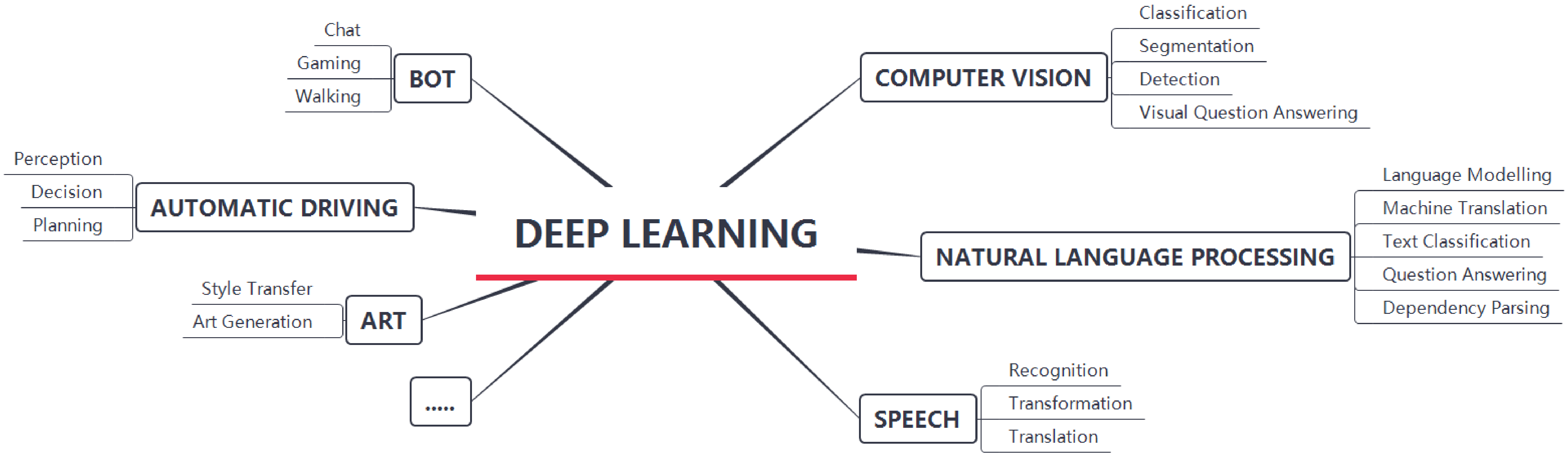
Deep500 BOF 2018

# Thinking about An HPC Oriented Deep Learning Benchmark

Jidong Zhai
Tsinghua University

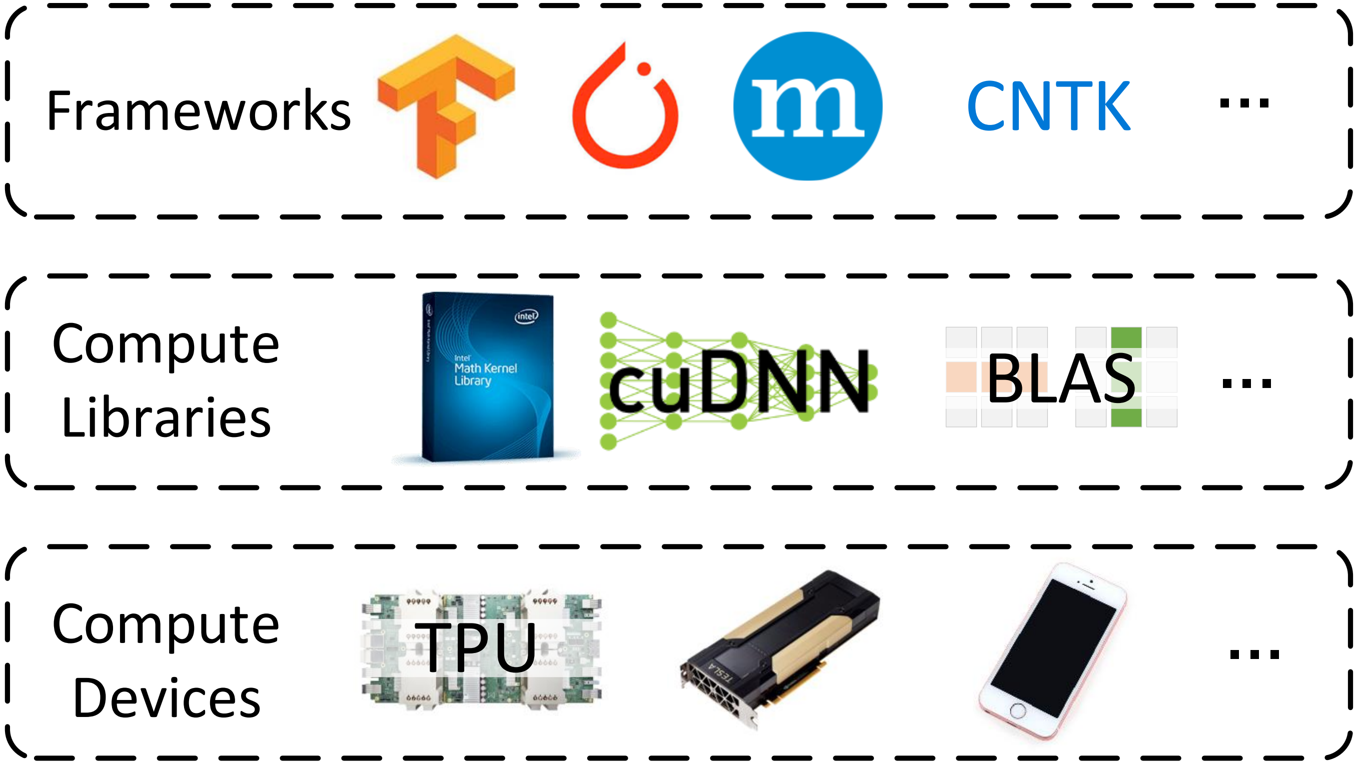
# Introduction

- **Deep learning has widely used in lots of areas**



# Introduction

- **A lot of deep learning frameworks, compute libraries and acceleration devices**



