搜狗语音交互技术中心 杨伟光

Mixed-Precision Programming with CUDA 8 - INT8

Menu

HPC&IP

- Concept && Meaning
- NVIDIA-INT8: dp4a()
- NVIDIA-INT8: cuDNN
- NVIDIA-INT8: TensorRT 2.0
- Greatest Challenge: Calibration

Concept && Meaning

HPC&IP

- Concept

32-bit: float, int

16-bit: short → FP16

8-bit: char → INT8

- Meaning

DL：在损失较小精度的前提下，减少模型大小，加速计算

- Success

Google：TPU

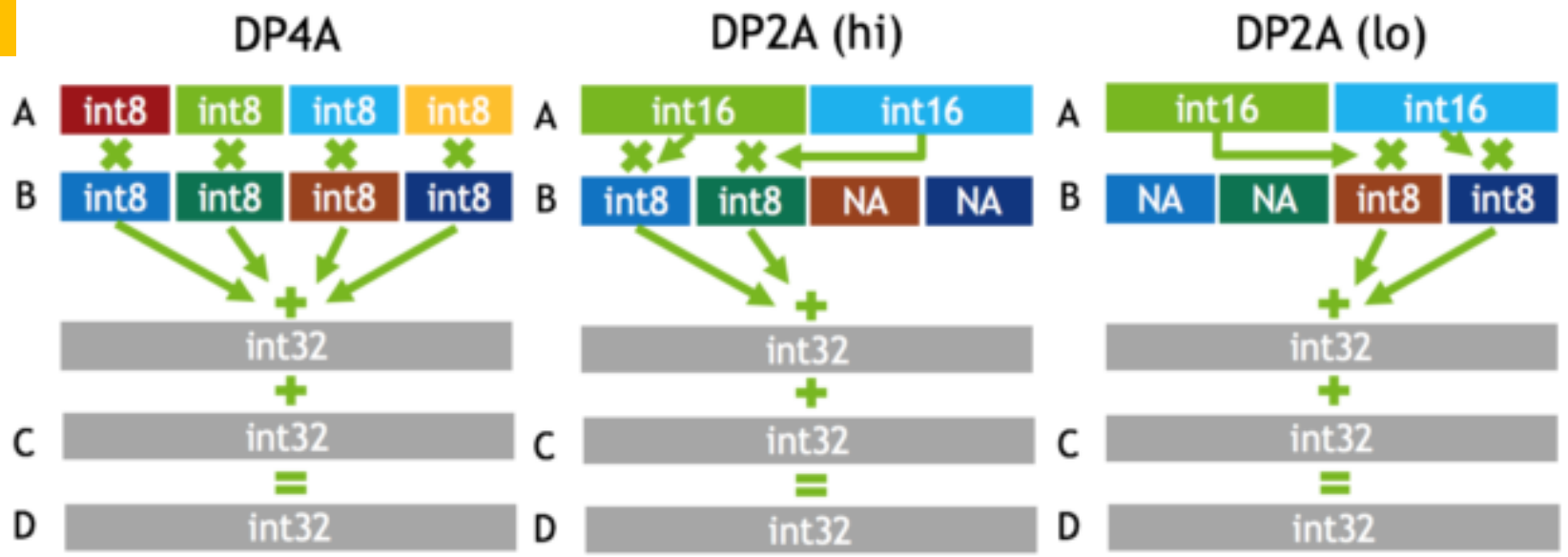
Intel： SSE

NVIDIA：P4，P40，P100

NVIDIA-INT8

GP102, GP104, or GP106 GPU

Method-1



New DP4A and DP2A instructions in Tesla P4 and P40 GPUs provide fast 2- and 4-way 8-bit/16-bit integer vector dot products with 32-bit integer accumulation.

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

NVIDIA-INT8

HPC&IP

GP102, GP104, or GP106 GPU

Method-2 CUDNN

3.16. cudnnDataType_t

cudnnDataType_t is an enumerated type indicating the data type to which a tensor descriptor or filter descriptor refers.

