Making LLMs Faster with Diffusion

Final Project Proposal – Group 1

20879 - Language Technology

Davide Beltrame, Giacomo Cirò, Luca Gandolfi, Vittorio Rossi

Autoregressive Language Models (ARMs) have demonstrated remarkable capabilities in text generation, but face significant bottlenecks during inference due to their sequential generation process. Diffusion Language Models (DLMs) offer a promising alternative with their parallel generation capabilities [4]. However, DLMs currently require predetermining the output length before generation begins, limiting their practical utility in real-world applications.

Broad research has been done on DLMs, but the issue of fixed output length remains mostly unexplored [1].

In this project, we want to design, develop and test a variable-length generation pipeline with for DLMs that integrates seamlessly with existing architectures in a plug-and-play manner. Our idea is to find a mechanism that predicts an upper bound for the output sequence length during the initial generation step, thereby constraining the number of tokens to be unmasked in subsequent stages without requiring fixed-length outputs.

Moreover, we think that exploiting the vast amount of resources already poured into ARMs training is fundamental, and training DLMs entirely from scratch should not be the preferred approach [3]. Interesting work has already successfully been done to adapt ARMs (e.g. GPT-2) for discrete diffusion generation [2].

Therefore, we want to reproduce some of these results by trying to adapt GPT-2 for discrete diffusion and test our dynamic length generation approach on this adapted model.

References

[1] Marianne Arriola, Aaron Gokaslan, Justin T Chiu, Zhihan Yang, Zhixuan Qi, Jiaqi Han, Subham Sekhar Sahoo, and Volodymyr Kuleshov. Block diffusion: Interpolating between autoregressive and diffusion language models. 2025.

[2] Shansan Gong, Shivam Agarwal, Yizhe Zhang, Jiacheng Ye, Lin Zheng, Mukai Li, Chenxin An, Peilin Zhao, Wei Bi, Jiawei Han, et al. Scaling Diffusion Language Models via Adaptation from Autoregressive Models. 2024.

[3] Zhengyang Liang, Hao He, Ceyuan Yang, and Bo Dai. Scaling laws for diffusion transformers. 2024.

[4] Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, Jun Zhou, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. Large language diffusion models. 2025.