Efficient Quantized Inference on CUDA with TVM

Wuwei Lin

TVM Conference, Dec 5, 2019
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
Optimizing Quantized Operators

• Utilizing hardware intrinsics via tensorization (DP4A, Tensor Cores)
• Packed layout (NCHW -> NCHW4c, OIHW -> OIHW4o4i)
• Automatic optimization with AutoTVM
Benchmark on NVIDIA 1080ti

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!