Quantization for TVM

Ziheng Jiang

TVM Conference, Dec 12th 2018

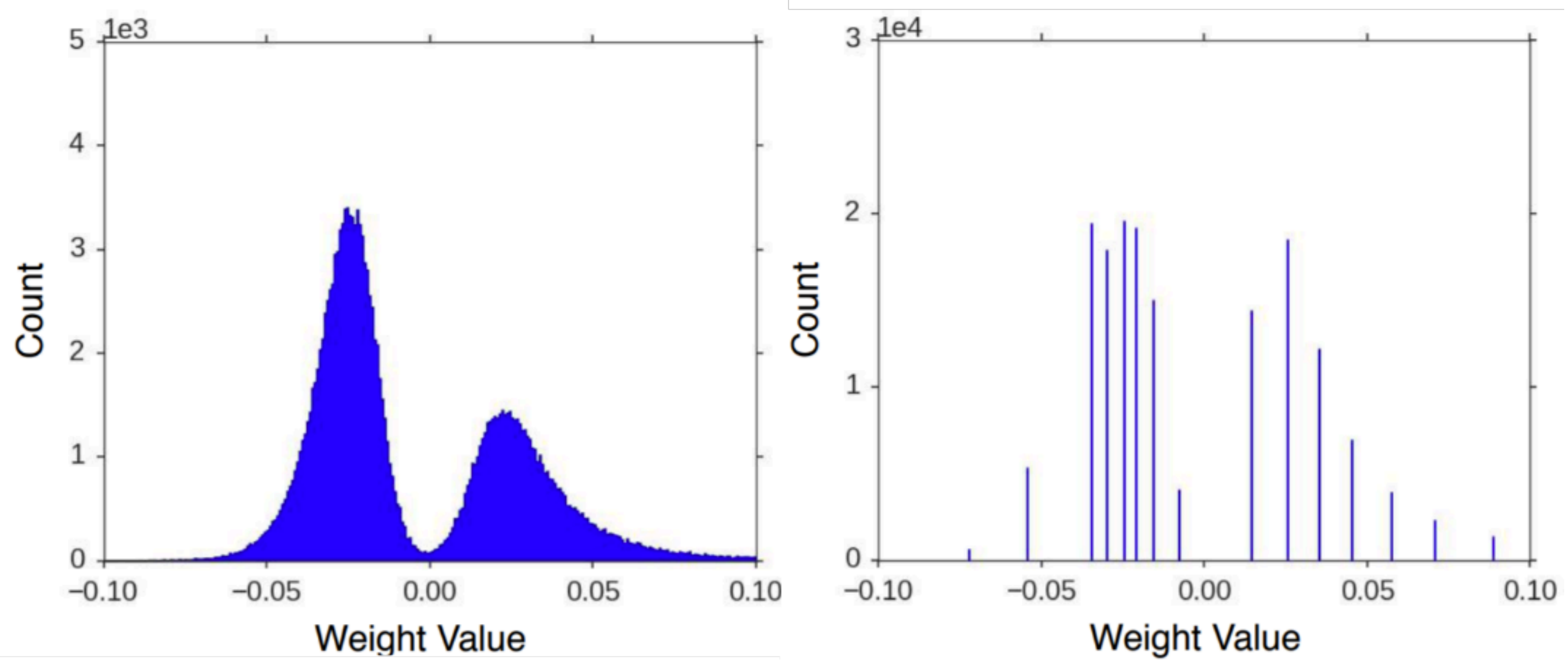
sampl

PAUL G. ALLEN SCHOOL
OF COMPUTER SCIENCE & ENGINEERING

tvm.ai

Quantization for TVM

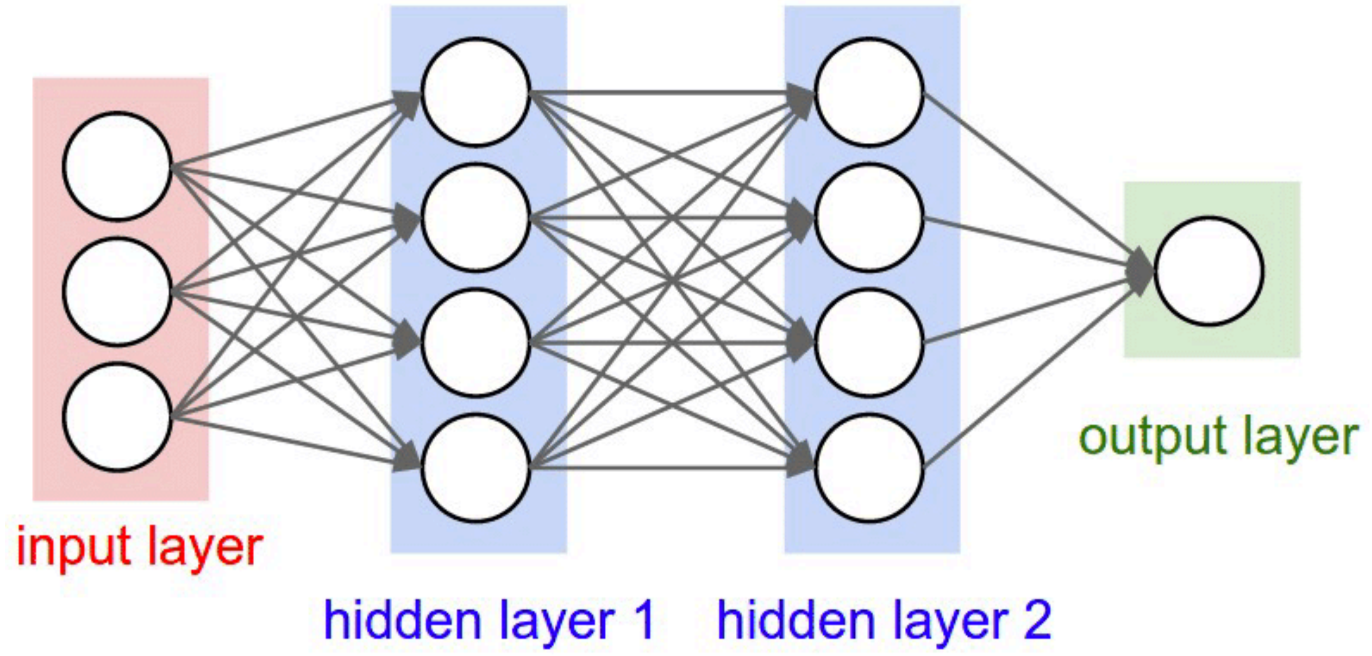
What is Quantization?



source: Han et al

Converting weight value to low-bit integer like 8bit precision from float-point without significant accuracy drop.

Quantization for TVM



Train →

Frontend DL Framework

↓ Convert

Relay: High-Level Graph IR

↓ Apply

Quantization

↓ Deploy

Gain Compression & Acceleration:

- Less storage space
- Faster arithmetic operation
- Friendly to accelerator and ultra low-power embedded devices

