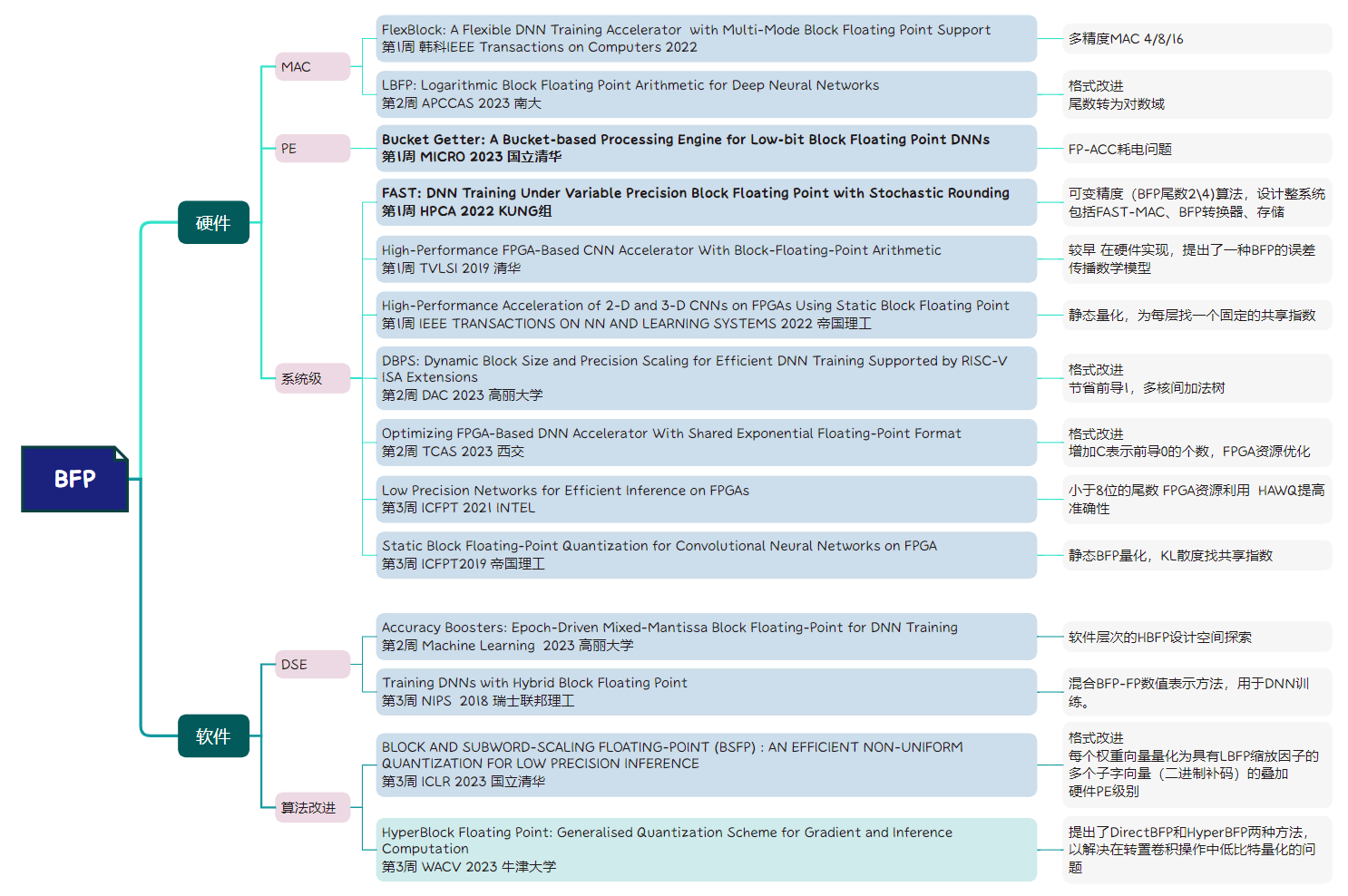
**文章1：BLOCK AND SUBWORD-SCALING FLOATING-POINT (BSFP) : AN EFFICIENT NON-UNIFORM QUANTIZATION FOR LOW PRECISION INFERENCE---**

