余锋伟

forwil@foxmail.com · (+86) 18810 676 076 · forwil.xyz · **工作地点：深圳**

TL;DR

行业背景：曾在多个 AI 领军企业的核心部门担任技术管理岗位，对 AI + System 领域具有深刻和独特的见解，具有从 0 到 1 搭建模型训练框架、压缩框架、部署框架、计算调度等深度学习基础设施的成功案例，坚持以业务价值为导向，对算法规模化落地的各类痛点有切身体会和实战经验。
管理能力：具备在高科技上市公司组建和管理 60+ 人研发团队的经验，曾与多个 985 高校教授建立长期产研合作关系。
学术背景：NOIp 两年一等奖，ASC15 世界一等奖，保送北航计算机本科 + 研究生，综合排名 top5%。于 ICCV / ECCV / CVPR / ICLR / Neurip / ICPP / MLSys / IPDPS /T-PAMI 等 CCF-A/B 会议/期刊上发表 20 多篇论文，Google Scholar 引用量 3500+，H-index 18。

教育背景

北京航空航天大学, *计算机学院, 软件工程 (双一流 A+)*, 硕士研究生 2015.9 – 2018.3
学位课程平均分：90.0/100，排名：6/241，北京市优秀毕业生，国家奖学金。

北京航空航天大学, *计算机学院, 创新实验班（双一流 A+）*, 本科生 2012.9 – 2015.6
核心课程平均分：88/100，综合排名：7/228，北航优秀毕业生，获研究生推免资格。

北京航空航天大学, *数学与系统科学学院, 华罗庚实验班*, 本科生 2011.8 – 2012.6
NOIp 保送入学，大二转系进入计算机学院，加入创新实验班。

工作经历

Momenta 初速度科技, DDInfra（数据驱动基础设施组，团队人数 15+） 2023.1–至今
研发高级总监 深圳

- 整体负责模型训练和推理优化。
- 训练效率优化：提升 D 模型在 A100 上的训练效率接近 100%，适配小显存平台 GTX3090; 提升 O 模型在 3090 上的训练效率达 400%+，D 模型在 3090 上训练效率提升 30%，端到端加速 3 倍以上。
- 模型量化：自研模型 PTQ 框架，推动感知模型支持 INT8 PTQ 交付，生产效率从 N 天缩短到 N 小时，机器需求从 100GPU 缩短到 1GPU，精度满足发版需求，同时支持多个 NPU 硬件，包括 orin/高通/瑞萨/mdc/自研芯片。
- 编译推理加速：使用 TVM 优化提升多个核心模型在 orin/xaiver 上的推理效率，提速 10 到 100%。
- Transformer 算子优化：使用 cutlass/FlashAttention/tvm 等多种手段，优化 Transformer 在 orin 上的推理效率，平均提速 30%。

SenseTime 商汤科技, 研究院, 模型工具链（二级部门负责人，团队 60+） 2020.9–2022.12
研究副总监 北京/深圳

- 团队工作介绍：https://zhuanlan.zhihu.com/p/268154983，开源项目组：https://github.com/ModelTC
- **整体负责公司内部研发平台 - SenseCore 模型工厂**，推动模型生产效率提升，2020：10000+，2021：21000+，2022 预计 40000+，年复合增长率 100%。模型工厂包含：
  - **任务调度**：集群任务调度 SpringScheduler，覆盖 7000+ 个 GPU，12+ 个集群；
  - **训练引擎**：研发 linklink 训练引擎，支持公司 10 亿 ~ 300 亿参数规模的视觉大模型高效训练；
  - **算法框架**：研发联合感知模型生产框架 UP，覆盖检测、属性、3D、分割、跟踪等算法的高效训练；

  - **模型压缩**：负责模型通用压缩技术体系，包含在线/离线量化，模型稀疏，基于硬件真实速度的的网络结构设计平台；
  - **模型部署**：负责多平台模型部署评测系统 Adela，支持 200 多类不同硬件平台自动部署和评测，其中一半以上为国产化硬件。
- 模型工厂的研发体系覆盖大部分公司业务，包括 90% 以上智慧城市 toG/toB 业务，70% 以上智慧汽车业务等。
- **ICCV-LPCV 2021 低功耗计算机视觉，FPGA 赛道冠军**。
- **团队获得 2021 院长创新奖第一名、小荷尖奖，Qdrop 获得 2022 年最佳论文奖**。
- **同时参与 3 个公司级别的商汤团队奖：通用模型，上市项目，国产化芯片适配**。
- **个人获得 2021 年商汤奖提名**。
- **NART、POD 获得 2021 年度最受欢迎开源项目一等奖、二等奖**。

