

# A Quick Survey on Deep Learning Engines

Chensong Zhang

with Zheng Li and Ronghong Fan

DL Seminar — April 27, 2017

# Outline

Introduction

Comparison

Python

TensorFlow

MXNet

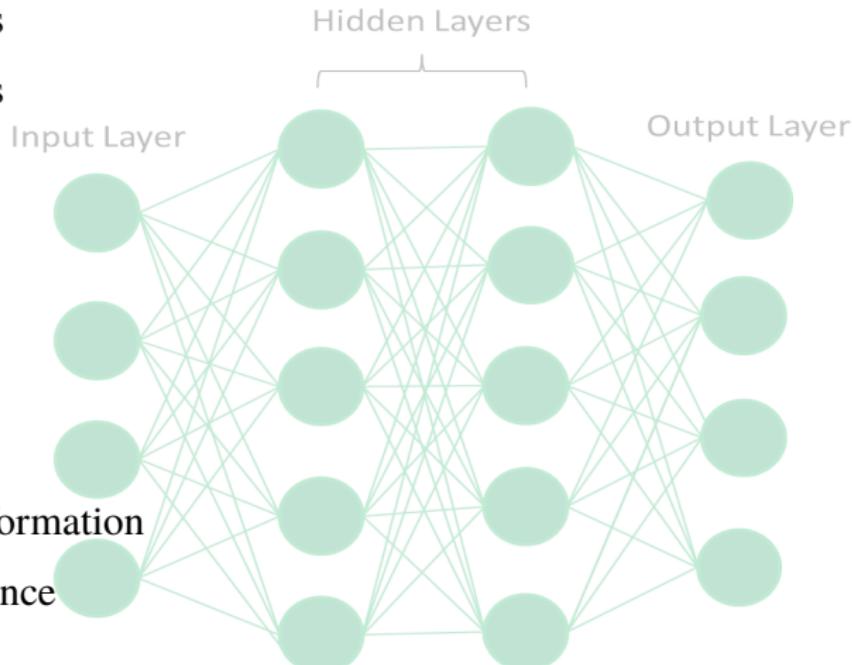
Torch

Caffe



# §1. Introduction

- 1 Machine Learning
- 2 ML Software Packages
- 2 ML Software Packages
- 3 Deep Learning
- 3 Deep Learning
- 4 Goal of This Study
- 4 Goal of This Study
- 4 Goal of This Study
- 5 DL Engines: Basic Information
- 6 DL Engines: Performance
- 7 Hardware Platforms
- 8 Neural Networks and Test Data Sets



# Machine Learning

## What is ML?

- Unlike traditional numerical simulation, “ML gives computers the ability to learn without being explicitly programmed” [Samuel 1959]
- As a research field, ML explores the study and construction of algorithms that can learn from and make predictions on **data**
- Fourth paradigm, big data, Internet of things, artificial intelligence, ...

## General Tasks of ML:

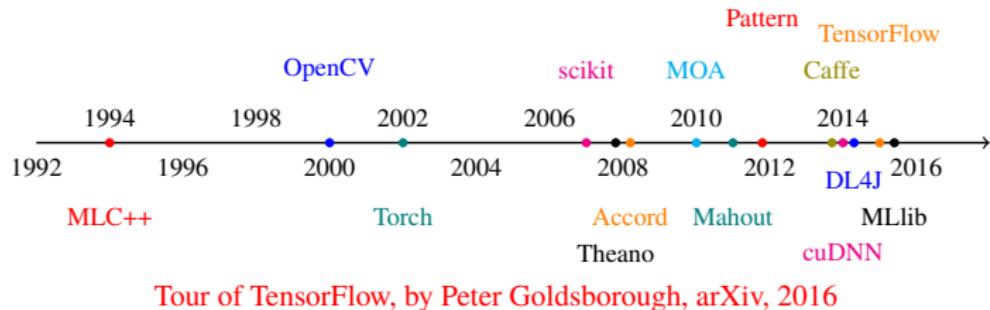
- ☞ Classification: Inputs are divided into two or more classes, and the learner must produce a model that assigns unseen inputs to one or more of these classes
- ☞ Regression: Similar to classification, but the outputs are continuous
- ☞ Clustering: Inputs are divided into several groups (Unlike in classification, the groups are not known beforehand, making this an unsupervised task)
- ☞ Density estimation, dimension reduction, ...

