推理引擎-模型压缩

训练后量化与部署

ZOMI

Talk Overview

1. **推理系统介绍**
   - 推理系统与推理引擎区别
   - 推理工作流程
   - 推理系统介绍
   - 推理引擎介绍
2. **模型小型化**
   - 基础参数概念
   - CNN小型化结构
   - Transform小型化结构
3. **离线优化压缩**
   - 低比特量化
   - 二值化网络
   - 模型模型剪枝
   - 模型模型蒸馏
4. **部署和运行优化**
   - 图转换优化（算子融合/重排/替换）
   - 并发执行与内存分配
   - 动态batch与bin Packing

Talk Overview

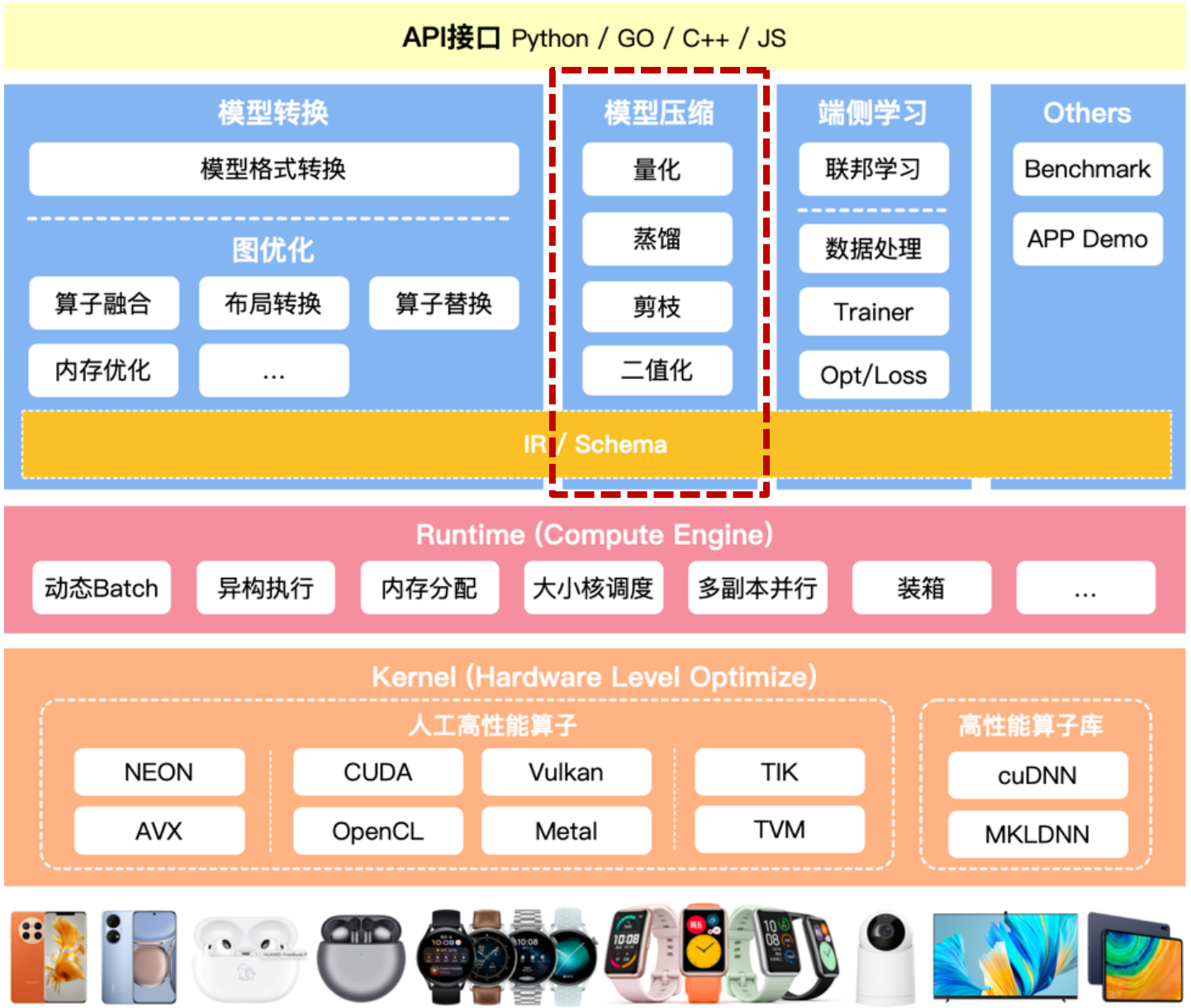
I. 低比特量化

- Base Concept of Quantization - 量化基础
- Quantization principle - 量化原理
- Quantization Aware Training - 感知量化（QAT）
- Post-Training Quantization - 训练后量化（PTQ）
- Deployment of Quantization - 量化部署

