杨程旭

手机：(+86) 18811586058 · 邮箱：yangchengxu@pku.edu.cn

教育背景

北京大学，信息科学技术学院，理学学士学位 2015.09 - 2019.06

北京大学，计算机学院，计算机科学与理论，在读博士（预计 2024 年毕业） 2019.09 - 2024.06

导师：刘譞哲。**研究方向：**机器学习系统，数据处理合规性保证。

奖励/奖学金

华为奖学金，华为，2023
优秀科研奖，北京大学，2022
国家奖学金，中国教育部，2021
三好学生标兵，北京大学，2021
校长奖学金，北京大学，2021
三好学生，北京大学，2020&2018
王胜地奖学金，北京大学，2018
TP-Link 奖学金，北京大学，2016

科研项目 - 数据处理合规性保证

设备异构性可知的联邦学习平台 [WWW'21][TMC'22]

- **用于描述设备间异构性的十万量级数据集。**通过商用输入法，收集了世界各地 13 万用户移动设备的设备间异构性信息，为联邦学习系统的仿真提供真实数据支持。
- **首个异构性可知的联邦学习仿真平台。**基于上述数据集和 Google 公司公开的联邦学习系统和协议，设计并实现了首个异构性可知的联邦学习平台，被斯坦福、密歇根、北邮等高校的研究团队使用。

面向 Spark 的细粒度隐私审计框架 [FSE'21] – 微软亚洲研究院实习期间完成

- **基于静态重写和动态翻译的污点分析方法。**可以在不修改现有大数据分析引擎（如 spark）的前提下，实现灵活且细粒度的数据流追踪。在真实商用数据处理脚本上达到了 93% precision 和 100% recall。
- **数据处理合规性检查框架。**框架支持灵活的污点标签定义以支持不同的隐私审计任务。在真实的隐私审计任务上，相比现有实现减少了 80% 的代码审查开销。

基于系统级跟踪的黑盒程序的控制流恢复工具 [TOSEM'23]

- **基于 LD_PRELOAD 的自动化系统级跟踪收集。**自动化的识别应用使用的库函数，并为每一个库函数生成 detour function 用于收集库函数的调用、返回信息，用于指导控制流恢复。
- **基于系统级跟踪的黑盒程序控制流恢复方法。**基于前一个工具收集的信息，设计并实现了一套黑盒程序控制流恢复方法。该方法不依赖插桩（低运行开销），不依赖于硬件特性，应用开发者和系统也无需做任何额外支持（低侵入性），恢复路径的准确率显著高于传统方法。

工程项目 - 机器学习系统

基于 Kubernetes 的 MLOps 平台

- **深度学习任务部署平台。**设计了平台的架构，并带领了 7 人团队完成了平台开发和部署。该平台自动完成算力用量评估、运行环境配置、调度、部署等流程，简化了深度学习任务的部署流程。目前平台在北京大学计算中心内部部署测试。(平台 Demo)
- **弹性流式训练调度算法的落地实现。**平台集成了学术界领先的弹性流式训练调度算法，后续还将为更多实验室技术提供落地验证平台。

大规模分布式训练 Profiler – 字节跳动 AML 组实习期间完成

- **PyTorch Profiler 功能优化。**大模型训练对 Profiler 提出了更高的要求。优化了 PyTorch Profiler 在 profile 计算、通信、内存时的功能，提供了更多的信息，提升了可扩展性。其中一些功能已向官方提交 pull request，正在审核中。

深度学习编译器 – 字节跳动 AML 组实习期间完成

- **基于 mlir 的计算图优化。** Profile GPT-2 训练过程，分析计算图，寻找性能瓶颈，通过算子融合等方式（主要为消除不必要的 transpose），优化计算图，最终每个 step 训练时间缩短为原本的 36%。
- **编译加速。**通过 profile 算子编译过程，寻找瓶颈，通过复用、cache、提高并行度等方式显著提高了 E2E 的编译时间（9-122 倍加速，平均约 40 倍加速）

实习经历

微软亚洲研究院，异构计算组，实习生 2020.06 - 2021.03
面向 Spark 的细粒度隐私审计框架

字节跳动，AML 组，实习生 2023.04 - Now
大规模分布式训练 profiler，深度学习编译器

学生工作

防疫委员，北京大学计算机学院 2022.5 - 2023.3
常任班长，北京大学计算机学院软件 1 班（230 名研究生） 2020.12 - Now
团支书，北京大学计算机学院软件 1 班团支部 2020.10 - 2021.10
带班辅导员，北京大学信息科学技术学院 19 级 5 班（52 名本科生） 2019.09 - 2023.06
课程助教，北京大学信息科学技术学院计算机导论课 2017, 2018
课程助教，北京大学信息科学技术学院计算概论课 2019, 2021
课程助教，北京大学计算机学院分布式机器学习：理论与系统课 2020

发表论文

[TOSEM'23] Xuanzhe Liu, **Chengxu Yang**, Ding Li, Yuhan Zhou, Shaofei Li, Jiali Chen, Zhenpeng Chen. "*Adonis: Control Flow Recovery through OS-Level Traces*" **[CCF A]** (导师一作，学生二作)

[TMC'22] **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.*" IEEE Transactions on Mobile Computing, 2022, 10.1109/ TMC.2022.3214234, 18 pages. **[CCF A]**

[FSE'21] **Chengxu Yang**, Yuanchun Li, Mengwei Xu, Zhenpeng Chen, Yunxin Liu, Gang Huang, Xuanzhe Liu. "*TaintStream: Fine-grained Taint Tracking for Big Data Platforms through Dynamic Code Translation.*" Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering. 2021. **[CCF A]**

[WWW'21] **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.*" Proceedings of the Web Conference 2021. **[CCF A]**