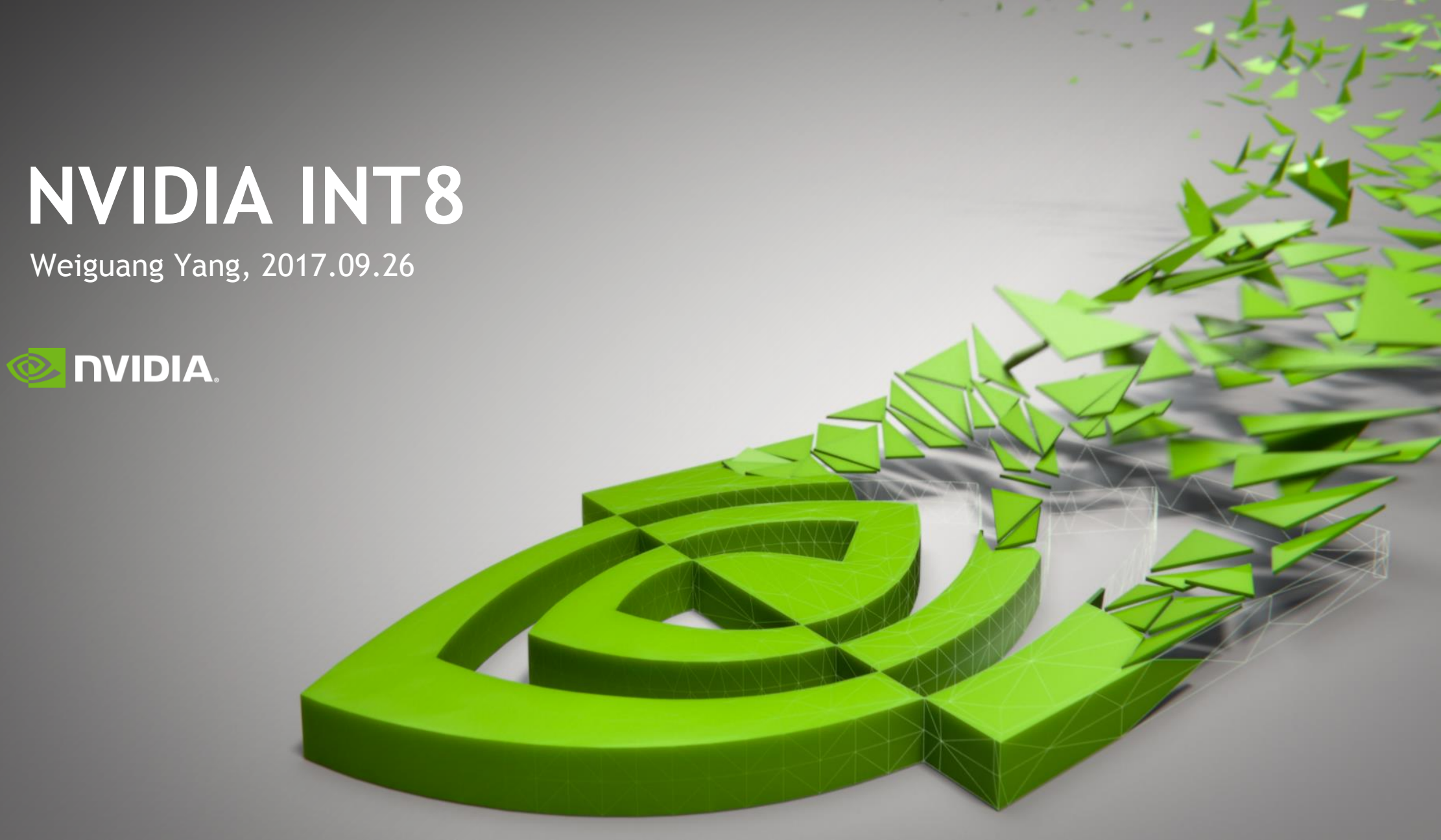
NVIDIA INT8

Weiguang Yang, 2017.09.26



NVIDIA INT8

GTC China 2017 Community Corner

什么是 NVIDIA INT8？

为什么要使用INT8加速？

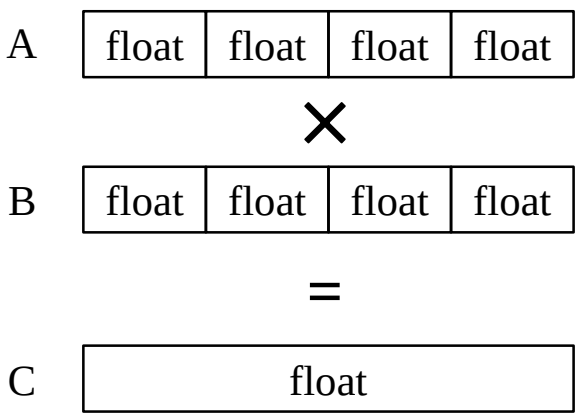
如何使用 NVIDIA INT8？

INT8 最大的挑战是什么？

