華中科技大學

研究生（参加校内外公开学术活动）报告

题 目：**端到端深度学习模型稀疏化报告**

学　　号　**M202273662**

姓　　名　金灿果

专　　业　计算机科学与技术

指 导 教 师　左琼

院（系、所）　计算机学院

华中科技大学研究生院制

填 表 注 意 事 项

一、本表适用于攻读硕士学位研究生选题报告、学术报告，攻读专业硕士学位研究生实践环节报告，攻读博士学位研究生文献综述、选题报告、论文中期进展报告、学术报告等。

二、以上各报告内容及要求由相关院（系、所）做具体要求。

三、以上各报告均须存入研究生个人学籍档案。

四、本表填写要求文句通顺、内容明确、字迹工整。

报告内容

2023 年 4 月 12 日，微软亚洲研究院高级研究院杨凡博士在华中科技大学计算机学院与微软亚洲研究院联合研讨会上，就深度学习模型稀疏化作报告。

随着信息技术的迅猛发展，大型分布式系统在现代计算环境中扮演着愈发关键的角色。与此同时，深度学习领域的快速演进使得模型规模不断扩大且复杂度日益攀升。在这样的背景下，研究深入探索大型分布式系统所衍生的新型计算机系统原理、设计与实现，并聚焦于深度学习模型稀疏性这一提升模型效率与规模的核心要素，尤其关注其在端到端模型稀疏化过程中的重要作用。通过提出创新性的系统抽象以及构建相应的编译框架，期望为推动相关领域的技术进步提供理论依据与实践参考。

深度学习模型的规模和复杂度的增长对计算资源和存储资源提出了严峻挑战。模型稀疏化作为一种有效的解决方案，通过减少模型中不必要的参数或连接，在不显著降低模型性能的前提下，显著提升模型的计算效率和存储效率。当前，已有多种稀疏化方法被提出，但大多集中在局部优化或特定场景下的应用，缺乏一种能够实现端到端模型稀疏化且具有广泛适应性的系统级解决方案。在端到端的模型稀疏化过程中，如何确保稀疏属性在模型的各个层级和不同硬件平台上有效传播，并实现高效的算子实现和优化，仍然是亟待解决的关键问题。

杨凡团队研究的核心目标是提出一种全新的系统抽象 —— 具有稀疏属性的张量 TeSA，以此为基础构建端到端的支持模型稀疏化的编译框架 SparTA，实现深度学习模型在大型分布式系统中的高效稀疏化处理。具体而言，期望通过 TeSA 抽象使得张量的稀疏属性和稀疏模式能够在整个深度学习模型中无缝传播，进而创建出针对不同稀疏模式和硬件平台的高效、专用模型算子实现。

最终，使 SparTA 在推理延迟和内存占用等关键性能指标上显著优于现有最先进的稀疏方案。

TeSA 抽象的提出是本研究的重要创新点之一。它突破了传统张量抽象的局限，首次将稀疏属性和稀疏模式纳入张量的核心定义中，为端到端的模型稀疏化提供了统一且灵活的基础框架。通过 TeSA，能够在模型构建和训练的早期阶段就充分考虑稀疏性因素，实现稀疏信息在整个模型生命周期中的有效传递和利用。针对不同稀疏模式和硬件特性的高效模型算子实现。这些算子将充分挖掘硬件的并行计算能力和存储层次结构，通过优化数据访问模式、计算调度策略等手段，实现模型稀疏化后的高效执行。例如，在 GPU 等硬件平台上，针对特定稀疏模式设计专门的线程分配和数据加载方案，以最大限度地提高计算效率和减少内存带宽压力。作为一个统一的平台，能够容纳各种稀疏模式和优化技术，实现从模型源代码到高效可执行代码的自动化转换。与传统的编译框架相比，SparTA 不仅关注代码的生成效率，更注重在编译过程中对模型稀疏性的深度优化。通过在编译阶段对模型结构、算子实现和硬件平台进行综合分析和优化，SparTA 能够在不增加用户过多编程负担的前提下，显著提升模型的稀疏化效果和执行性能。

