DSSD: Efficient Edge-Device Deployment and Collaborative Inference via Distributed Split Speculative Decoding

Anonymous Authors[1]

Abstract

In order to maintain satisfactory inference accuracy while improving overall inference speed, speculative decoding (SD) has been integrated into the edge-device inference frameworks [10]–[12]. Following the idea of SD, the SLM on an end device generates a multi-token draft sequence, which undergoes parallel verification by the in-edge LLM with a single inference step (also called a verification step). The parallel verification significantly accelerates inference compared to traditional sequential generation while preserving high accuracy through the rejection of divergent draft tokens and subsequent content. However, as the input and output tokens are shared between devices and the cloud, this framework still inevitably raises significant privacy concerns.

1. Introduction

Large language models (LLMs) have revolutionized natural language processing, enabling powerful applications such as conversational agents, machine translation, and code generation. Despite their capabilities, deploying LLM faces significant challenges across both devices and cloud environments. On devices, stringent constraints such as limited memory capacity, restricted battery life, and insufficient computational power hinder the adoption of traditional LLM frameworks. Cloud-based deployments, while benefiting from scalable computational resources, suffers from unpredictable network latency and jitter. In addition, the inherent mobility of end-users can lead to frequent connectivity disruptions, making continuous access to cloud-based services unreliable.

To address these challenges, researchers have proposed a collaborative edge-device architecture that strategically deploys a small language model (SLM) on the device while offloading the large language model (LLM) to a base station (BS) or edge server (Ding et al., 2024; Hao et al., 2024). In (Ding et al., 2024), a router trained to predict query difficulty and desired quality level enables cost-efficient assignment of queries to either the small or large model. In (Hao et al., 2024), a cost-aware draft-verification approach was employed. By tuning a predefined threshold p_t for the generated token probability, a controllable performance-cost trade-off was achieved.

However, these studies improves efficiency with a compromise of LLM inference accuracy. Therefore, **speculative decoding (SD)** was taken into account, where a small "draft" model to generate γ candidate tokens autoregressively, and then a big "target " model to verify these draft tokens in parallel (Leviathan et al., 2023; Chen et al., 2023). In this way, the inefficiency of autoregressive token generation was mitigated without sacrificing the quality of inference. Furthermore, a **distributed speculative decoding (DSD)** architecture was first introduced in (Zhao et al., 2024) with the draft model for token generation on the device and the target model for verification at the edge. The author trys to optimize the number of tokens generated by SLM to minimize delay and power consumption, taking uplink and downlink transmission into consideration.

Nevertheless, this hybrid deployment approach is constrained by communication bottlenecks: for each token, the device must transmit a full vocabulary distribution to the BS/edge server for LLM verification, resulting in a communication payload linearly dependent on vocabulary size. In (Oh et al., 2024), the author proposed to skip uplink transmissions and LLM inference on tokens likely to be accepted. This improves token throughput but still at the expense of inference accuracy.

Building on previous works, we offer our solution: **a distributed split speculative decoding (DSSD)** framework. Specifically, SLM and LLM are still deployed on device and at edge, respectively. But the verification phase of SD was further split and distributed across the device and edge.

[1] Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. **Correspondence to**: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

References

Chen, C., Borgeaud, S., Irving, G., Lespiau, J.-B., Sifre, L., and Jumper, J. Accelerating large language model decoding with speculative sampling. *arXiv preprint arXiv:2302.01318*, 2023.

Ding, D., Mallick, A., Wang, C., Sim, R., Mukherjee, S., Ruhle, V., Lakshmanan, L. V., and Awadallah, A. H. Hybrid llm: Cost-efficient and quality-aware query routing. *arXiv preprint arXiv:2404.14618*, 2024.

Hao, Z., Jiang, H., Jiang, S., Ren, J., and Cao, T. Hybrid slm and llm for edge-cloud collaborative inference. In *Proceedings of the Workshop on Edge and Mobile Foundation Models*, pp. 36–41, 2024.

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

Leviathan, Y., Kalman, M., and Matias, Y. Fast inference from transformers via speculative decoding. In *International Conference on Machine Learning*, pp. 19274–19286. PMLR, 2023.

Oh, S., Kim, J., Park, J., Ko, S.-W., Quek, T. Q., and Kim, S.-L. Uncertainty-aware hybrid inference with on-device small and remote large language models. *arXiv preprint arXiv:2412.12687*, 2024.

Zhao, W., Jing, W., Lu, Z., and Wen, X. Edge and terminal cooperation enabled llm deployment optimization in wireless network. In *2024 IEEE/CIC International Conference on Communications in China (ICCC Workshops)*, pp. 220–225. IEEE, 2024.

A. You *can* have an appendix here.