导师介绍

张文涛

研究员、博士生导师

北京大学大数据科学研究中心

北京大学国际机器学习研究中心

联系方式

邮箱：wentao.zhang@pku.edu.cn

电话：13269585797

主页：https://zwt233.github.io/

个人介绍

张文涛，北京大学国际机器学习研究中心助理教授、研究员、博士生导师，曾任职于腾讯机器学习平台部、Apple AIML 和加拿大 Mila 人工智能实验室。研究兴趣为以数据为中心的机器学习、大模型数据准备和 AI4Science。近 5 年在机器学习（ICML/NeurIPS/ ICLR）、数据挖掘（SIGKDD/WWW）和数据管理（SIGMOD/VLDB/ICDE）等领域发表 CCF-A 类论文 60 （一作/通讯 40+）余篇，也担任多个国际顶会的 PC Member/Area Chair。他获得多个最佳论文奖（如 WWW’22-第一作者，APWeb’23-通讯作者，CIKM’24），领导或参与开源了多个机器学习系统。曾获 Apple Scholar、世界人工智能大会云帆奖、北京大学/北京市/中国人工智能学会优秀博士学位论文奖、未名青年学者、世界互联网大会领先科技成果奖、华为火花奖、中国电子学会科技进步一等奖等多项荣誉。

研究领域

➢ 以数据为中心的机器学习（Data-centric ML，DCML）

- 近些年来 AI 模型发展遇到了瓶颈，性能收益来源由模型→数据。本研究旨在优化 Data quality, quantity 和 efficiency，以较低成本高效获得大量高质量数据，如针对大模型（LLM）的数据获取、处理、质量评估和利用等。

➢ DCML 算法和应用

- **大模型和多模态大模型**：从数据层面优化强逻辑推理大模型（Math 和 Code）、大模型的后训练（SFT 和 RL）、RAG 知识库等
- **图学习**：图结构优化、图数据增强和图异常处理等
- **AI4Science**：以数据为中心，研究和设计高效的 Science 数据（如蛋白质和分子）构建和预处理方式，以及分子建模与生物制药等交叉应用。

➢ DCML 系统

- 针对大模型数据侧，开源 DCML 工具和系统，支持多种类型的数据格式，大规模数据的处理，为大模型准备大量高质量数据。