Frontier of Artificial Network

Department of Data Science
香港城市大學
City University of Hong Kong

大模型压缩与推理加速

特邀讲者
张淼
哈尔滨工业大学（深圳）

日期: 2026年6月11日(星期四)
时间: 10:30am (香港时间)
Zoom会议号: 801 137 0362

讲者信息

张淼，哈尔滨工业大学（深圳）教授，博士生导师，国家级青年人才，河套学院客座教授，计算机科学与技术学院院长助理。曾就职于丹麦奥尔堡大学计算机学院助理教授，澳大利亚莫纳什大学研究员，博士毕业于澳大利亚悉尼科技大学。主要从事端侧人工智能和高效机器学习方向，涉及领域包括大模型压缩与推理加速，高效模型架构设计、高效多智能体系统等。一作/通讯发表CCF A类会议或期刊论文40余篇，如IEEE TPAMI，ICML，NeurIPS，CVPR等。目前担任IEEE TCSVT Associate Editor，ACL AE，IJCAI SPC等，主持6项国家级和省部级项目和课题。在产业应用与成果转化方面，相关研究成果已在大模型推理优化与多智能体系统等方向与多家头部企业开展合作（如华为、美团、小米等）。

报告内容

随着大模型技术的快速发展，模型规模和计算开销持续增长，如何在保证模型能力的同时降低部署成本、提升推理效率，已成为大模型落地应用中的关键问题。大模型压缩与推理加速旨在通过量化、稀疏、高效注意力、KV Cache 优化等技术，减少模型参数量、显存占用和计算延迟，从而提升大模型在云端、边缘端和端侧设备上的部署可行性。本报告将围绕大模型压缩与推理加速的典型方法展开介绍，重点分析低比特量化、激活稀疏、Token 剪枝、KV Cache压缩等方向的基本思想与团队相关近期工作。