**ICLR 2023 国立清华CCF-A**

* 研究问题： 用于神经网络中权重向量的偏斜和非均匀分布的BSFP，将每个权重向量量化为具有缩放因子（低位浮点，LBFP）的多个子字向量（二进制补码）的叠加。PTQ
* 研究层次：中相关，BFP格式改进，PE级别
* 研究重要性：提出的BSFP方法在保持高计算效率的同时，能够有效地适应神经网络中权重向量的偏斜和非均匀分布。BSFP比MSFP的准确率高，模型小。吞吐量高，能效高。
* 研究同异：

同：基于BFP低浮点量化，与这篇论文最相关的工作是MSFP in NeurIPS 2020

异：BSFP设计与MSFP相比有根本的不同和创新，原因如下：1)BSFP使用多个子字缩放向量的叠加来逼近全精度权重向量，这是MSFP和其他以前的工作没有探索的。2)设计了支持不同配置的BSFP位串口处理引擎。相比之下，MSFP工作考虑的是位并行处理引擎，不支持格式更改。

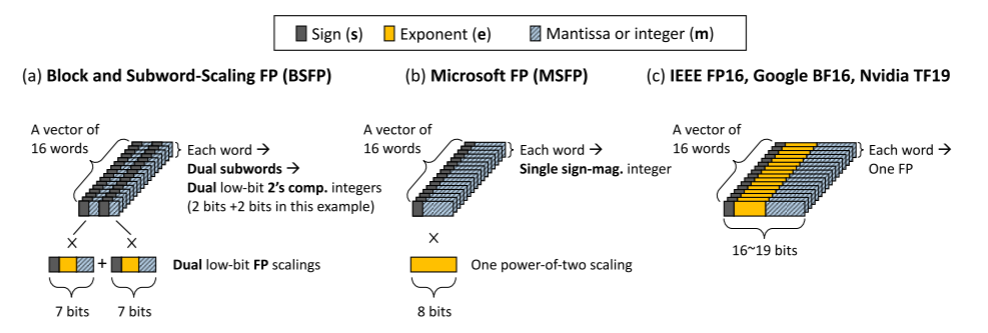
* 挑战：

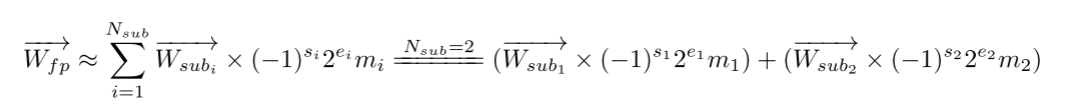
为深度神经网络（DNN）设计一种硬件算法，该算法可以以比现有数据类型更好的计算吞吐量和能效实现更高的精度。

* 解决：

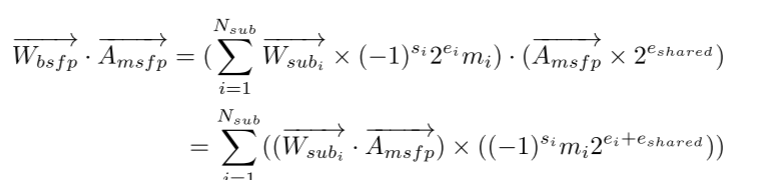
1）BSFP格式，两个缩放的两个子字向量之和来近似每个全精度权重向量，每个子字是低位宽的有符号整数(2补码），缩放因子是低位宽浮点（7b 1s3m3e)。使用粗向量和细向量来近似期望的权重向量。具有大缩放的一个子字向量捕获大权重，并且具有小缩放的另一个子字向量减轻剩余偏差。

好处：二补码3b表示8个数而MSFP7个（-/+0)；二补码不对称（-2,-1,0,1)更适应权值的不对称分布而MSFP对称；组合两个子字缩放提供非均匀量化而MSFP均匀；量化步长由LBFP决定而MSFP 缩放是2幂次。