Deep Learning, by I. Goodfellow, Y. Bengio, and A. Courville, Chinese version available online

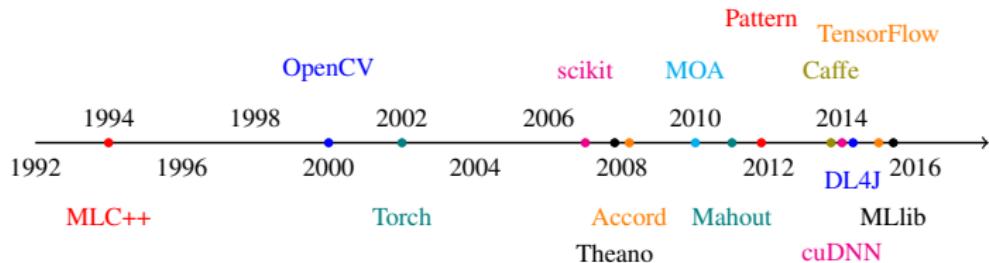
<https://github.com/exacity/deeplearningbook-chinese>



## ML Software Packages



# ML Software Packages



Tour of TensorFlow, by Peter Goldsborough, arXiv, 2016

## Other worth-noting ML (DL) packages:

- PyTorch (still in **beta** stage)
- Theano (Lasagne/Keras): Slow in graph compilation
- ▣ MXNet (DMLC): First released on Jan 2015, scalable distributed computing
- ▣ CNTK/DMTK (Microsoft): First released on April 2015
  - Windows/Linux, no official OS X support thought
  - C++/Python front-end
- Neon (Nervana & Intel): First released on May 2015, professional support
- Caffe2 (Google, Facebook, ...): Release on late 2016
- Digits (Nvidia, powered by Caffe): Web interface

# Deep Learning

Deep Learning has been introduced with the objective of moving ML closer to one of its original goals—**Artificial Intelligence**. Main motivations:

- The brain has a deep architecture
- Cognitive processes seem deep
- Insufficient depth can hurt in practice

# Deep Learning

Deep Learning has been introduced with the objective of moving ML closer to one of its original goals—**Artificial Intelligence**. Main motivations:

- The brain has a deep architecture
- Cognitive processes seem deep
- Insufficient depth can hurt in practice

## Pros:

- conceptually simple
- nonlinear
- highly flexible and configurable
- learned features can be extracted
- can be fine-tuned with more data
- efficient for multi-class problems
- good at pattern recognition

## Cons:

- hard to interpret
- theory not well understood
- more parameters
- slow to train and score
- overfits (needs regularization)
- inefficient for categorical variables
- data hungry, learns slowly

# Goal of This Study

What is our purpose?

- Theoretical analysis for machine learning, ...
- ➡ Developing algorithms and optimizing implementation (development)
- Solving practical problems from various applications (user interface)

# Goal of This Study

What is our purpose?

- Theoretical analysis for machine learning, ...
- ☞ Developing algorithms and optimizing implementation (development)
- Solving practical problems from various applications (user interface)

How to use a ML package?

- Train models with large data set on high-performance computing platforms
- Deploy applications on various (computing) platforms, like smart phones

# Goal of This Study

What is our purpose?

- Theoretical analysis for machine learning, ...
- ↖ Developing algorithms and optimizing implementation (development)
- Solving practical problems from various applications (user interface)

How to use a ML package?

- Train models with large data set on high-performance computing platforms
- Deploy applications on various (computing) platforms, like smart phones

What we want for a ML package?

