丁羿然

yiran.ding2@gmail.com | github.com/Yiyi-philosophy | 网站 | 博客 | 谷歌学术 | 丁羿然

教育背景

杭州电子科技大学 (电子与科技) 2021/09 - 2024/06
工学学士，电子与信息工程 中国，浙江，杭州

- GPA: **3.8**/4.0 (**90**/100, 前 3%)
- 一等奖学金（四个学期），获奖率 5% | 省政府奖学金，获奖率 2%

研究兴趣与技能

- 大语言模型 (LLM):
  - 自然语言处理 (NLP): 评估，数据工程，SFT
  - 机器学习系统 (MLSys): 推理优化，微调
  - 架构: Transformer, Mamba
- 技能
  - Python(Pytorch), C/C++, Matlab
  - OpenMP, MPI, CUDA,
  - Git, Shell, Docker, Conda | Verilog

发表论文

- **LongRoPE: 超越 200 万 token 的大语言模型上下文扩展. Y. Ding**, L. L. Zhang, C. Zhang, Y. Xu, N. Shang, J. Xu, F. Yang, M. Yang. (2024). 第 *41* 届国际机器学习大会 (**ICML**). [论文]

研究经验

大语言模型序列扩展: LongRoPE 2023/06 - 2024/07
实习生，微软亚洲研究院 *(MSRA)*，导师 张丽娜 中国，北京

- 将预训练的大语言模型（Llama，Mistral）的上下文窗口扩展至 **2048k** tokens，在 256k 训练长度下仅需 **1k** 微调步骤，保持了原有性能。
- 利用**位置插值中的非均匀性**进行更好的微调初始化，采用渐进扩展策略，并**重新调整** LongRoPE 以**恢复短上下文**窗口性能。
- 支持微调 **Phi-3** (mini, small) 至 **128k** 上下文：Phi-3 模型, Phi-3 报告
  - 从不同来源准备和清理 128k 长度的数据集进行微调，并研究恢复短上下文（4k）性能的方法。

大语言模型推理优化，导师 2023/03 - 2023/07
杭州电子科技大学，导师 缪正 中国，杭州

- 开发了一种新颖的**块调度**方法，将批处理细粒化为层，与当前最优块调度相比，理论上能提高 **2 倍**吞吐量和延迟性能。
- 通过**聚类**、**重排**和使用**稀疏注意力**，在不显著损失精度的情况下将权重、KV 缓存和激活压缩为 **4 位**，减少内存消耗。

医学影像处理 2023/03 - 2023/07
杭州电子科技大学，导师 朱力 中国，杭州

- 领导并设计了自动评估帕金森病患者手指敲击视频的项目。item 开发了基于 **LSTM-FCN** 的模型来分类患者。结果准确率为 83.7%，在该论文的数据集上超越了文献中的最先进结果。 item **使用技术**: 姿态估计 (Mediapipe Hands), RIFE 算法 (时间序列插值), LSTM, FCN.

其他经验

边缘设备上的大语言模型推理 2023/07 - 2023/09

- 基于 **7B Alpaca 模型** 开发了一种**离线**大语言模型。实现了**中文问答**和对话功能，并部署在拥有 16Tops 计算能力的 8GB 边缘设备上，采用 int8。扩展了中文词汇表，**微调**模型以适应中文指令数据，并使用 **int4** 量化对模型进行压缩，显著提升了对中文指令的理解与执行。

DGEMM (报告) 2023/07 - 2023/09

- 实现并优化了各种矩阵乘法技术，包括**分块式**、**递归**和**缓存无关**方法，将计算时间减少了 **82%**。通过 **Z 字型莫顿序列**重排矩阵数据，改善缓存利用率。