评价 NVIDIA INT8的两种使用方式

什么是 NVIDIA INT8 ?

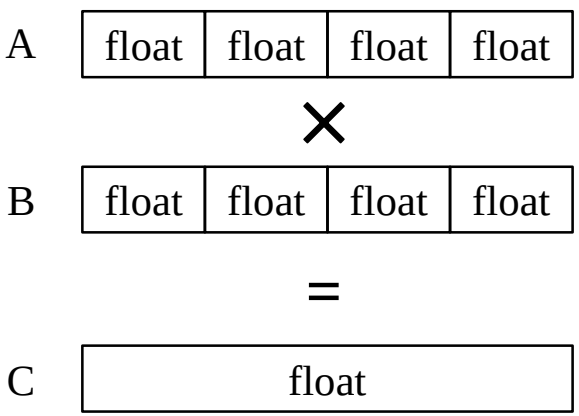
什么是 NVIDIA INT8 ?

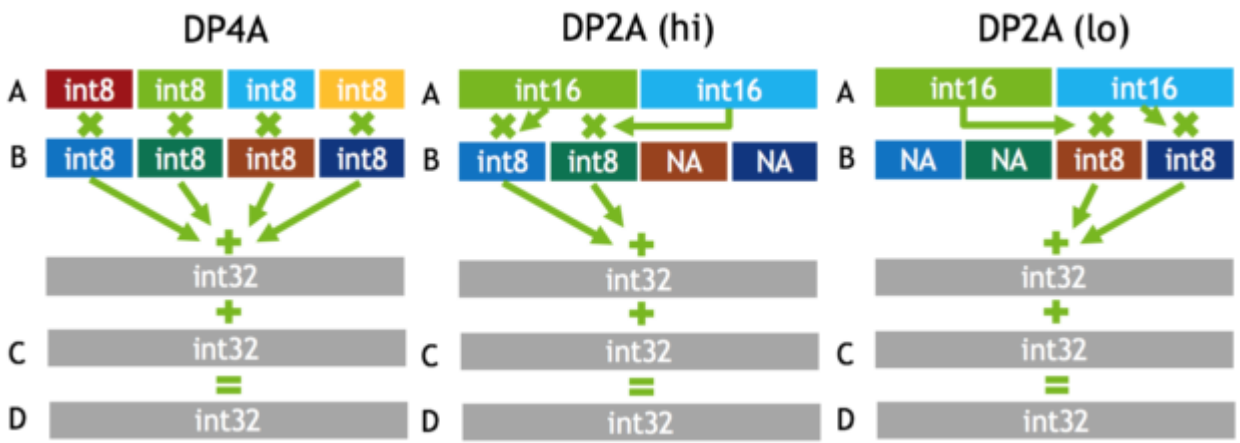


四次乘法+四次加法

https://devblogs.nvidia.com/parallelforall/mixed-precision-programming-cuda-8/

什么是 NVIDIA INT8 ?