- Easy for debugging (with good **documentation, support**, and active community)
- Good flexibility
  - Easy to add new tasks
  - Easy to build new network structures
- Good performance and scalability
  - Multicore CPU
  - **Single or multiple GPU(s)**
  - Cluster

# DL Engines: Basic Information

Viewpoint	Torch	Caffe	TensorFlow	MXNet
First Released	2002	2013	2015	2015
Main Developers	Facebook, Twitter, Google, ...	BAIR BVLC	Google	DMLC
Core Languages	C/Lua	C++	C++ Python	C++
Supported Interface	Lua	C++/Python Matlab	Python/C++/R Java/Go/...	C++/Python/R Matlab/Julia/...
License	BSD	BSD	Apache	Apache

- BAIR, Berkeley Artificial Intelligence Research Lab
- BVLC, Berkeley Vision and Learning Center
- DMLC, Distributed (Deep) Machine Learning Community, supported by Amazon, Intel, Microsoft, nVidia, Baidu, ...

# DL Engines: Performance

Viewpoint	Torch7	Caffe	TensorFlow	MXNet
Pretrained Models	Yes	Yes	No	Yes
High-level Support	Good	Good	Good	Good
Low-level Operators	Good	Good	Fairly good	Increasing fast
Speed One-GPU	Great	Great	Good	Good
Memory Management	Great	Great	Not so good	Excellent
Parallel Support	Multi-GPU	Multi-GPU	Multi-GPU	Distributed
Coding Style	Imperative	Declarative	Declarative	Mixed
GitHub Watching	649/268	1856	4939	887

# Hardware Platforms

Four test platforms were used in the following tests

- CPU: one quad-core desktop CPU (Intel i7-3820 CPU @3.60GHz) and two 8-core server-grade CPUs (Intel Xeon CPU E5-2630 v3 @2.40GHz)
- GPU: GTX 1080 @1607MHz with Pascal architecture, and Tesla K80 @562MHz with Kepler architecture

Computational Unit	Cores	Memory	OS	CUDA
Intel CPU i7-3820	4	64 GB	Ubuntu 14.04	–
Intel CPU E5-2630x2	16	128 GB	CentOS 7.2	–
GTX 1080	2560	8 GB	Ubuntu 14.04	8
Tesla K80 GK210	2496	12 GB	CentOS 7.2	8

Figure: The experimental hardware settings for numerical tests

Benchmarking State-of-the-Art Deep Learning Software Tools, by S.-H. Shi, et al., arXiv, 2017

# Neural Networks and Test Data Sets

- A large fully-connected neural network (**FCN-S**) with around 55 million parameters is used to evaluate the performance of FCN
- The classical AlexNet (**AlexNet-S**) is used as an representative of CNN
- A smaller FCN (**FCN-R**) is constructed for MNIST data set
- An AlexNet (**AlexNet-R**) architecture is used for Cifar10 data set
- For RNNs, considering that the main computation complexity is related to the length of input sequence, 2 LSTM layers with input length of 32.

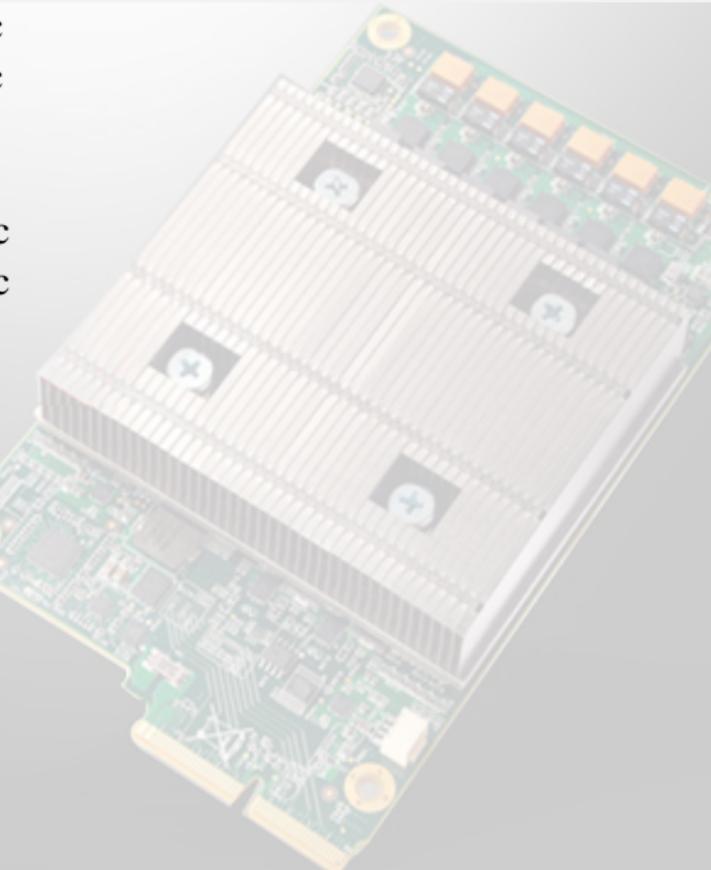
Networks		Input	Output	Layers	Parameters
FCN	FCN-S	26752	26752	5	~55 millions
FCN	FCN-R	784	10	5	~31 millions
CNN	AlexNet-S	150528	1000	4	~61 millions
CNN	AlexNet-R	3072	10	4	~81 thousands
RNN	LSTM	10000	10000	2	~13 millions