SenseTime 商汤科技, 研究院, 工具链, 链接与编译（部门负责人，团队 15+） 2016.12–2020.9
研究经理 北京

- 算法部署框架、模型量化工具、深度学习编译器、智能端边 SDK
- **获得研究院 2018 年度杰出员工称号**
- **SCG/研究院开源技术中台团队获 2019 年商汤团队奖（全公司共两个）**
- **前端相机团队获得商汤 2017 年度优秀团队**
- **获得商汤 2017 年度未来之星称号**。

实习/项目经历

基于深度学习的中文文本查错, 北京航空航天大学, 软件所 2016.12–2017.7
毕业设计

- 使用基于 LSTM 的语言模型进行中文文本查错。

SenseTime 商汤科技, 研究中心, 检测跟踪组 2016.3–2016.12
见习研究员

- 将动态人脸检测跟踪识别系统从单卡 4 路优化至 16 路，至 TX1 实时处理，再到 CPU 实时处理，最后移植到了嵌入式 IPC（HI3519）中。
- 使用行人检测和 ReID 特征优化了多目标跟踪系统，在 MOT16 榜单上取得包括 MOTA 指标（68.2 和 66.1）在内的多项第一。
- **SenseFace 动态人脸布控系统获 2016 年安防展优秀奖**

ASC15 世界大学生超级计算机竞赛 2014.12–2015.5
北航代表队队长

- 在初赛中：负责将 4 台浪潮服务器组成超算小集群的软硬件搭建和维护，对 HPCC 的多个测试子项目（包括 Linpack、FFT、DGEMM）进行深入分析和编译优化，撰写英文 proposal，队伍以初赛大陆第一，世界第二进入全球总决赛。
- 在总决赛中：负责集群软硬件平台搭建、功耗控制、HPL、HPCG 调优、WRF-CHEM。应用优化和集群运行策略调度，最终队伍以全球第五名获得一等奖。

Microsoft ARD 微软亚太研发集团，CEC , IoT Group 2014.7–2014.12
研发实习生

- 在智能插座项目中，为 STM32F 上的.Net Micro Framework 固件添加高级 ADC 操作。
- 在基于低功耗蓝牙的室内定位项目中，设计并实现在 51MCU 上的 RS-485 总线多对一通信协议。

- 在自动化测试项目中，提取测试程序调用外部库的依赖关系，存入数据库并对外提供 WCF 接口。

♡ 在校获奖情况

研究生国家奖学金	2017 年
华为奖学金	2016 年
硕士研究生学业奖学金, 一等奖	2015、2016 年
ASC15 世界大学生超级计算机竞赛, 一等奖, 第五名	2015 年
蓝桥杯全国软件大赛, 全国二等奖	2014 年
高教社杯全国大学生数学建模竞赛, 全国二等奖	2013 年
第十一届“福建省小科学家”称号	2011 年
全国信息学奥林匹克联赛 *(NOIp)*, 一等奖, 分数：310/400	2010 年
全国信息学奥林匹克联赛 *(NOIp)*, 一等奖, 第七名, 分数：325/400	2009 年