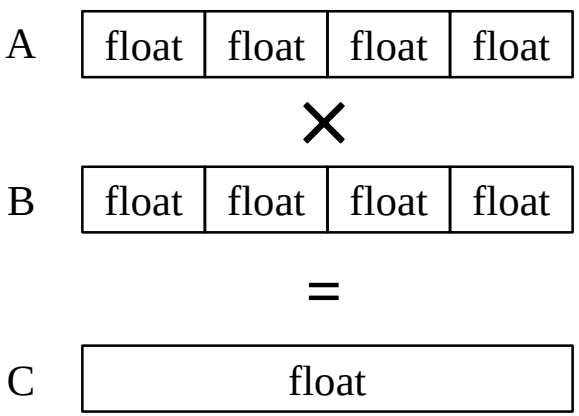
High Performance with Low-Precision Integers

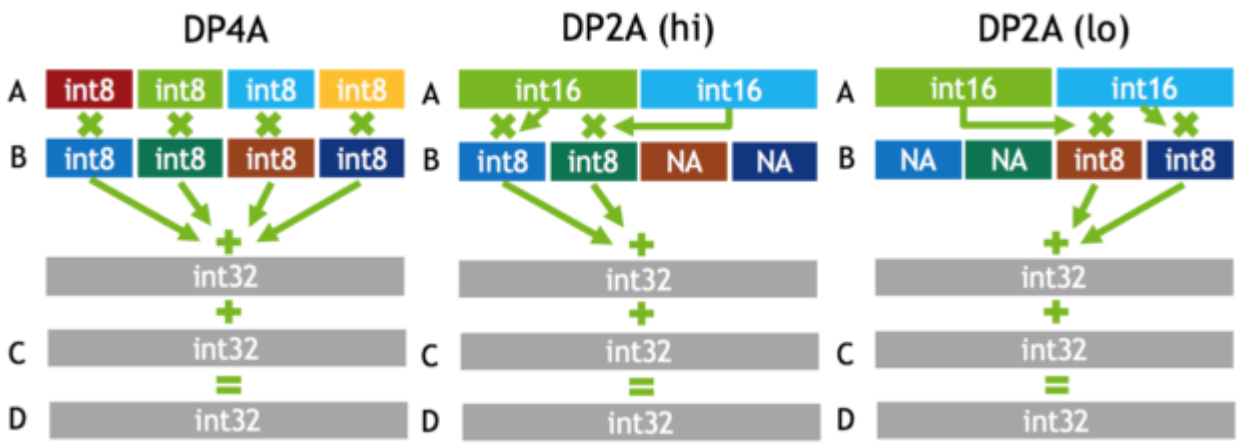
四次乘法+四次加法

一次dp4a

https://devblogs.nvidia.com/parallelforall/mixed-precision-programming-cuda-8/

什么是 NVIDIA INT8 ?





High Performance with Low-Precision Integers

四次乘法+四次加法

理论上4倍加速

模型压缩75%

一次dp4a

https://devblogs.nvidia.com/parallelforall/mixed-precision-programming-cuda-8/

为什么要使用INT8加速？

为什么要使用INT8加速？

线上推理的计算压力随着用户群体的增大而增大

移动端、嵌入式设备内存和计算资源不足

模型越来越大

为什么要使用INT8加速？

线上推理的计算压力随着用户群体的增大而增大

移动端、嵌入式设备内存和计算资源不足

模型越来越大

INT8加速技术可以有效缓解这些问题

使用INT8加速技术对深度学习模型进行加速是非常必要的

INT8 加速技术成功案例

硬件

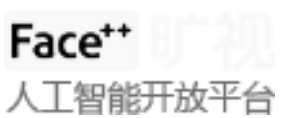
Google TPU

NVIDIA INT8

Intel SSE

.....

深度学习应用



微软亚洲研究院

.....

如何使用 NVIDIA INT8 ?

如何使用 NVIDIA INT8 ?