Figure: The experimental setup of neural networks for synthetic and real data

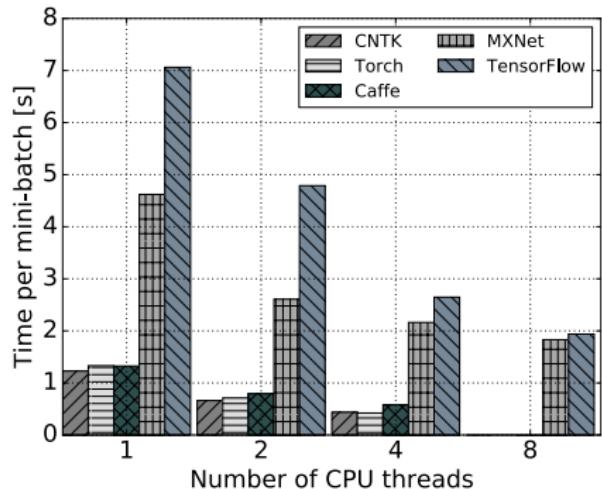
## §2. Comparison

- 9 CPU Scalability: FCN Synthetic
- 9 CPU Scalability: FCN Synthetic
- 10 CPU Scalability: FCN Real
- 10 CPU Scalability: FCN Real
- 11 CPU Scalability: CNN Synthetic
- 11 CPU Scalability: CNN Synthetic
- 12 CPU Scalability: CNN Real
- 12 CPU Scalability: CNN Real
- 13 CPU Scalability: RNN LSTM
- 13 CPU Scalability: RNN LSTM
- 14 GPU Speed: FCN Synthetic
- 14 GPU Speed: FCN Synthetic
- 15 GPU Speed: FCN Real
- 15 GPU Speed: FCN Real
- 16 GPU Speed: CNN Synthetic
- 16 GPU Speed: CNN Synthetic
- 17 GPU Speed: CNN Real
- 17 GPU Speed: CNN Real

