## 主题：多精度INT加速器

## 论文1：

* **论文名称**: **Bit Fusion: Bit-Level Dynamically Composable Architecture for Accelerating Deep Neural Networks** (ISCA’18, A会，替代计算技术（ACT）实验室, 乔治亚理工学院, 美国)
* **作用**：作为对比组，强相关。
* **研究问题：**为了防止精度损失，数据的比特宽度在DNN引用之间显著变化，也可以针对每个层单独调整。固定位宽加速器要么提供有限的好处来适应最坏情况下的位宽要求，要么不可避免地导致最终精度的下降。
* **研究层次：**低精度计算->硬件设计->系统->多精度INT加速器：**动态比特级融合/分解**，**INT2/4/8**
* **研究重要性：**高性能计算机体系结构设计的进步一直是深度神经网络（DNN）快速发展的主要驱动力。研究热点。
* **研究差异：**
* **同**：多精度组合的INT加速器
* **异**：提出动态比特级融合/分解的INT加速器设计
* **挑战：**DNN内和DNN之间的数据位宽各不相同，以确保不损失准确性。为了利用这一特性，可编程加速器需要在运行时提供**比特级灵活性**。
* **解决方案：**本文将**动态比特级融合/分解**作为DNN加速器设计中的一个新维度。通过设计比特融合来探索这一维度，比特融合是一种比特灵活的加速器，它构成了一个比特级处理元件阵列，动态融合以匹配各个DNN层的比特宽度。该体系结构中的这种灵活性使得能够在不损失准确性的情况下以尽可能细的粒度最小化计算和通信。

1. **动态比特级融合和分解**。介绍并探索了**位级柔性DNN加速器架构**Bit-Fusion的维度，该架构将位级可组合处理引擎与DNN层所需的不同位宽动态匹配。通过提供这种灵活性，Bit Fusion旨在最大限度地减少DNN在每层的比特粒度上所需的计算和通信。
2. **微结构设计用于位级复合**。比特块的2D阵列构造了可融合的处理引擎，其可以以各种比特宽度执行DNN计算。该微体系结构还配有存储逻辑，允许向BitBrick提供不同的位宽操作数。
3. **用于位灵活加速的硬件-软件抽象。**为了使DNN应用程序能够利用这些独特的比特级融合能力，提出了一种称为**fusion ISA**的块结构指令集架构。为了分摊可编程性的成本，Fusion ISA将DNN层的操作表示为具有迭代语义的位灵活指令块。

## 论文2：

* **论文名称**：**BitBlade: Area and Energy-Efficient Precision-Scalable Neural Network Accelerator with Bitwise Summation** (DAC’19, A会，浦项科技大学，韩国)
* **作用**：（1）作为对比实验组，**强相关；** （2）**写法值**得借鉴在INT SA中。
* **研究问题：**支持位可重新配置的逻辑往往会在加速器中造成较大的面积和功率开销。
* **研究层次：**低精度计算->硬件设计->系统->多精度INT加速器：多精度的硬件优化, INT2/4/8
* **研究重要性：**为了最大限度地提高不同DNN应用的硬件加速器的能效，加速器需要支持各种位宽配置。 研究热点。
* **研究差异：**
* **同**：支持激活/权重的多种精度组合的INT加速器
* **异**：优化可变位宽配置的**移位-加法逻辑**的开销。
* **挑战：**即使有了DNN专用的硬件加速器，执行DNN任务中大量的MAC操作仍然需要大量的能量。比特融合（Bit Fusion）论文中可重新配置位宽的移位-加法逻辑占用大面积（67%）并消耗大量功耗（79%），这迫切需要减少可变化的位宽配置的**移位-加法逻辑**的开销。
* **解决方案：**本文介绍了BitBlade，面积和能效高精度可扩展神经网络硬件加速器。BitBlade通过**逐位求和的方法最小化了与支持各种位宽相关的逻辑。**首先利用循环嵌套优化方法来最小化计算中的移位操作。然后，介绍基于循环嵌套优化的逐位求和方法，并将其应用于BitBlade架构。