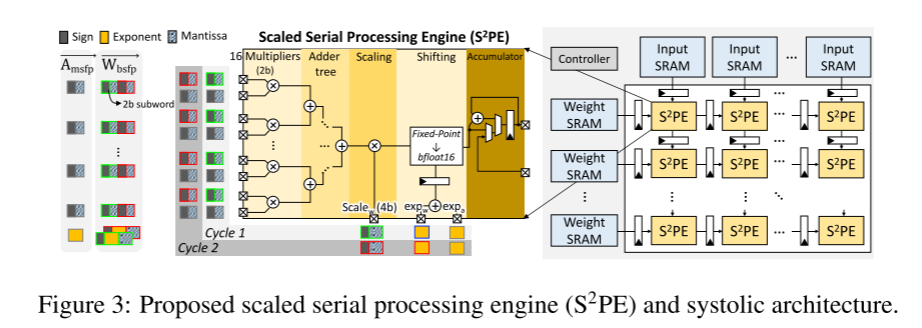
其中Nsub=2个子字（二补码）缩放向量（LBFP）



BSFP权重（离线）与MSFP激活的点积



2）可伸缩串行处理引擎(S2PE)，乘法器2b，位串行多个周期实现不同位宽，将缩放的部分和转换回BF 16进行累加。BSFP比MSFP多的硬件开销是将部分和乘以缩放因子的4b尾数



* 作用：

对于非均匀分布，调整数据格式，可以做横向对比

**文章2：Static Block Floating-Point Quantization for Convolutional Neural Networks on FPGA ---**

**ICFPT 2019 帝国理工 CCF-B**

* 研究问题：
* 研究层次：强相关，系统级的硬件设计与量化的软件工具
* 研究重要性：本文的静态量化方法提高了量化的性能和硬件效率，并保持了较高的准确性。
* 研究同异：

同：都关注了CNN的量化问题，并提出了相应的解决方案。

引用[**High-Performance FPGA-Based CNN Accelerator With Block-Floating-Point Arithmetic**] 片外和片内存储器之间传输的数据仍然基于16位FP。

异：本文的共享指数是静态的，减少了格式的转换；所有数据在推断期间使用预先固定的共享指数值，因此不需要用于指数值的额外存储器和计算资源。

* 挑战：

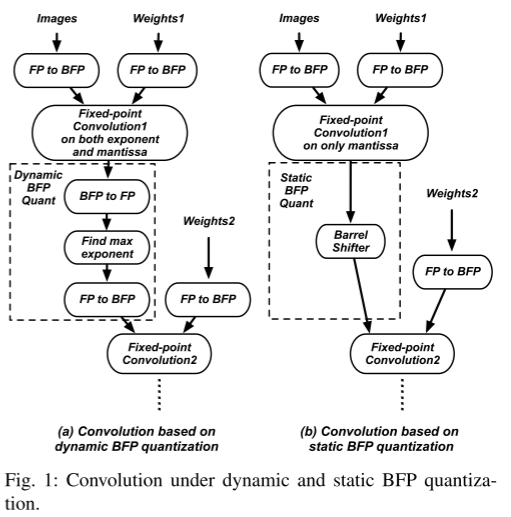
1）降低计算复杂性的量化方法涉及的再训练过程需要在分析和计算中做出重大努力以恢复精度。此外，在大型数据集上无法保证其准确性。

2）之前的研究涉及BFP和FP之间的频繁转换，对资源和性能造成很大的开销。

* 解决：

1）本文提出了一种静态BFP量化方法，代替在推理期间动态地确定最大指数，该恒定值用Kullback-Leibler发散确定。因为共享指数在推理之前被确定，所以层间结果不需要被转换回FP来找共享指数，而是直接移位。