Frameworks

CNTK

…

Compute Libraries

cuDNN

BLAS

…

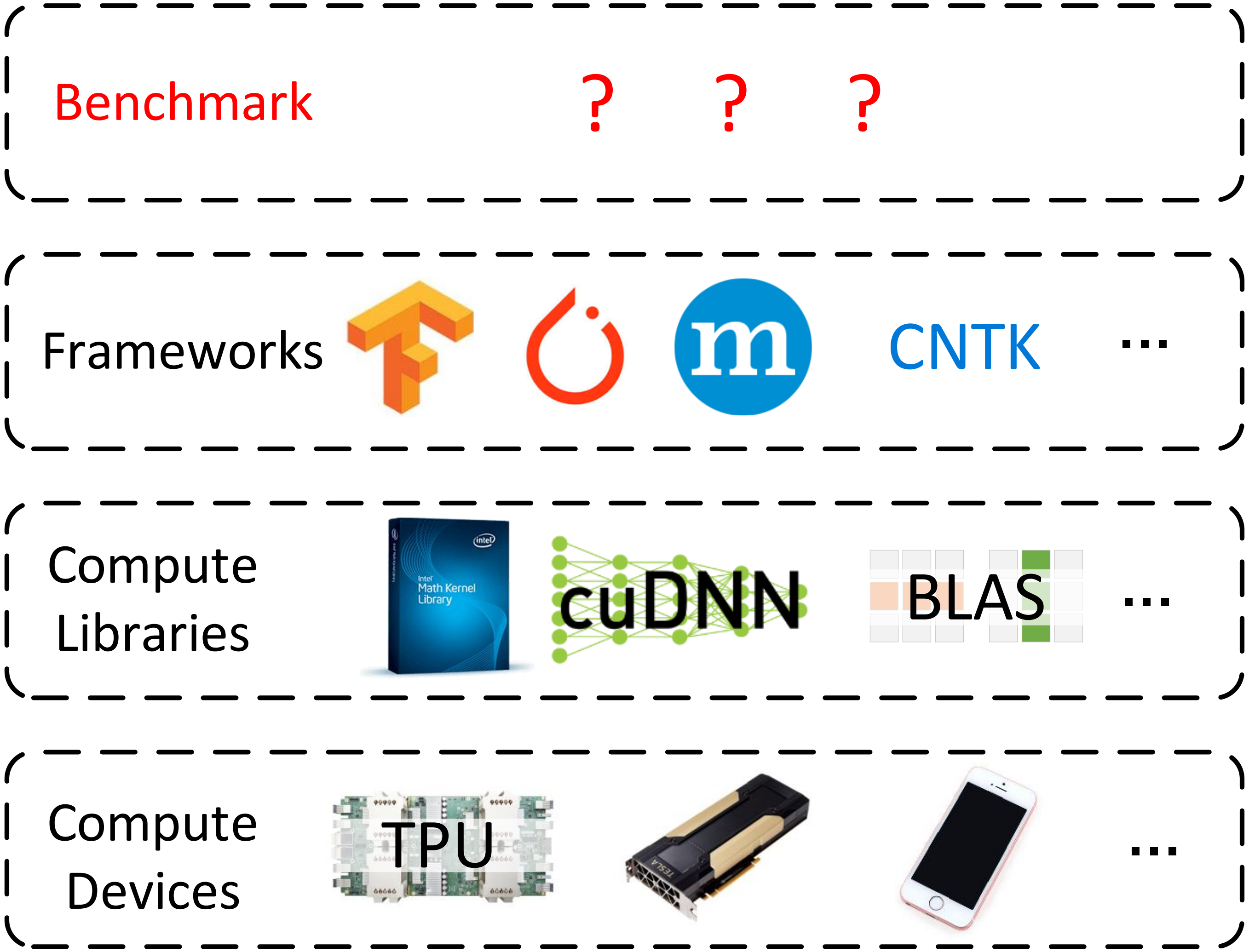
Compute Devices

TPU

…

# Introduction

- **However, how to evaluate?**



Benchmark ? ? ?

