# 王启鹏


/ (+86)158-1155-3886

wangqipeng@stu.pku.edu.cn

https://qipengwang.github.io


## 研究兴趣

机器学习系统

- 边缘智能：联邦学习，边缘深度学习部署等
- 深度学习系统：训练及推理系统优化

## 教育背景

**北京大学** **博士研究生** 计算机软件与理论 **2020.9 ~ 现在**

- **导师：** 刘譞哲

**北京航空航天大学** **学士** **2016.9 ~ 2020.6**

- 软件工程（主修） **2016.9 ~ 2020.6**
- 数学（辅修） **2017.9 ~ 2020.6**

## 校内科研项目

**终端训练内存优化** **MobiSys'2022 (CSRanking)**

**唯一获得全部代码评估（Artifact Evaluation）徽章的论文**

- **终端训练内存优化框架**。根据张量生命周期等信息设计了内存管理优化技术，减少了内存碎片；根据给定的内存预算，利用内存池和重计算技术节约训练时内存。至多提升了 6.5× 批大小，同等批大小训练下提升了 4.1× 吞吐，节约 49% 能耗。
- **快速训练上下文切换**。在内存预算变化时快速训练策略切换，避免了至少 45.5% 训练结果的浪费。

## 校内工程项目

**基于 K8S 的深度学习任务部署平台 (Demo)**

- **平台开发和部署**。该平台支持包括 Nvidia GPU 和 Ascend NPU 在内的多种硬件上运行深度学习。平台的特色包括：(1) 自动完成算力用量评估；(2) 自动完成运行环境配置；(3) 自动完成部署。目前平台在北京大学计算中心完成部署，50 余名用户参与测试。
- **集成先进的任务调度算法**。该平台集成了 ElasticFlow (ASPLOS'2023) 调度器，实现资源高效管理、优化任务完成时间。

## 实习经历

**微软亚洲研究院 | 异构计算组** 全职实习生 导师：姜世琦 **2022.4 ~ 2023.7**

- **荣誉：明日之星 *Star of Tomorrow***
- **科研项目：浏览器中深度学习推理性能度量与优化** TOSEM'2024
  - **分析浏览器中推理与本地推理之间的性能差异及根本原因**。从模型和算子层面分析了浏览器中推理的性能，指出了其与本地推理之间性能差异的根本原因，包括不匹配/缺失的 SIMD 指令、线程管理与通信的开销等。
  - **首次提出了浏览器中推理场景下的 QoE**。提出的 QoE 指标包括网页响应性、网页流畅度和推理准确率三个指标，并定量分析了浏览器中推理带来的 QoE 下降以及原因。
  - **基于 v8 和 Eigen 搭建了浏览器中算子执行引擎**。基于 v8 和 Eigen 在 v8 的 js 解释层加入了矩阵乘法执行引擎，可在 JavaScript 层面调用，并实现了与本地推理相同的速度。
- **科研项目：终端扩散模型推理优化** 在投
  - **无训练中心点学习和策略生成**。无需端到端训练即可获得查表的中心点，并得到不同的加速下的推理策略。比现有的工作在中心点学习的效率上提升了 3000 倍以上。
  - **并行推理查表引擎**。基于 ARM/X86 CPU、OpenCL 中的 *shuffle*、*tbl*、*bitwise select* 等指令实现了高效的并行查表，加速内存访问。

**腾讯 | TEG| 训练加速组** 全职实习生 **2022.6.17 ~ 现在**

- **工程项目: GPU 训练显存优化** 进行中
  - **TensorFlow BFC 分配器优化**。利用模型训练具有周期性的特点，通过 warmup 阶段的信息指导后续迭代的 GPU 显存分配。目前支持按层堆叠的模型，比如能够减少 ResNet50 20.3% 训练显存。

## 论文

- [MobiSys'2022] **Qipeng Wang**, Mengwei Xu, Chao Jin, Xinran Dong, Jinliang Yuan, Xin Jin, Gang Huang, Yunxin Liu, Xuanzhe Liu. Melon: Breaking the memory wall for resource-efficient on-device machine learning. **[CCF-B; CSRanking]**
- [TOSEM'2024] **Qipeng Wang**, Shiqi Jiang, Zhenpeng Chen, Xu Cao, Yuanchun Li, Aoyu Li, Ying Zhang, Yun Ma, Ting Cao, Xuanzhe Liu. Anatomizing Deep Learning Inference in Web Browsers. **[CCF-A]**
- [在投] **Qipeng Wang**, et al. Lookup Table Optimized Efficient On-Device Diffusion Model Inference. **[CCF-A; CSRanking]**
- [WWW'2021] Chengxu Yang, **Qipeng Wang**, Mengwei Xu, Zhenpeng Chen, Kaigui Bian, Yunxin Liu, Xuanzhe Liu. Characterizing Impacts of Heterogeneity in Federated Learning upon Large-Scale Smartphone Data. **[CCF-A; CSRanking]**
- [TMC' 2022] Chengxu Yang, Mengwei Xu, **Qipeng Wang**, Zhenpeng Chen, Kang Huang, Yun Ma, Kaigui Bian, Gang Huang, Yunxin Liu, Xin Jin, Xuanzhe Liu. FLASH: Heterogeneity-aware Federated Learning at Scale. **[CCF-A]**
- [EMDL'2021] Dongqi Cai, **Qipeng Wang**, Yuanqiang Liu, Yunxin Liu, Shangguang Wang, Mengwei Xu. Towards Ubiquitous Learning: A First Measurement of On-Device Training Performance. **[EMDL@MobiSys]**
- [MobiCom'2022] Daliang Xu, Mengwei Xu, **Qipeng Wang**, Shangguang Wang, Yun Ma, Kang Huang, Gang Huang, Xin Jin, Xuanzhe Liu. Mandheling: Mixed-Precision On-Device DNN Training with DSP Offloading. **[CCF-A; CSRanking]**
- [arXiv:2401.08092] Mengwei Xu, Wangsong Yin, Dongqi Cai, Rongjie Yi, Daliang Xu, **Qipeng Wang**, et al. A Survey of Resource-efficient LLM and Multimodal Foundation Models.
- [MobiSys'2024] Fucheng Jia, Shiqi Jiang, Ting Cao, Wei Cui, Tianrui Xia, Xu Cao, Yuanchun Li, **Qipeng Wang**, et al. Empowering In-Browser Deep Learning Inference on Edge Through Just-In-Time Kernel Optimization.

## 学生工作

| | | |
|---|---|---|
| **课程助教** | 北京大学计算概论 A 本科生 120 人 | 2020，2021 |
| **课程助教（负责人/leader）** | 北京大学计算概论 A 本科生 120 人 | 2022，2023 |