Quantization for TVM

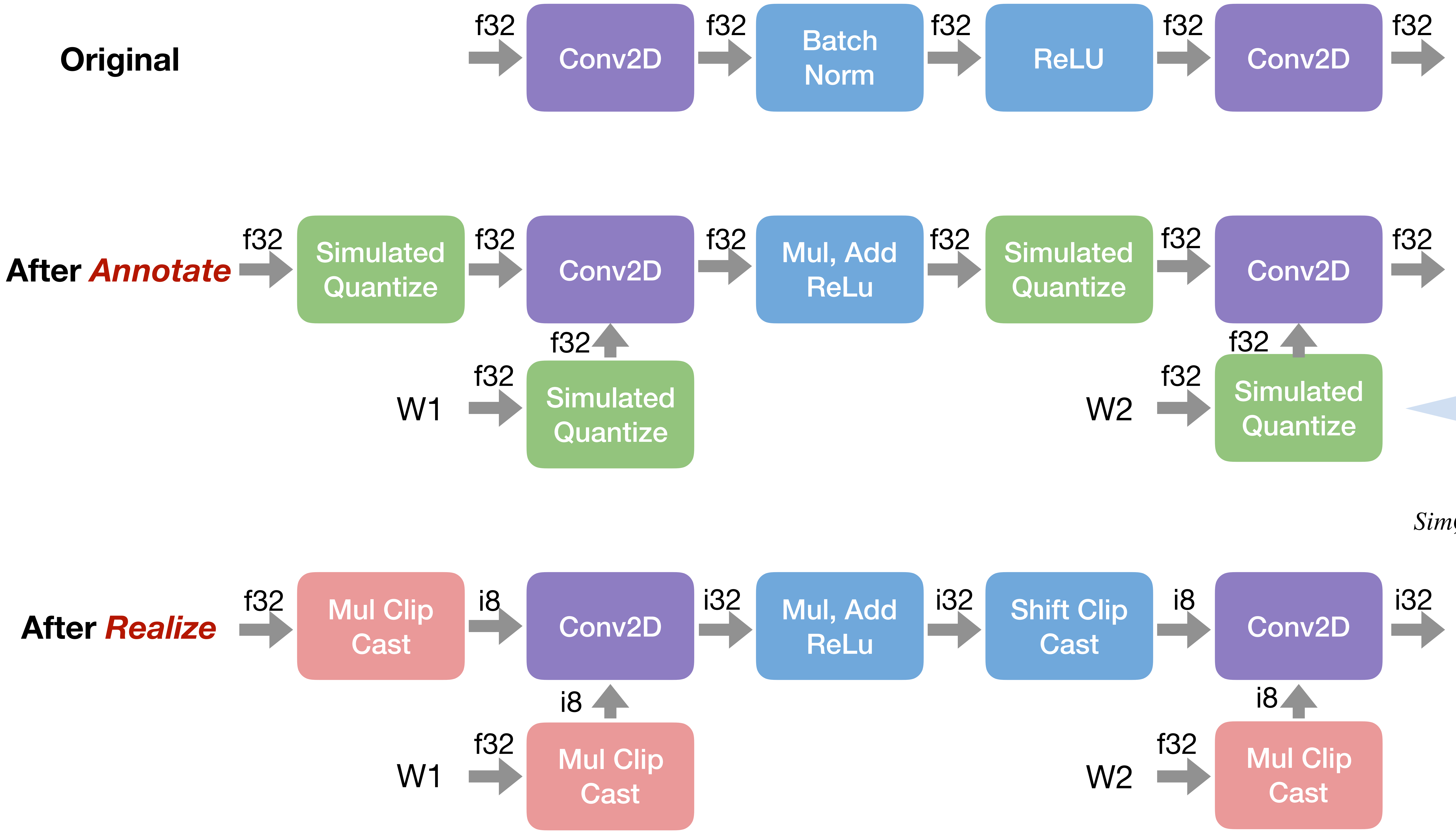
Choice Spaces for Quantization

- number of bit
  - 4bit, 8bit, 16bit
- quantization scheme:
  - symmetric, asymmetric, etc.
- hardware constraint:
  - e.g. prefer integer shift instead of float multiplication

Goal

Instead of proposing "the only right way to achieve quantization in TVM", we would like to build ***a quantization workflow which can be customized flexibly***.

Quantization for TVM



SimQ simulates the rounding error and saturating error during quantizing. Its argument will get tuned during ***calibrate***.

$$SimQ(nbit, range, sign) = \frac{Clip(Round(\frac{x}{r} * 2^{nbit-sign})) * r}{2^{nbit-sign}}$$

Quantization for TVM

Code Sample

```python
# user can override the annotate function
@register_annotate_function("nn.conv2d", override=True)
def annotate_conv2d(ref_call, new_args, ctx):
    lhs, rhs = new_args
    lhs = attach_simulated_quantize(lhs, sign=False, rounding='round')
    rhs = attach_simulated_quantize(lhs, sign=False, rounding='stochastic_round')
    return expr.Call(ref_call.op, [lhs, rhs], ref_call.attrs)

# assuming we have an existed mxnet model, convert it to relay graph
graph, params = relay.frontend.from_mxnet(mxnet_model)

# quantize the relay graph with all kinds of configure
with qconfig(nbit_dict={QFieldKind.ACTIVATION: 24}, global_scale=8.0, skip_k_conv=1):
    qgraph, qparams = quantize(graph, params)

# ...build and deploy it locally or remotely with tvm
```

Quantization for TVM

Demonstration with 8bit Symmetric Quantization

Global Scale	Accuracy
2.0	64.1%
4.0	68.1%
8.0	69.5%
16.0	69.6%

Accuracy Drop with ResNet18 (original 70.8%)

Time/ms	Cortex A53	VTA
ResNet18	307.09	64.87
MobileNet	131.14	51.96

End to End Performance