推理引擎架构

对模型进行压缩

- 减少模型大小
- 加快训练速度
- 保持相同精度



训练后量化

Post-Training Quantization, PTQ

Static/Dynamic

PTQ Dynamic

动态离线量化（Post Training Quantization Dynamic, PTQ Dynamic）

- 仅将模型中特定算子的权重从FP32类型映射成 INT8/16 类型
- 主要可以减小模型大小，对特定加载权重费时的模型可以起到一定加速效果
- 但是对于不同输入值，其缩放因子是动态计算，因此动态量化是几种量化方法中性能最差的

- 权重量化成 INT16 类型，模型精度不受影响，模型大小为原始的 1/2。
- 权重量化成 INT8 类型，模型精度会受到影响，模型大小为原始的 1/4。

PTQ Dynamic 算法流程

PTQ Static

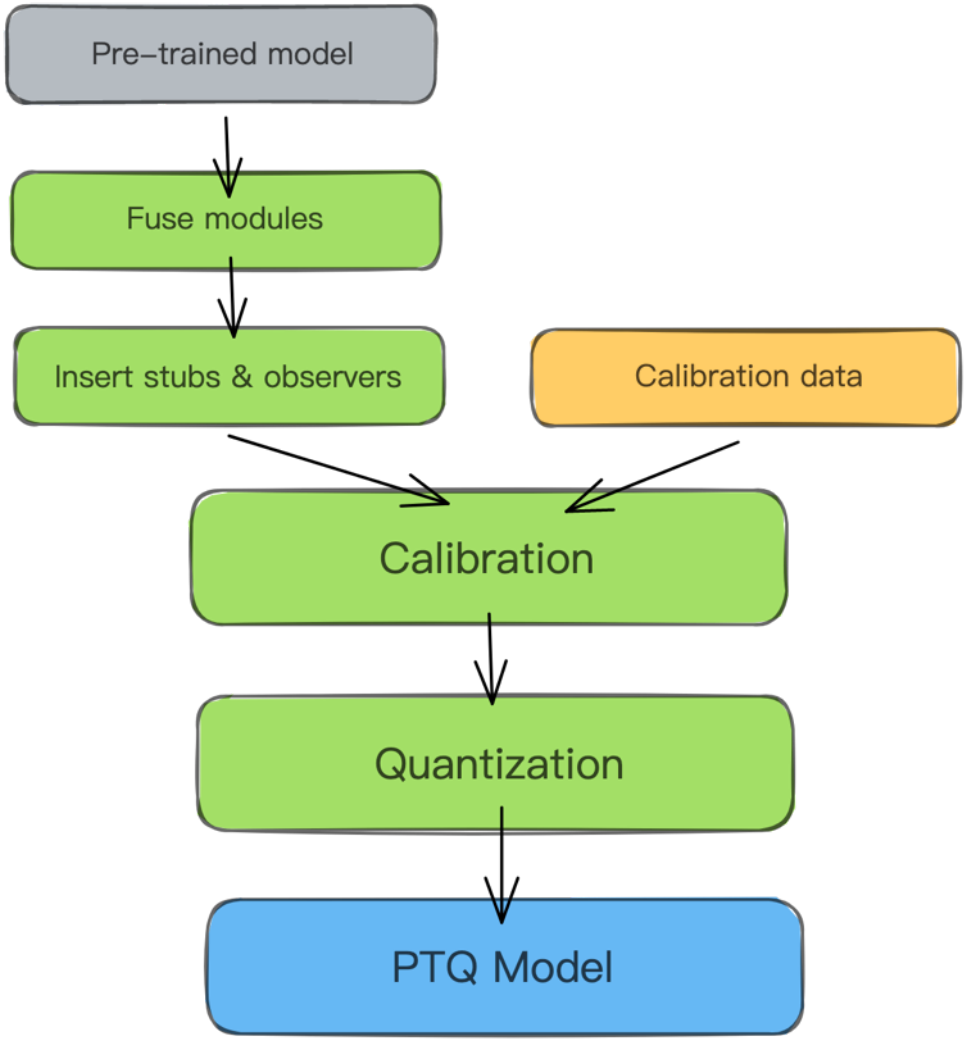
静态离线量化（Post Training Quantization Static, PTQ Static）

