An Exploration of Sampling strategies for Diffusion-Large Language Models

Topic outline

Zhang, Haoqi (hz3223) - *Data Science AI Concentration*

Supervision

– Lu, Yucheng (lu.yucheng@nyu.edu)

Context

Aside from Model architecture, data processing, optimization methods, decoding strategies also play an important role on the quality of the output of a Large Language Model [1]. Sampling strategies are an important component and subset of decoding strategies. Investing sampling strategies help us to utilize the maximum performance of a model, without having to make changes to the model [1].

Diffusion-Large Language Models (dLLMs) are an emerging alternative to Autoregressive Large Language Models (ARMs) [2] [3]. Autoregressive models are constrained to decode and generate sequentially, whereas dLLMs generate by decoding the full sequence in parallel from a fully masked sequence, allowing the potential for bidirectional generating and sampling acceleration [4].

The sampling process for dLLM is very different from Autoregressive models, it is the process of denoising from a fully masked sequence [4] [5]. Sampling strategies play an important role in achieving the fast-sampling potential of dLLMs. An exploration of sampling strategies for dLLMs would be helpful in unlocking the potential of dLLMs under high level of parallel decoding [5].

Objectives

By the end of the capstone project, I hope to have done a somewhat thorough exploration of the role and performance of different sampling strategies for Diffusion-Large Language Models. I will conduct experiments on a variety of sampling strategies with some different diffusion Language Models, and benchmark the results, and after doing so, reach some naive conclusion on the performance of sampling strategies, document results and report findings.

Below is my work plan (there could be a lot of changes to this)

1. Conduct thorough literature Review on diffusion language models, and then select base models used for running experiments.
2. Conduct thorough literature review on first, sampling strategies for LLMs in geenral, then second, study with the focus of sampling strategies for dLLMs. After finishing these two parts, the select the sampling strategies used for conducting experiments.
3. Find appropriate datasets to run the experiments on, these datasets should reflect a wide range of topics, such as context, math, coding, etc. This is to help find the ideal sampling strategy for different specific scopes, and also to see which strategies work well in general for all tasks.
4. Conduct the experiments. Try to use the best engineering practices to get ideal results.(also to avoid engineering reasons influencing the selection and comparison of the strategies) Carry out very good engineering to make sure the experiments go smoothly, and get ideal results.
5. Benchmark the results of the experiments, compare the performance of different sampling strategies, select the ideal sampling strategies, reach some naive conclusion, conduct analysis on potential findings.
6. Document the results, and report findings. Organize the results into well-formatted slides and reports for better further discussing and presentation. Conclude the project.

References

[1] C. Shi, H. Yang, D. Cai, Z. Zhang, Y. Wang, Y. Yang, and W. Lam, “A thorough examination of decoding methods in the era of llms,” 2024. [Online]. Available: https://arxiv.org/abs/2402.06925

[2] S. S. Sahoo, M. Arriola, Y. Schiff, A. Gokaslan, E. Marroquin, J. T. Chiu, A. Rush, and V. Kuleshov, “Simple and effective masked diffusion language models,” 2024. [Online]. Available: https://arxiv.org/abs/2406.07524

[3] R. Yu, Q. Li, and X. Wang, “Discrete diffusion in large language and multimodal models: A survey,” 2025. [Online]. Available: https://arxiv.org/abs/2506.13759

[4] M. Arriola, A. Gokaslan, J. T. Chiu, Z. Yang, Z. Qi, J. Han, S. S. Sahoo, and V. Kuleshov, “Block diffusion: Interpolating between autoregressive and diffusion language models,” 2025. [Online]. Available: https://arxiv.org/abs/2503.09573

[5] Q. Wei, Y. Zhang, Z. Liu, D. Liu, and L. Zhang, “Accelerating diffusion large language models with slowfast sampling: The three golden principles,” 2025. [Online]. Available: https://arxiv.org/abs/2506.10848