论文阅读记录

1. 标题

Accelerating Deep Convolutional Networks Using Low-precision And Sparsity

2. 发表会议/期刊（年份）

ICASSP(2017)

3. 作者及单位

Ganesh Venkatesh
Accelerator Architecture Lab, Intel Corporation

4. 所提出方法的名称

5. 所提出方法的流程图

我们探索在不影响深度卷积网络精度的前提下，显著提高其计算效率和性能的技术。为了提高计算效率，我们专注于用极低精度(2位)权值网络实现高精度，并加速执行时间，我们积极跳过对零值的操作。在低精度网络的Imagenet对象分类挑战中，我们获得了76.6%的Top-1/93%的Top-5的最高报告精度，同时与达到类似精度的全精度网络相比，减少了约3×的计算需求。此外,充分利用我们的网络精度的好处,我们建立一个深入学习加速器核心,D虫胶,可以实现1 TFLOP /毫米2等效为单精度浮点操作(half-precision~2 TFLOP /毫米2),这是~5×比线性代数核心[16]和~4×比先前的深度学习加速器建议[8]。

索引术语-深度神经网络，三元卷积，加速器

介绍

深度卷积神经网络(DNNs)为许多计算机视觉和图像分析任务提供了最先进的精度。DNNs的精度正在迅速提高(例如，Imagenet object classification challenge[9]的Top-5 error在5年内提高了>6×)，在某些情况下已经接近人类水平的精度[12]。

DNN s通过建立由更多层(网络深度)组成的更强大的模型来达到更高的精度。然而，网络深度的增加导致计算和内存需求的急剧增加。因此，这些网络需要更长的时间来培训;即使使用多个GPU卡，多周的训练时间也是很常见的。此外，更大的计算需求使得DNNs更难部署，这导致最近对专用硬件解决方案的大量兴趣，无论是商业上的[3,5]，还是学术界的[10,6,8]。

在本文中，我们基于最近提出的低精度卷积网络[17,15,14]来降低这些网络的计算要求。虽然之前的工作牺牲了精度来获得计算效率，但我们的目标是在较低的计算复杂度下实现类似(或稍好一点)的精度。特别地，我们训练了一个34层深剩余网络(Resnet[13])的低精度变体，它比普通的18层深剩余网络(~3×lower)精度更高，同时需要更少的浮点操作(~3×lower)和更小的模型大小(7×fewer)。此外，为了充分利用这种低精度网络的优点，我们提出并评估了一个深度学习加速器核心D LAC，它可以通过跳过对零值的操作，实现高达~1 Teraflop/mm 2的等效性能。我们在本文中做出了以下贡献

使用低精度的权重来实现高精度
使用低精度的2位权值网络[14]，我们在Imagenet[9]上实现了76.6%的前1/93%的前5名的高精度，据我们所知，这是低精度网络的最高报告，与2015年Imagenet获奖者[13]的准确率仅为1.3%。此外，我们还证明了在这些低精度网络中，大多数浮点运算都是在零值上运行的——既包括训练，也包括推理。

定义一个深度学习加速器核心，D LAC
我们提出了一种深度学习加速器核心D LAC，它利用这些网络的可用稀疏性来实现高效的性能，并可应用于训练和推理。

实现~1 tflops /mm 2的高效性能密度
我们的评估，基于我们的设计综合在14nm，表明D LAC可以维持极高的性能密度，达到1 tflops /mm 2等效性能的许多层在目前最先进的剩余网络[13]。这比之前提出的深度学习加速器[8]的性能密度要高一个数量级。

其用的：

$$W_{ter}(i,j,k,l) = \begin{cases} 1 & : \text{if W(i,j,k,l)} > w_{th} \\ -1 & : \text{if W(i,j,k,l)} < -1 * w_{th} \\ 0 & : \text{otherwise} \end{cases} \quad (1)$$

最近的深度压缩[11]方案着眼于通过剪枝来诱导深度卷积网络中的稀疏性，在剪枝中某些权值/激活量被夹到零值。相反，我们探索了在网络中引入动态稀疏性的技术，其中零权值/激活可以在输入样本和训练阶段发生变化。通过这样做，我们实现了更高的有效稀疏性(>2×)，并展示了它在最先进的网络上的适用性。