- 同时也称为校正量化或者数据集量化。使用少量无标签校准数据，核心是计算量化比例因子。使用静态量化后的模型进行预测，在此过程中量化模型的缩放因子会根据输入数据的分布进行调整。

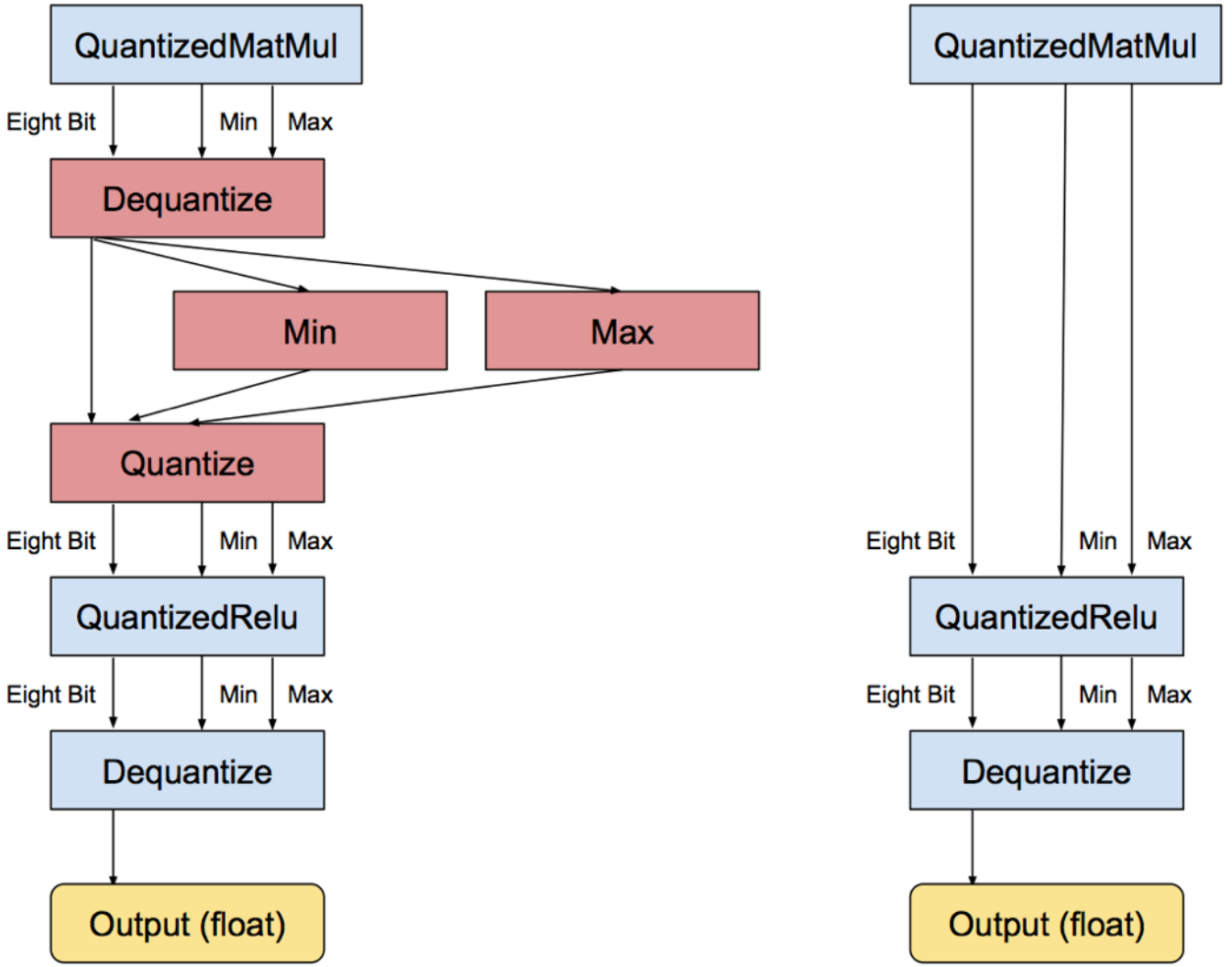
$$uint8 = round(float/scale) - offset$$

- 静态离线量化的目标是求取量化比例因子，主要通过对称量化、非对称量化方式来求，而找最大值或者阈值的方法又有MinMax、KLD、ADMM、EQ等方法

PTQ Static 算法流程



PTQ Static 算法流程



KL散度校准法：原理

- KL散度校准法也叫相对熵，其中p表示真实分布，q表示非真实分布或p的近似分布：

$$D_{KL}(P_f \parallel Q_q) = \sum_{i=1}^{N} P(i) * log_2 \frac{P_f(i)}{Q_q(i)}$$

- 相对熵，用来衡量真实分布与非真实分布的差异大小。目的就是改变量化域，实则就是改变真实的分布，并使得修改后得真实分布在量化后与量化前相对熵越小越好。

KL散度校准法：流程

1. 选取validation数据集中一部分具有代表的数据作为校准数据集 Calibration
2. 对于校准数据进行FP32的推理，对于每一层
    1. 收集activation的分布直方图
    2. 使用不同的threshold来生成一定数量的量化好的分布
    3. 计算量化好的分布与FP32分布的KL divergence，并选取使KL最小的threshold作为saturation的阈值

通俗地理解，算法收集激活Act直方图，并生成一组具有不同阈值的8位表示法，选择具有最少kl散度的表示；此时的 kl 散度在参考分布（FP32激活）和量化分布之间（即8位量化激活）之间。

