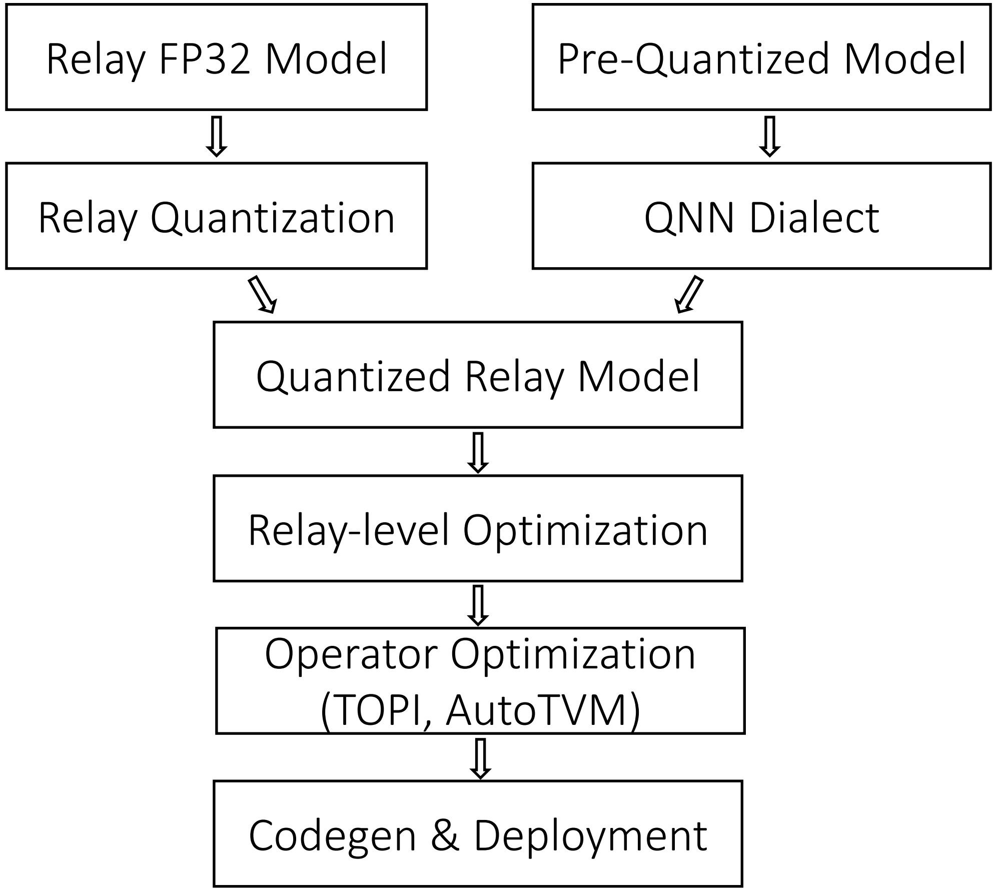
Efficient Quantized Inference on CUDA with TVM

Wuwei Lin

TVM Conference, Dec 5, 2019

Carnegie Mellon University

Quantization in TVM



Two modes of quantization

- Relay quantization pass to convert FP32 model in Relay IR
- QNN dialect to import pre-quantized models from other framework

Unified optimizations for quantized models

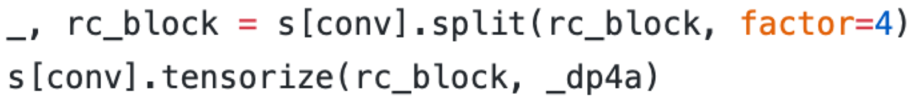
- Relay-level optimization
- Tensor-level operator optimization

Optimizing Quantized Operators

- Utilizing hardware intrinsics via tensorization (DP4A, Tensor Cores)
- Packed layout (NCHW -> NCHW4c, OIHW -> OIHW4o4i)
- Automatic optimization with AutoTVM

Optimizing Quantized Operators

- Utilizing hardware intrinsics via tensorization (DP4A, Tensor Cores)
- Packed layout (NCHW -> NCHW4c, OIHW -> OIHW4o4i)
- Automatic optimization with AutoTVM