评估了两个数据集- Cifar10和Imagenet

训练过程中用的方法：

预初始化从全精度网络
我们在前几个迭代(15个迭代)中以完全精确的方式训练网络，然后切换到低精度模式进行其余的训练(~75个迭代)。这使得精确度提高了近2%(第4.2节)。

跳过降低精度的部分网络
我们不降低网络第一层的精度，以减少输入图像的信息损失。这使我们的Top-1精度提高了约0.5%(第4.1节)。

积极降低学习效率
我们维护了一个列车错误的历史记录，并在列车错误在几次迭代中都没有下降时降低了学习率。在某些情况下，这可以将我们排名前1的准确度提高1%以上。

正则化
我们对激活进行正则化，以降低噪声和诱导更多的稀疏性。作为一个副作用，这种技术趋向于平滑我们的收敛曲线(图3)。

ReLU阈值
我们改变我们的整流单元的阈值，以诱导更高的稀疏性和降低噪声。

通过跳过对零值的操作来提高3-6×的性能（利用稀疏性）

DEEP LEARNING ACCELERATOR

D　LAC是一个二维的处理元素网格，其中有用于网络权重和输入特征图的缓冲区(图2)。处理元素有算术单元、用于输出特征图的缓冲区和用于跳过零值操作的控制逻辑，以利用低精度网络中的稀疏性。为了能够跳过零操作，我们为每个处理元素分配多个输出缓冲区(在几个算术单元上交换更多的缓冲区)，并且在调度更新这些输出缓冲区的操作时，我们跳过对零值进行操作的缓冲区，以加快性能。

我们的加速器支持深层网络的常见操作如下:

mmOp:矩阵相乘
我们使用它来执行卷积和全连接层。

ptWiseOp:逐点的操作
加速器支持在输出缓冲区的某些或所有元素上执行形式的点态操作。op可以是算术运算(add/sub、muli -add)，也可以是三元控制表达式(?:)。我们使用它来执行非线性(Relu)和批处理规范化(推理)。

培训网络的D LAC:
在训练模式中，我们用512个单精度浮点单元实例化每个D LAC，所有的数据路径都是32位宽的。这使得D LAC能够支持当前使用单精度操作的标准培训技术。此外，我们的加速器能够通过利用这些网络中的动态稀疏性来维持更高的有效性能(3-4 TeraOps/cycle)。

D LAC推理:
在推理模式下，我们用256个半精度浮点单元、256个半精度加法器实例化每个D LAC，所有的数据路径都是16位宽的。这为我们的加速器提供了~2×性能密度的提升。与培训阶段类似，由于有效的归零，D LAC能够维持更高的吞吐量。除了低精度网络，D LAC还可以加速修剪后的网络[11]。

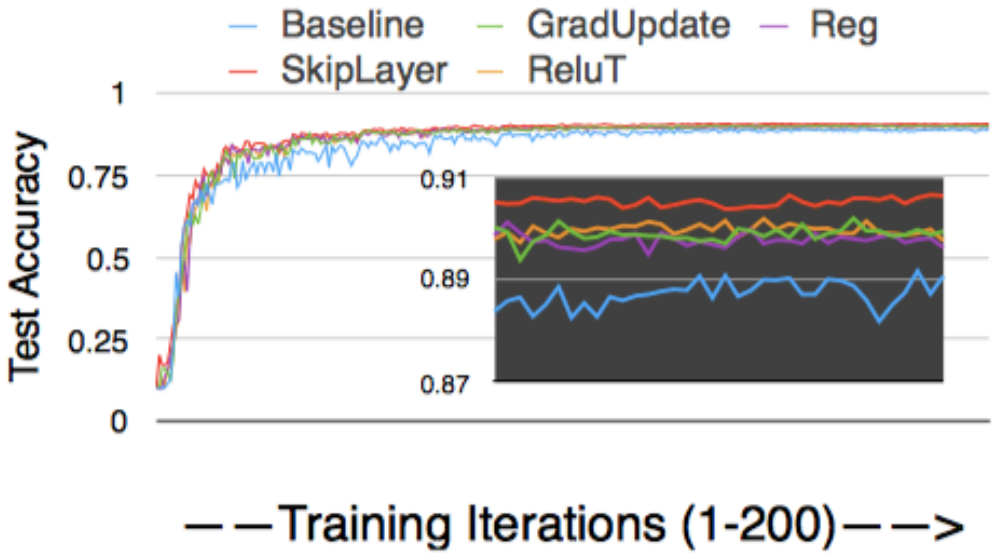

Fig. 3. VGG Convergence Graph for Cifar10 The dark graph inside is zoomed-in version of the final-few epochs of training.