KL散度校准法：流程

- 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_divergence, quant_divergence).

- 需要准备小批量数据（500~1000张图片）校准用的数据集；
- 使用校准数据集在FP32精度的网络下推理，并收集激活值的直方图；
- 不断调整阈值，并计算相对熵，得到最优解

KL散度校准法：实现

Input: FP32 histogram H with 2048 bins: bin[0], …, bin[2047]

For i in range(128, 2048):
    reference distribution P = [bin[0], …, bin[i-1]]
    outliers count = sum(bin[i], bin[i+1], …, bin[2047])
    reference distribution P[i-1] += outliers count
    P /= sum(P)
    candidate distribution Q = quantize [bin[0], …, bin[i-1]] into 128 levels
    expand candidate distribution Q to I bins
    Q /= sum(Q)
    divergence[i] = KL divergence(reference distribution P, candidate distribution Q)
End For

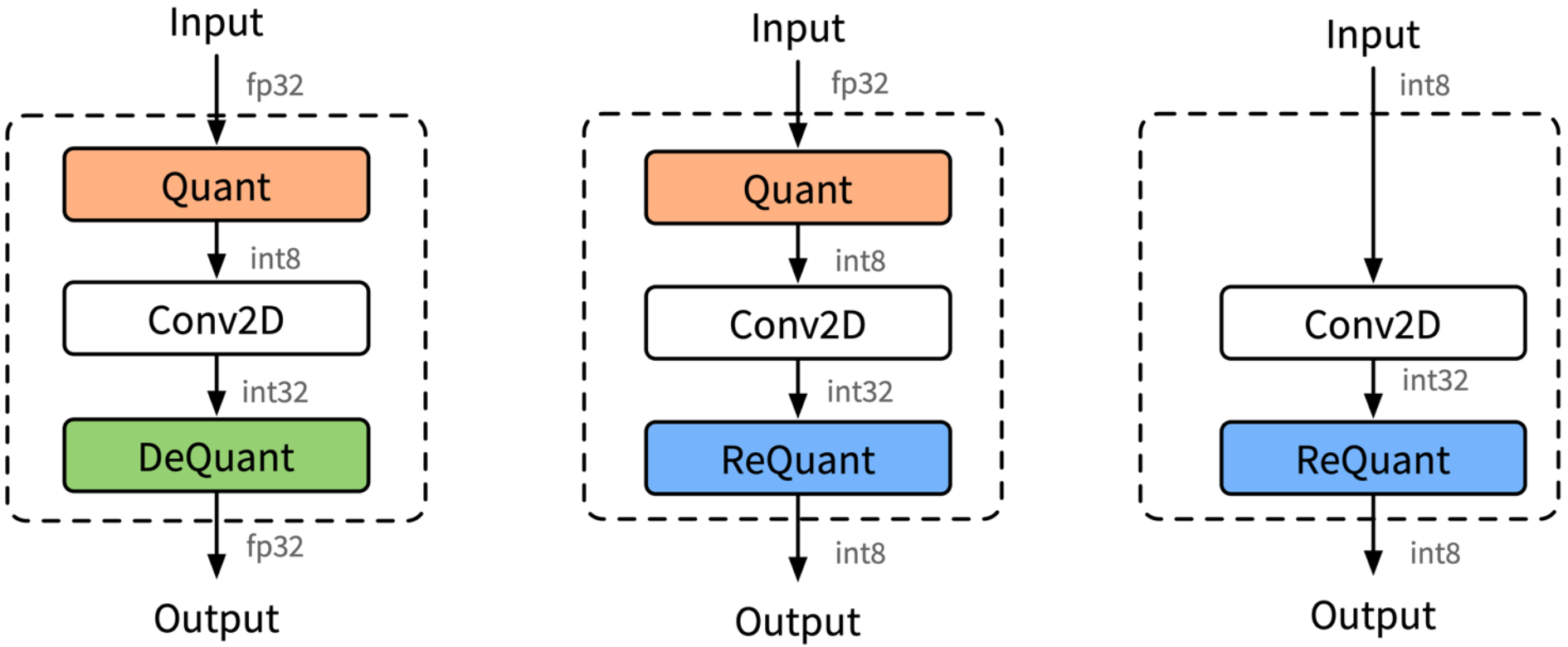
Find index m for which divergence[m] is minimal

