华中科技大学

研究生参加学术会议情况记载单

学　　号	M202273662	姓　名	金灿果
学术会议名称	华中科技大学计算机科学与微软亚洲研究院联合研讨会		
举办会议的时间及地点	2023 年 4 月 12 日 8：00-12：00 华中科技大学梧桐语问学中心问道厅		
学术报告题目	具有稀疏属性的张量实现端到端的深度学习模型稀疏化 报告人：杨凡		
内容概述 摘要	报告主题关注系统方向的研究规划，包括大型分布式系统所产生的新型计算机系统原理、设计与实现。随着深度学习模型规模变大且愈加复杂，深度学习模型的稀疏性作为提升模型效率和规模的关键因素，在端到端的模型稀疏化中起到重要作用。 研究提出一种新的系统抽象，具有稀疏属性的张量 TeSA，来实现端到端的模型梳化。TeSA 这一抽象扩展了传统的张量抽象，使得张量抽象的稀疏属性和稀疏模式能够在整个深度学习模型中传播。TeSA 可以用于创建高效、专门的模型算子实现，在实现中充分考虑到了各种稀疏模式在不同硬件上的执行效率。研究基于 TeSA 构建了 SparTA（端到端的支持模型稀疏化的编译框架），SparkTA 可以容纳各种稀疏模式和优化技术，在推理延迟方面比七种最先进的稀疏方案快 1.7 至 8.4 倍，同时内存占用更小。作为编译框架，其有助于利用最新的稀疏算法更快地探索更好的稀疏化深度学习模型。		
评 语	会议主持人签名：		

注：考核完后十天内，由会议主持人（或导师）将此表和会议论文一并送交学生所在院系研究生（教务）科，存入学生个人学籍档案。