i 论文与专利

Exploiting Subgraph Similarities for Efficient Auto-tuning of Tensor Programs.(ICPP2023)
*Mingzhen Li, Hailong Yang, Shanjun Zhang, **Fengwei Yu**, Ruihao Gong, Yi Liu, Zhongzhi Luan, Depei Qian*
SysNoise: Exploring and Benchmarking Training-Deployment System Inconsistency.(MlSys2023)
*Yan Wang, Yuhang Li, Ruihao Gong, Aishan Liu, Yanfei Wang, Jian Hu, Yongqiang Yao, Yunchen Zhang, Tianzi Xiao,**Fengwei Yu**, Xianglong Liu*
Exploiting Input Tensor Dynamics in Activation Checkpointing for Efficient Training on GPU.(IPDPS2023)
*Jianjin Liao, Mingzhen Li, Hailong Yang, Qingxiao Sun, Biao Sun, Jiwei Hao, Tianyu Feng, **Fengwei Yu**, Shengdong Chen,Ye Tao, Zicheng Zhang, Zhongzhi Luan, Depei Qian*
NNLQP: A Multi-Platform Neural Network Latency Query and Prediction System with An Evolving Database.(ICPP2022)
*Liang Liu, Mingzhu Shen, Ruihao Gong, **Fengwei Yu** , Hailong Yang*
QDrop: Randomly Dropping Quantization for Extremely Low-bit Post-Training Quantization .(ICLR2022)
*Xiuying Wei, Ruihao Gong, Yuhang Li, Xianglong Liu, **Fengwei Yu***
Supervision Exists Everywhere: A Data Efficient Contrastive Language-Image Pre-training Paradigm.(ICLR2022)
*Yangguang Li, Feng Liang, Lichen Zhao, Yufeng Cui, Wanli Ouyang, Jing Shao, **Fengwei Yu**, Junjie Yan*
Real World Robustness from Systematic Noise.(ADMM2021)
*Yan Wang, Yuhang Li, Ruihao Gong, Tianzi Xiao, **Fengwei Yu***
MQBench: Towards Reproducible and Deployable Model Quantization Benchmark.(NeurIPS2021)
*Yuhang Li, Mingzhu Shen, Jian Ma, Yan Ren, Mingxin Zhao, Qi Zhang, Ruihao Gong,**Fengwei Yu**, Junjie Yan*
Incorporating Convolution Designs into Visual Transformers.(ICCV2021)
*Kun Yuan, Shaopeng Guo, Ziwei Liu, Aojun Zhou, **Fengwei Yu**, Wei Wu*
Differentiable Dynamic Wirings for Neural Networks.(ICCV2021)
*Kun Yuan, Quanquan Li, Shaopeng Guo, Dapeng Chen, Aojun Zhou, **Fengwei Yu**, Ziwei Liu*
MixMix: All You Need for Data-Free Compression Are Feature and Data Mixing.(ICCV2021)
*Yuhang Li, Feng Zhu, Ruihao Gong, Mingzhu Shen, Xin Dong, Shaoqing Lu, **Fengwei Yu**, Shi Gu*
Towards High Performance Extremely Low-bit Neural Networks.(ICCV2021)
*Mingzhu Shen, Feng Liang, Ruihao Gong, Yuhang Li, Chuming Li, Chen Lin, **Fengwei Yu**, Junjie Yan, Wanli Ouyang*
Diversifying Sample Generation for Accurate Data-Free Quantization.(CVPR2021 oral)
*Xiangguo Zhang, Haotong Qin, Yifu Ding , Ruihao Gong, Qinghua Yan, Renshuai Tao , Yuhang Li, **Fengwei Yu**, Xianglong Liu*
BRECQ: Pushing the Limit of Post-Training Quantization by Block Reconstruction.(ICLR2021)
*Yuhang Li, Ruihao Gong, Xu Tan, Yang Yang, Peng Hu, Qi Zhang, **Fengwei Yu**, Wei Wang, Shi Gu*
Extremely Low-bit Convolution Optimization for Quantized Neural Network on Modern Computer Architectures.(ICPP2020 oral)
*Qingchang Han, Yongmin Hu, **Fengwei Yu**, Hailong Yang, Bing Liu, Peng Hu, Ruihao Gong, Yanfei Wang, Rui Wang, Zhongzhi Luan, Depei Qian*
DMS: Differentiable Dimension Search for Binary Neural Networks.(ICLR2020 NAS workshop)
*Yuhang Li, Ruihao Gong, **Fengwei Yu**, Xin Dong, Xianglong Liu*
Towards Unified INT8 Training for Convolutional Neural Network.(CVPR2020)

*Feng Zhu, Ruihao Gong, **Fengwei Yu**, Xianglong Liu, Yanfei Wang, Zhelong Li, Xiuqi Yang, Junjie Yan*
Forward and Backward Information Retention for Accurate Binary Neural Networks.(CVPR2020)
*Haotong Qin, Ruihao Gong, Xianglong Liu, Mingzhu Shen, Ziran Wei, **Fengwei Yu**, Jingkuan Song*
Differentiable Soft Quantization: Bridging Full-Precision and Low-Bit Neural Networks.(ICCV2019)
*Ruihao Gong, Xianglong Liu, Shenghu Jiang, Tianxiang Li, Peng Hu, Jiazhen Lin, **Fengwei Yu**, Junjie Yan.*
POI: Multiple Object Tracking with High Performance Detection and Appearance Feature.(ECCV2016 workshop)
***Fengwei Yu**, Wenbo Li, Quanquan Li, Yu Liu, Xiaohua Shi, Junjie Yan*