Efficient Inference for Large Language Models – Algorithm, Model, and System

Xuefei Ning, Guohao Dai, Haoli Bai, Lu Hou, Yu Wang and Qun Liu

The inference of LLMs incurs high computational costs, memory access overhead, and memory usage, leading to inefficiencies in terms of latency, throughput, power consumption, and storage.

To this end, this tutorial focuses on the increasingly important topic of ***Efficient Inference for LLMs*** and aims to *provide a systematic understanding of key facts and methodologies from a designer's perspective*. We start by introducing the basic concepts of modern LLMs, software and hardware. Following this, we define the efficiency optimization problem. To equip the audience with a designer's mindset, we briefly explain how to diagnose efficiency bottlenecks for a given workload on specific hardware.

After introducing the basics, we will introduce our full-stack taxonomy of efficient inference methods for LLMs. We will walk through each category of methodology, using one to three representative methods as examples for each leaf subcategory, elaborating on the design logic behind each method and which inefficiency factors they primarily address. Finally, we will wrap up with a takeaway summary, and future research directions.

Xuefei Ning, Research-Track Assistant Professor, Tsinghua University
email: foxdoraame@gmail.com
website: https://nics-effalg.com/ningxuefei/
Xuefei Ning is a research-track assistant professor with the Department of Electronic Engineering at Tsinghua University. She obtained her Ph.D. at Tsinghua University in 2021. Her research focuses on efficient deep learning. She has published 30+ papers in leading AI conferences and journals. She has published a Chinese book on efficient deep learning. She serves as a senior area chair for ACL 2025, an area chair for CVPR 2025 and ICLR 2026.

Guohao Dai, Associate Professor, Shanghai Jiao Tong University
email: daiguohao@sjtu.edu.cn
website: https://dai.sjtu.edu.cn/pepledetail.html?id=218
Guohao Dai is an associate professor with the Department of Electronic Information and Electrical Engineering at Shanghai Jiao Tong University.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts, pages 1–3

His research focuses on sparse computing, heterogeneous hardware computing, emerging hardware architecture, etc. He served as Co-Chair for the Ph.D. Forum at DAC 2024, TPC member for DAC 2024/DAC 2023/VLSID 2024. He received the Best Paper Award in ASP-DAC 2019, and Best Paper Nominations in DATE 2024/DATE 2023/DAC 2022/DATE 2018. He is the winner of the NeurIPS Billion-Scale Approximate Nearest Neighbor Search Challenge in 2021, and the recipient of the Outstanding PhD Dissertation Award of Tsinghua University in 2019.

Haoli Bai, Researcher, Huawei Technologies Co. Ltd.
email: baihaoli@huawei.com
website: https://haolibai.github.io/
Haoli Bai is a researcher at Huawei Noah's Ark Lab. He obtained his Ph.D. at the Chinese University of Hong Kong in 2021. His research focus is efficient deep learning with the purpose to minimize memory and computational requirements, particularly for large language models. He has published multiple research works on network quantization, pruning, and relevant topics, with applications on Huawei Ascend Chips and products. He obtained the ACML Best Student Paper Runner-up Award (2016), and has served as the PC member for top AI conferences (e.g., NeurIPS, ICML, ICLR).

Lu Hou, Researcher, Huawei Technologies Co. Ltd.
email: houlu3@huawei.com
website: https://houlu369.github.io/
Lu Hou is a researcher at Huawei Noah's Ark Lab. She obtained her Ph.D. from Hong Kong University of Science and Technology in 2019. Her research focuses on developing efficient deep learning models with lower memory and computation costs, especially for large pre-trained language and multimodal models. Her researches have been published at leading conferences (e.g., NeurIPS, ICML, ICLR, ACL, EMNLP) as well as been applied to various chips, products and LLMs at Huawei.

Yu Wang, Full Professor, Tsinghua University
email: yu-wang@tsinghua.edu.cn
website: https://nicsefc.ee.tsinghua.edu.cn/people/YuWang
Yu Wang is a professor, an IEEE fellow, the chair of the Department of Electronic Engineering in Tsinghua University, the dean of the Institute for Electronics and Information Technology in Tianjin, and the vice dean of the School of Information Science and Technology in Tsinghua University. His research interests include the application specific heterogeneous computing,

2

processing-in-memory, intelligent multi-agent system, and power/reliability aware system design methodology. He has published more than 90 journals (64 IEEE/ACM journals) and 270 conference papers in the areas of EDA, FPGA, VLSI Design, and Embedded Systems, with the Google Scholar citation over 22,000. He has received four best paper awards and 12 best paper nominations. He has been an active volunteer in the design automation, VLSI, and FPGA conferences. He is the co-founder of Deephi Tech (a leading deep learning solution provider), which is acquired by Xilinx (AMD) in 2018. He is also the promoter of Infinigence AI Tech (a leading AI infrastructure solution provider), which achieves industry-leading large language model inference performance on more than 10+ different chips.

Qun Liu, Huawei Technologies Co. Ltd.
email: `qun.liu@huawei.com`
website: `https://liuquncn.github.io/`
Qun Liu is the chief scientist of Speech and Language Computing of Huawei Noah's Ark Lab. He is formerly a professor of Dublin City University, the Theme Learder of NLP at the ADAPT Centre, Ireland, a professor & researcher & the leader of NLP research group in the Institute of Computing Technology , Chinese Academy of Sciences (ICT-CAS). He obtained his B.Sc., M.Sc. and Ph.D. degrees in the University of Science and Technology of China, ICT-CAS, and Peking University respectively. His research interests cover natural language processing, language modeling, machine translation, question answering, dialog, etc. His academic achievements include ICTCLAS Chinese word segmentation and POS tagging system, syntax-based statistical machine translation, neural machine translation, machine translation evaluation, etc. He has been the leader or a participant in several large-scale projects funded by Chinese government, Irish government or European Union. He has published 300+ papers in academic conferences or journals, with 20,000+ citations. He has supervised 50+ Master or Ph.D. students into completion. He has obtained Google Research Award (2012), first prize of Qian Weichang Award for Chinese Information Processing Science and Technology (2010), and second prize of China National Award for Science and Technology Progress (2015), ACL Best Long Paper Awards (2019), and ACL Outstanding Paper Awards (2022, 2024).