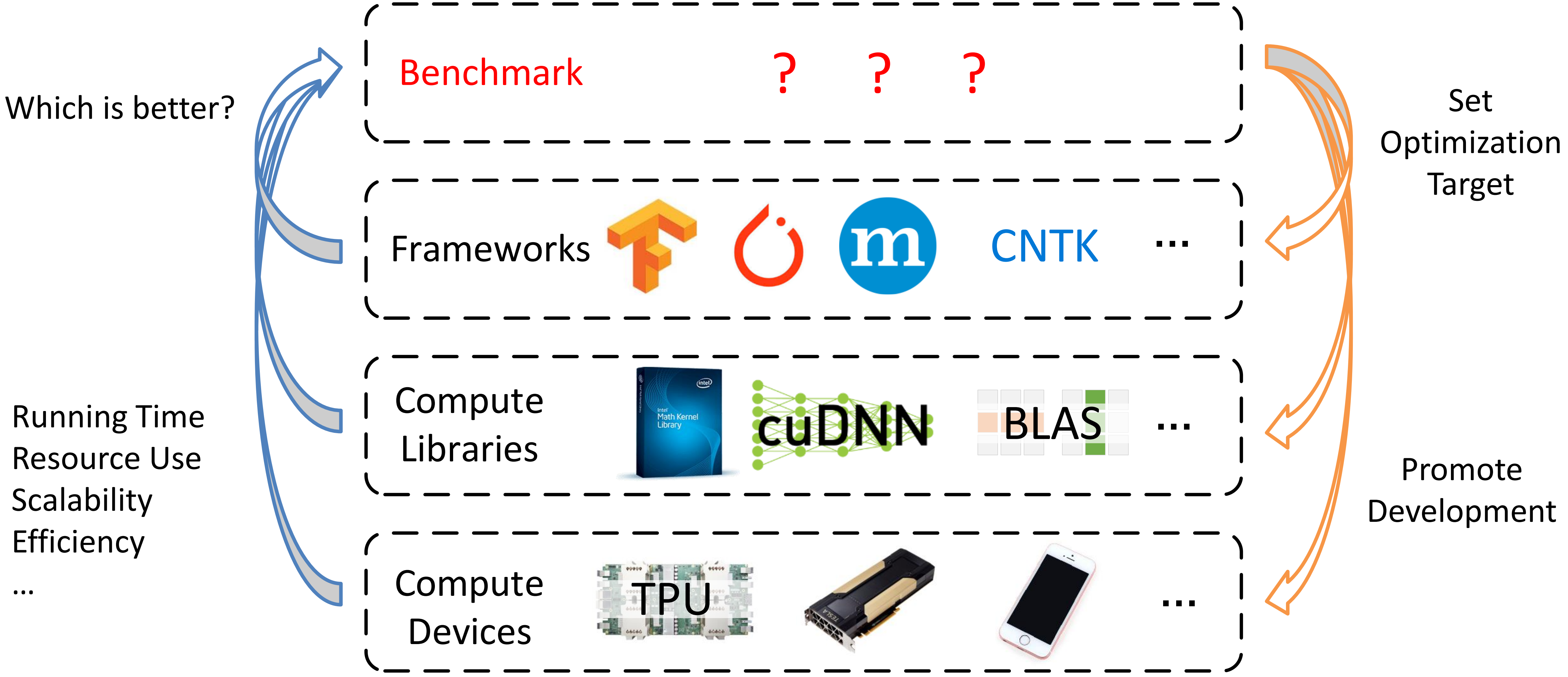
Frameworks CNTK …

Compute Libraries cuDNN BLAS …

Compute Devices TPU …

# Introduction

- **However, how to evaluate?**



# Related Deep Learning Benchmarks

| | | convnet-benchmarks[1] | DeepBench[2] | DAWNBench[3] | TensorFlow Benchmark[4] |
|---|---|---|---|---|---|
| **Target** | | Framework<br>Compute Library | Compute Library<br>Compute Device | Compute Library<br>Framework | Framework |
| **Models** | Granularity | Neural Network | Basic Operation | Neural Network | Neural Network |
| | Diversity | | | | |
| **Dataset** | | | | | |
| **Metrics** | | T | | Training Time and<br>Accuracy | |

Low Diversity

Limited Dataset

Single Metric

[1] convnet-benchmarks: https://github.com/soumith/convnet-benchmarks

[2] Baidu DeepBench: https://github.com/baidu-research/DeepBench

[3] Cody A. Coleman et al. *DAWNBench: An End-to-End Deep Learning Benchmark and Competition*. NIPS 2017

[4] TensorFlow Benchmark https://www.tensorflow.org/performance/benchmarks

MLPerf

