专利名称：一种全整型8-bit卷积神经网络训练感知量化方法

发明人：刘阳、鲁金铭

1. 发明所要解决的技术问题是什么？或者本方案能产生什么好的效果？

解决的技术问题：

卷积神经网络中涉及大量的数据量和计算量，采用传统的标准单精度浮点数会造成过高的处理复杂度，从而导致端侧设备难以支撑其功耗和速度开销。

通过低比特量化能够有效地降低计算复杂度和存储开销，从而满足端侧设备实时处理的需求。

1. 介绍技术背景，并描述已有的与本发明效果最相近似的现实方案

背景技术：

近十年来，随着大数据、云计算、互联网、物联网等信息技术的发展，泛在感知数据和图形处理器等计算平台推动以深度神经网络为代表的人工智能技术飞速发展，大幅跨越了科学与应用之间的“技术鸿沟”，诸如图像分类、语音识别、知识问答、人机对弈、无人驾驶等人工智能技术实现了从“不能用、不好用”到“可以用”的技术突破，迎来爆发式增长的新高潮。

主流人工智能的解决方案，目前大多是采用CNN来实现。但是目前的CNN网络模型都比较大，需要海量的计算来完成任务。因此出于应用部署的需要，低比特量化能够显著降低计算和存储开销，从而满足实际应用的需求。

现有技术方案：

量化的基本思想是将浮点数(FP32)表示的数据转换为均匀整数(INT IP1\_A8bit)表示的形式，从而实现计算的简化和存储需求的降低。在神经网络中，每一层的输入、输出、权重均以张量(tensor)的形式表示。由于其中的数据范围往往不在原始的INT表示范围内，且不同tensor的数据范围各有不同，通常采用一种共享scale的方式，进行数据范围的调节，从而使整数数值表示满足相应的精度需求。具体如下:

设原始浮点数据为，整数位宽为, scale值为, 量化后的数据为, 则：

本方案与现有方案的区别是什么？

本方案与现有方案的区别：

现有方案中，scale值依旧为浮点数，并且各层之间的scale值有较为明显地差异，依旧不可避免地需要浮点计算，以及复杂的重(chong)量化过程。

本方案中，scale选择为2的整数次幂，因此在硬件计算中可以通过简单地移位操作完成重量化。同时对于存在残差块(residual block) 和拼接(concat)操作，进行预先对齐，而不需要重新对齐的操作。

1. 发明方案详细阐述，应结合结构图、流程图、原理框图、电路图、时序图进行说明

本发明的技术方案：本发明提出了全整型8-bit卷积神经网络训练感知量化方法。

该方法的优势1：所有计算可以用8-bit乘累加完成，极大降低了计算复杂度和存储需求。

该方法的优势2：可以自适应实现scale对齐，不需要复杂地重量化过程。

下面对技术方案进行具体描述。

设原始浮点数据为，整数位宽为, scale值为, 量化后的数据为, 则：

由于量化后的数据是离散表示的，且量化中的floor函数是不可求导的，会导致量化模型无法训练的问题。本方案采用了straight through estimate (STE) 的方案近似，实现梯度的反向传播。即：

为实现全整型量化，引入以下约束条件:

1. 引入scalem，来体现对scale的限制为2的整数次幂， 即: scalem = -log(scale)/log(2)
2. Conv, BN, FC层中，bias的scalem应小于输入的scalem和weight的scalem之和，大于该和的值减去15；
3. Conv, BN, FC层，存在截位问题，输入tensor的scalem和weight的scalem的和应大于输出tensor的scalem，且小于该值加上15；
4. Concat层需要对应的两个tensor的scalem相等；
5. Residual Block中，需要对应的两个tensor的scalem相等；

为了实现上述约束，训练过程中，每一层均增加input\_scalem的输入，从而将各层的scalem的值沿模型前向执行过程传播，并根据约束关系自适应调整各层数据的scalem范围。

在一个卷积层内，约束关系如下:

对于Residual Block与concat层，采取固定的scale值，令其scalem=4, 为了在满足这种约束条件下，而不需要额外的重量化/移位操作，对于Residual Block与concat的上一层输出，即采取强制约束的方式， 令其，固定其scale值。

1. 发明的关键点和保护点是什么？

关键点和保护点：

* 1. 一种8-bit量化训练方法。
  2. 一种scale自适应对齐的方法。

1. 针对上述第四项中的技术方案，是否还有别的替代方案同样能完成发明目的？

无

1. 其他有助于专利代理人理解本技术的资料

无