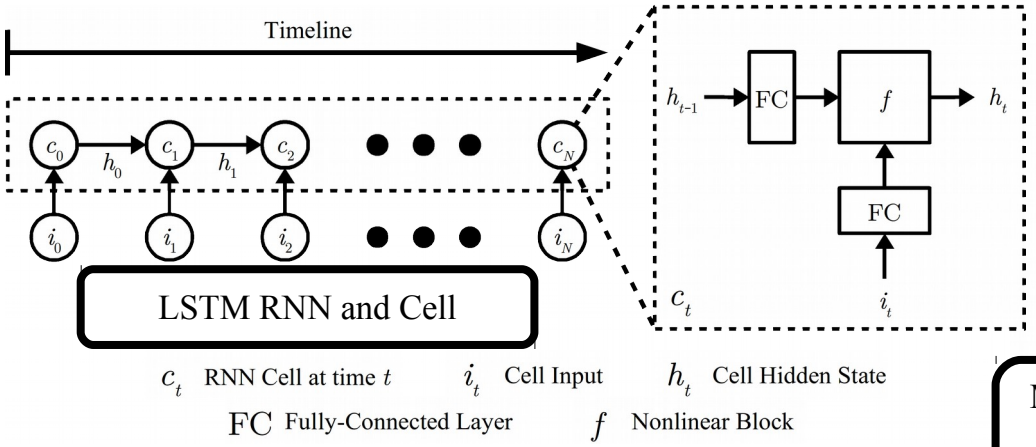
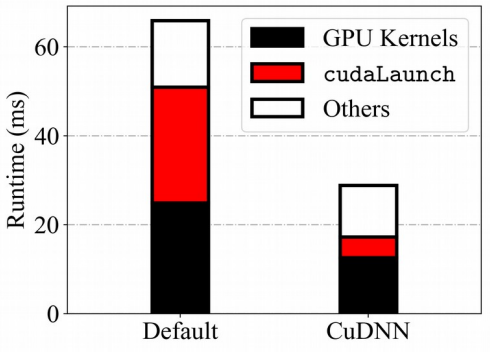
EcoRNN: Fused LSTM RNN Implementation with Data Layout Optimization

Bojian Zheng[1], Akshay Nair[1], Qiongsi Wu[1], Nandita Vijaykumar[2], Gennady Pekhimenko[1]
[1]University of Toronto, [2]Carnegie Mellon University

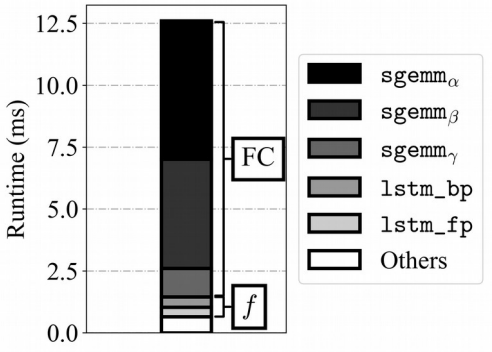
Background





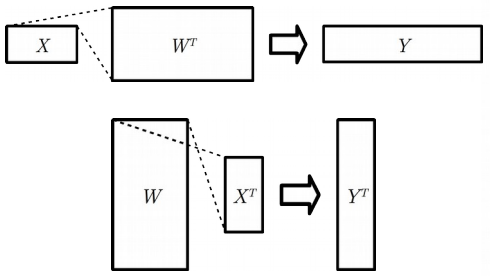
MXNet default implementation has **`cudaLaunch`** overhead. cuDNN implementation is **closed-source**.

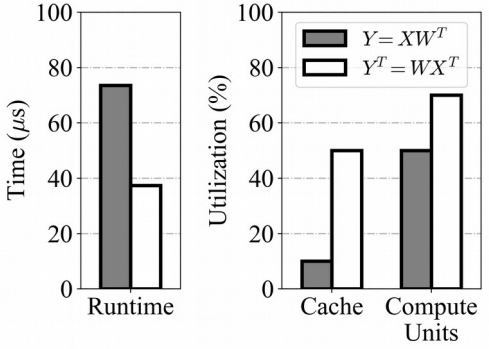
Observation



The runtime bottleneck of LSTM RNN is FC layers.

Data Layout Optimization

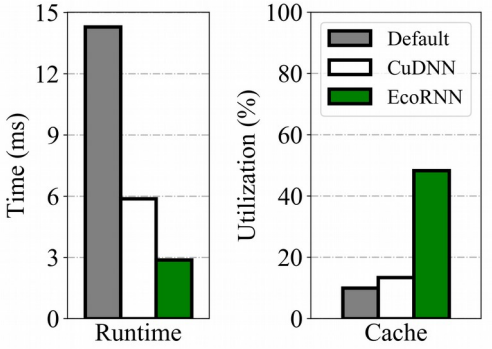




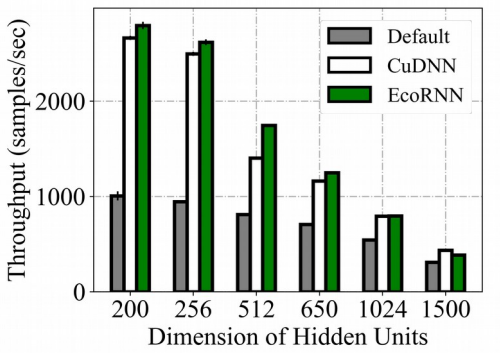
$Y = XW^T$ vs. $Y^T = WX^T$

Optimizing data layout improves **cache utilitzation** of FC layers.

Results



Performance comparison of LSTM RNN with different backends



Training throughput on MXNet language modeling benchmark