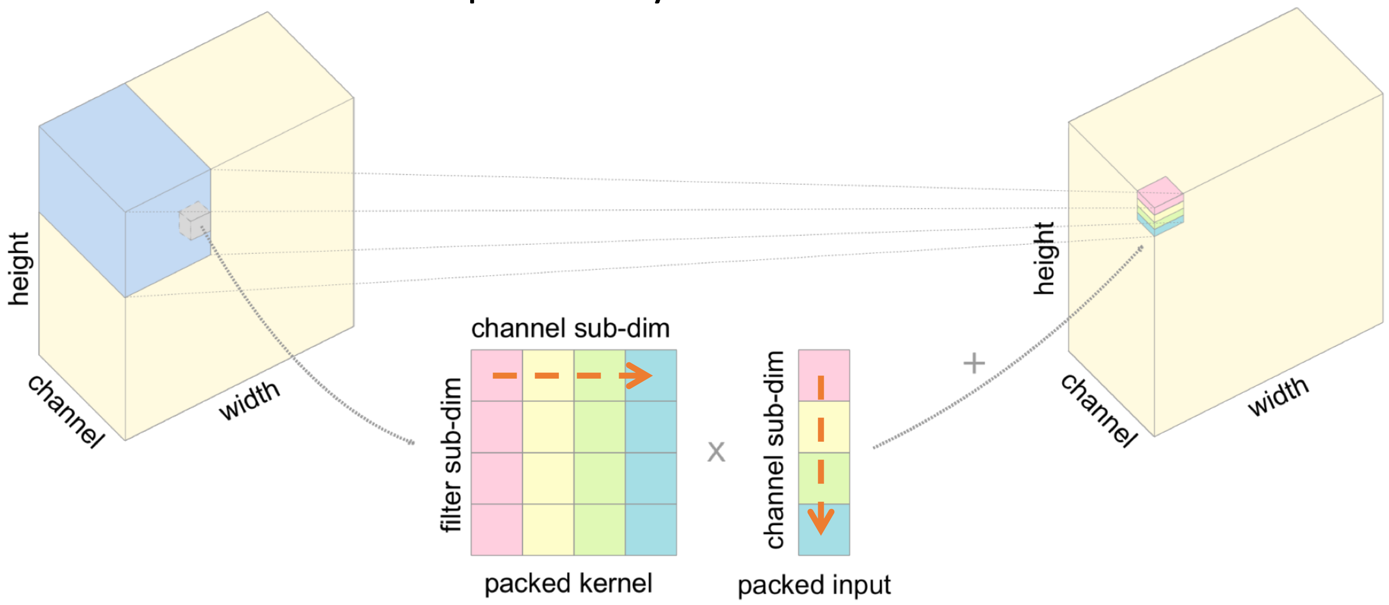
```
_, rc_block = s[conv].split(rc_block, factor=4)
s[conv].tensorize(rc_block, _dp4a)
```