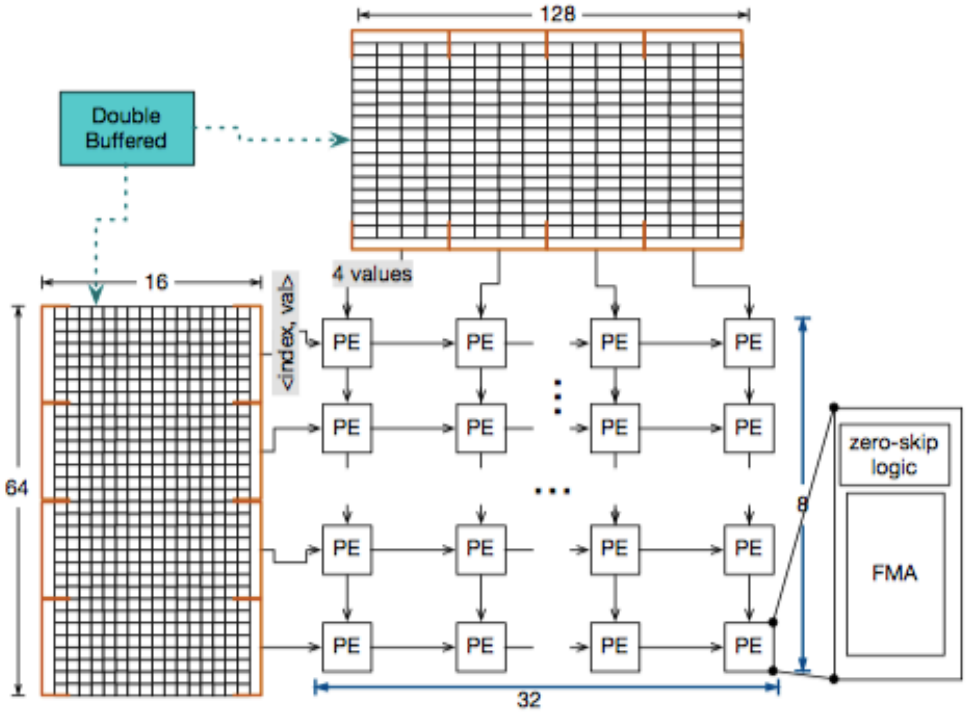

Fig. 2. DLAC **Architecture** It is 2-D grid of processing elements where each one has floating-point units and logic to perform zero-skipping

性能评估：

4.1. Accuracy of Low-precision Networks on Cifar10
图3

4.2Accuracy of Low-precision Networks on Imagenet
在本节中，我们展示了使用多个Resnet网络(2015年Imagenet大赛冠军)对Imagenet数据集的精度结果。我们使用了在Cifar10数据集上显示出希望的技术(正则化、ReLU阈值、降低第一层的跳过精度)，以及在训练精度停止提高时积极降低学习率的额外技巧。
我们从不同Resnet网络的训练模型出发，降低除第一层外的所有层的精度，并对生成的网络进行训练。我们在表1中报告了全精度网络和低精度2位变体的精度数字。数据表明，随着网络深度的增加，低精度网络具有更好的精度。因此，与全精度网络一样，低精度网络也可以通过扩展深度来提供更高的精度。根据表1中的数据，我们还观察到1的低精度变体!较大的网络比常规的全精度网络提供更好的精度(低精度的Resnet-34/Resnet-50/Resnet-152分别比全精度的Resnet-18/Resnet-34/Resnet-50具有更高的精度)。这是一个重要的结果，因为与常规的全精度网络相比，大网络的低精度变种需要更少的计算，而且模型尺寸更小。因此，实际上，使用较大网络的较低精度变体可以获得更好的精度，并且比原来的全精度网络所需的计算量更少。我们用34层和2位权值来训练Resnet。我们尝试以下两种技术来提高精度(i)积极降低学习速度(ii)在前几次迭代中进行全精度训练，并在此之后切换到低精度。我们的结果如图4所示——使用低精度的[14]，我们的精度比以前的工作(约4.8%)高，并且比常规的18层Resnet全精度变体的精度稍好一些。

Network Depth	Full-precision		2-bit precision	
	Top-1 [4]	Top-5 [4]	Top-1	Top-5
Resnet-18	69.56	89.24	-	-
Resnet-34	73.27	91.26	71.6	90.37
Resnet-50	76	93	73.85	91.8
Resnet-152	77.84	93.84	76.64	93.2

Table 1. Accuracy of Resnet network on Imagenet dataset for different depths (first column suffix), regular full-precision, and the extremely low-precision 2-bit version.

