Efficient Model Specialization via Training-time and Test-time Adaptation

Huanrui Yang

Department of Electronic and Computer Engineering, University of Arizona
1230 E. Speedway Blvd., P.O. Box 210104
Tucson, AZ 85721, USA
huanruiyang@arizona.edu

Abstract

In this talk, we discuss efficient model specialization algorithm to adapt the pretrained model towards downstream tasks while improving its efficiency, efficiently generalizing to multiple tasks via dynamic architectures, and improving inference-time efficiency utilizing the diversity within model block functionalities. These research directions serve as the foundation towards co-designing models, tasks, systems, and hardware for a reconfigurable efficient intelligence future.

Introduction

The scaling up of single model size is stalling due to the inadequate high-quality data to pretrain on and the diminishing return in performance gain. Inspired by the trend of specialization and heterogeneity explored in the semiconductor's Moore's Law, we discuss effective methods to specialize a pretrained model to heterogeneous variants for both improved efficiency and performance.

Work Being Surveyed

Compression-aware adaptation. Supervised finetuning (SFT) is used for specialization. However, model compression before SFT leads to reduced learnability, yet compression after SFT leads to performance drop. We discuss two compression-aware adaptation work: PAT (Liu et al. 2025), in AAAI 2025, proposes a novel multiplicative PEFT module that can achieve 30% structural sparsity in LLMs without performance loss through an instruction tuning process. NPFT (WANG and Yang 2025), in CPAL 2025, proposes noise-perturbed finetuning that can make the model more robust against quantization noises through a quick PEFT process, improving PTQ performance for all quantizers.

Efficient generalization to diverse tasks. LLM requires different capabilities to fulfill diverse tasks. MoE allows dynamic switching of model capability but requires significant memory. Our work T-REX (Zhang et al. 2024) introduces a rank-1 mixed-and-match MoE that can enable quadratic amount of experts with linear memory overhead. A cluster-based router guidance method is also proposed to remove expert redundancy and improve model generalizability. T-REX achieves up to 1.78% mean accuracy improvement with30%-40% less trainable parameters across 14 datasets.

Internal diversity across model blocks Foundation models are typically designed with uniform blocks, which benefits training but cause redundancy at inference. We explore diversifying model architecture and KV cache at inference time. MSQ (Han et al. 2025), in ICCV 2025, propose LSB sparsification to achieve mixed-precision quantization scheme across model layers. FIER (Wang et al. 2025), in EMNLP 2025 findings, explores query-aware dynamic KV cache retrieval that leads to an 1.2-1.5$\times$ lossless speedup.

In summary, in this talk we will discuss efficient specialization towards downstream task, improved generalization, and diversified intermediate modules. We will advocate "***don't just squeeze it, specialize it***" as the take-home message to achieve efficient foundation model applications.

Proposed Future Work

Building on top of the surveyed work, we will discuss two potential directions to explore in the near future. One is the specialization for multi-round reasoning, where we can tradeoff model capability with thinking length or to specialize multiple model variants for different thinking stages. The other is the specialization for multi-agent interactions, where the model can be specialized towards different agent functionalities and can be co-designed with the task assigned to each agent for overall system efficiency.

Acknowledgments

Work surveyed in this talk are supported in part by the research collaboration grant from Panasonic and TetraMem, Inc. Part of the work surveyed are based on the High Performance Computing (HPC) resources supported by the University of Arizona TRIF, UITS, and Research, Innovation, and Impact (RII), and maintained by the UArizona Research Technologies department.

References

Han, S.; Yoon, S.; Kim, J.; Wang, D.; Jeon, K. E.; Yang, H.; and Ko, J. H. 2025. MSQ: Memory-Efficient Bit Sparsification Quantization. *arXiv preprint arXiv:2507.22349*.

Liu, Y.; Yang, H.; Chen, Y.; Zhang, R.; Wang, M.; Du, Y.; and Du, L. 2025. PAT: Pruning-Aware Tuning for Large Language Models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 24686–24695.

Wang, D.; Liu, Z.; Wang, S.; Ren, Y.; Deng, J.; Hu, J.; Chen, T.; and Yang, H. 2025. FIER: Fine-Grained and Efficient KV Cache Retrieval for Long-context LLM Inference. *arXiv preprint arXiv:2508.08256*.

WANG, D.; and Yang, H. 2025. Taming Sensitive Weights: Noise Perturbation Fine-tuning for Robust LLM Quantization. In *The Second Conference on Parsimony and Learning*.

Zhang, R.; Liu, Y.; Yang, H.; Zheng, S.; Wang, D.; Du, Y.; Du, L.; and Zhang, S. 2024. T-REX: Mixture-of-Rank-One-Experts with Semantic-aware Intuition for Multi-task Large Language Model Finetuning. *arXiv preprint arXiv:2404.08985*.