李腾

电子与计算机工程博士候选人
伊利诺伊大学芝加哥分校 · 美国 芝加哥
+86 18639320104 | ninmiao@gmail.com | tli81@uic.edu | linkedin.com/in/tengli95

个人网站
与完整简历

LinkedIn

个人总结

电子与计算机工程博士候选人，研究**边缘人工智能**与**分布式学习与推理**，面向通信高效、具有韧性的机器学习系统。研究重点为无线与去中心化网络的**优化**与资源调度。使用 **Python/PyTorch** 在 **NVIDIA Jetson** 上部署**视觉模型和大语言模型**，开发 **RF 信号检测**系统，并使用 **ONNX Runtime / TensorRT** 优化 GPU 推理延迟与吞吐量。无线系统实践涵盖 **O-RAN**、**OpenAirInterface (OAI)**及 **SDR (USRP)**。

教育背景

伊利诺伊大学芝加哥分校（University of Illinois Chicago） 美国 芝加哥
电子与计算机工程博士候选人 · GPA: 4.0/4.0 2019年9月 - **预计2026年12月毕业**
导师：Prof. Hulya Seferoglu
研究兴趣：分布式学习、边缘人工智能、网络优化、韧性系统、大语言模型端侧部署

电子与计算机工程硕士 · GPA: 4.0/4.0 2017年8月 - 2019年4月
专业方向：计算机通信网络、深度学习、无线通信、计算机体系结构

东北大学 中国 沈阳
工学学士，自动化专业 · GPA: 80.8/100 2013年8月 - 2017年5月
专业方向：自动控制系统、智能优化、嵌入式系统

研究与项目

伊利诺伊大学芝加哥分校 · 网络研究实验室

Priority-Aware Model-Distributed Inference at Edge Networks
第一作者 · IEEE ICMLCN 2025

面向多数据源、不同任务优先级的边缘网络，设计 PA-MDI 算法，优化**分布式推理**的模型分区与优先级感知分配，并将任务失败概率纳入分配，提升节点移动与动态网络下的推理韧性。在 **NVIDIA Jetson 上验证 ResNet**，在 **Colosseum 上评估 GPT-2**，相比基线方法**缩短推理时间**。

REDIT: Resilient Distributed Text-to-Speech at Edge Networks
第一作者 · IEEE Globecom 2022

设计并实现分布式**文本转语音（TTS）**框架 REDIT，将文本摘要作为冗余，提高资源受限边缘设备对**慢节点的容错能力**。在 **NVIDIA Jetson** 上完成评估，相比无容错机制的分布式 TTS **缩短任务完成时间**。

Communication-Aware Model-Distributed Inference via LRC 共同作者 · ACM MobiHoc 2026（已录用）

我们通过 Top-k 稀疏化、量化和 LLM.int8 进行动态激活压缩，降低模型分布式推理的通信开销。本人为项目贡献了**激活压缩器**的探索与实现，以及 **T5（Flan-T5）和 ResNet** 在 **NVIDIA Jetson** 上的推理实现、测试与多数据集评估。我们在实验中达到未压缩推理无法满足的**吞吐量目标**，并相比基线压缩策略**提高准确率**；我们的 Jetson 测试在相近准确率下**缩短推理时延**。

YOLO-Based CBRS Detection and Distributed Edge Inference (under submission)
第一作者 · 投稿中

构建**基于 YOLO 的目标检测流程**，从 CBRS 频谱图中识别并定位**雷达和 LTE 信号**；改进检测器 CEDAR-YOLO，减少模型大小与激活传输量。实现 CEDAR-MDI，通过模型分区、节点分配和激活压缩，在多节点 **NVIDIA Jetson** 上加速分布式推理，取得**高于本地执行的吞吐量**，同时保持雷达与 LTE 信号重叠时的检测质量。另基于原项目模型探索本地 NVIDIA GPU 部署，搭建 **ONNX Runtime / TensorRT** 工作流，对比多个模型配置的 FP32/FP16、经校准的 INT8 推理与 PyTorch 基线，评估**延迟、吞吐量和检测准确率**。

Open6G+AI Hackathon: 5G Multi-Service Scheduler 美国东北大学 · 2026年3月

- 在六人团队中使用 Lua 共同设计并实现**5G MAC 调度器**，在物理 **O-RAN** 测试床兼顾**eMBB 高吞吐量**与**URLLC UE 低时延、低丢包率**。

技能专长

编程语言	Python, C/C++, MATLAB, Java, Bash
机器学习与部署	PyTorch, TensorFlow, Hugging Face Transformers, ONNX / ONNX Runtime, TensorRT
平台与工具	NVIDIA Jetson, Linux, Docker, Git, LaTeX, socket、HTTP、Wi-Fi，O-RAN、OAI、SDR (USRP)
语言能力	中文 (母语), 英语 (流利)

学术服务

审稿人：IEEE MILCOM、IEEE Globecom、IEEE Transactions on Mobile Computing