## 主题：在DNN INT4应用场景支撑

## 论文1：

**论文名称**：**Ultra-Low Precision 4-bit Training of Deep Neural Networks** (NeurIPS 2020, A会, IBM T.J.Watson研究中心，美国)

**作用**：（1）对**INT4-FP4**格式的支撑，中相关，（2）可以尝试部署到INT-SA上

* **研究问题：**

本文探索了一种新的自适应**梯度缩放技术（GradScale**），解决了量化梯度的范围和分辨率不足的挑战，并探索了在模型训练过程中观察到的**量化误差**的影响。本文从理论上分析了**偏差在梯度量化**中的作用，并提出了减轻这种偏差对模型收敛影响的解决方案。使训练系统的精度能够从8位大幅缩**减到4位**。

**研究层次：**应用级别：软件优化-> FP4-e格式, GradScale

**研究重要性：**在低精度算术的一些关键进展以及吞吐量对精度的二次方依赖性的推动下，低精度训练已成为提高深度学习硬件性能和功率效率的事实技术。研究热点。

**研究差异：**

* **同**：INT的DNN训练
* **异**：对所有张量（权重、激活和梯度）使用4位进行深度学习模型训练收敛。以4位整数（WA-INT4）量化权重和激活，**FP4-e（1-3-0）梯度**格式和量化方法相结合。
* **挑战：**

1. 从量化误差（舍入）、精度和动态范围的角度来看，**4位梯度**表示似乎具有挑战性。这导致了显著的模型优化和泛化困难；
2. 在正向和反向传播中使用的张量可能需要显著不同的数值表示和范围。4bit训练可能需要**不同的数字格式**来表示训练的前向和反向阶段中的不同张量。

* **解决方案：**

1. 展示了一种端到端解决方案，该解决方案使用**4位来完成DNN训练期间所需的绝大多数计算**。本文提出了一种**新的4位浮点格式(FP4-e:1-3-0)、舍入方案以及新的梯度缩放技术（提出GradScale，一种新的逐层梯度缩放技术）**，以最大限度地减少梯度表示、精度和范围挑战。
2. 没有使用相同的4bit梯度格式来表示权重和激活**，而是将最先进的4位推理方法与本文的技术相结合**。

## 论文2：

**论文名称**：**Training Transformers with 4-bit Integers** (arXiv, 22 Jun 2023,清华大学,**代码开源**)

**作用**：(1) 对INT4格式的支撑，实验部署，强相关；（2）INT4格式的大模型的训练部署

* **研究问题：**

将激活、权重和梯度量化为4位有望加速神经网络训练。然而，现有的4位训练方法需要**定制的数字格式**，而现代硬件**不支持**这种格式。本文提出了一种用INT4算法实现所有**矩阵乘法**的Transformer的训练方法。

**研究层次：**应用级别：软件优化-> 阿达玛量化器、比特分割

**研究重要性：**训练神经网络在计算上要求很高。低精度算术训练（也称为全量化训练或FQT）有望提高计算和记忆效率。**研究热点**。

**研究差异：**

* **同**：低精度定点的DNN训练
* **异**：**阿达玛**（Hadamard）量化器抑制异常值,用**比特分割**和利用**分数采样**技术来准确量化梯度；本文算法可以在GPU上实现。
* **挑战：**

在极低的4位水平上训练神经网络存在显著的优化挑战。首先，前向传播中的不可微量化器使**损失**景观变得崎岖不平，其中基于梯度的优化器很**容易陷入局部最优**。其次，梯度仅以低精度近似计算。这种**不精确的梯度减缓了训练过程**，甚至导致训练**不稳定或偏离**

* **解决方案：**

1. 对于正向传播，本文发现**激活中的异常值**是精度下降的主要原因。为了抑制异常值，本文提出了一种**阿达玛（Hadamard）量化器**，它量化激活矩阵的变换版本。
2. 对于反向传播，本文利用激活梯度的结构稀疏性，提**出比特分割**，将每个令牌的梯度分割为**较高的4比特和较低的4比特**。然后，本文通过**杠杆得分抽样**来选择信息量最大的梯度，这是RandNLA的一种重要抽样技术。
3. 结合前向和反向传播的量化技术，本文提出了一种算法，该算法将**INT4 MM**用于Transformers中的所有线性运算。本文的算法与GPU等现代硬件兼容。

## 论文3：

**论文名称**：**Is Integer Arithmetic Enough for Deep Learning Training?** (NeurIPS 2022, 华为诺亚方舟实验室，蒙特利尔研究中心)

**作用**：（1）对INT8+INT16格式的支撑，强相关, （2）可以作为研究背景

* **研究问题：**

用低位整数运算代替浮点运算是一种很有前途的方法，可以节省深度学习模型的能量、内存占用和延迟。因此，近年来量化引起了研究人员的注意。然而，使用整数来形成包括前向传递、反向传播和随机梯度下降在内的**全函数整数**训练流水线并没**有得到详细的研**究。

**研究层次：**应用级别：软件优化🡪计算的数值格式变换，纯整数推理

**研究重要性：**INT训练，研究热点。

**研究差异：**

* **同**：INT的DNN训练方法
* **异**：与其他工作提出的量化反向传播不同，本文直接**更改浮点值的**数字表示，且从**经验和理论**上证明了。
* **挑战：**

