曾文轩

zwx.andy@stu.pku.edu.cn | 15388161786 | https://wenxuanzeng.netlify.app

北京市 | 博客: xuanland.blog.csdn.net | 谷歌学术

教育经历

北京大学 软件与微电子 硕士 2023 年 9 月–至今

人工智能研究院 学生研究员、软件与微电子学院 学术科创部副部长

电子科技大学 信息与软件工程 本科 2019 年 9 月–2023 年 6 月

学生科创竞赛工作室主席、机器学习组组长

研究兴趣

(所列参考文献指我所参与的相关工作)

- **大模型优化算法**：探索 LLM 的稀疏注意力以及高效思维链推理
  - 用于长上下文 LLM 推理的稀疏注意力和 KV cache 压缩：研究高效的 KV cache 优化，包括重计算、低秩分解、分块稀疏注意力、静态和动态 KV cache 剪枝 [4][2][1]
  - 高效思维链（CoT）推理：研究当前高效的 CoT 压缩算法，包括潜在空间 CoT、提示词引导的 CoT 压缩、基于训练的 CoT 压缩
- **高效模型压缩/加速**：探索如何针对不同模型（包括 CNN、ViT 和 LLM）优化高效模型压缩算法
  - 模型架构设计和神经架构搜索：设计并搜索高效注意力机制 [11]、自动剪枝 ReLU 和 GeLU [10][3][11]
  - 知识蒸馏：采用基于输出的蒸馏和基于特征的蒸馏来提高小模型的性能 [10][8][11][6]
  - 低精度量化：将激活和权重压缩到低精度，以实现高效推理，并探索不同层的最优精度分配 [8][6]
  - 高效隐私保护深度学习：研究如何将 CNN、ViT 和 LLM 部署到隐私推理场景，并协同优化高效的 AI 算法和推理协议，目的是在提高系统效率和性能的同时，保护用户私有数据和服务商专有模型参数
- **多模态大模型（MLLM）**：近期主要研究如何赋予 MLLM 人类的思考模式，包括多模态思维链推理（MCoT），关注 MLLM 的“慢思考”推理能力

精选论文

(* 表示贡献相同, † 表示通讯作者)

[1] UniCAIM: A Unified CAM/CIM Architecture with Static-Dynamic KV Cache Pruning for Efficient Long-Context LLM Inference
Weikai Xu*, **Wenxuan Zeng***, Qianqian Huang, Meng Li†, Ru Huang†
DAC 2025

[2] MPCache: MPC-Friendly KV Cache Eviction for Efficient Private Large Language Model Inference
Wenxuan Zeng, Ye Dong, Jinjin Zhou, Junming Ma, Jin Tan, Runsheng Wang, Meng Li†
Preprint 2025, arxiv.org/abs/2501.06807

[3] EQO: Exploring Ultra-Efficient Private Inference with Winograd-Based Protocol and Quantization Co-Optimization
Wenxuan Zeng, Tianshi Xu, Cheng Hong, Meng Li†, Runsheng Wang
Preprint 2024, https://arxiv.org/abs/2404.09404

[4] CoPriv: Network/Protocol Co-Optimization for Communication-Efficient Private Inference
Wenxuan Zeng, Meng Li†, Haichuan Yang, Wen-jie Lu, Runsheng Wang, Ru Huang
NeurIPS 2023

[5] MPCViT: Searching for Accurate and Efficient MPC-friendly Vision Transformer with Heterogeneous Attention
Wenxuan Zeng, Meng Li†, Wenjie Xiong, Tong Tong, Wen-jie Lu, Jin Tan, Runsheng Wang, Ru Huang
ICCV 2023

工作经历

人工智能研究院 北京大学 2022 年 6 月–至今

- 角色：学生研究员
- 导师：李萌教授、王润生教授
- 方向：高效人工智能算法、高效隐私推理（主要关注协议-算法的协同优化）
- 项目：高效注意力机制设计 [11]、ReLU 和 GeLU 剪枝 [10][11][3]、高效卷积算法 [8][10]、大模型 KV cache 压缩 [4][2][1], 低精度量化 [6][8]、隐私推理协议层面优化 [3][5][10][8][6]
- 企业合作：与蚂蚁集团合作，录制并发布“隐私保护机器学习”课程

知识工厂 复旦大学 2022 年 4 月–2022 年 7 月

- 角色：科研实习生
- 导师：肖仰华教授
- 方向：自然语言生成的事实性和忠实性、人工智能可解释性
- 项目：事实性纠错 [12]

四川省网络与数据安全重点实验室 电子科技大学 2021 年 6 月–2022 年 6 月

- 角色：科研实习生
- 导师：周帆教授
- 方向：图神经网络、街道级 IP 地理定位系统、多元时序建模
- 项目：基于图注意力神经网络的街道级 IP 地理定位系统 [13]

