Tensor Core技术解析（下）_吴建明wujianming_110117-CSDN博客

https://blog.csdn.net/wujianing_110117/article/details/106442372

None Mon Jul, 05 01:24

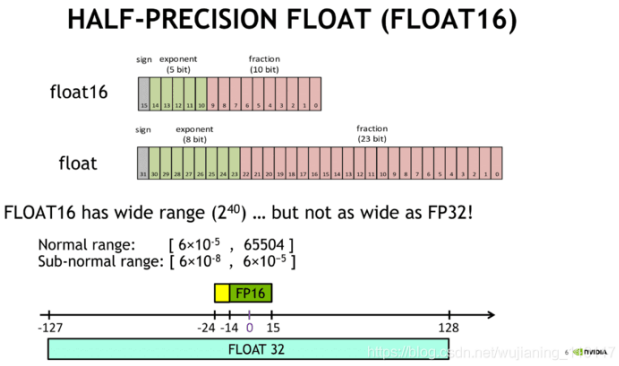
Tensor Core技术解析（下）

让FP16适用于深度学习

Volta的深度学习能力是建立在利用半精度浮点（IEEE-754 FP16）而非单精度浮点（FP32）进行深度学习训练的基础之上。

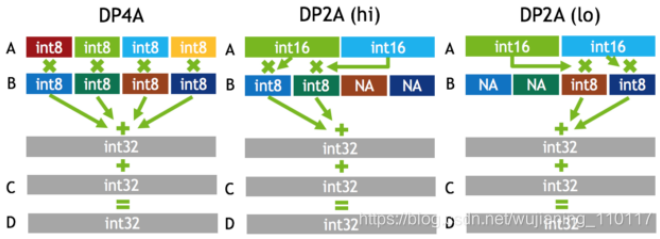
该能力首先由cuDNN 3支持并在Tegra X1的Maxwell架构中实现，随后原生半精度计算被引入Pascal架构并被称为“伪FP16”，即使用FP32 ALUs处理成对的FP16指令，理论上可以使每个时钟的FP16吞吐量增加一倍。这一特性实际上已经在Tensor Core处理寄存器中矩阵片段的过程中得到体现，其两个FP16输入矩阵被收集在8个FP16*2或16个FP16元素中。

就FP32与FP16而言，由于单精度浮点所包含的数据多于半精度浮点，因此计算量更大，需要更多的内存容量和带宽来容纳和传输数据，并带来更大的功耗。因此，在计算中成功使用低精度数据一直是穷人的圣杯，而目标则是那些不需要高精度数据的应用程序。

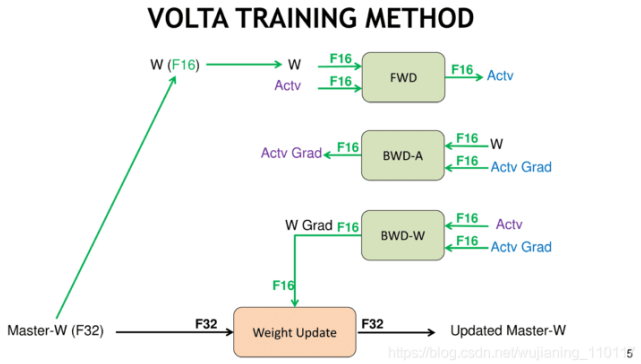


除了API/编译器/框架的支持之外，深度学习一直都有在使用FP16数据类型时损失精度的问题，这会让训练过程不够准确，模型无法收敛。

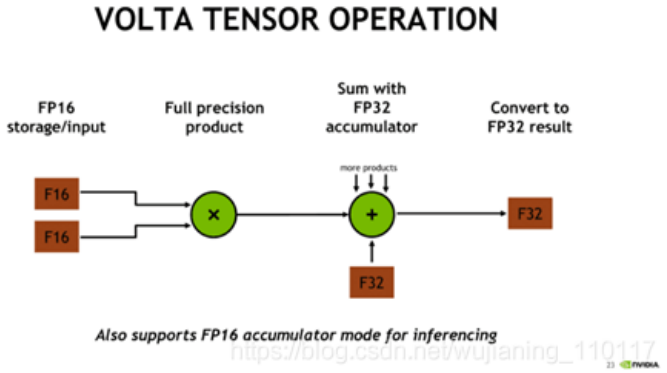
NVIDIA以前也曾在类似的情况下推出过“混合精度”这一概念，在Pascal的快速FP16（针对GP100）和DP4A/DP2A的整数点积操作（针对GP102、GP104和GP106 GPU）中，就曾提出过类似的说法。



