我国学者在深度学习处理器体系结构研究方面取得新进展

日期 2023-05-29 来源：信息科学部 作者：廖清 吴国政 赵瑞珍 谢国 肖斌 【大 中 小】 【打印】 【关闭】

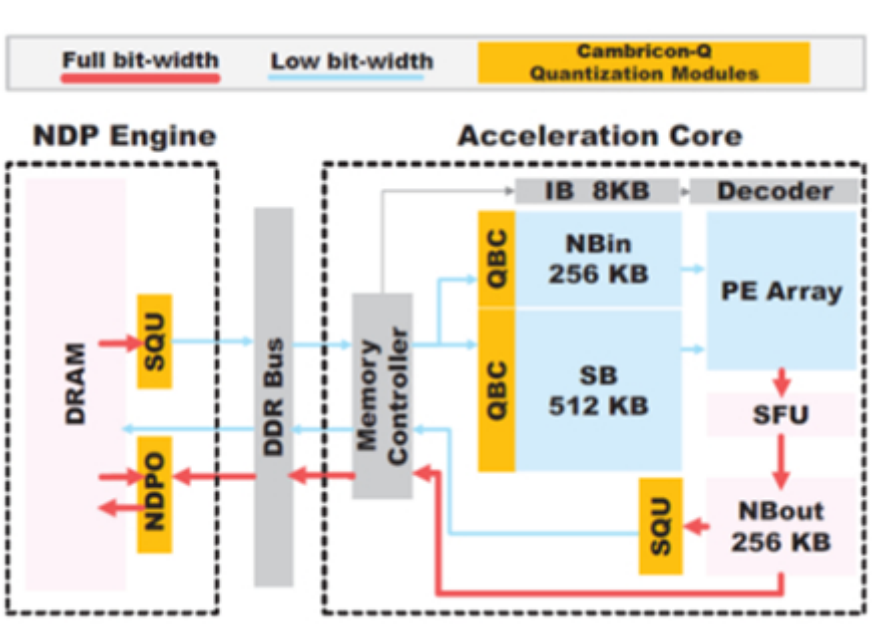

图1 支持量化训练的深度学习处理器架构Combricon-Q

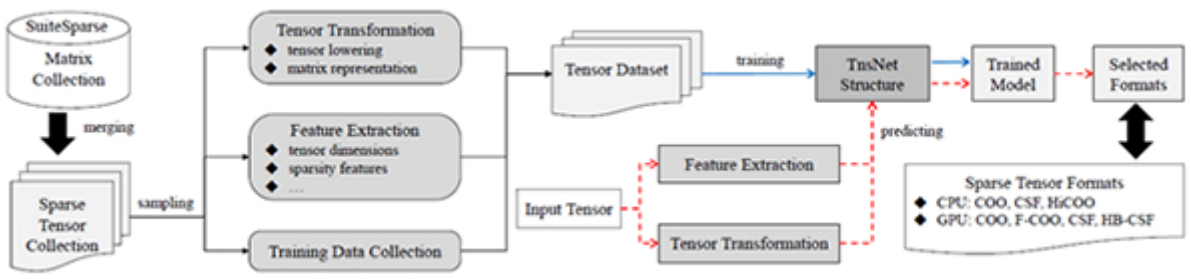

图2 用卷积神经网络预测稀疏张量最优存储格式的方法SpTFS

在国家自然科学基金项目（批准号：61732002）资助下，北京航空航天大学钱德沛教授团队与中国科学院计算技术研究所陈天石研究员团队合作，在深度学习处理器体系结构研究方面取得新进展。

近年来深度学习研究取得了巨大进步，已经开始影响社会的生产和生活，与此同时，算力需求也呈指数上升趋势，现有的智能计算硬件难以同时满足高性能和高能效的需求。尽管包括稀疏化和低比特量化在内的多种非精确计算方法等已被证明可以降低神经网络的理论计算量，但现有的计算硬件无法充分发挥上述算法的优势。因此，如何从处理器体系结构、算法、软件等多个层面系统地支持非精确性优化方法，理解非精确计算过程与结果准确性的关系，解决深度学习处理器性能和能效两者的矛盾，是亟待解决的问题。

钱德沛和陈天石团队以充分利用深度学习对非精确计算的容忍特性为核心思想，围绕深度学习处理器设计，跨越多个系统层次开展了研究。提出了一种用于深度学习处理器的领域专用指令集架构，集成了标量、向量、矩阵、逻辑、数据传输和控制指令以及支持非精确计算的扩展指令，为微体系结构明确了设计目标。基于此，团队取得以下进展：

（1）提出了一系列支持稀疏神经网络的深度学习处理器体系结构Cambricon-X、Cambricon-S和Cambricon-SE，能够有效利用神经网络中的权值稀疏性和神经元稀疏性，提升神经网络算法的计算效率。与早期的深度学习处理器DianNao相比，最新的Cambricon-SE处理器在性能和能效方面分别提高了10倍和20倍。团队提出的支持量化训练的处理器Cambricon-Q，可以以微小的精度损失为代价达到相比GPU的 4.2倍性能提升和相比TPU的1.7倍性能提升，如图1所示。

（2）为了解决算法到硬件的适配问题，提出了一种支持多种深度学习处理器体系结构的编译技术，可以通过张量抽象机抽象出多种深度学习处理器的共同架构特征。为了从工具角度支持非精确计算，提出了一种张量计算冗余零值分析工具ZeroSpy，可以有效识别由于数据结构使用不当和无用计算而造成的冗余零并指导相应的代码优化。提出了一种利用神经网络预测不同硬件平台下稀疏张量最优存储格式的方法SpTFS，如图2所示。为了从基础算法层面支持非精确计算，提出了一种可分解Winograd卷积计算轻量化算法，可突破原始算法对卷积核大小的限制，提高卷积计算效率。

项目研究工作覆盖了指令集、微体系结构、基础算法库、系统软件和应用等多个层次，探索了从支持非精确计算的角度同时提升深度学习处理器性能和能效的新途径，形成了非精确深度学习处理器体系结构的系统性方法，有助于进一步推动智能芯片研究和应用。

项目研究成果发表在多个计算机体系结构领域重要会议和期刊，包括国际计算机体系结构会议（International Symposium on Computer Architecture）、国际微体系结构会议（The International Symposium on Microarchitecture）、国际超算大会（The International Conference for High Performance Computing, Networking, Storage, and Analysis）、《美国计算机学会·计算机系统汇刊》（ACM Transactions on Computer Systems）、《电气电子工程师协会·计算机汇刊》（IEEE Transactions on Computers）等。