threshold = (m+0.5) * (width of a bin)

端侧量化推理部署

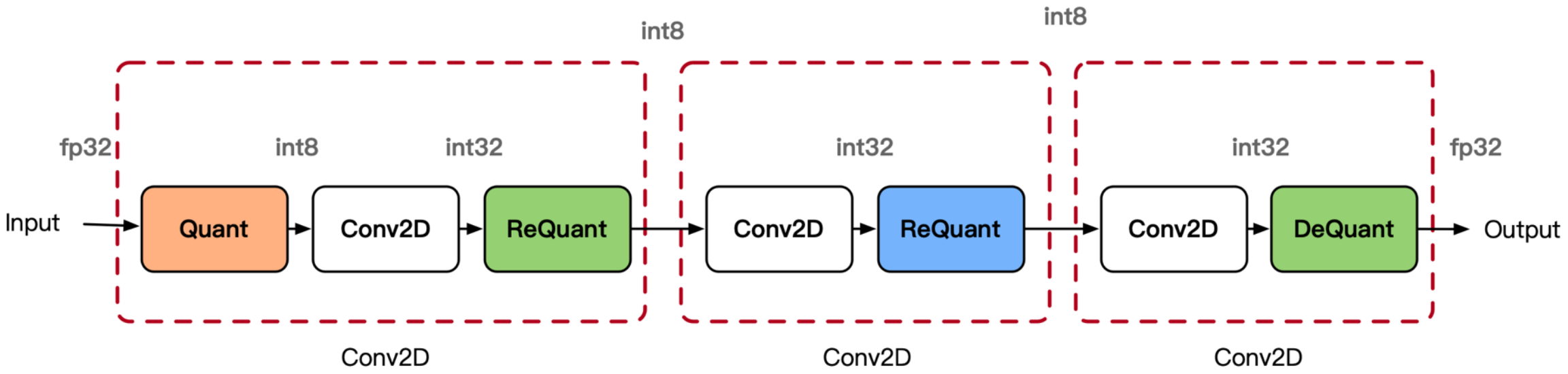
端侧量化推理部署

- 端侧量化推理的结构方式主要由3种，分别是下图 (a) Fp32输入Fp32输出、(b) Fp32输入int8输出、(c) int8输入int32输出



端侧量化推理部署

- INT8卷积示意图，里面混合里三种不同的模式，因为不同的卷积通过不同的方式进行拼接。使用INT8进行inference时，由于数据是实时的，因此数据需要在线量化，量化的流程如图所示。数据量化涉及Quantize，Dequantize和Requantize等3种操作：