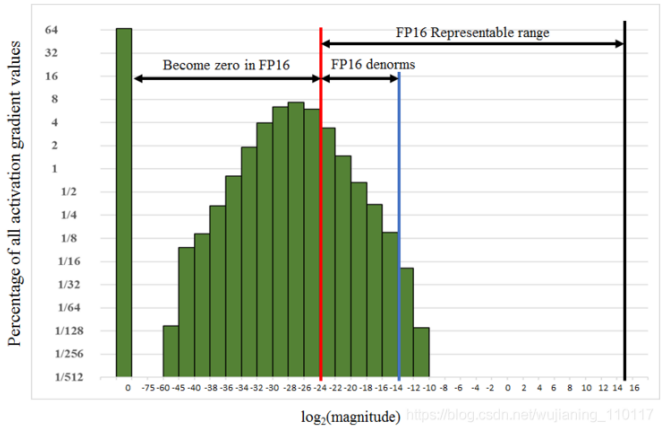
当时人们关注的是推理能力，就像Titan V的“深度学习TFLOPS”一样，Titan X（Pascal）推出了“44
TOPS（新型深度学习推断指令）”。新的指令对4元8位向量或2元8位/16位向量执行整数点积，从而得到一个32位整数积，可以与其他32位整数一起累积。



对于Volta的混合精度而言，重要的精度敏感数据（如主权重）仍然会存储为FP32；而Tensor Core的混合精度训练则会将两个半精度输入矩阵相乘得到一个精度乘积，然后累积成一个精度和。NVIDIA表示，在将结果写入内存之前，Tensor Core会将结果转换回半精度，这样当使用半精度格式时，寄存器和存储器中需要的数据更少，这有助于处理超大矩阵中的数据。



FP16与FP32所包含的数据空间并不相同，归一化方法可以解决FP32格式数据超出FP16可表示范围的问题。举个例子，许多激活梯度的值都落在FP16的范围之外，但由于这些值聚集在一起，因此将损耗乘以缩放因子便可以移动FP16范围内的大部分值。在完成最终的权重更新之前，将梯度重新缩放到原始范围，便可以维持其原始的精度。

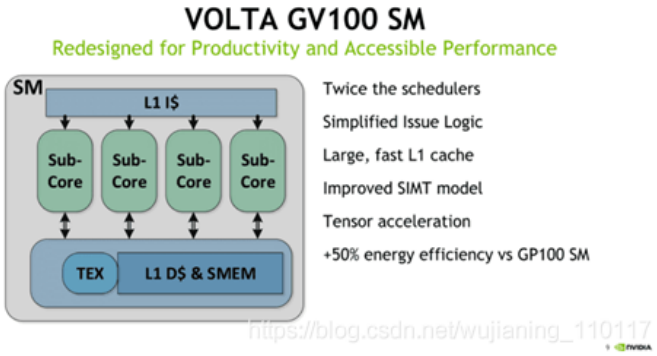


不过，并非所有的数学、神经网络和层都适用于FP16，通常FP16和Tensor Core的混合精度最适合卷积和RNN重图像处理等，而对于不适合的神经网络框架或类型，FP16将默认禁用或不推荐使用。

内存改进，SM变化

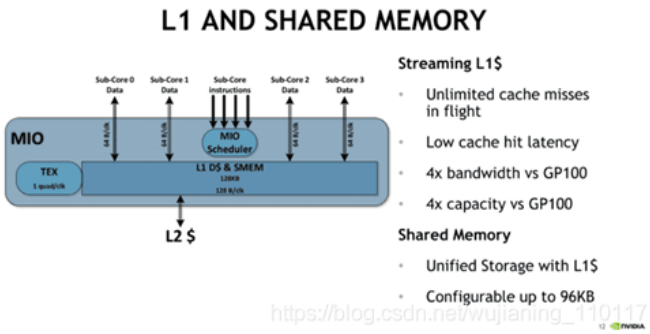
使用Tensor Core处理混合精度数据似乎可以减轻内存带宽问题，但事实证明，尽管Volta在几乎所有方面都得到了内存子系统的增强，但幅度并不明显。

首先，Volta有一个12 KB的L0指令缓存，虽然Pascal及其他之前的GPU核心也有指令缓存，但Volta更高效的L0是子核心SM分区私有的，因此它对warp调度器来说也是私有的，这对Volta架构更大的指令大小是一种补偿，并有可能为支持Tensor Core通道的框架做出贡献。同时Volta指令延迟也要低于Pascal，特别是核心FMAs从6个周期减少到了4个周期。



随着每个SM中调度器比例的增加，砍掉第二个调度端口似乎是对具有独立数据路径和数学调度单元的子核心的权衡。而具备FP32/INT32执行能力，也为其他低精度/混合精度模型打开了大门。这些子核方面的增强，都是为了优化Tensor Core阵列。

另一个重大变化是合并L1缓存和共享内存。在同一个块中，共享内存可配置为每SM 最高96 KB。HBM2控制器也进行了更新，其效率提高了10~15%。



深度学习基准测试

深度学习从框架到模型，再到API和库，AI硬件的许多部分都是高度定制化的，这样的新领域有时会让人非常难以理解。

俗话说“光说不练假把式”，实践永远是检验真理的唯一标准。对计算机来说，介绍的再详细也不如真刀真枪跑一下测试，没有什么比benchmark更能提现硬件的实际表现了。

随着ImageNet和一些衍生模型（AlexNet、VGGNet、Inception、Resnet等）的影响，ILSVRC2012（ImageNet大规模视觉识别挑战）中的图像数据集训练逐渐被行业所认可。现在基本上所有深度学习框架都支持CUDA和cuDNN，对于Volta而言，支持FP16存储的框架也都支持Tensor Core加速，启用FP16存储后Tensor Core加速会自动启用。