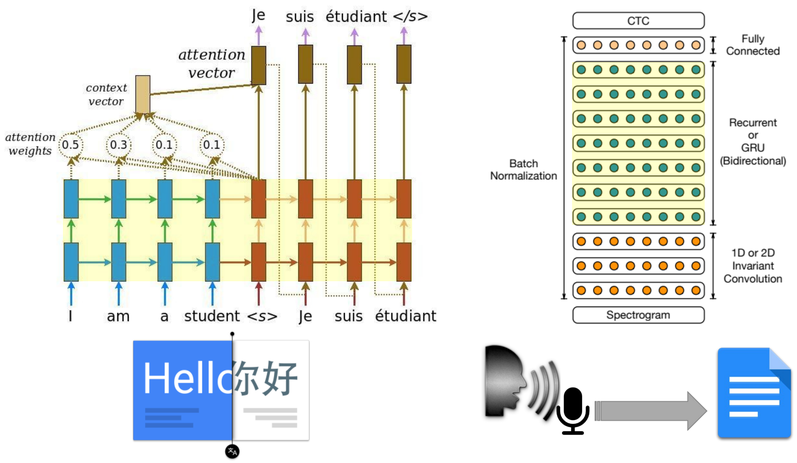
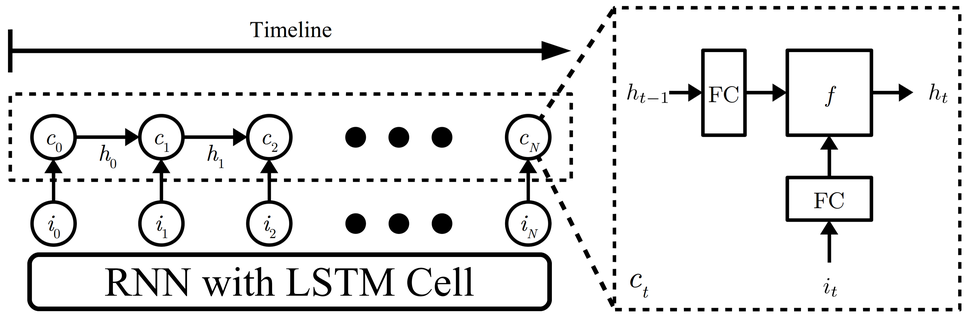
EcoRNN: Efficient Computing of LSTM RNN on GPUs

Bojian Zheng (M.Sc. Student), **Gennady Pekhimenko** (Advisor)
EcoSystem Research Group, Department of Computer Science, University of Toronto

EcoSystem
www.cs.toronto.edu/ecosystem

Background: Long-Short-Term-Memory Recurrent Neural Network

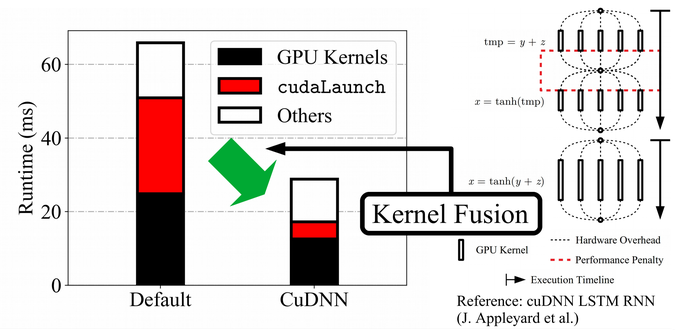




c_t RNN Cell at time t i_t Cell Input h_t Cell Hidden State
FC Fully-Connected Layer f Nonlinear Block

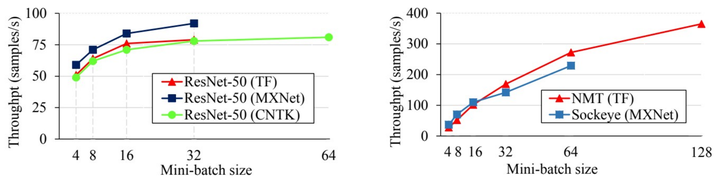
Tensorflow NMT (Luong et al.), *Baidu DeepSpeech2* (Amodei et al.).
Graphs demonstrating usage are from *Google Images*.