将输入输出分成很多块，得到FP分布 得到emax 用i（=3足够）个emax附近值作为共享指数量化，构造了i个不同量化分布，计算FP分布与不同量化分布的KL偏差，选择最小KL的指数值。分块，整个激活作为一个块；滤波器维数作为权值的分块数。



2）提出了一种基于FPGA的静态BFP量化的硬件设计。软件找好共享指数，BFP转换器，只进行移位，数据翻转的桶形移位器 双向。

3）开发了一个在PyTorch框架中实现的工具，该工具接受BFP的配置作为输入，例如尾数位和块大小，然后产生优化的共享指数和BFP量化精度，无需再训练，整个优化过程在 CPU上只需几分钟。

* 作用：

有github源码可以参考下

**文章3：Low Precision Networks for Efficient Inference on FPGAs---**

**ICFPT 2021 INTEL CCF-B**

* 研究问题：
* 研究层次：中相关， FPGA上资源利用优化
* 研究重要性：本文提供了一种在FPGA上进行神经网络推理的高效解决方案，并通过重新训练和精度调整等方法，克服了量化对准确性的影响。
* 研究同异：

同：与FAN（文章2）、LIAN（第一周 文章4）等人研究类似，在FPGA上探索尾数位宽。

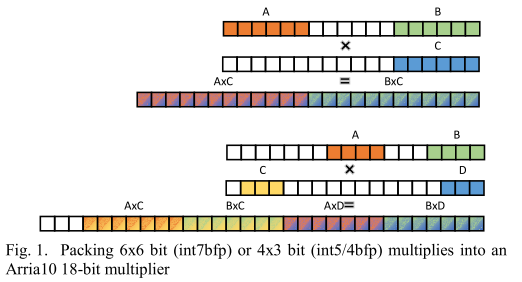
异：本文提供了一个完整的端到端的BFP重新训练实现，研究8b以下低尾数影响。 INTEL的Stratix10 NX支持BFP（[【芯片论文】英特尔-Altera：Stratix 10 NX FPGA架构和应用 - 知乎 (zhihu.com)](https://zhuanlan.zhihu.com/p/658703146) FPGA2021 [Stratix10 NX：超越GPU的人工智能时代“最强”FPGA？ - 知乎 (zhihu.com)](https://zhuanlan.zhihu.com/p/345302527)）

* 挑战：

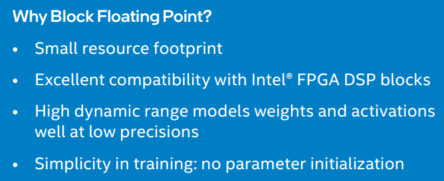
其他工作将BFP用于深度学习，保持较大的尾数大小以避免重新训练，但对8位以下的精度进行了肤浅的研究。对于MobileNet和EfficientNet等更紧凑的网络，在非常低的精度水平（int5/4bfp）下，精度明显下降。

* 解决：

研究小于8位尾数大小对硬件的影响。本研究的最大量化值是int 8bfp，由7b尾数位 1b符号 5b指数。低精度是int 5 bfp 激活和int 4 bfp 权值，int 5 bfp由5b尾数（包括符号）5b指数。



量化为int 8bfp 比FP16使用的DSP数减半，int 5 bfp再减半



2）由于量化而导致的准确性损失，可以通过低成本的几次重新训练方法恢复了部分准确性。在精度低于int7bfp的情况下，可以对关键层中的权值激活的位宽加倍。

**文章4：HyperBlock Floating Point: Generalised Quantization Scheme for Gradient and Inference Computation---**

**WACV 2023 牛津大学**

* 研究问题：
* 研究层次：中相关，BFP量化方法
* 研究重要性：解决使用低精度计算中梯度的挑战，同时在训练和推理时使用相同的量化构建块。块浮点的扩展，以达到更少的量化转换操作。
* 研究同异：