通信与信息安全实验室 北京大学 2021 年 8 月 - 2022 年 5 月

- 角色：科研实习生
- 导师：朱跃升教授
- 方向：细粒度图像识别和虹膜识别

荣誉成就

北京大学三等奖奖学金	2024
北京大学三好学生	2024
电子科技大学优秀毕业生和荣誉研究证书	2023
电子科技大学优秀学生奖学金	2020 - 2022
电子科技大学创新创业训练计划	2020 - 2022
“腾讯”特等奖学金（全校共 3 个名额）	2022
“世强”一等奖学金（全校共 5 个名额）	2021

竞赛获奖

北京大学人工智能研究院“AI 杯”羽毛球赛 2024
第二名

北京大学人工智能研究院年度研讨会暨科技日活动 2023
论文海报最受欢迎奖

中国高校计算机大赛 2023
全国特等奖（第一名）

“泛珠三角”全国大学生计算机作品竞赛 2022
全国三等奖

中国高校计算机大赛 2021
全国一等奖（第一名）

“中公杯”四川省大学生计算机作品竞赛 2021
省级特等奖

论文工作

(* 表示贡献相同, † 表示通讯作者)

完整论文列表详见谷歌学术（130+ 次引用）：请点击这里

[1] H^2EAL: Hybrid-Bonding Architecture with Hybrid Sparse Attention for Efficient Long-Context LLM Inference
Zizhuo Fu, Xiaotian Guo, **Wenxuan Zeng**, Shuzhang Zhong, Yadong Zhang, Peiyu Chen, Runsheng Wang, Le Ye, Meng Li†

[2] UniCAIM: A Unified CAM/CIM Architecture with Static-Dynamic KV Cache Pruning for Efficient Long-Context LLM Inference
Weikai Xu*, **Wenxuan Zeng***, Qianqian Huang, Meng Li†, Ru Huang
DAC 2025

[3] OptiPrime: Efficient Private Inference at ImageNet Scale
Jiangrui Yu, Ye Yu, Si Chen, **Wenxuan Zeng**, Junfeng Fan, Runsheng Wang, Ru Huang, Meng Li†
Work in Progress

[4] MPCache: MPC-Friendly KV Cache Eviction for Efficient Private Large Language Model Inference
Wenxuan Zeng, Ye Dong, Jinjin Zhou, Junming Ma, Jin Tan, Runsheng Wang, Meng Li†
Preprint 2025, arxiv.org/abs/2501.06807

[5] FlexHE: A Flexible Kernel Generation Framework for Homomorphic Encryption-Based Private Inference
Jiangrui Yu, **Wenxuan Zeng**, Tianshi Xu, Renze Chen, Yun (Eric) Liang, Runsheng Wang, Ru Huang, Meng Li†
ICCAD 2024

[6] PrivQuant: Communication-Efficient Private Inference with Quantized Network/Protocol Co-Optimization
Tianshi Xu, Shuzhang Zhong, **Wenxuan Zeng**, Meng Li†, Runsheng Wang, Ru Huang
ICCAD 2024

[7] BAT: Behavior-Aware Human-Like Trajectory Prediction for Autonomous Driving
Haicheng Liao, Zhenning Li, Huanming Shen, **Wenxuan Zeng**, Dongping Liao, Guofa Li, Shengbo Eben Li, Chengzhong Xu†
AAAI 2024

[8] EQO: Exploring Ultra-Efficient Private Inference with Winograd-Based Protocol and Quantization Co-Optimization
Wenxuan Zeng, Tianshi Xu, Cheng Hong, Meng Li†, Runsheng Wang
Preprint 2024, https://arxiv.org/abs/2404.09404

[9] Kuaiji: the First Chinese Accounting Large Language Model
Jiayuan Luo, Songhua Yang, Xiaoling Qiu, Panyu Chen, Yufei Nai, **Wenxuan Zeng**, Wentao Zhang†, Xinke Jiang
Preprint 2024, https://arxiv.org/abs/2402.13866

[10] CoPriv: Network/Protocol Co-Optimization for Communication-Efficient Private Inference
Wenxuan Zeng, Meng Li†, Haichuan Yang, Wen-jie Lu, Runsheng Wang, Ru Huang
NeurIPS 2023

[11] MPCViT: Searching for Accurate and Efficient MPC-friendly Vision Transformer with Heterogeneous Attention
Wenxuan Zeng, Meng Li†, Wenjie Xiong, Tong Tong, Wen-jie Lu, Jin Tan, Runsheng Wang, Ru Huang
ICCV 2023

[12] Factual Error Correction via Iterative Constrained Editing
Jiangjie Chen*, Rui Xu*, **Wenxuan Zeng**, Changzhi Sun†, Lei Li, Yanghua Xiao†
AAAI 2023

[13] Connecting the Hosts: Street-Level IP Geolocation with Graph Neural Networks
Zhiyuan Wang*, Fan Zhou*†, **Wenxuan Zeng***, Goce Trajcevski, Chunjing Xiao, Yong Wang, Kai Chen
KDD 2022

学术服务

- NeurIPS 2025 会议审稿人