整数梯度计算存在一**些挑战**。

（1）梯度必须正确缩放，以适应整数的有限动态范围。

（2）梯度的数值误差必须是无偏的，以便保持训练算法的收敛轨迹。

（3）整数训练必须与分布无关，即训练方法不应取决于梯度、权重或训练数据的分布。

* **解决方案：**

本文提出了一种新的整数训练方法，旨在同时解决上述挑战（i）、（ii）和（iii）。将线性动态定点映射与非线性逆映射相结合是全整数训练的关键。

1. 基于提取张量的**最大浮点指数作为尺**度，提出了一种硬件友好的整数训练方法。本文提出张量的**线性不动点映射**，而浮**点的逆映射是非线性的**。该方法解决了前面提到的所有挑战：（a）它动态计算量表，而且在训练过程中不需要调整量表；（b）它提供梯度的无偏估计，因此，它的收敛轨迹紧密遵循浮点版本；（c）本文提出的表示映射不依赖于训练数据、权重或梯度的分布。
2. 研究了提出的整数训练算法的**最优性差距**，并**证明**它在训练过程中类似于浮点训练算法（定理1）。
3. 提出的方法使用**纯整数算法**有效地执行了现代神经网络中所需的所有运算。

## 论文4：

**论文名称**：**I-BERT: Integer-only BERT Quantization** (ICML 2021,**A会**，加州大学伯克利分校，**代码开源**)

**作用**： （1）INT的大模型推理；强相关。（2）实验方法：准确性、延迟，消融实验。

**研究问题：**

基于Transformer的模型内存占用、推理延迟和功耗**阻碍**了边缘甚至数据中心的高效**推理**。虽然量化可能是一种可行的解决方案，但之前关于量化基于Transformer的模型的工作在推理过程中使用**浮点运算（如非线性函**数），这**不能**有效地利用**仅整数的逻辑单元**。

**研究层次：**应用级别：软件优化-> **全量化**-大模型**推理**->软件优化:非线性函数定点化

**研究重要性：**量化以低精度存储参数/激活来减少内存占用。使用低精度整数乘法和累加代替浮点运算，也可以受益于更快的推理速度。研究的热点。

**研究差异：**

* **同**：定点的大模型推理。
* **异**： I-BERT为基于Transformer的模型引入了一系列新颖的**纯整数量化**方案，关注的核心为：**非线性函数**，仅使用**二阶多项式**就可以得到近似值。
* **挑战：**

与ReLU不同，由于GELU和Softmax的**非线性**，使用纯整数算法计算它们并不简单。此外， LayerNorm需要**动态计算**每个输入的方差平方根，这不能简单地用纯整数运算来计算。另一个挑战是，以低精度处理GELU、Softmax和LayerNorm可能会导致显著的**精度下降**。

* **解决方案：**

I-BERT结合了一系列新颖的仅限整数的量化方案，用于基于Transformer的模型。

（1）用轻量级的**二阶多项式**来近似GELU和Softmax，这可以用纯整数算法来评估；

（2）通过提供已知的**平方根整数计算算法**来执行仅整数计算的LayerNorm；

（3）使用GELU、Softmax和LayerNorm的这些近似来设计基于Transformer 的模型的**纯整数量化。用INT8乘法和INT3**2累加处理嵌入和矩阵乘法（MatMul）。然后对INT32累加结果计算非线性运算（GELU、Softmax和LayerNorm），然后重新**量化回INT8。**本文用整数表示整个计算图中的所有参数和激活，并且从不将它们转换为浮点。

## 论文5：

**论文名称**：**HADAMARD DOMAIN TRAINING WITH INTEGERS FOR CLASS INCREMENTAL QUANTIZED LEARNING** (arXiv, 22 Oct 2023,圣安德鲁斯大学, 英国)

**作用**：（1） INT的背景；弱相关。 （2）作为相关工作，新的应用场景

**研究问题：**

对于具有隐私和低延迟要求的应用程序，持续学习（Continual learning，CL）施加的计算和内存要求对于资源受限的边缘平台来说**可能成本过高**。通过全量化训练（FQT）降低计算精度，同时减少内存占用，提高训练和推理的计算效率。然而，激进量化，尤其是整数的FQT，通常会将模型**精度降低到不可接受的水平**。

**研究层次：**应用级别：软件优化-> 持续学习(CL)

**研究重要性：**通过量化以降低精度的张量对NN进行操作，同时提高了计算的能量效率，并减少了内存占用和带宽要求。研究热点。

**研究差异：**

* **同**：低精度定点的训练
* **异**： 先前关于全量化训练（FQT）的工作，在非CL场景（NoCL）中使用自定义浮点格式、随机舍入或自适应尺度，本文将**FQT**用于CL以使其能够在资源受限的环境中工作。
* **挑战：**

CL必须适应数据、权重和梯度的**变化分布**，遗忘可能对少数参数中的小扰动**非常敏感。**FQT应用于CL的**可行性**和向**上兼容性**。

* **解决方案：**

本文提出了一种技术，该技术利用廉价的**阿达玛（Hadamard）变换**来实现**仅使用整数矩阵乘法**的低精度训练。本文进一步确定哪些张量需要随机舍入，并提出**拼接矩阵乘法**以实现低位宽累加器。当本文使用**8位累加器**将所有矩阵乘法**输入量化为4位**时，本文实现了小于0.5%和3%的精度下降。