Quantize量化

- 将float32数据量化为int8。离线转换工具转换的过程之前，根据量化原理的计算出数据量化需要的scale和offset：

$$scale = (x_{\max} - x_{\min}) / (Q_{\max} - Q_{\min})$$

$$offset = Q_{\min} - round(x_{\min} / scale)$$

$$uint8 = round(float / scale) - offset$$

$$float = scale \times (uint + offset)$$

Dequantize反量化

- INT8相乘、加之后的结果用INT32格式存储，如果下一Operation需要float32格式数据作为输入，则通过Dequantize反量化操作将INT32数据反量化为float32。Dequantize反量化推导过程如下：

$$
\begin{aligned}
y &= x \cdot w \\
&= x_{scale} \cdot \left(x_{int} + x_{offset}\right) \cdot w_{scale} \cdot \left(w_{int} + w_{offset}\right) \\
&= \left(x_{scale} * w_{scale}\right) \cdot \left(x_{int} + x_{offset}\right) \cdot \left(w_{int} + w_{offset}\right) \\
&= \left(x_{scale} \cdot w_{scale}\right) \cdot \left(x_{int} \cdot w_{int} + x_{int}x_{offset} + w_{int}x_{offset} + w_{offset}\, x_{offset}\right) \\
&= \left(x_{scale} \cdot w_{scale}\right) \cdot \left(INT32_{result} + x_{int}x_{offset} + w_{int}x_{offset} + w_{offset}\, x_{offset}\right) \\
&\approx \left(x_{scale} \cdot w_{scale}\right) \cdot INT32_{result}
\end{aligned}
$$

Requantize重量化

- INT8乘加之后的结果用INT32格式存储，如果下一层需要INT8格式数据作为输入，则通过Requantize重量化操作将INT32数据重量化为INT8。重量化推导过程如下：

$$
\begin{aligned}
y &= x \cdot w \\
&= x_{scale} \cdot (x_{int} + x_{offset}) \cdot w_{scale} \cdot \left(w_{int} + w_{offset}\right) \\
&= (x_{scale} \cdot w_{scale}) \cdot (x_{int} + x_{offset}) \cdot (w_{int} + w_{offset}) \\
&= (x_{scale} \cdot w_{scale}) * INT32\_result
\end{aligned}
$$

Requantize重量化

- 其中 y 为下一个节点的输入，即 $y = x_{next}$：

$$y_{int} = y_{scale} * (y_{int} + y_{offset})$$

- 有：

$$x_{next\ int} = \left(x_{scale} \cdot w_{scale} / x_{next\ scale}\right) \cdot INT32_{result} - x_{next\ offset}$$

- 因此重量化需要本Operation输入input和weight的scale，以及下一Operation的input输入数据的scale和offset。

Question?

1. 为什么模型量化技术能够对实际部署起到加速作用？
2. 为什么需要对网络模型进行量化压缩？
3. 为什么不直接训练低精度的模型？（大模型呢？）
4. 什么情况下不应该/应该使用模型量化？

参考文献

1. Learning Accurate Low-Bit Deep Neural Networks with Stochastic Quantization
- Differentiable Soft Quantization: Bridging Full-Precision and Low-Bit Neural Networks（ICCV 2019）
- IR-Net: Forward and Backward Information Retention for Highly Accurate Binary Neural Networks（CVPR 2020）
- Towards Unified INT8 Training for Convolutional Neural Network（CVPR 2020）
- Rotation Consistent Margin Loss for Efficient Low-bit Face Recognition（CVPR 2020）
- DMS: Differentiable diMension Search for Binary Neural Networks（ICLR 2020 Workshop）
- Nagel, Markus, et al. "A white paper on neural network quantization." *arXiv preprint arXiv:2106.08295* (2021).
- Krishnamoorthi, Raghuraman. "Quantizing deep convolutional networks for efficient inference: A whitepaper." arXiv preprint arXiv:1806.08342 (2018)
- Gholami, A., Kim, S., Dong, Z., Yao, Z., Mahoney, M. W., & Keutzer, K. (2021). A survey of quantization methods for efficient neural network inference. arXiv preprint arXiv:2103.13630.
- Wu, H., Judd, P., Zhang, X., Isaev, M., & Micikevicius, P. (2020). Integer quantization for deep learning inference: Principles and empirical evaluation. arXiv preprint arXiv:2004.09602.
- 全网最全-网络模型低比特量化 https://zhuanlan.zhihu.com/p/453992336