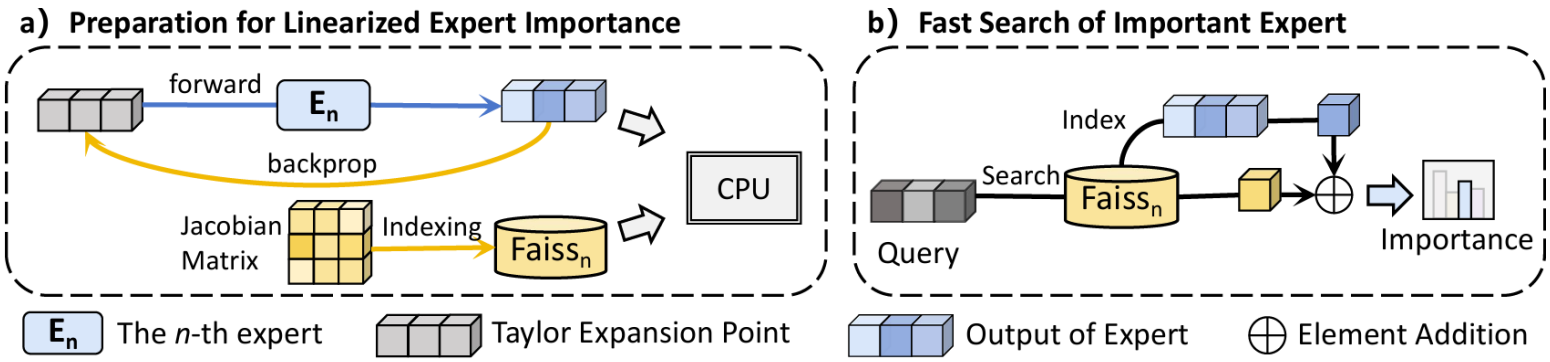
PREP: Input-Aware Expert Pruning for Efficient MoE Deployment

Chaoran ZHANG, Lixin ZOU, Xixun LIN, Wen ZOU

Frontiers of Computer Science, DOI: 10.1007/s11704-026-52030-x

Problems & Ideas

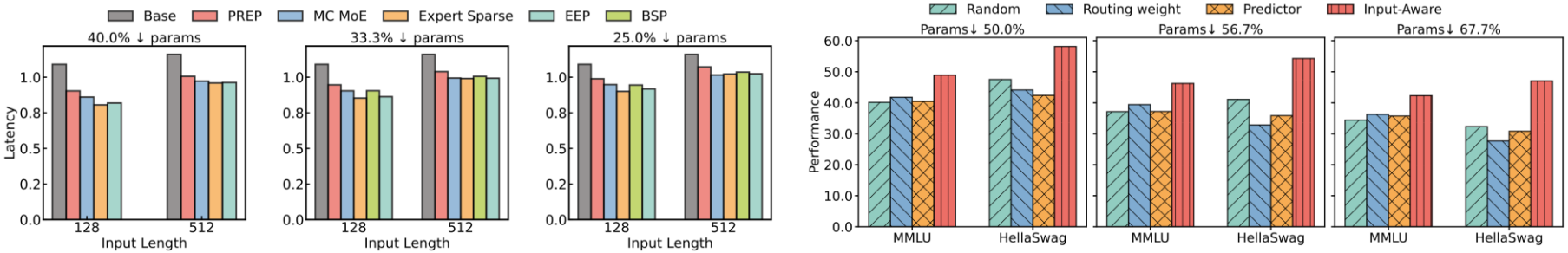
- Problems of conventional MoE compression & deployment:
  - MoE-LLM inference requires storing full expert parameters on GPU, making deployment heavily constrained by GPU memory.
  - Existing MoE pruning/approximation methods rely on static calibration data, lacking input flexibility as expert importance is highly input-dependent.
- Ideas: Propose PREP, a training-free and input-aware expert pruning method for efficient MoE inference, which dynamically evaluates input-conditioned expert importance, retains only critical experts in GPU memory and offloads the rest to CPU.



The process of evaluating an expert's importance. Subfigure (a) illustrates the process of collecting the output and Jacobian matrix of the expert at the expansion point during the \textbf{offline} stage. Subfigure (b) shows the \textbf{online} stage, where the importance of the expert is computed based on queries through fast search.

Main Contributions

- Contributions:
  - Theoretically derive an effective activation-based expert importance metric, and propose an efficient input-aware expert evaluation strategy to identify critical experts without additional retraining;
  - Introduce a hardware-friendly adaptive layer-wise expert loading strategy, which reduces peak GPU memory consumption while preserving model accuracy;
  - Extensive experiments verify that PREP achieves favorable accuracy-efficiency trade-offs and consistently outperforms static pruning baselines across multiple compression ratios.



Per-token inference latency and expert evaluation strategy performance under dynamic pruning.Left: Per-token inference latency on Mixtral-8×7B Instruct across input lengths under different expert-parameter reduction ratios;Right: Comparison of expert evaluation strategies for dynamic pruning in terms of task performance.