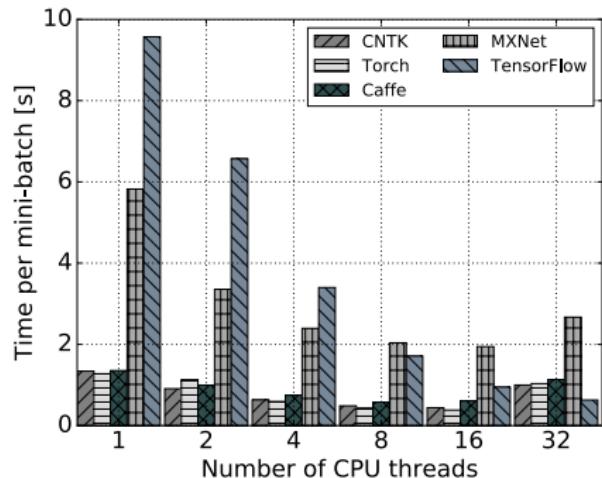
Tensor Processing Unit



# CPU Scalability: FCN Synthetic



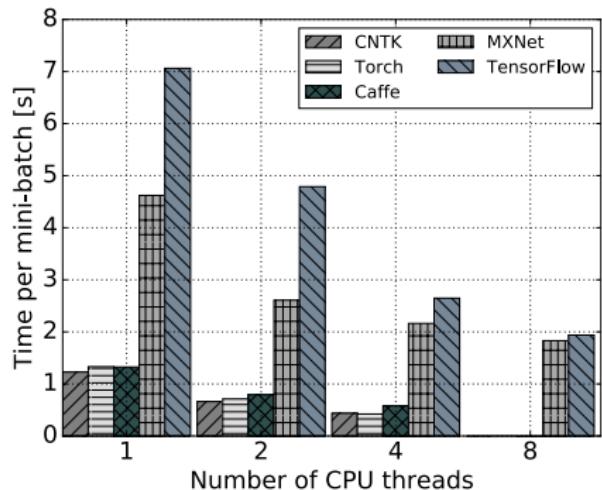
(a) Results on i7-3820.



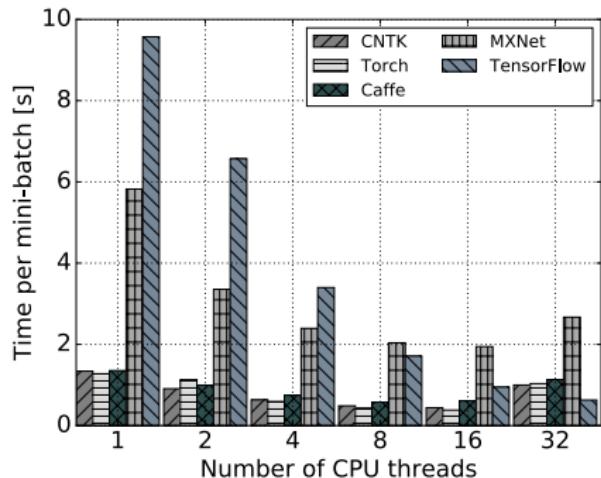
(b) Results on E5-2630.

Figure: FCN-S performance on CPU platform with a mini-batch size of 64

# CPU Scalability: FCN Synthetic



(a) Results on i7-3820.

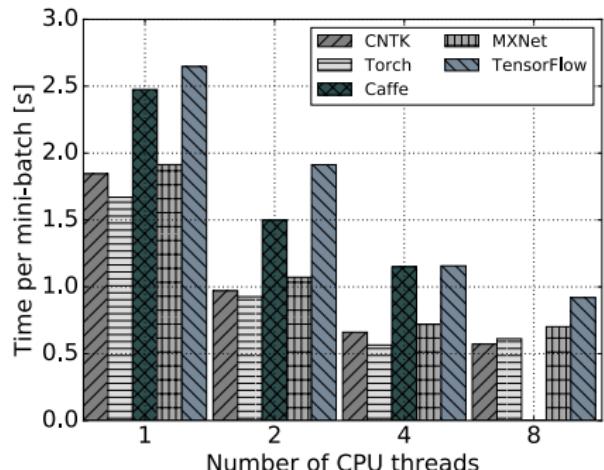


(b) Results on E5-2630.

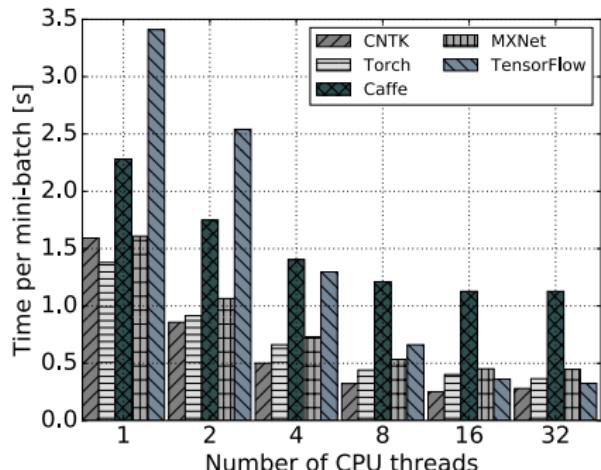
Figure: FCN-S performance on CPU platform with a mini-batch size of 64