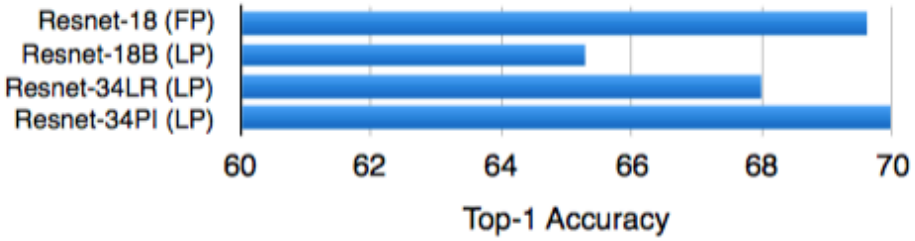

Fig. 4. Training Resnet-34 with low-precision weights The graph shows the accuracy we obtain by training Resnet-34 using low-precision 2-bit weights by aggressively lowering the learning rate (*Resnet-34LR*) as well as pre-initializing with a full-precision network(*Resnet-34PI*). We obtain higher accuracy than previous work on a low-precision network [14] (Resnet-18B) as well as the full-precision Resnet-18 while requiring fewer computations than either.

4.3. Performance of D LAC on Low-precision Networks

图5显示了我们的加速器在Resnet-34中可以支持不同卷积层的性能。该加速器可以维持高达5K的触发器/周期(平均2.78K触发器/周期)，在500MHz时转换为2.5 Teraflops/秒(1.34 Teraflops/秒)。此外，从图中可以看出，我们的加速器为网络的更深层提供了更好的性能，因为这些层的稀疏性更强。因此，当我们将更深层次的网络映射到加速器时，我们期望得到更好的性能，因为它能够利用更深层次的稀疏性。在单精度模式下，D LAC在14nm范围内合成2.2 mm 2(16位模式下为1.09 mm 2)的单元面积，所有缓冲区和算术单元(没有优化的宏块)均采用纯ASIC流。因此，D LAC计算IP的计算密度为0.6 Teraflop/s/mm 2，可以超过1 Teraflop/s/mm 2 !网络的更深层。

与以往工作的比较

与DaDianNao超级计算机[8]相比，一个D LAC实例的性能略好于DaDianNao的一个节点，但>4×更小。此外，随着网络稀疏性的增加，我们的加速器可以提供更高的性能。与最近关于零跳[6]的工作相比，我们利用激活和权重的稀疏性实现了更高的加速。与EiE[10]不

同，我们主要关注卷积层，并支持训练阶段的单精度。

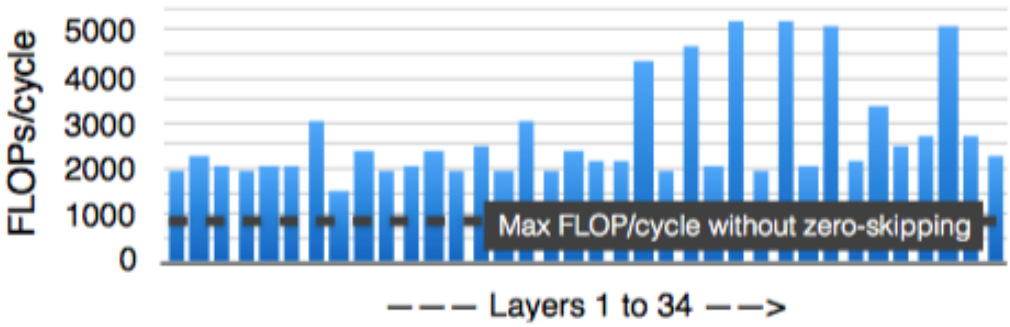

Fig. 5. Performance of DLAC on Resnet-34 The data shows the performance our accelerator can sustain for each layer in 34-layer deep Resnet. The graph shows that our accelerator gets significant performance boost (1.8 - 5×) by skipping operations on zero-values and that our accelerator provides greater speed-up as we go deeper in the network because the layers get more and more sparse.