Optimization Mapping for Deep Learning

Wencong Xiao†*, Cheng Chen*, Youshan Miao*, Jilong Xue*, Ming Wu*
†Beihang University, *Microsoft Research

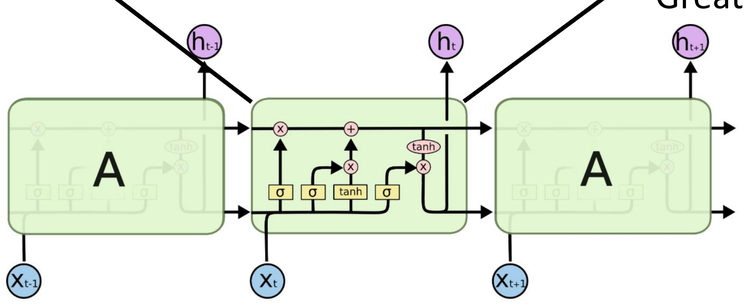
Motivating Scenario

Common process for deep learning

Conv. Matmul Sigmoid Mul Tanh ReLu TensorFlow

Model prototype

```
…
new_c = c * sigmoid(f + 1.0) +
        sigmoid(i) + tanh(j)
new_h = tanh(new_c) * sigmoid(o)
…
```



Great model for industry!

Bottlenecks!

resolve it with highly optimized operator!

Hardware trend: heterogeneous devices with various accelerated libs

- Custom scenarios: cloud, mobile
- Custom algorithms: CNN, RNN

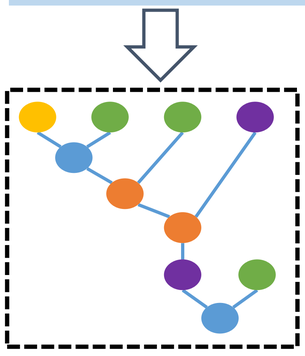
GPU-based: nvidia cuDNN

FPGA-based: Microsoft Brainwave

ASIC-based: Google TPU, Intel Nervana, Cambricon, …

Optimization Mapping

```
…
new_c = c * sigmoid(f + 1.0) +
        sigmoid(i) + tanh(j)
new_h = tanh(new_c) * sigmoid(o)
…
```

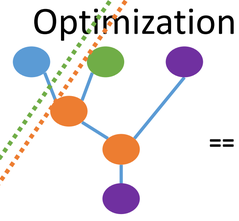


Original dataflow graph

Data scientist

Hardware/System engineer

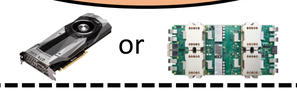
Optimization Mapping Module



==

OptimizedOp.

or



OptimizedOp.

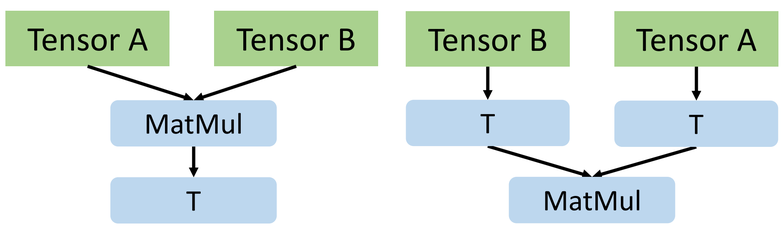
New dataflow graph

Key insights

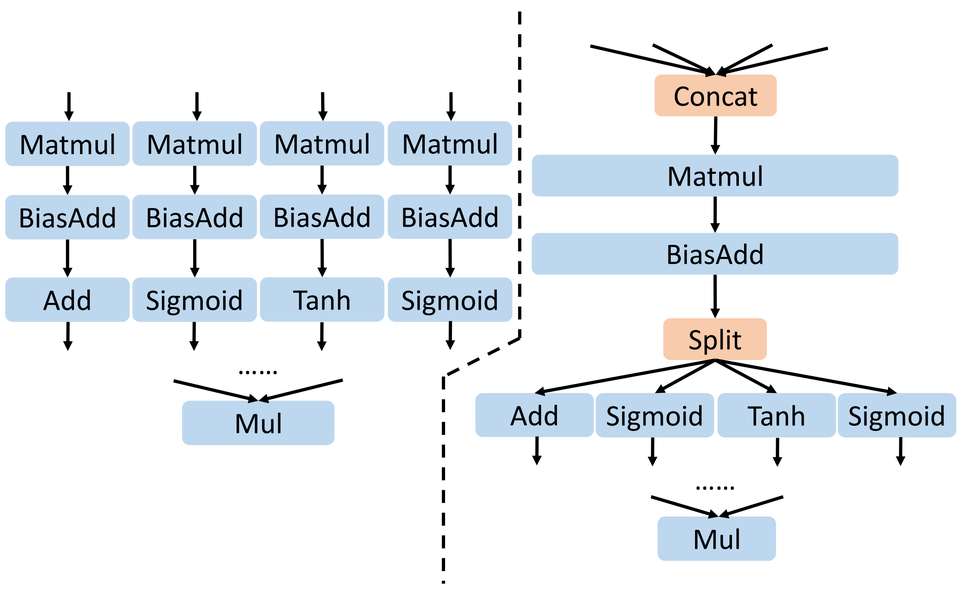
- Isolate "*deep learning expressions*" from "*customized optimizations*" for ad-hoc hardware and scenarios
- Refactor-free to apply customized optimization for performance

Key challenges, technologies, results

Intermediate representation without ambiguity

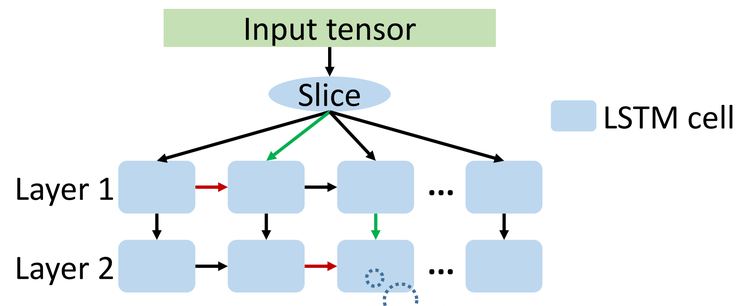


Math expression: $(AB)^T = B^TA^T$

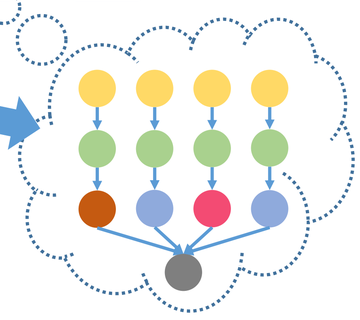


Non-math expression: Native LSTM = Concat-split LSTM

Dynamic subgraph matching




```
cuDNN_LSTM_pattern{
  "cell": { … },
  "op_type": {
    cuDNN_LSTM,
    #layer,
    #step},
  "cell_input": [
    {external_input | other_cell},
    {other_cell}]
}
```



Search space optimization

- Heterogeneous vertices
- Bottom-up search
  - Outputs of operator can be used by unlimited operators
  - Inputs of operator are limited

Preliminary results

- Implemented as an optimizer in Tensorflow r1.3
- Leverage defined cuDNN LSTM pattern to automatically map cuDNN LSTM operator to native LSTM
- Improve performance by 4.12x with refactor-free