智源学者成果展示——智源青年科学家

作者　梁云（北京大学）

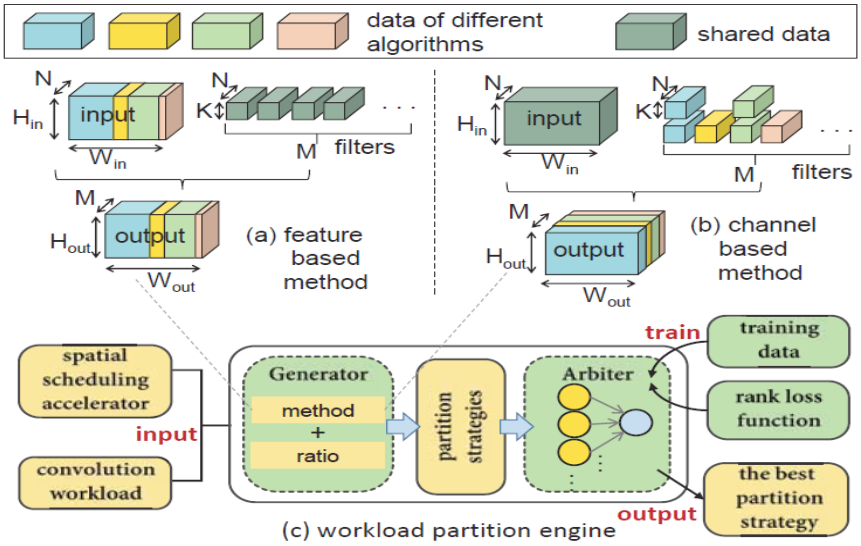
2020年6月

项目负责人：梁云

单位 / 团队：北京大学

成果名称：面向敏捷硬件设计的高效算子库（Efficient Library for Agile Hardware Design）

成果简介：人工智能应用种类众多，需要庞大算力的支持。此外，人工智能应用表现出前所未有的多样性，需要定制化的芯片满足不同应用场景在性能、准确度、功耗等方面的多样化需求。为了让芯片设计的流程更加敏捷，我们设计了高效的硬件算子库，降低人工智能芯片开发的门槛，提高硬件模块的复用。首先，我们设计了基于 FPGA 平台的高效卷积算子库 FCNNLib, 支持多种不同计算复杂度的卷积实现 (spatial, gemm, fft, winograd)。FCNNLib 提供三种调度方式来支持不同卷积算法的组合：分时调度、分域调度、混合调度。FCNNLib 还设计了简单的用户接口，并且集成到 Pytorch 中，以帮助软件工程师探索不同的硬件调度方式和资源分配，并自动生成最终的硬件实现。其次，我们设计一种高效的脉动阵列 (systolic array) 硬件加速器的生成器。脉动阵列架构广泛应用于各类硬件加速器中，适合多种向量运算。然而，目前的设计完全依赖人工，不易于性能调优和复用。通过分析脉动阵列的设计空间，我们设计了模块化、参数化的硬件结构模板。这些模板可以通过互相组合，生成不同的脉动阵列架构，这显著提升了开发及优化硬件架构的效率。在硬件模板的基础上，我们设计了一个脉动阵列结构的生成器，通过计算模式和数据流、参数的定义实例化硬件模板，从而生成完整的脉动阵列加速器架构。相关论文即将发表在 DAC 2020 和 IEEE MICRO 2020 上。



[1] “FCNNLib: An Efficient and Flexible Convolution Algorithm Library on FPGAs,” to appear in the proceedings of *the Design Automation Conference (DAC)*, July 2020.

[2] Linacheng Kia, Liqiang Lu, Xuechao Wei, Yun Liang. “Generating Systolic Array Accelerators with Reusable Blocks,” to appear *IEEE MICRO Special Issue on Agile and Open-Source Hardware, 2020*.

Beijing Academy of Artificial Intelligence

微信关注

北京智源人工智能研究院