BLIMEY: TOWARDS BETTER ROUTING METHODS IN SPARSE MIXTURE OF EXPERTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Mixture of Experts (MoE) architectures offer a promising avenue for scaling neural networks by facilitating parameter-efficient model expansion while optimizing FLOP utilization. However, several challenges persist, including sub-optimal expert utilization, vanishing gradients, sub-par expert specialization, and inconsistent routing decisions. We introduce BLIMEY, a novel diagnostic framework that systematically quantifies these critical issues, enabling design and implementation of more robust and efficient expert routing algorithms. BLIMEY equips researchers with granular, interpretable metrics on MoE dynamics, explaining both the performance advantages over monolithic architectures and the persistent algorithmic bottlenecks, thereby offering a comprehensive diagnostic framework for MoE optimization. Leveraging BLIMEY, we establish new scaling laws, that surpass established benchmarks like Chinchilla, achieving 3x reductions in FLOPs and 5% performance gains. Furthermore, our framework unveils significant optimization potential in routing algorithms, revealing sub-optimal expert specification and load imbalances in current methodologies. To accelerate innovation in MoE routing and computational efficiency, we have open-sourced BLIMEY framework, including its diagnostic tools and implementation libraries.