A broad ML benchmark suite for measuring performance of ML software frameworks, ML hardware accelerators, and ML cloud platforms.

# Related Deep Learning Benchmarks

| | | MLPerf[1] |
|---|---|---|
| Evaluation Target | | Framework<br>Compute Device |
| Characteristics | Granularity | Neural Network |
| | Diversity | 1. Image(Classification, Detection)<br>Various Applications<br>3. Speech(Recognition)<br>4. Reinforcement Learning & Recommendation |
| Dataset | | Various Datasets |
| Evaluation Metrics | | Training Time, Power Use and Cost to certain Accuracy |

[1] https://mlperf.org/

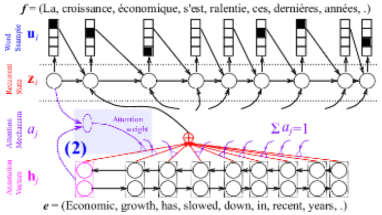
# How to evaluate HPC systems for machine learning?

# Our Work on Workload Analysis for Deep Learning

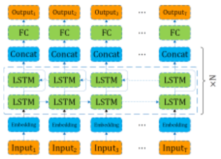
- Preliminary workload analysis

Applications

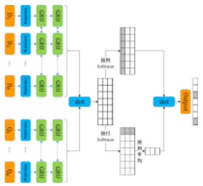
Image Classification


f = (La, croissance, économique, s'est, ralentie, ces, dernières, années, .)
$u_i$
$z_i$
$a_j$
$h_j$
Attention weight
(2)
$\Sigma a_j=1$
e = (Economic, growth, has, slowed, down, in, recent, years, .)

