

教育经历

东北大学, 软件工程专业, 本科, 中国

部分必修课程: 操作系统, 计算机网络, 计算机体系结构

2021 年 9 月 – 2025 年 6 月

工作经历

腾讯科技有限公司 WXG 微信事业群

技术架构部, 基座大语言模型 pre-train/post-train 工程师

2024 年 12 月 – 至今

- 负责微信内部基座大语言模型预训练/后训练的框架开发工作, 包括模型训练加速, 显存优化等。
- 参与 DeepSeek Infra 的复现工作, 使用 ThunderKittens 重构 DeepGEMM, 负责框架内 80% 的 Kernel 开发工作。
- 参与 RL 框架的搭建工作, 搭建 Parameter Server, 使用 RDMA 实现参数高性能通信。

项目经历

大模型分布式预训练引擎

2024 年 12 月 – 至今

大模型分布式预训练引擎是为了微信内部新一代基座 WELM 自研的大规模预训练引擎, 支持多机多卡训练万亿级别参数的语言模型, 实现业界领先的模型大规模集群分布式训练, 完美适配微信生态。

- 高性能并行策略:** Dense 模型实现 3D 并行 TP+DP+PP 以及 ZeRO, 同时对 MoE 模型支持 EP 并行, 实现多机多卡分布式训练。
- 定制化内核:** 使用 Triton/ThunderKittens 实现大部分算子的前向和反向计算集成到框架内, 同时实现多种算子的 Fusion, 优化显存占用, 加速训练效率。
- FP8 精度支持:** 支持 FP8 低精度训练, 同时也可以灵活选择训练精度, 通过 Quant/Dequant 实现低精度训练。
- 前沿技术集成:** 支持 Dualpipe/DeepEP, 实现计算和通信的 overlap。

技能

- ACM 竞赛经历** 熟悉算法与数据结构, 代码能力强, 曾获得 **ICPC 金牌**, ICPC/CCPC 银牌若干, **省赛亚军**, codeforces 2300+
- 编程语言:** 泛语言 (编程不受特定语言限制), 且尤其熟悉 C/C++, Python, Go, 较为熟悉 Java, Rust (排名均不分先后)。
- LLM 算法:** 熟悉大模型预训练相关算法, 熟悉 Dense, MoE 前沿技术, 深入了解 Llama, DeepSeek, Qwen 等模型结构, 了解 Rotary, RMSNorm, Swiglu 等原理和设计. 熟悉大模型后训练相关算法, 熟悉 DPO, PPO, GRPO 等原理和设计。
- LLM Infra:** 阅读过 Megatron-LM, DeepSpeed, FSDP, Verl 等框架源码, 了解其原理和设计, 了解 Ray, NCCL 等分布式计算/通信库, 熟练掌握 PyTorch.
- System:** 了解 Linux 底层原理, 了解 GPU/CPU 硬件知识, 掌握 X86-64, ARM64, RISC-V 等架构知识, 熟练掌握 NVIDIA GPU 底层架构以及 PTX 汇编, 熟练编写 CUDA 相关代码。

其它

- 个人博客: [Crates.io](https://crates.io), 记录学习心得与技术分享。
- 语言: English - 熟练(通过四六级), 汉语 - 母语水平