- CNTK/Torch/Caffe have similar CPU performance
- TensorFlow has excellent scalability

# CPU Scalability: FCN Real



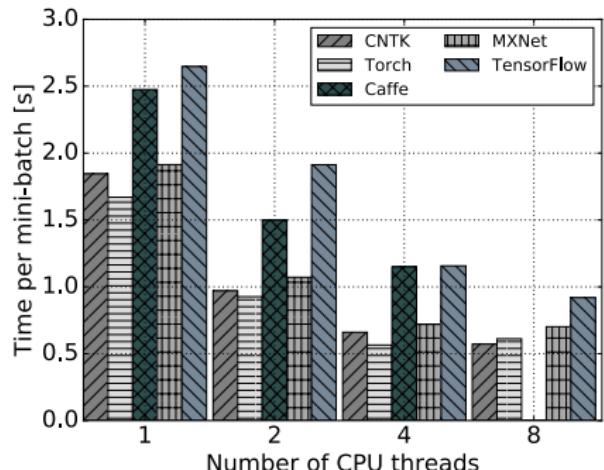
(a) Results on i7-3820.



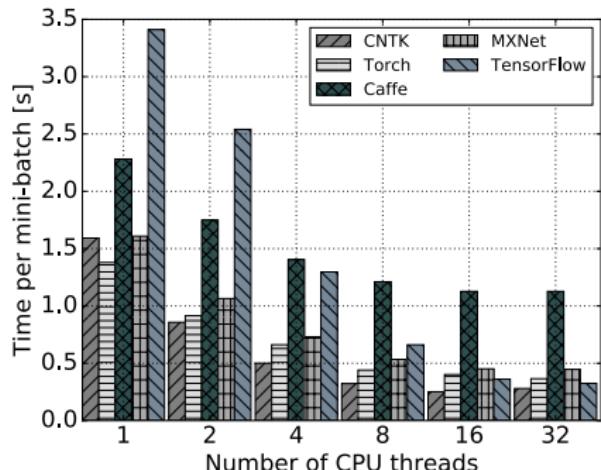
(b) Results on E5-2630.

**Figure:** The FCN-R performance on CPU platform with a mini-batch size of 1024

# CPU Scalability: FCN Real



(a) Results on i7-3820.

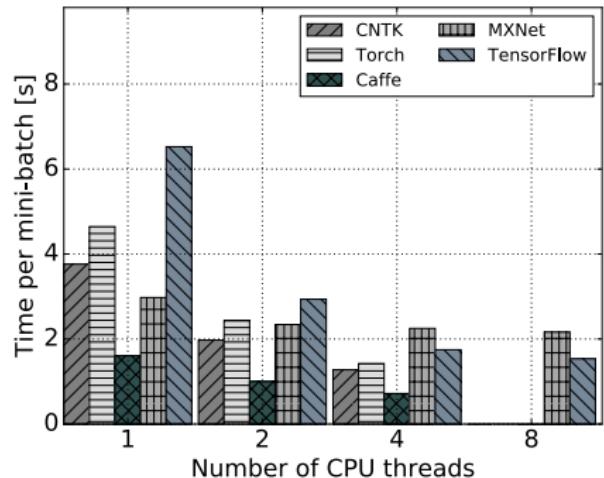


(b) Results on E5-2630.

