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

Logistics:

- Plan & Schedule:
    - **20 March:** Examine operators such as Softmax for opportunities for fission
    - **27 March:** Explore techniques and tools for ONNX graph transformation to implement operator fission
    - **3 April:** additional week for operator fission and developing transformation pass with ONNX graph tools
    - **[Milestone] 10 April:** Apply graph transformations for efficient operation fusion (Leverage existing work from TASO)
        - We expect to hit 75% goal here
        - Expect to have working infrastructure to perform operator fission + op fusion with TASO by this time
    - **17 April:** Leverage TVM for kernel generation and benchmarking on GPU
    - **24 April:** Further benchmarking, ablation studies and evaluation
    - **Extension/ Reach Goal:**
        - Investigate AOT compilation of TVM kernels for efficient runtime GPU resource utilization
        - Examine TVM code-gen for host and device side for data persistence on GPU. Investigate if TVM kernel stitching for binary generation is doing any further optimizations. Profile kernels on GPU for HW utilization and performance numbers.
        - Explore techniques for parallel kernel orchestration, especially for memory-bound sub-graphs. Select a couple interesting cases.
- Literature Search: See References below
- Resources Needed:
    - Example test case subgraphs with reduction nodes on which to perform operator fission + fusion
    - Utilize existing Catalyst AWS server for this project
- Getting Started:
    - Leveraging existing research work in Catalyst Lab for this class

References

- Tianqi Chen, Thierry Moreau, Ziheng Jiang, Haichen Shen, Eddie Q. Yan, Leyuan Wang, Yuwei Hu, Luis Ceze, Carlos Guestrin, and Arvind Krishnamurthy. TVM: end-to-end optimization stack for deep learning. CoRR, abs/1802.04799, 2018
- Zhihao Jia, Oded Padon, James Thomas, Todd Warsza-wski, Matei Zaharia, and Alex Aiken. Taso: optimizing deep learning computation with automatic generation of graph substitutions. In Proceedings of the 27th ACM Symposium on Operating Systems Principles, pages 47–62, 2019.
- Wei Niu, Jiexiong Guan, Yanzhi Wang, Gagan Agrawal, and Bin Ren. Dnnfusion: accelerating deep neural networks execution with advanced operator fusion. In Proceedings of the 42nd ACM SIGPLAN International Conference on Programming Language Design and Implementation, pages 883–898, 2021.
- Junru Shao, Xiyou Zhou, Siyuan Feng, Bohan Hou, Ruihang Lai, Hongyi Jin, Wuwei Lin, Masahiro Masuda, Cody Hao Yu, and Tianqi Chen. Tensor program optimization with probabilistic programs, 2022.