Problem Statement



Reference: cuDNN LSTM RNN (J. Appleyard et al.)

✘ ***Default*** has **cudaLaunch overhead**.
✘ ***CuDNN*** is **closed-source**, **limits innovation**.

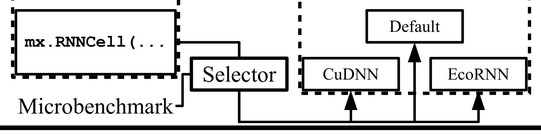


✘ RNN Training is **Memory Capacity**-bounded.

TBD DNN Training Benchmark Suite (Zhu et al., IISWC'18).
tbd-suite.ai.

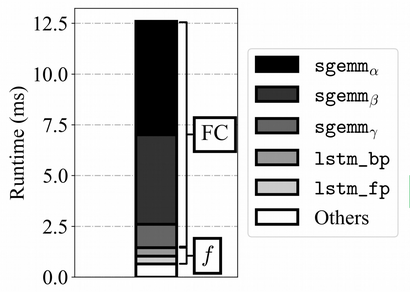
EcoRNN Full Vision

EcoRNN is an **open-source** implementation that has runtime performance **comparable with or even better than *CuDNN***, but **consumes less memory** and **supports auto-tuning**.

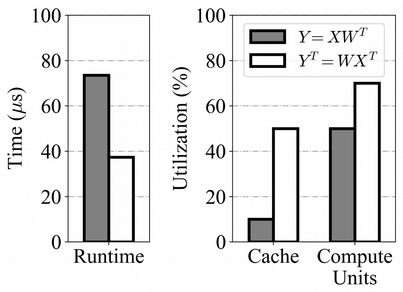


All changes are **transparent** to programmers.

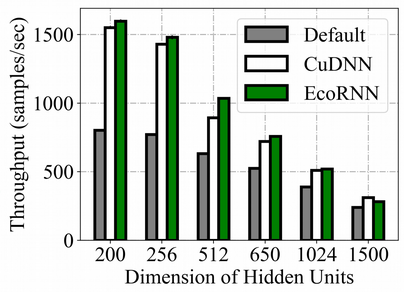
Preliminary Results: Performance



The runtime bottleneck is **FC layers**.

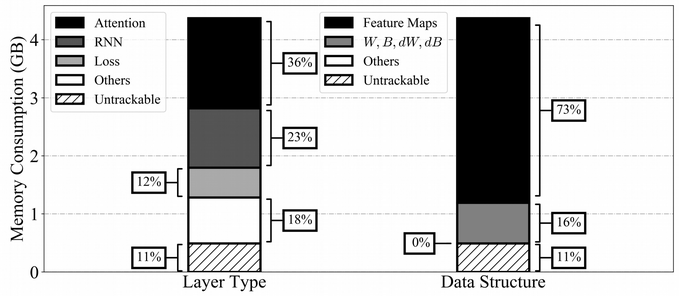


Data layout optimization **improves cache hit rate**.



Training Throughput on MXNet Language Modeling benchmark

Memory Consumption



The memory bottleneck is **Feature Maps** of **Attention** and **RNN**.

Future Work

Baidu persistent RNN

✔ Weight Parameter Reuse
✔ High Performance when Batch Size is Small

✘ **Inflexible**
(hard to be ported to new GPUs and cell types)

ML Compilers
(e.g., *TVM*, *XLA*)

Gist (Jain et al., ISCA'18)

The 51st Annual IEEE/ACM International Symposium on Microarchitecture®, 2018, Fukuoka, Japan