TENSORRT

a. 神经网络线上推理加速库

b. TensorRT 1支持FP16加速，TensorRT 2支持INT8加速

CUBLAS

使用cublasGemmEx函数的CUDA_R_32I计算模式对矩阵乘进行INT8加速

CUDNN

使用cudnnConvolutionForward函数的INT8,INT8_EXT,INT8x4,INT8x4_EXT配置对卷积操作进行INT8加速

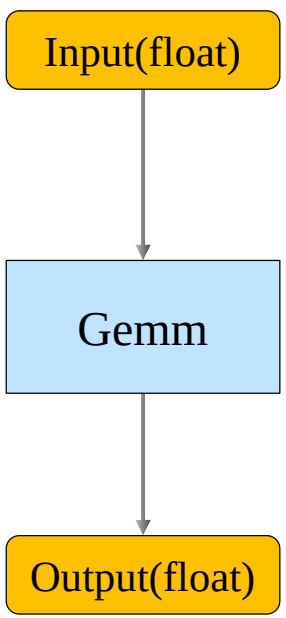
现成的库

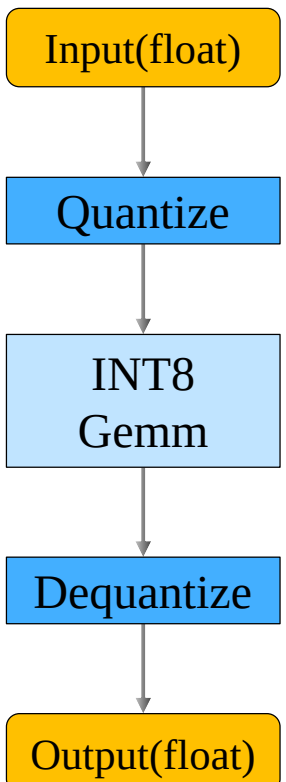
自己动手，丰衣足食

INT8 最大的挑战是什么？

INT8 最大的挑战是什么?

精度问题





量化(quantize)

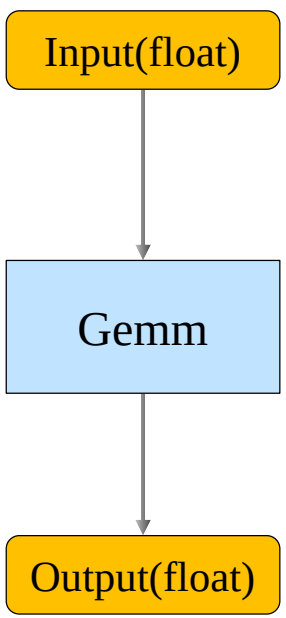
将32位浮点压缩成INT8 (char or uchar)

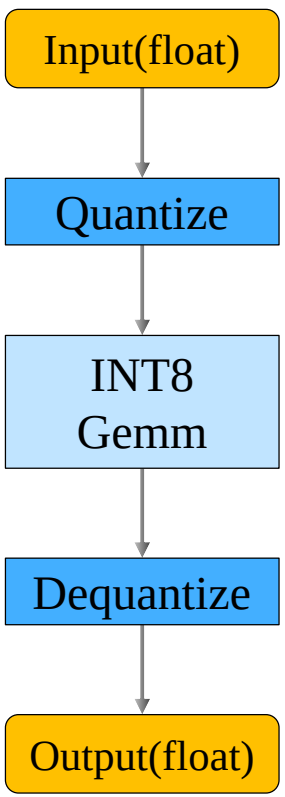
反量化(dequantize)