Value	Meaning
CUDNN_DATA_FLOAT	The data is 32-bit single-precision floating point (float).
CUDNN_DATA_DOUBLE	The data is 64-bit double-precision floating point (double).
CUDNN_DATA_HALF	The data is 16-bit floating point.
CUDNN_DATA_INT8	The data is 8-bit signed integer.
CUDNN_DATA_INT32	The data is 8-bit signed integer.
CUDNN_DATA_INT8x4	The data is 32-bit element composed of 4 8-bit signed integer. This data type is only supported with tensor format CUDNN_TENSOR_NCHW_VECT_C.

https://developer.nvidia.com/cudnn

NVIDIA-INT8

GP102, GP104, or GP106 GPU

Method-2 CUDNN

3.16. cudnnDataTy

cudnnDataType_t is an enumer
descriptor or filter descriptor refe

Value
CUDNN_DATA_FLOAT
CUDNN_DATA_DOUBLE
CUDNN_DATA_HALF
CUDNN_DATA_INT8
CUDNN_DATA_INT32
CUDNN_DATA_INT8x4

4.45. cudnnConvolutionForward

```
cudnnStatus_t
cudnnConvolutionForward( cudnnHandle_t                      handle,
                         const void                        *alpha,
                         const cudnnTensorDescriptor_t      xDesc,
                         const void                        *x,
                         const cudnnFilterDescriptor_t      wDesc,
                         const void                        *w,
                         const cudnnConvolutionDescriptor_t convDesc,
                         cudnnConvolutionFwdAlgo_t          algo,
                         void                              *workSpace,
                         size_t
 workSpaceSizeInBytes,
                         const void                        *beta,
                         const cudnnTensorDescriptor_t      yDesc,
                         void                              *y )
```

.....

https://developer.nvidia.com/cudnn

NVIDIA-INT8

HPC&IP

Method-3 **TensorRT 2.0**

Concept

- C++ library that facilitates high performance inference
- a network definition and optimizes it
- changed the name in version 2 from GIE to TensorRT

Features

- only for execution
- friendly to caffe
- friendly to Ubuntu

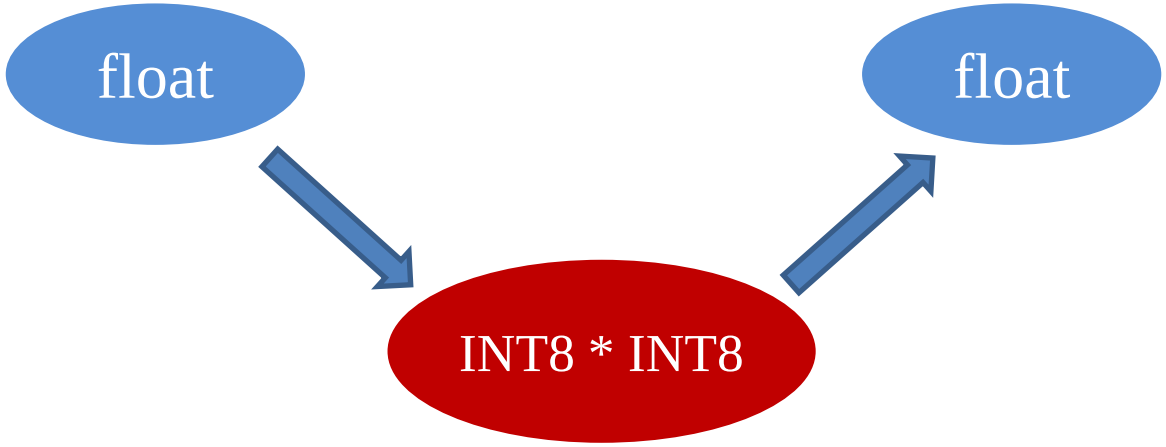
TensorRT has the following layer types:

- **Convolution**, with or without bias. Cu
  supported. **Note**: The operation this la
  are formatting weights to import via Te
- **Activation**: ReLU, tanh and sigmoid.
- **Pooling**: max and average.
- **Scale**: per-tensor, per channel or per-v
  Batch Normalization can be implemen
- **ElementWise**: sum, product or max of t
- **LRN**: cross-channel only.
- **Fully-connected** with or without bias
- **SoftMax**: cross-channel only
- **Deconvolution**, with and without bias

https://developer.nvidia.com/nvidia-tensorrt-20-download

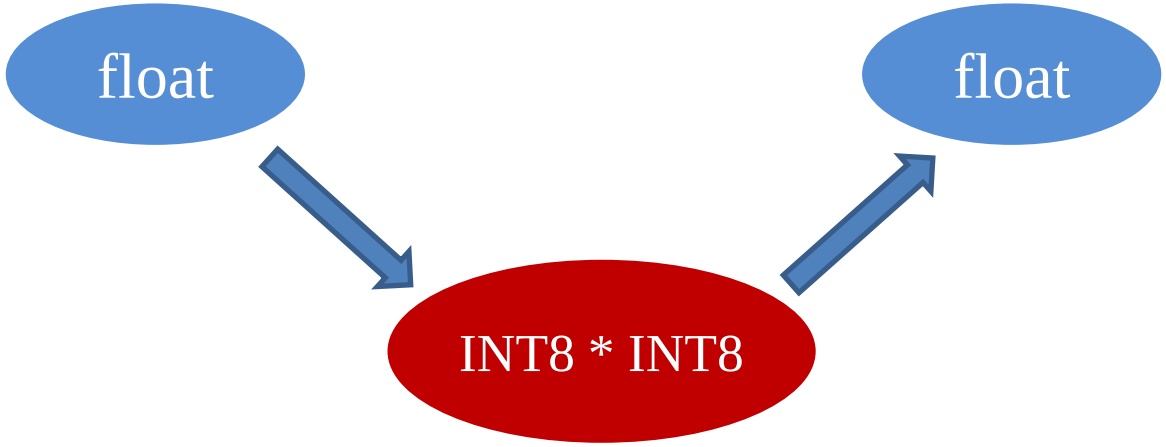
Greatest Challenge: Calibration

As small as possible the loss of precision



