LayerSkip for Mixture of Experts (MoE) Architecture

Tom Jeong Nicholas Papciak
Kavya Golamaru Vishal Maradana Rishi Bandi
Georgia Tech

March 9, 2025

Abstract

This project proposal explores both the implementation of LayerSkip in the Mixture of Experts (MoE) architecture for large language models. MoE has emerged as a promising approach for scaling LLMs efficiently, as demonstrated by models like DeepSeek V3. However, current MoE implementations still face challenges in load balancing and expert utilization. We propose applying LayerSkip, a novel technique that allows selective bypassing of expert layers based on input complexity, to add even more sparsity to an MoE model. By dynamically adjusting the depth of processing, LayerSkip aims to improve computational efficiency and model performance. We hypothesize that this approach will lead to more balanced expert utilization, reduced inference times, and the possibility of enhanced generalization across many different tasks. Our research involves designing the LayerSkip mechanism, integrating it into an existing MoE framework (likely Mixtral 8x7B), and conducting extensive experiments to evaluate its impact on model efficiency and effectiveness across various benchmarks.

Contents

1 Introduction

Recent LLMs advancements have leveraged MoE architectures to improve computational efficiency while maintaining performance. They do this by selectively activating only a subset of expert layers per input, significantly reducing the number of active parameters per forward pass. However, there are some challenges. Traditional MoE models suffer from inefficiencies in utilization, high computational overhead, challenges in fine-tuning, and suboptimal load balancing.

LayerSkip has been proposed as an architecture to dynamically bypass certain layers based on input complexity, allowing models to adjust processing depth adaptively. This model has proven its efficiency in transformers, specifically Llama 17B [1]. LayerSkip in MoE has not been explored; we aim to integrate LayerSkip into a Mixture of Experts model,thereby optimizing computational efficiency while preserving—or even enhancing—model accuracy.

2 Proposed Framework: LayerSkip for MoE

Our approach to implementing LayerSkip for Mixture of Experts (MoE) consists of four main stages:

2.1 Training using Expert Dropout and Early Exit Loss

2.1.1 Expert Dropout

Layer dropout, first proposed by [3], has been used to improve model robustness and increase training/inference speeds. LayerSkip, uses dropout in a novel way to improve "early exit inference". We will integrate layer skip dropout with MoE techniques like a dynamic mechanism to randomly deactivate a subset of experts during training, similar to layer dropout in traditional architectures. The high level implementation is:

- Randomly deactivate a subset of experts in each forward pass.
- Adjust the gating network to redistribute probabilities among active experts.
- Implement a dynamic routing mechanism that adapts to the available experts.
- Apply LayerSkip within the different experts.
- Adding intermediate output layers after groups of experts.
- Computing loss at each intermediate layer
- Implementing a confidence-based early exit mechanism.

2.2 Inference using Speculative Decoding

Building upon the LayerSkip mechanism, we will try to enhance inference efficiency using speculative decoding. We're going to utilize a smaller draft model to propose multiple candidate tokens.

Traditional speculative decoding methods have proven to be inefficient with MoE models. However, research by [4] suggests that we can combine autoregressive decoding for initial tokens and parallel decoding to boost speculative decoding via MoE.

2.3 Integration with MoE models

Our framework will be integrated into a smaller MoE model, we are currently thinking about Mixtral 8x7B:

- Modifying the model architecture to include LayerSkip components.
- Ensuring auxiliary loss-free balancing by monitoring expert utilization during training.
- Adapting the routing algorithm to consider both input complexity and expert specialization.

3 Related Reading

Our proposed LayerSkip for MoE architecture builds upon several recent advancements in the field of efficient language model inference and training. Here's an overview of relevant research that informs our approach:

3.1 LayerSkip for Transformer Models

[1] introduced LayerSkip, an "end-to-end solution for speeding up inference in large language models".

3.2 Mixture of Experts Optimization

Recent work has focused on optimizing MoE architectures:

- [2] proposed "Switch Transformers," which use a sparse gating mechanism to route tokens to a small number of experts, enabling efficient scaling to trillion-parameter models.
- The "Sparsely Gated Mixture of Experts" approach introduced by [5] uses top-k gating to activate only a subset of experts per input, reducing computational overhead.

3.3 Speculative Decoding for MoE

[4] developed "Jakiro," a method that combines autoregressive decoding for initial tokens with parallel decoding to enhance speculative decoding in MoE models. This approach addresses the inefficiencies of traditional speculative decoding methods when applied to MoE architectures.

By integrating these concepts, our LayerSkip for MoE framework aims to leverage the strengths of both LayerSkip and MoE architectures, potentially leading to significant improvements in inference speed and model efficiency.

4 Compute requirements

Based on the LayerSkip proposal, we will likely need multiple (2-4) H100 GPUs. We will rent these GPUs from Hyperbolic.

5 References

References

[1] M. Elhoushi, Z. Li, A. Srinivas, et al. Layerskip: Enabling early exit inference and self-speculative decoding. arXiv:2404.16710, 2024.

[2] W. Fedus, B. Zoph, and N. Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. arXiv:2101.03961, 2021.

[3] Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Weinberger. Deep networks with stochastic depth, 2016.

[4] H. Huang, F. Yang, Z. Liu, et al. Jakiro: Boosting speculative decoding with decoupled multi-head via moe. arXiv:2502.06282, 2024.

[5] N. Shazeer, A. Mirhoseini, K. Maziarz, et al. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. arXiv:1701.06538, 2017.