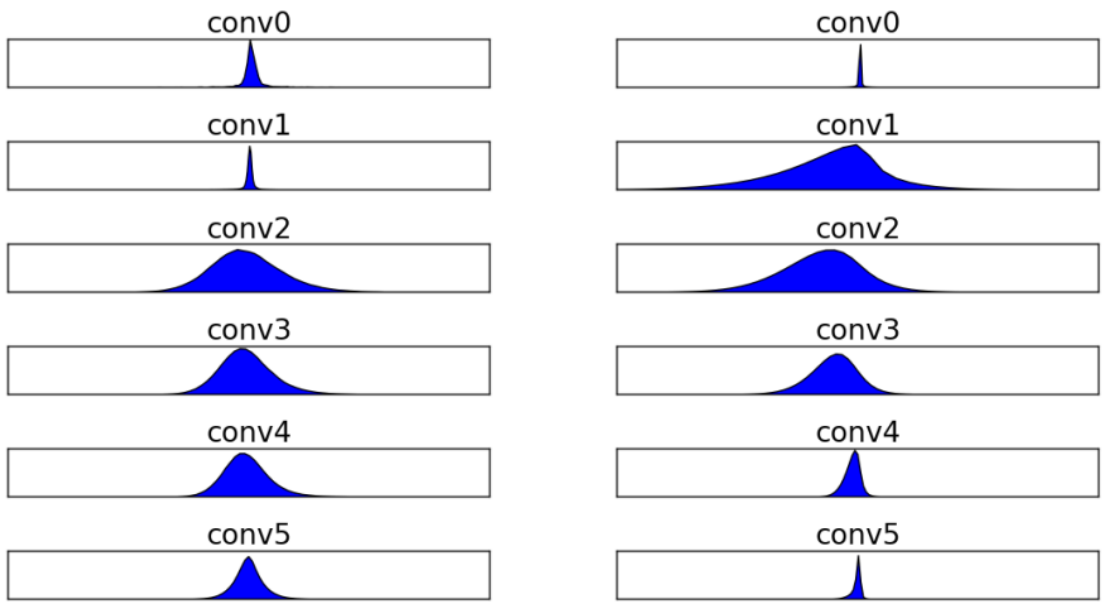
将INT8还原成32位浮点

INT8 最大的挑战是什么?

精度问题







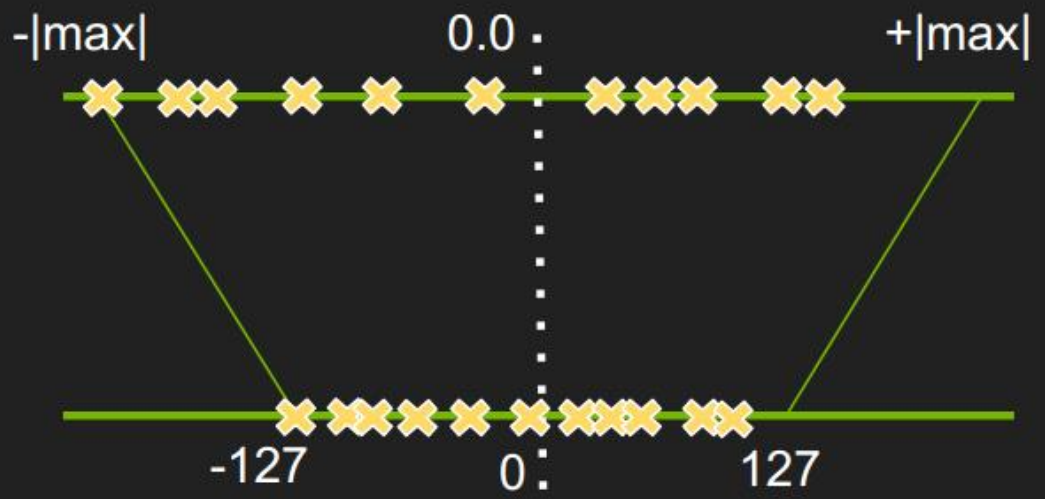
(a) Histogram of weights (b) Histogram of activations

Figure 2. Distribution of weights & activations in a DCN design for CIFAR-10 benchmark.

Lin D D, Talathi S S, Annapureddy V S. Fixed Point Quantization of Deep Convolutional Networks[J]. Computer Science, 2015.

TENSORRT INT8量化方式

- No saturation: map |max| to 127



real_value = scale * (quantized_value)