Machine Translation



Language Model



Question Answering

Models VGG ResNet Seq2seq RNN LM AoA Reader

Dataset

WikiText-2

Cifar

Real Data

Tatoeba

CBTest

Real time

Easy to obtain

Dummy Data

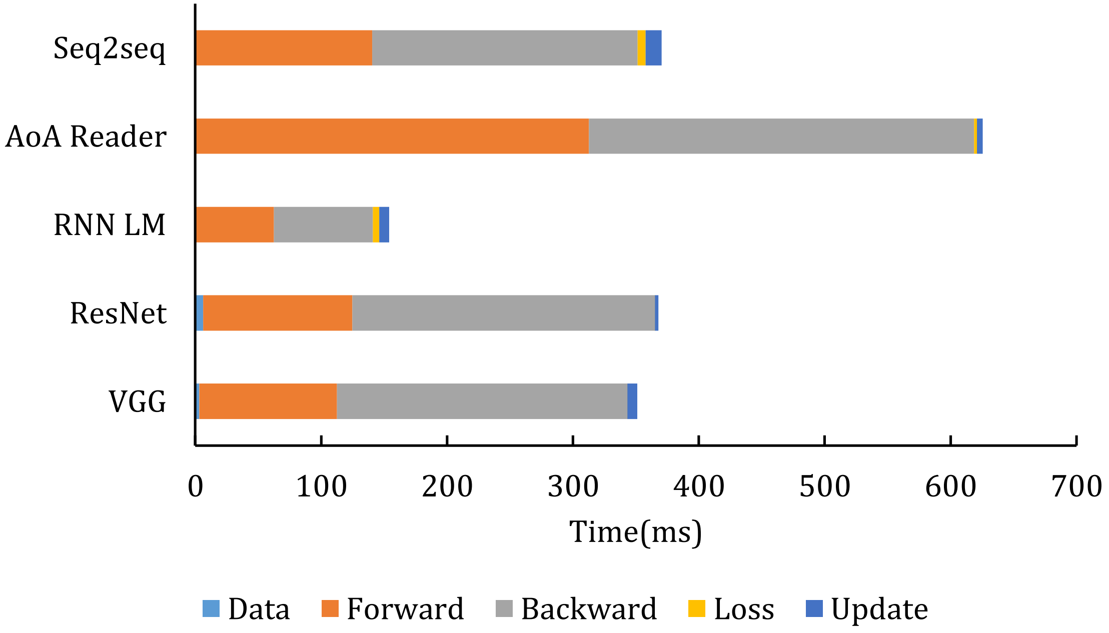
Generative

Controllable

# Our Work

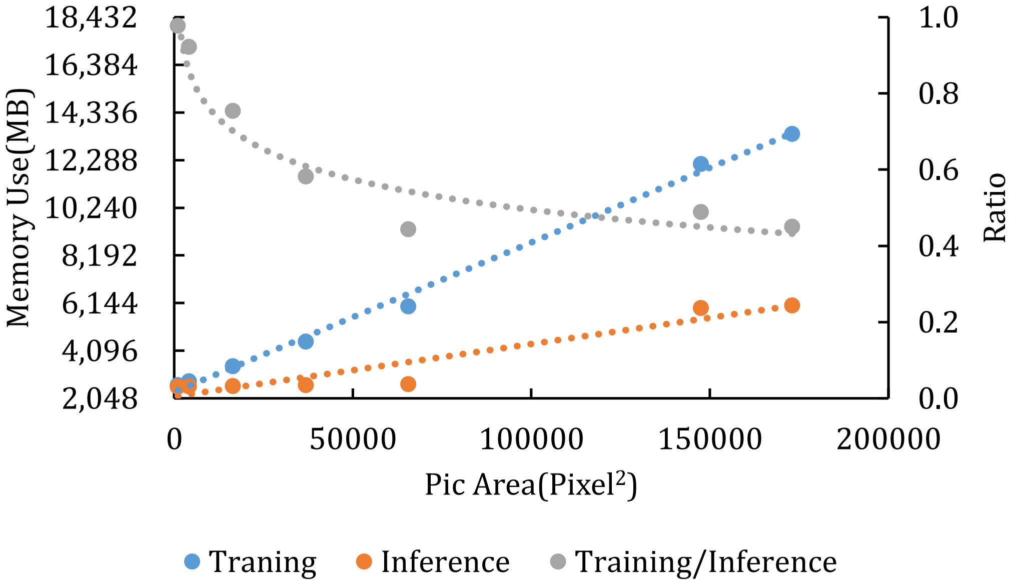
- Time
  - Time of every operation type within one iteration
  - Time of phases within one iteration