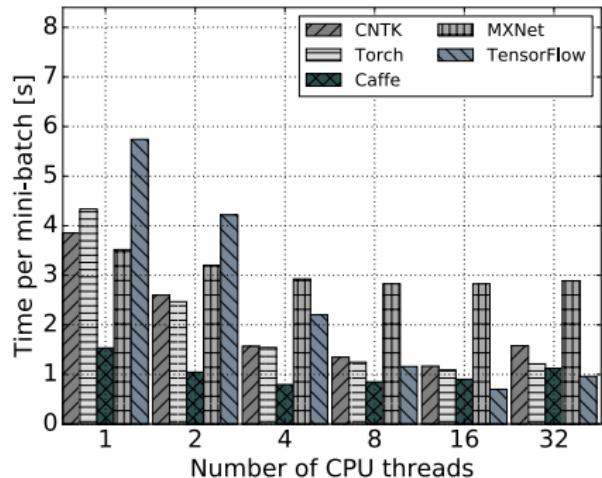
**Figure:** The FCN-R performance on CPU platform with a mini-batch size of 1024

- All engines have good CPU performance
- TensorFlow has good scalability but considerably slower

# CPU Scalability: CNN Synthetic



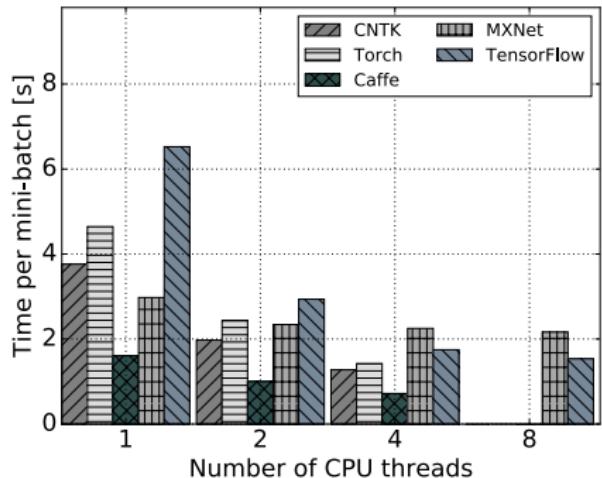
(a) Results on i7-3820.



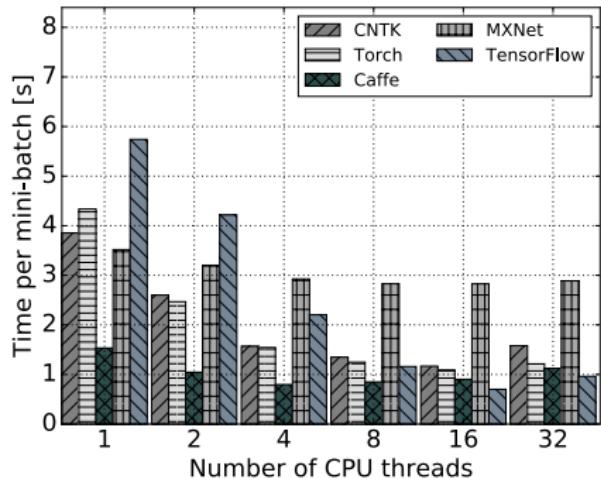
(b) Results on E5-2630.

Figure: AlexNet-S performance on CPU platform with a mini-batch size of 16

# CPU Scalability: CNN Synthetic



(a) Results on i7-3820.

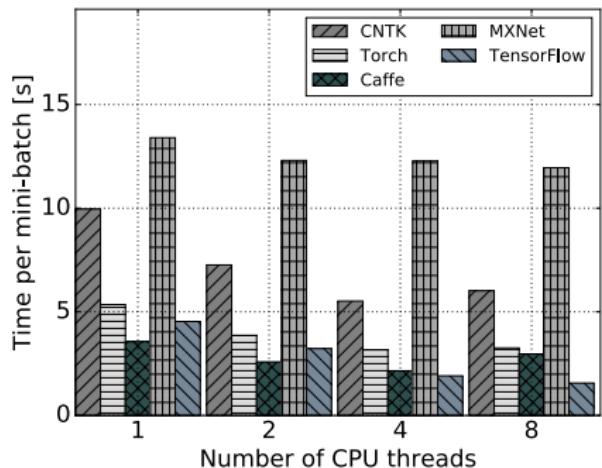


(b) Results on E5-2630.

