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

Model prototype

Hardware trend: heterogeneous devices with various accelerated libs
- Custom scenarios: cloud, mobile
- Custom algorithms: CNN, RNN

Optimization Mapping

Hardware/System engineer

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

Dynamic subgraph matching

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

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

Valid Perplexity

Training Epoch

Native LSTM
Replaced cuDNN LSTM

20 epochs runtime (s)

Native LSTM
Replaced cuDNN LSTM

4.12x