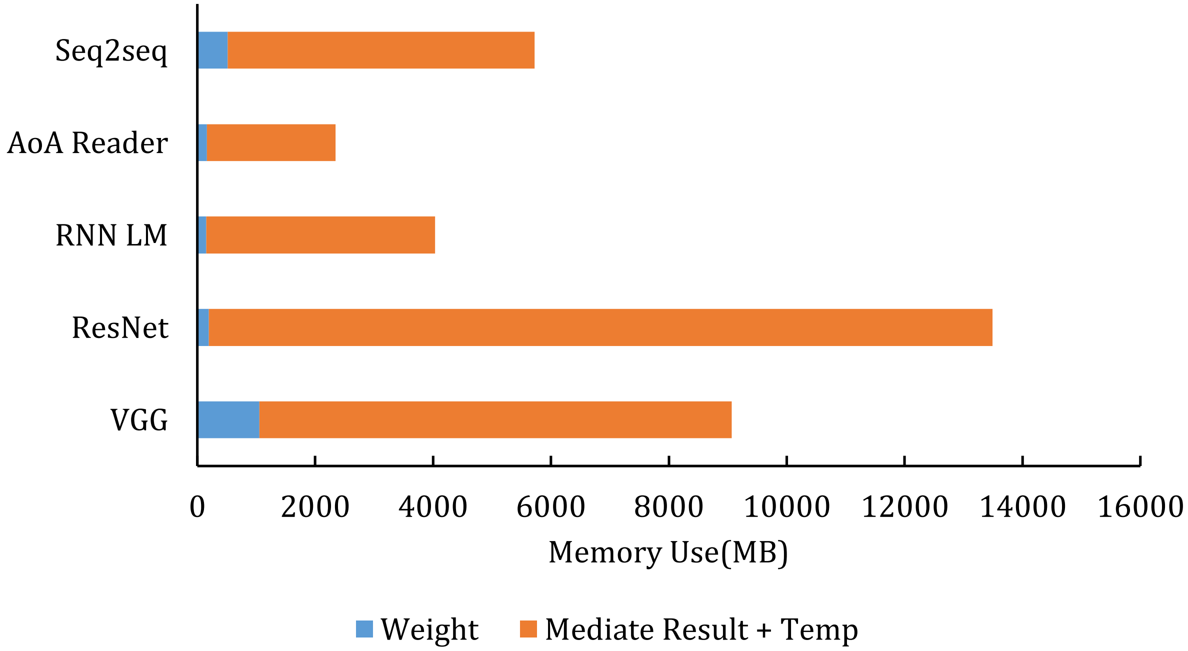
| | Bmm | Linear | Conv | Pool | BN | Mul | Div | ReLU | Dropout | Sum | Softmax | LSTM/GRU | Embedding | Total(ms) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| VGG | | 1 | 71 | 3 | 16 | | | 8 | | | | | | 109.05 |
| ResNet | | | 66 | 1 | 21 | | | 12 | | | | | | 107.54 |
| RNN LM | | 28 | | | | | | | | | | 72 | | 62.05 |
| AoA Reader | | | | | | | | | | 73 | | 26 | | 312.18 |
| Seq2seq | | 73 | | | | 4 | | | | 3 | 5 | 14 | 1 | 108.95 |

