EcoRNN: Fused LSTM RNN Implementation with Data Layout Optimization

Bojian Zheng¹, Akshay Nair¹, Qiongshi Wu¹, Nandita Vijaykumar², Gennady Pekhimenko¹
¹University of Toronto, ²Carnegie Mellon University

**Background**

EcoRNN: Fused LSTM RNN Implementation with Data Layout Optimization

Bojian Zheng¹, Akshay Nair¹, Qiongshi Wu¹, Nandita Vijaykumar², Gennady Pekhimenko¹
¹University of Toronto, ²Carnegie Mellon University

**Data Layout Optimization**

- Optimizing data layout improves cache utilization of FC layers.

**Results**

- Performance comparison of LSTM RNN with different backends
- Training throughput on MXNet language modeling benchmark

**Observation**

- MXNet default implementation has cudaLaunch overhead. cuDNN implementation is closed-source.