Optimizing Quantized Operators

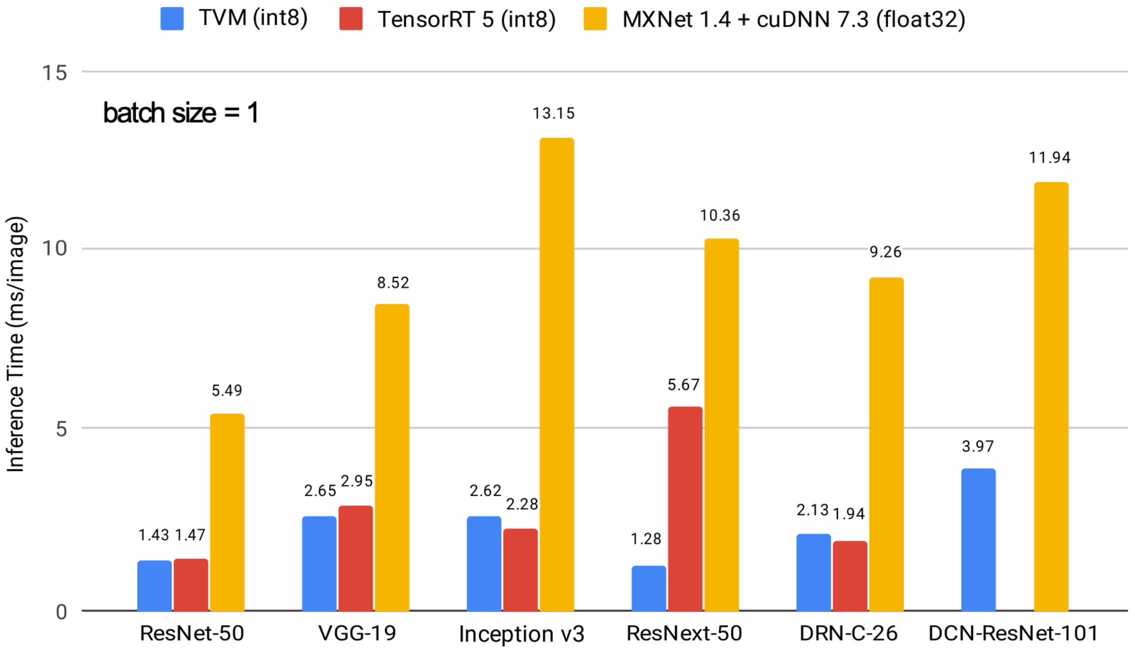
- Utilizing hardware intrinsics via tensorization (DP4A, Tensor Cores)
- Packed layout (NCHW -> NCHW4c, OIHW -> OIHW4o4i)
- Automatic optimization with AutoTVM

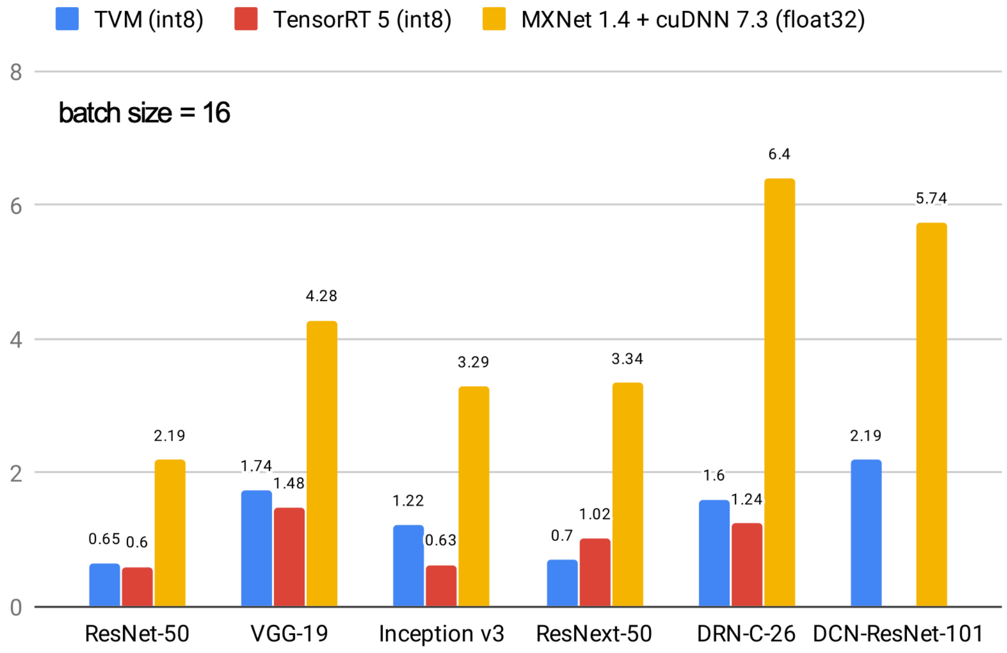
```
_, rc_block = s[conv].split(rc_block, factor=4)
s[conv].tensorize(rc_block, _dp4a)
```

Conv2d with DP4A and packed layout



Benchmark on NVIDIA 1080ti





https://tvm.apache.org/2019/04/29/opt-cuda-quantized

Summary and Future Work

- We achieved competitive performance with joint optimizations from Relay and tensor expression level.
- Working on improving model coverage and calibration schemes.
- Feedback and contribution are welcomed!