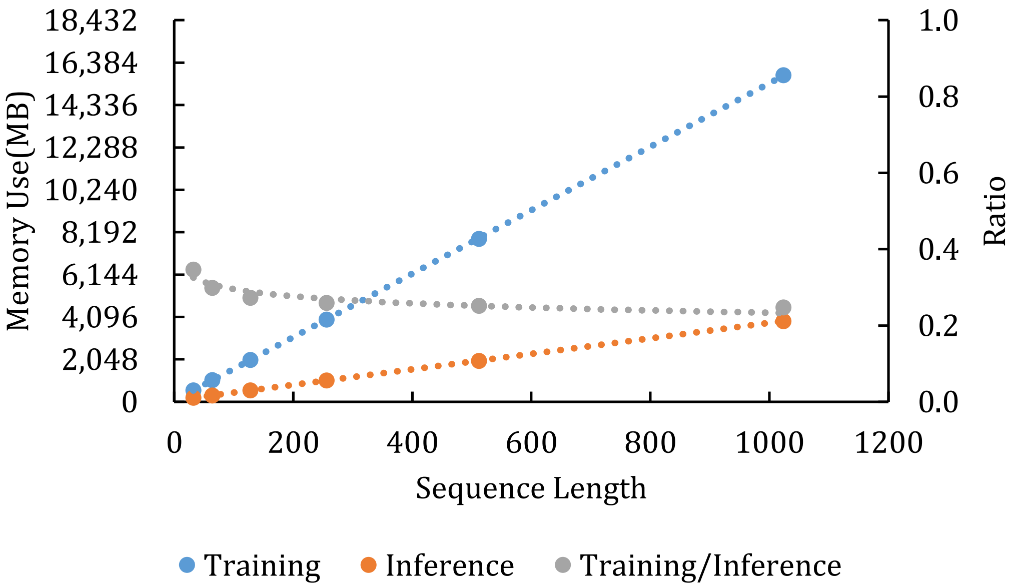


# Workload Analysis

- **Memory Usage**
  - Memory Usage Break Down
  - Memory Usage – Input Size

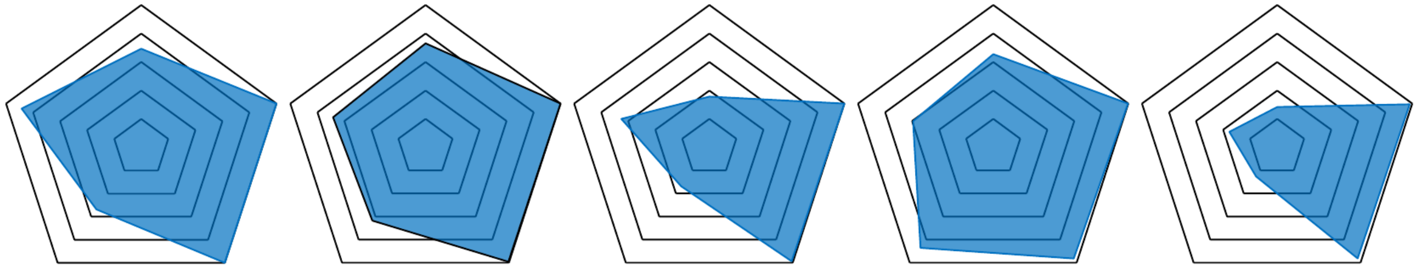






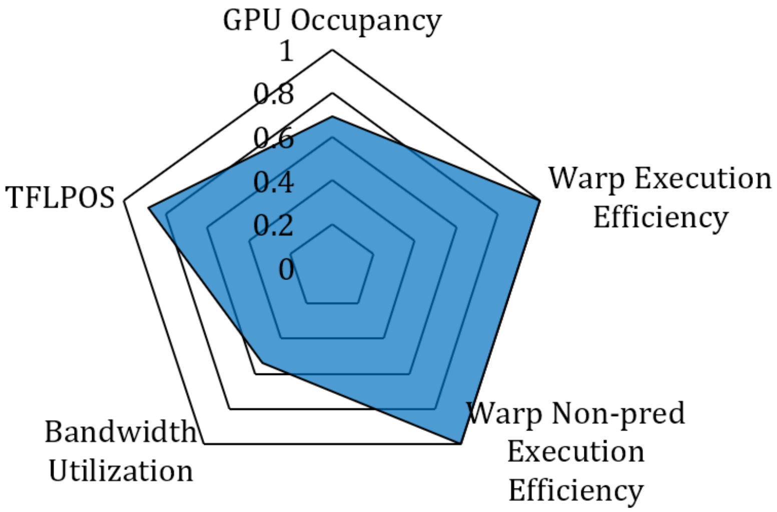
# Workload Characterization

- **Hardware Counters**
  - **For GPU**



(a) VGG (b) ResNet (c) RNN LM (d) Seq2seq (e) AoA Reader

| | GPU Occupancy | Warp Execution Efficiency | Warp Non-Pred Execution Efficiency | Bandwidth Utilization | TFLPOS |
|---|---|---|---|---|---|
| Normalized 1 | 0.46 | 1.00 | 1.00 | 4.02 | 5.65 |



# Questions about an HPC Oriented Deep Learning Benchmark

- **Questions we need to think:**
  - **Model Selection**
    - **Various application areas?**
    - **A synthetic model with main features?**
  - **Dataset**
    - **Fixed data set (Imagenet)?**
    - **A Generative Data?**
  - **Metrics**
    - **Time for training?**
    - **Gflops?**
    - **AI operations per second?**

# Thanks!