余锋伟

forwil@foxmail.com · (+86) 18810 676 076 · forwil.xyz

TL;DR

NOIP 连续两年一等奖，ASC15 世界一等奖，保送北航计算机本科 + 硕士。于 ICCV / ECCV / CVPR / Neurip / ICPP 等系统/AI 会议上发表 20 多篇论文，引用量 1300+。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经验，对算法落地与大规模支持业务有实战经验。具备组建和管理 50+ 人研发团队的经验，与多个 985 高校教授建立长期产研合作关系。

教育背景

北京航空航天大学, 计算机学院, 软件工程, 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, 计算机学院创新实验班, 本科生 2012.9 – 2015.6

核心课程平均分：88/100，本科综合排名 7/228，北航优秀毕业生，获得研究生推免资格。

北京航空航天大学, 数学与系统科学学院, 华罗庚数学实验班, 本科生 2011.8 – 2012.6

NOIp 保送入学，大一结束后转系进入计算机学院。

工作经历

SenseTime 商汤科技, 研究院, 模型工具链（二级部门负责人，团队人数约 40 +）2020.9–至今

研究副总监

- 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 亿参数规模的大模型训练；
  - **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。
- 公司大部分业务都依赖于模型工厂的研发体系，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**。
- **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**。

SenseTime 商汤科技, 研究院, 工具链, 链接与编译（部门负责人，团队 15+） 2019.4–2020.9

研究经理

- 算法部署框架、模型量化工具、训练加速框架、深度学习编译器、智能端边 SDK
- **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）**

SenseTime 商汤科技, 研究院, 智能视频, 基础技术与工具组 2018.4–2019.4
研究员

- gpu，arm、tx1、movidius 多个架构的前端视频分析算法落地及优化、自研 movidius 深度学习加速库，汇编级优化、自研 arm-quant 定点化加速库，CNN 部署框架 NART，模型量化框架 DRCL。
- **获得研究院 2018 年度杰出员工称号**

SenseTime 商汤科技, 研究院, 智能视频, 基础技术与工具组 2016.12–2018.4
全职见习研究员

带领小团队，负责智能视频组安防相关所有算法落地和升级，基础工具链搭建。

- 搭建 SenseVideo-GPU-SDK 高性能视频结构化分析系统、推动包括人脸跟踪识别服务器阵列（TX1）、前端人脸抓拍相机 (海思 3519)、前端人脸识别芯片（Movidius 芯片）等前端产品落地。
- **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

- 背景：传统基于分词和规则错词表的中文查错系统已经被用到实际的中文查错系统中去，但是其依赖于错词表
- 数据：收集了超过 10 亿字的语料。
- 模型和策略：设计了基于 word-embedding、char level、2-stack-LSTM、dropout 的语言模型，并使用双向模型、拼音相似性等策略来提升识别精度。
- 成果：系统在 SIGHAN2013 上取得了很好的错误检测精度 (F1：0.69)。

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

- SenseFace-GPU/CPU-SDK 高性能动态人脸检测跟踪识别系统，维护并优化 GPU 版本 SDK，使其处理速度从单卡 4 路到单卡 16 路。移植到 TX1 嵌入式平台，优化使其支持 2 路高清视频分析。从零编写并维护 CPU 版本，可单核处理一路实时视频流。
- **SenseFace 动态人脸布控系统获 2016 年安防展优秀奖**
- 将人脸检测跟踪算法移植到性能受限的网络监控相机中，并优化至产品可用速度 (12fps)。
- 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。

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。

♡ 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

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*

专利：目标跟踪方法、系统及电子设备（CN201710124025.6） 2017 年
专利：目标对象的检测方法、装置和电子设备（CN201710059806.1） 2017 年
专利：目标对象识别方法、装置、存储介质和电子设备 (CN2017111812995) 2017 年

学生工作经历

编译原理/形式语言与自动机, 本科课/研究生课, 助教 2016.9—2017.1

- 负责批改作业、小测验、实验课习题课讲解

北航高等工程学院高等代数, 助教 2014.9—2015.1

- 负责批改作业、讲授习题课

北航计算机学院创新实验班, 班长 2013.9—2015.7

- 负责通知学生各类事宜，组织班会、聚餐等班级活动

全国信息学奥林匹克联赛（*NOIp*）, 北京赛区, 监考员 2013.11

- 监考普及组/提高组，负责解决考生遇到的编译/调试等问题