Project Title: ML Computation Graph Operator Fission + Fusion

Group Info: Ashwin Venkatram [ashwinve@andrew.cmu.edu], Gabriele Oliaro [goliaro@andrew.cmu.edu]

Project Description:

Research Question: Can operator fission before fusion optimization pass expose new fusion opportunities that can reduce overall number of kernels required for E2E Deep Learning Computation Graph execution on GPU?

Measurement metric: Latency benchmark of sub-graph test case execution on GPU

What You Have Accomplished So Far

- Selected Softmax and InstanceNorm for operator fission
- Implemented op-fission using ONNX graphsurgeon using python api for simplicity of testing
    - Required special handling of ReduceMax by transforming as a Matmul
    - Enable fusion of Matmuls due to above
- Integrated with TASO for graph transformations on sub-graph/ E2E deep learning model optimization
- TVM is able to now reduce number of tasks for kernel codegen due to above transformations
- Initial results:
    - Linear attention sub-block: 2X speed-up over TensorRT (already optimized for attention)
    - Softmax op fission: 1.11X speed-up over TensorRT (without TASO graph optimization)

Meeting Your Milestone:

- Yes we are on track. Upcoming tasks:
    - Benchmarking and ablation study

Surprises:

- No major surprises

Revised Schedule:

- We will be working on benchmarking and finding more test-cases to test the innovation

Resources Needed:

- Yes no change