TVM Operator Fusion

2025-12-04 김규진

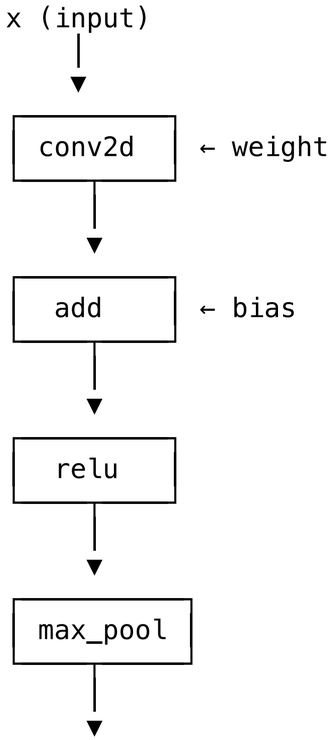
Tuning Pipeline

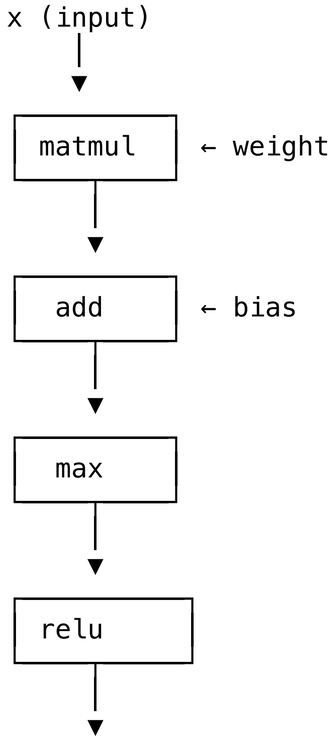
- tune_TIR (relax, tir)
- tune_task (task)
- tune_schedule (tune_context 조율)
- tune_context (op tune)

Fusion PIPELINE

- Create Relax IR
- DAG Structure
- Segment Operation
- Select Fusion Group
- TIR Code gen (Middle IR)
- Meta Tuning (MetaSchedule Option)

DAG Structure





State

- Injective
- Reduction
- Complex-out-fusible
- Opaque

Rule

- Injective -> Injective : 항상 fusion 가능
- Complex-out-fusible -> Injective : Fusion 가능 (conv2d + relu)
- Reduction -> Injective : Fusion 가능 (pooling + relu)
- Injective -> Reduction : Fusion 가능

Group

- Group 1 (covn2d + add + relu)
- Group 2 (max_pool2d)
- Group3 (conv2d + add + relu) <- Conv2 Block
- Group4 (reshape) <- flatten
- Group 마다 TIR 코스 생성

Schedule

- Tiling (cache)
- Vectorization (sims)
- Parallelization (multi-core)
- Memory hierarchy (shared memory, local memory)

- -> Targeting (LLVM , CUDA)

TODO

- Create Relax Model Graph
- Weight Override
- File Structure