李杰

南京大学 计算机学院
计算机软件新技术全国重点实验室
导师 李武军教授
jie-li@smail.nju.edu.cn
13373291169
中共党员

教育背景

南京大学，计算机学院，计算机科学与技术专业，*硕士* 2023.9 - 至今

哈尔滨工业大学，计算学部，计算机科学与技术专业，*本科* 2019.9 - 2023.7

- 研究生一等学业奖学金 * 1、本科校级优秀毕业生、本科生国家励志奖学金 * 3

过往经历

科研工作：自动并行方法 UniAP 2023.12 - 至今

- *描述*：该工作提出了一种新的自动并行方法，联合优化层内并行和跨层并行。前人方法中，一部分方法只优化了这两类并行策略中的某一类；另一部分方法虽然同时考虑了两类并行策略，但是层次化地优化这一问题，即先单独考虑层内并行求出最优解，再固定层内并行策略，求出跨层并行的最优解。而我们的方法将层内并行和跨层并行统一建模成一个混合整数二次规划问题，可以直接求解全局最优解。方法分为 3 部分：性能评估、代价模型和搜索算法。
- *我的职责*：实现性能评估中的模型显存评估和代价模型中的显存代价模型。添加整个框架对 Llama 模型的支持并在大集群（最大 64 卡）上实验。完成框架在国产卡（海光 DCU）上的适配。
- *成果*：经在 4 种不同集群、5 种不同的基于 Transformer 结构的模型上的实验验证，相对于前人最好的方法，我们的方法搜出的策略的吞吐量最大有 3.8 倍的提升，策略搜索速度最大有 107 倍的提升。该方法已总结成文并被 CVPR2025 正式录用（**Oral**），本人为**共同一作**。该论文已被选为 **Award Candidate**（仅有 14 篇入选，占所有投稿的 0.1%）。该论文已公开在 arXiv。该方法支撑了本人导师的**科技部国家重点研发计划课题**“云计算环境下的机器学习优化基础算法及判别分析算法研究”。

研发课题：面向算力网的任务求解核心算法及基础理论研究 2024.9 - 至今

- *描述*：该课题由徐宗本院士牵头，是南京大学、复旦大学、鹏城实验室三方的合作课题。我方主要负责分布式训练算法。本人是我方的主要学生负责人。
- *我的职责*：完成我们的自动并行方法 UniAP 在鹏城实验室提供的华为昇腾 NPU 集群上的适配。
- *成果*：在最大 32 卡昇腾 910a 集群上进行实验，使用 Megatron 框架训练 Llama2-7B，用适配后的 UniAP 搜出来的并行策略，相对于经验设置，最大有 1.61 倍的吞吐量提升。

科研工作：4D 并行的自动并行方法 UniAP-4D 2024.7 - 至今

- *描述*：为适应大模型的训练数据的序列长度的不断增长，研究者们提出了一类新的并行策略：序列并行。序列并行可与 3D 并行结合，组成 4D 并行。现有的工作都是在 3D 并行上做自动并行，还没有工作在 4D 并行的自动并行算法上做出探索。UniAP-4D 将 UniAP 拓展到 4D 并行上，实现一个支持 4D 并行的自动并行的训练系统。
- *我的职责*：该工作基本由本人独立完成。工作内容主要分为两部分：训练框架的拓展和自动并行算法的拓展。训练框架的拓展包括接入 RingAttention 和 DeepSpeed Ulysses 两种序列并行策略，并支持它们和其他并行策略的任意结合。自动并行算法的拓展包括性能评估的拓展，代价模型的拓展，搜索算法的拓展。
- *成果*：完成训练框架的拓展，在 4 卡 a6000 上使用较小模型做实验，验证了 loss 的稳定下降和不同并行策略下的 loss 曲线的一致，初步确认实现的正确性。基本完成自动并行算法的拓展，但仍有待在更大集群上进一步实验。

项目：中文医疗大模型 2023.10 - 至今

- *描述*：选取开源的中文大模型 Baichuan2-13B 做为基座模型，使用 Llama-factory 和 Deepspeed 框架，在医学教材、开源医疗数据上做增量预训练，最后在下游任务上微调和评测。
- *我的职责*：开源中文大模型的医学能力调研。在 DISC-Med-SFT 数据集上的微调。在 PromptCBLUE 和 CMB 评测基准上评测。使用 UniAP 算法加速训练。
- *成果*：基于该基础模型，在南京鼓楼医院的真实数据上微调获得的医学影像报告生成模型已在该医院落地应用。同时，使用 Llava 做为基座模型的多模态模型的训练也正在进行中。

项目：电网异常流量检测大模型 2024.4 - 至今

- *描述*：与国网智能电网研究院合作训练了一个异常流量检测大模型，目标是该模型能够识别该公司日常流量中的恶意攻击流量。选取 CodeQwen1.5-7B 做为基座模型。目前该项目主要完成了两项工作：在该公司的大量日常流量上做增量预训练，然后在公开数据集 data_capec_multilabel 上微调并评测。
- *我的职责*：调研相关领域的开源大模型并选取效果最好的作为基座模型。对该公司的日常流量数据做预处理，得到大模型可以处理的格式。调研相关公开数据集并选取最终的微调和评测数据集。使用 UniAP 算法加速训练。
- *成果*：目前该项目仍在进行中。初步结果表明，训练得到的大模型在 data_capec_multilabel 上的 F1 分数达到近 0.9，优于传统机器学习方法 (0.8 左右)。该公司认为该模型有落地应用价值。

竞赛：ACM 竞赛 2021.6 - 2022.4

- 描述：在大三参与一个赛年的 ACM 竞赛。
- 我的职责：本人所在队伍实力比较均衡，三人贡献基本均等。
- 成果：ICPC 亚洲区域赛银牌

专业技能

分布式训练：熟悉 Megatron 框架的使用，了解其底层代码实现，我们的 UniAP 训练框架是基于 Megatron 实现的。熟悉 DeepSpeed 框架的使用，用其做过大模型的分布式训练。

大语言模型：熟悉主流的大语言模型的结构。熟悉 Llama-factory 框架的使用，用其做过大模型的预训练、微调、推理。

机器学习与深度学习：掌握机器学习与深度学习相关算法。

国产显卡的使用：有使用昇腾 NPU 和海光 DCU 做大模型分布式训练的经验。

编程语言：掌握 Python、C++。

奖励荣誉

论文方面:

UniAP: Unifying Inter- and Intra-Layer Automatic Parallelism by Mixed Integer Quadratic Programming（共同一作，CVPR2025, Oral, Award Candidate） 2025

比赛方面:

第 46 届 icpc 亚洲昆明区域赛银牌 (667 支参赛队伍中排名第 56 位) 2022

第十四届 iCAN 国际创新创业大赛中国总决赛三等奖 2020

2020 物联网大赛华为杯东北赛区一等奖 2020

全国信息学奥林匹克联赛 NOIP2017 提高组省一等奖 2017