同：都探索了在卷积神经网络训练中使用低比特量化的方法。

异：与DoReFa-Net等方法相比，本文的创新之处在于提出了DirectBFP和HyperBFP两种方法，以解决在转置卷积操作中低比特量化的问题。

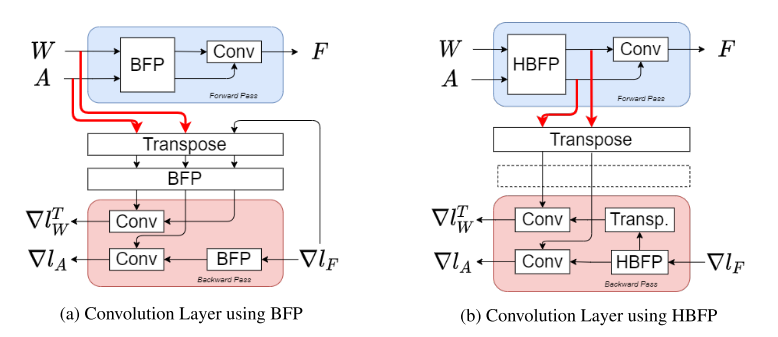
* 挑战：

BFP存在的问题是在batch和通道维度的转置之后，指数不再共享。因为具有共享指数的数字在不同的批次中

* 解决：

1. 直接梯度块浮点: 将BFP应用于梯度意味着在转置之后、卷积之前添加BFP量化器。前向传播和反向传播中分别应用BFP量化，几乎没有准确性损失。

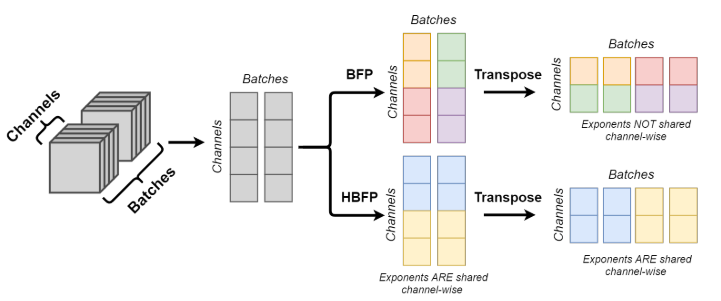
需要6次量化



2）HyperBlock提出了一种更简单的方法来避免多次重复相同的张量。

HBFP在转置后将共享的指数保持在相同的批次中，因为它们在2 × 2的2D块中共享指数。

需要3次量化



* 作用：

在计算过程找到的优化点，如何选择共享指数的粒度。

**文章5：Training DNNs with Hybrid Block Floating Point---**

**NIPS 2018 瑞士洛桑联邦理工**

* 研究问题：
* 研究层次：中相关，软件层面设计空间探索
* 研究重要性：使用混合BFP-FP来提高计算效率和精度。
* 研究同异：

同：窄精度的训练和推理。

异：相对于纯BFP计算，提出混合BFP和FP的DNN训练。

* 挑战：

之前旨在信号处理用BFP, BFP代替浮点数进行DNN训练面临三个重大挑战：

1. 尽管BFP点积的面积效率很高，但其他BFP操作可能效率不高，从而导致硬件具有类似浮点的算术密度。
2. 如果指数值太大或太小，指数共享可能导致数据丢失，指数选择很重要。
3. 如果张量的值分布太宽而不能被尾数位捕获，BFP可能会导致数据丢失。

* 解决：

1. 所有基于点积的操作中使用BFP，并在所有其他操作(即激活、正则化等)中使用FP32。权重BFP，激活FP。
2. 为量化点选择合适的尺度：BFP产生的数据丢失量由两个因素决定:共享指数的张量大小和尾数的宽度。本文设计了两种优化方法:分块，将权重矩阵划分为预定义大小块 块内共享指数。限制了共享指数的值的数量，从而减少数据丢失；；；宽权值存储：只有权重更新使用更宽的尾数，其他操作如前向 反向传递仍然使用原始尾数执行，。
3. 只在点积之前使用最大张量值的指数将张量转换为BFP，其他操作将浮点数作为输入，从而能够精确表示任意值分布。

* 作用：

引用本文作为相关工作