SparTA 编译框架可直接应用于深度学习模型的开发和优化过程，帮助研究人员和开发者在不显著降低模型性能的前提下，显著提高模型的计算效率和存储效率。例如，在自动驾驶领域，通过使用 SparTA 对深度学习模型进行稀疏化处理，可以降低车载计算设备的功耗和成本，提高系统的实时响应能力。TeSA 系统抽象和 SparTA 编译框架的提出为新型计算机系统的设计提供了新的思路和方法。未来的计算机系统可以基于这些成果构建更加智能、高效的硬件架构和软件生态，以适应不断增长的人工智能计算需求。例如，在设计专门用于深度学习的芯片（如 TPU）时，可以借鉴 TeSA 的稀疏属性表示和 SparTA 的编译优化技术，提高芯片的性能和能效比。

学习感悟

不管是数据库还是系统底层，都随着研究的发展，面对大规模数据和计算量出现计算资源等性能瓶颈，从而延伸出新的研究方向和拓展思路。以会议报告中的大型分布式系统与深度学习模型稀疏化内容为例，精准地捕捉到了当下大型分布式系统与深度学习蓬勃发展过程中所面临的核心瓶颈 —— 计算资源与模型复杂度之间的矛盾，并将模型稀疏性作为突破这一困境的关键钥匙。这种敏锐的问题洞察力和明确的研究导向，使我深刻认识到在学术探索中，找准关键问题并确定清晰的研究目标是创新思维的首要目标。

包括数据库里的创新与改良，往往也随着分布式系统、流数据处理等方向的拓展而出现新的难点。类比到数据库领域，当前数据库对于复杂数据类型（如多媒体数据、半结构化数据等）的抽象表示能力仍有提升空间。可以探索设计一种新型的数据抽象方式，将数据的语义特征、访问模式以及存储特性等关键信息融入到数据的基础表示结构中。例如，对于频繁进行关联查询的数据集，可以在数据抽象层标记出其潜在的关联键信息，以便数据库管理系统在查询优化阶段能够更精准地制定执行计划，就如同 TeSA 让深度学习模型能提前感知稀疏性而优化计算一样，提高数据库操作的整体效率。考虑不同数据分布模式（类似深度学习中的稀疏模式）对算子执行的影响，如对于倾斜数据分布的数据集，开发专门的聚合算子来减少数据倾斜带来的性能瓶颈，从而实现数据库在不同硬件平台和数据场景下的高效运行。

在数据库中，数据冗余是一个常见的问题，尤其是在分布式数据库环境下，数据复制和存储开销较大。可以借鉴模型稀疏化的思想，探索数据压缩与精简存储的新方法，通过识别和去除数据中的冗余信息（如重复记录、无效字段等），在不影响数据完整性和查询结果准确性的前提下，降低存储需求。同时，在查询执行过程中，根据数据的重要性和访问频率动态调整资源分配，类似于深度

学习中根据模型稀疏模式优化计算资源分配，优先保障高频、关键查询的执行效率，提高数据库系统整体的资源利用率和性能表现。

在具体的创新点剖析方面，具有稀疏属性的张量 TeSA 中的算子并非孤立地追求计算效率，而是充分考虑了不同硬件平台的特性以及各种稀疏模式的差异。在技术的创新中，设计更具备适应性和功能性的技术或组件，往往是创新和技术改良的第一步。

从成果和应用前景来看，这项研究在深度学习模型的优化、大型分布式系统的性能提升，还是在新型计算机系统的设计方面，都与实际应用紧密联系。学术研究不应仅仅是理论上的自研，而应紧密结合社会需求和行业发展趋势，以解决实际问题为出发点和落脚点，从而实现学术价值与社会价值的有机统一。

大型分布式系统与深度学习模型稀疏化研究为数据库研究在数据抽象、算子优化、系统架构以及性能优化等多个关键方面提供了新颖的视角和可借鉴的方法路径。通过积极吸收这些启发并深入探索应用，数据库技术有望在未来实现创新性的突破与发展，更好地适应日益增长的数据处理需求和复杂多变的应用场景。

研　究　生　签　字_____

指　导　教　师　签　字_____

院(系、所)领导签字_____

年　　月　　日