suboptimal sequences because it does not account for future token probabilities.

Beam Search

Beam search improves upon greedy search by maintaining multiple candidate sequences (beams) at each step. By keeping track of the top sequences, beam search balances between exploring different possible token sequences and exploiting the most likely tokens. This method increases the chances of finding a more optimal sequence but can be computationally expensive.

Temperature Sampling

Temperature sampling introduces a randomness factor into the token selection process by adjusting the probabilities of the tokens. A lower temperature value (e.g., < 1) makes the model more confident, leading to more deterministic outputs, while a higher temperature value (> 1) increases randomness, promoting diversity in the generated sequences.

Top-p (Nucleus) Sampling

Top-p sampling, also known as nucleus sampling, selects tokens from the smallest possible set of tokens whose cumulative probability exceeds a threshold p. This method dynamically adjusts the number of tokens considered at each step, balancing between diversity and quality by including only the most probable tokens.

Top-k Sampling

Top-k sampling is a method where the model selects the next token from the top k most probable tokens at each step. This approach ensures that only the highest probability tokens are considered, providing a balance between deterministic and random token selection. The value of k determines the breadth of the token selection process; a smaller k makes the output more deterministic, while a larger k introduces more diversity.

Implementation of Top-k Sampling

In this study, we implemented top-k sampling with k set to 500. This choice aims to provide a broad enough selection to capture diverse linguistic patterns while maintaining high-quality outputs. The process involved the following steps:

Dataset Preparation:

We utilized the "Open Platypus"[2] dataset, a comprehensive collection of text datasets. The dataset was accessed from HuggingFace, ensuring the latest updates and ease of use. We tokenized our dataset with latest, state of the art, Llama – 3[3] Large Language Model which released by Meta.

Deduplication:

A deduplication rate of 0.95 was applied to filter out near-duplicate entries, enhancing dataset diversity and ensuring high-quality training data.

Top-k Sampling:

After deduplication, top-k sampling was employed with k set to 500. This method selected the 500 most probable tokens at each step, balancing diversity and quality in the dataset.

Training Procedure:

We fine-tuned two separate models using the "YTU-Cosmos"4 language model:

One with the sampled dataset.

One with the non-sampled dataset.

Both models were trained using 2xT100 GPUs, with identical parameters and for a fixed duration of 5 epochs. Key metrics such as training time, loss, GPU hour, and GPU power consumption were monitored throughout the training process.

By leveraging top-k sampling, our methodology aims to demonstrate how this technique can enhance model performance and sustainability, providing a more efficient and environmentally friendly approach to training language models.

Results

The implementation of top-k sampling and the subsequent training of models with sampled and non-sampled datasets yielded notable differences in various performance metrics. The key findings are summarized in Table 1.

Table 1

Results

Metric	Non – Sampled Dataset	Sampled Dataset
Training Time	1 hour 55 minutes	2 minutes
Mean Loss	6.72	0.8
Mean Gpu Usage per minute (W)	70	70
Total Power Consumption (W)	8050	140

Table 1 presents a comparative analysis of the training metrics for language models trained on non-sampled and sampled datasets using top-k sampling. The data highlights significant improvements in efficiency and performance for the sampled dataset. Specifically, the training time for the sampled dataset was reduced from 1 hour and 55 minutes to just 2 minutes. Additionally, the mean loss decreased from 6.72 to 0.8, indicating better model performance. Despite both models maintaining a mean GPU power usage of 70W, the total power consumption for the non-sampled dataset was 8050W, while the sampled dataset consumed only 140W. These findings underscore the effectiveness of top-k sampling in enhancing model training efficiency and sustainability.

Conclusion

This study highlights the dual benefits of employing top-k sampling in the training of language models. The findings show that top-k sampling not only improves model performance by reducing the mean loss but also significantly decreases training time and

power consumption. These improvements contribute to more efficient and sustainable AI practices.

The substantial reduction in GPU power consumption underscores the environmental benefits of top-k sampling. By minimizing the computational resources required, top-k sampling aligns with the goals of sustainable AI development, promoting lower energy usage and reduced carbon footprints.

Future research could further explore the application of top-k sampling in various language models and datasets, potentially extending these benefits across different domains and tasks. Additionally, comparing top-k sampling with other advanced sampling techniques could provide deeper insights into optimizing language model training for performance and sustainability.

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] Lee, A. N., Hunter, C. J., & Ruiz, N. (2023). Platypus: Quick, cheap, and powerful refinement of llms. arXiv preprint arXiv:2308.07317..

[3] Zhang, P., Shao, N., Liu, Z., Xiao, S., Qian, H., Ye, Q., & Dou, Z. (2024). Extending Llama-3's Context Ten-Fold Overnight. arXiv preprint arXiv:2404.19553.

[4] Dogan, E., Uzun, M. E., Uz, A., Seyrek, H. E., Zeer, A., Sevi, E., ... & Amasyali, M. F. (2024). T\" urk\c {c} e Dil Modellerinin Performans Kar\c {s}{\i} la\c {s} t {\i} rmas {\i} Performance Comparison of Turkish Language Models. arXiv preprint arXiv:2404.17010.