## 论文3：

* **论文名称**：**An Energy-Efficient Mixed-Bitwidth Systolic Accelerator for NAS-Optimized Deep Neural Networks** (TVLSI’22, B刊，南方科技大学，中国)
* **作用**：（1）作为对比实验，**强相关。** （2）**写法值**得借鉴在INT SA中。
* **研究问题：**现有的低比特宽度组合（LBC）方法以大的硬件成本改进了低比特宽度吞吐量。高比特宽度分割（HBS）方法最小化用于配置的附加逻辑门，但在低比特宽度模式下的吞吐量性能较差。LBC和HBS方法都没有同时针对吞吐量和功率性能进行优化。
* **研究层次：**低精度计算->硬件设计->系统->多精度INT加速器：多精度加速器优化和设计,INT2/4/8
* **研究重要性：**为了满足具有各种卷积核大小和数据位宽的网络的计算要求，实现的加速器应该是可配置的以提高硬件使用率和能效。研究热点。
* **研究差异：**
* **同**：多种精度的INT加速器
* **异**：混合位宽MAC操作的bit-拆分-组合(BSC)方法来优化加速器
* **挑战：**如何在硬件资源有限的终端设备上实现具有节能和高通量性能的**混合位宽DNN**。
* **解决方案：**本文提出了一种高效节能**的脉动加速器**，用于处理**混合位宽DNN计算**。使用一种新的**空间配置**方法来提高吞吐量和功率性能，特别是对于中间位宽操作。主要贡献概述如下：

1. **用于混合位宽MAC操作的BSC方法**：它支持拆分高位宽单元和组合低位宽单元进行混合位宽操作，提高了能效；
2. **混合位宽的单元（PE）内的脉动和PE间的并行数据流**：它可以支持PE阵列之间的混合位宽卷积。为了实现低功耗和高通量性能，采用了**数据重用**策略来减少数据传输。PE中**混合位宽模式的**并行性计算进一步提高了吞吐量。

## 论文4：

* **论文名称**：**Mix-GEMM: An efficient HW-SW Architecture for Mixed-Precision Quantized Deep Neural Networks Inference on Edge Devices** (HPCA ’23,A会, 加泰罗尼亚理工大学, 巴塞罗那超级计算中心)
* **作用**：（1） 中相关。 （2）作为实验对比研究背景
* **研究问题：**量化窄精度整数数据的DNN推理可以实现高效深度学习计算。但当前针对资源受限设备的CPU体系结构和指令集体系结构（ISAs）对计算DNN内核所支持的数据大小范围存在限制。
* **研究层次：**低精度计算->硬件设计->系统->多精度INT加速器：CPU上运行的窄精度INT2~8
* **研究重要性：**基于量化窄精度整数数据的深度神经网络（DNN）推理是在边缘和移动设备上实现高效深度学习计算的一个很有前途的研究方向。
* **研究差异：**
* **同**：窄精度INT的DNN推理的软、硬件设计
* **异**：Mix-GEMM的硬件微结构位于处理器执行阶段，基于**二进制分割**技术，在**窄整数**上计算高性能的SIMD操作，可**重新使用**处理器功能单元（FUS）。
* **挑战**：优化DNN在许多领域都是一个重大挑战，尤其是在针对边缘和移动领域设计的硬件架构进行部署时，需要高性能，但在面积、内存和能耗方面存在严格限制。
* **解决方案：**本文提出了**Mix-GEMM**，这是一种**硬件-软件结合**结构，能够有效地计算基于**字节和子字节数据大小的量化DNN卷积核。**Mix-GEMM加速了**通用矩阵乘法（GEMM）**，支持**从8位到2位**的所有数据大小组合，包括混合精度计算，并具有随着计算数据大小的减小而扩展的性能。贡献：

1. 设计了一种**面积少和功耗低的硬件加速器**，集成到边缘处理器管道中，能够计算基于窄整数的混合精度**GEMM内核**。所提出的架构称为**µ-引擎**，利用二进制分割技术执行3到7个MAC/周期，同时重用处理器乘法器；
2. 使用**自定义指令扩展了RISC-V ISA**，用于设计处理µ-引擎的高性能GEMM软件库，允许对数据大小进行细粒度选择，并在吞吐量、能效和内存占用方面平衡总体DNN性能；
3. 将µ**-引擎集成到基于RISC-V的边缘系统芯片（SoC**）中，并在六个卷积神经网络（CNNs）上对Mix-GEMM的性能进行了基准测试。；

