EcoRNN: Fused LSTM RNN Implementation with Data Layout Optimization

LSTM RNN and Cell

Timeline

- Default implementations in machine learning frameworks have **cudaLaunch** overhead.
- cuDNN solves the issue, but is **closed-sourced**.

Runtime Breakdown of MXNet Default and cuDNN

Bojian Zheng¹, Akshay Nair, Qiongsi Wu, Nandita Vijaykumar², Gennady Pekhimenko
¹ University of Toronto, ² Carnegie Mellon University
cuDNN Runtime Breakdown

Data Layout Optimization (FC)

Optimize

Data Layout Optimization (RNN)

Integrate

Language Modeling Benchmark

Employ

(1)

(2)

(3)

Employ