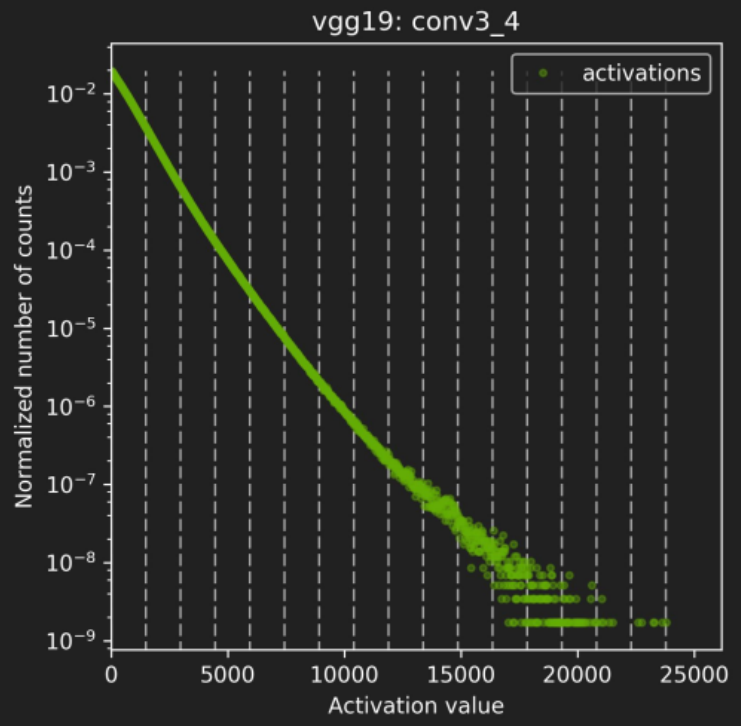
- Significant accuracy loss, in general

- Saturate above |threshold| to 127

- Weights: no accuracy improvement
- Activations: improved accuracy
- Which |threshold| is optimal?

TENSORRT INT8量化方式-解决精度问题

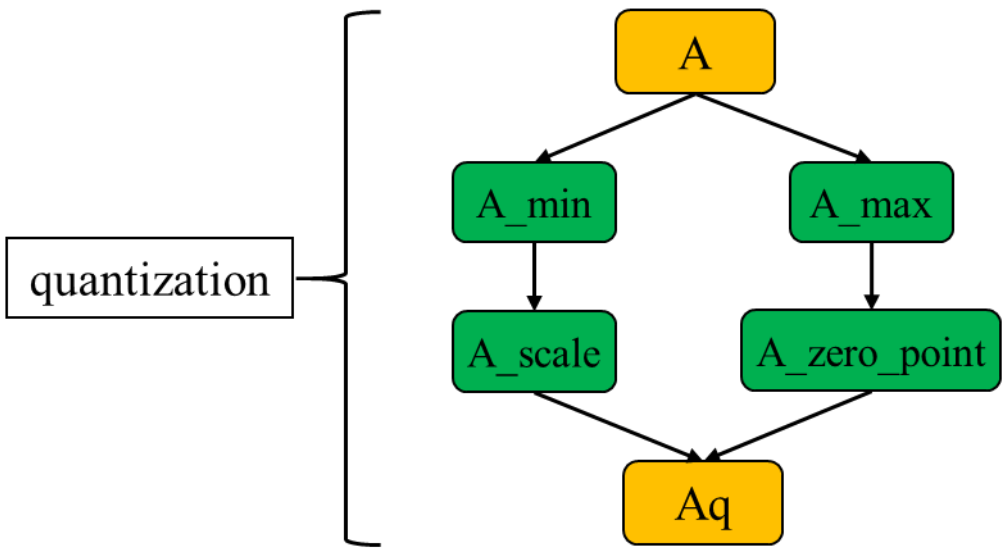
Solution: Calibration



- Run FP32 inference on Calibration Dataset.
- For each Layer:
  - collect histograms of activations.
  - generate many quantized distributions with different saturation thresholds.
  - pick threshold which minimizes KL_divergence(ref_distr, quant_distr).
- Entire process takes a few minutes on a typical desktop workstation.