## 论文5：

* **论文名称**：**FlexBlock: A Flexible DNN Training Accelerator With Multi-Mode Block Floating Point Support** (TC’23, A刊，大邱庆北科学技术研究所, 韩国高级科学技术研究院)
* **作用**：（1）多种精度BFP计算核研究（实质是多精度的INT计算部分研究），可作为对比实验，强相关（2）写法值得借鉴在INT SA中。
* **研究问题：**现有的基于BFP的DNN加速器只针对特定（单一）的精度，使其通用性较差。
* **研究层次：**低精度计算->硬件设计->系统->多精度INT加速器：多精度BFP的DNN训练加速器设计（核心是多精度INT计算核设计）
* **研究重要性：**BFP格式有效地减少了内存占用，并为乘法累加（MAC）操作提供了更便宜的固定点算法。 研究热点。
* **研究差异：**
* **同**：BFP的硬件加速器
* **异**：支持多种BFP的DNN训练加速器设计
* **挑战**：（1）高效利用乘累加（MAC）单元资源；（2）压缩深度(depthwise)运算的硬件利用率;（3）训练期间自动设置BFP模式。
* **解决方案：**本文提出了**FlexBlock**，一种具有**三种BFP模式**的DNN训练加速器，在激活、权重和梯度张量之间数据格式有所不同。通过将FlexBlock配置为较低的BFP精度，与16位模式相比，核心处理的MAC数量在8位模式下最多增加4倍，在4位模式下增加16倍。为了达到这一理论上限，**FlexBlock在各种精度级别或层类型下最大限度地提高了核心利用率**，并允许**动态精度控制**在**不牺牲训练精度的情况下将吞吐量保持在峰**值。贡献：

1. **多模式BFP支持**：开发了一种基于BFP的DNN训练加速器，支持多种BFP精度。有了这种支持，用户可以将加速器配置为每个张量所需的最小精度，以加快训练速度。
2. **高核心利用率**：在所有操作条件下最大限度地提高核心利用率，方法是：i）有效地**对乘法器进行分组**，以降低精度生成多个输出，以及ii）为深度操作放置单独的归约单元。
3. **低精度训练**：展示了FlexBlock的用例，该用例通过动态选择每个张量所需的最小精度来最大化训练性能。

## 论文6：

* **论文名称**：**A Precision-Scalable Deep Neural Network Accelerator With Activation Sparsity Exploitation** (TCAD’24, A刊，上海交通大学)

**作用**：（1）多精度INT+稀疏的加速器，中相关 （2）**写法值**得借鉴在INT SA中。

* **研究问题：**在不同的实际场景中，比特宽度可能在DNN之间变化，甚至在DNN内的层之间变化。因此，在低精度的情况下，固定位宽的DNN加速器将面临计算资源利用不足的问题，精度可扩展性可以来解决此问题。
* **研究层次：**低精度计算->硬件设计->系统->多精度INT加速器：多精度INT+稀疏，INT2/4/8
* **研究重要性：**DNN的不同层对于相同的量化技术具有不同的灵敏度，这意味着可以单独地推动每一层的比特宽度限制。为了进一步挖掘基于融合单元（fusion unit, FU）的DNN加速器的潜力，为其提供利用稀疏性的能力至关重要。
* **研究差异：**
* **同**：多精度 FU 的硬件加速器
* **异**：多精度 FU+稀疏 的加速器
* **挑战：**很难直接利用基于FU的DNN加速器的稀疏性。FU被设计用于计算低精度模式下的内积。即使大多数输入操作数为零，它仍然需要为非零操作数工作。该功能阻碍了现有的基于FU的DNN加速器。
* **解决方案：**本文提出了一种基于**FU的精确可扩展DNN加速**器，该加速器也可以利用稀疏性。在设计中利用了激活的固有稀疏性。**主要贡献**如下：

1. 修改了FU的传统架构，使其更适合于**零跳跃方案**。基于改进的FU（MFU），首次开发了一种基于FU的精确可扩展DNN加速器，该加速器也可以利用稀疏性。
2. 为了解决由非零激活的随机位置引起的**内存访问冲突问题**，提出了一种将输入特征图和滤波器划分为一些独立片段的方法。
3. 提出了一种**稀疏性感知映射方法**来缓解工作负载的不平衡。它根据特征图的平均稀疏性对输入激活进行重新排序，并显著提高了吞吐量方面的性能。
4. 为了进一步利用比特级的稀疏性，提出了一种**比特分割策略**，将每个激活划分为多组比特。它只需要对整个体系结构进行非常轻微的更改，就能显著提高吞吐量。