Greatest Challenge: Calibration

As small as possible the loss of precision



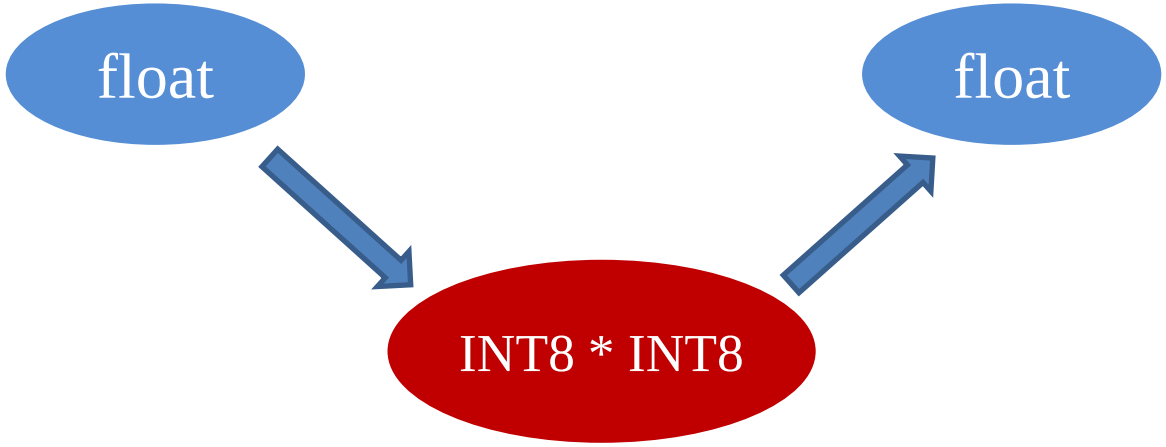
Floating-point to Fixed-point

pow(x, y)
sin/asin
log/exp

<</>>

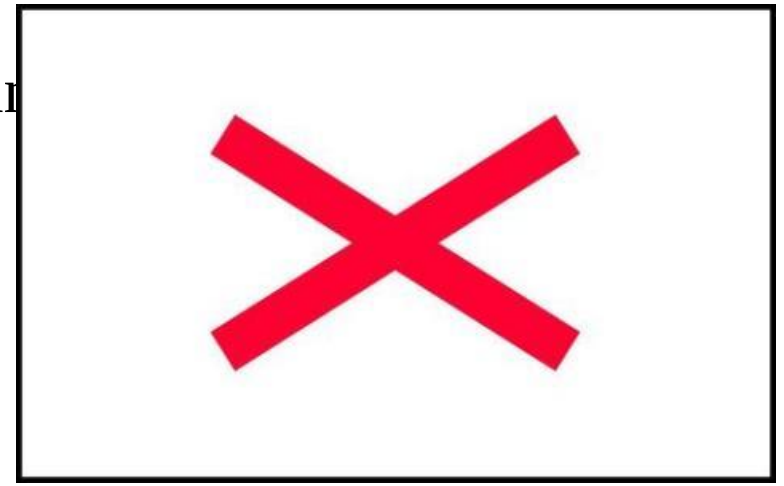
Greatest Challenge: Calibration

As small as possible the loss of precision



Floating-poi

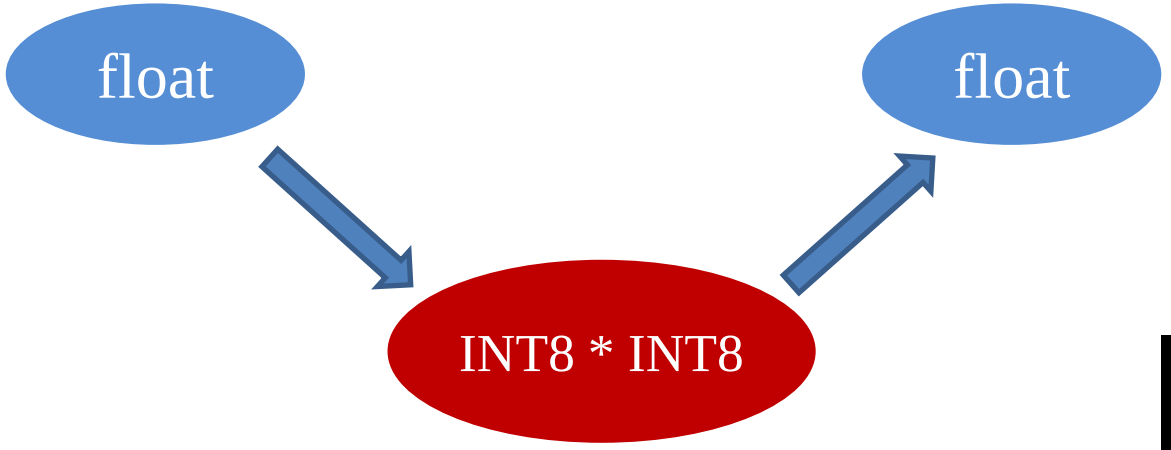
pow(x, y)
sin/asin
log/exp



<< / >>

Greatest Challenge: Calibration

As small as possible the loss of precision



<< / >>

However, INT8 only 8 bits

Retrain net using INT8

INQ问世，让深度神经网络百倍无损压缩美梦成真！
http://mp.weixin.qq.com/s/C7T1lBF-k2A_4pKUgmeJlQ

HPC&IP

Thanks