TENSORFLOW INT8量化方式

real_value = scale * (quantized_value - zero_point)



优点：精度损失更低，适用范围更广

缺点：流程较复杂，损失一定的性能提升

https://petewarden.com/2016/05/03/how-to-quantize-neural-networks-with-tensorflow/

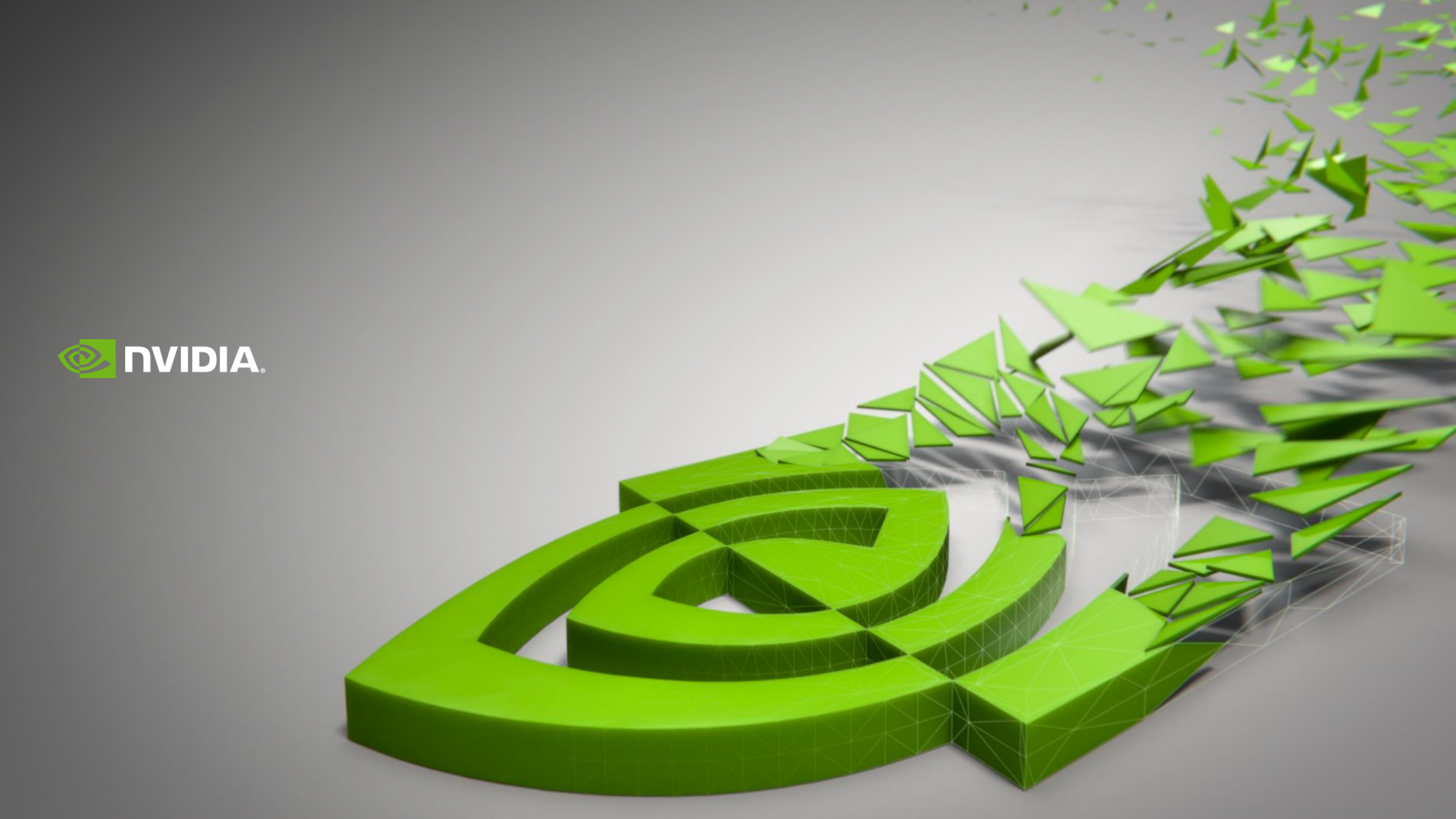
评价 NVIDIA INT8的两种使用方式

评价 NVIDIA INT8的两种使用方式

	优点	缺点
TensorRT	开发成本低 性能提升有保障	灵活性低 闭源软件且比较新，文档资料少
自己动手，丰衣足食	灵活性强，适用于复杂情况	开发成本高

TensorRT Tutorial：https://github.com/LitLeo/TensorRT_Tutorial

Source: Source information is 8 pt, italic


NVIDIA