Qeios

Peer Review

Review of: "Router-Tuning: A Simple and Effective Approach for Enabling Dynamic-Depth in Transformers"

Yan Pang[1]

1. University of Virginia, Charlottesville, United States

Summary:

The paper introduces Router-Tuning, a method for enabling dynamic-depth computation in transformers by fine-tuning only lightweight router networks. This approach significantly reduces computational costs while preserving high model performance. Additionally, the authors propose MindSkip, which selectively applies dynamic depth to attention layers—one of the most computationally expensive components in transformers. By focusing on attention layers, the method optimizes computational efficiency while minimizing performance degradation. Experimental evaluations on models such as Llama-3 and Mistral demonstrate substantial improvements in inference speed and memory usage, with minimal trade-offs in performance.

Detailed Comments:

Clarity and Organization:

The paper is generally well-organized, with clear sections outlining the motivation, methodology, and results. However, the writing quality requires improvement:

- In the introduction, “Despite its potential, current MoD methods MoD methods are still underexplored and face several critical challenges.” repeats “MoD methods.”

These issues affect readability and should be carefully revised for conciseness and clarity.

Declarations

Potential competing interests: No potential competing interests to declare.

doi.org/10.32388/FZMB63