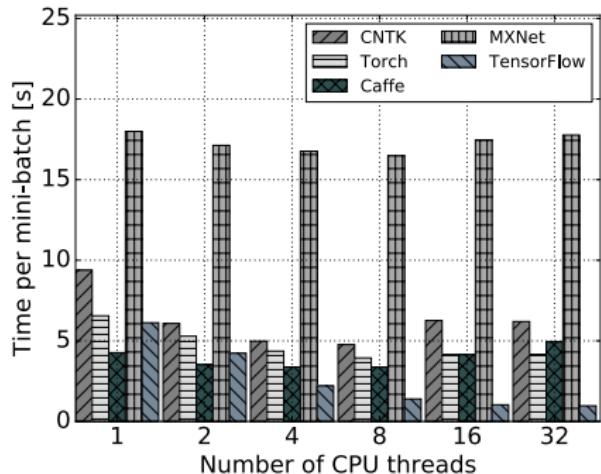
Figure: AlexNet-S performance on CPU platform with a mini-batch size of 16

- Caffe has best CNN performance as promised
- MXNet does not scale well for this test

# CPU Scalability: CNN Real



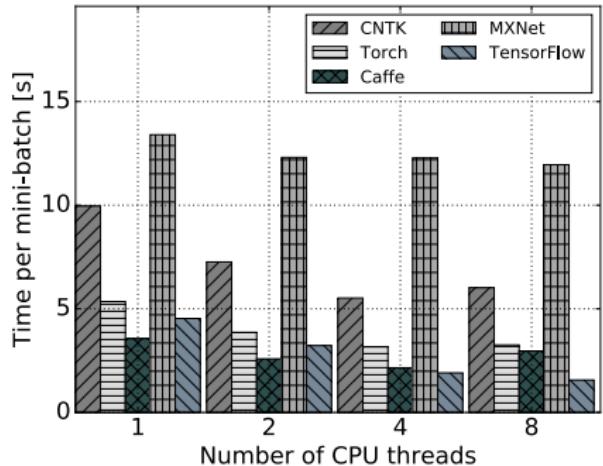
(a) Results on i7-3820.



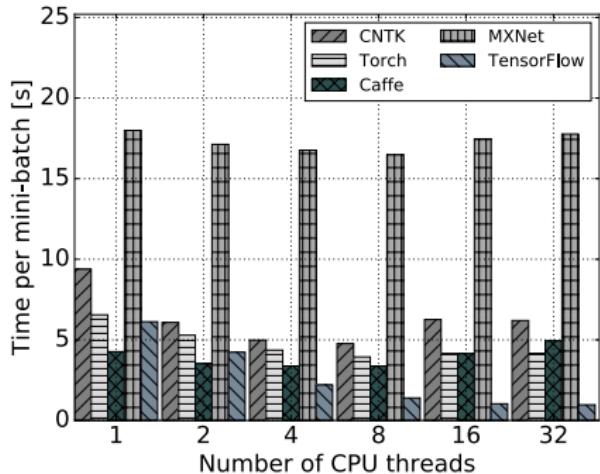
(b) Results on E5-2630.

**Figure:** AlexNet-R performance on CPU platform with a mini-batch size of 1024

# CPU Scalability: CNN Real



(a) Results on i7-3820.

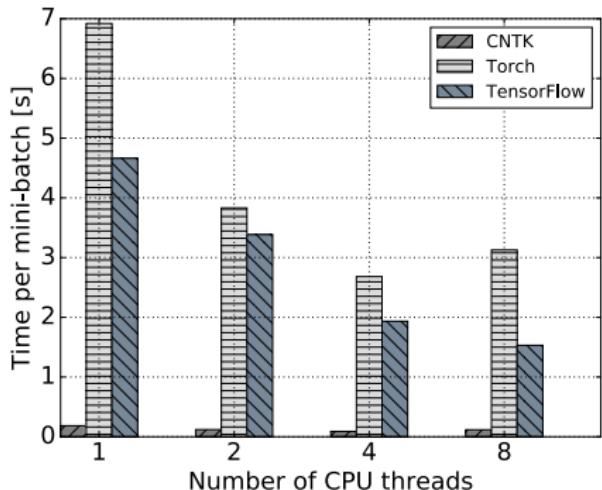


(b) Results on E5-2630.

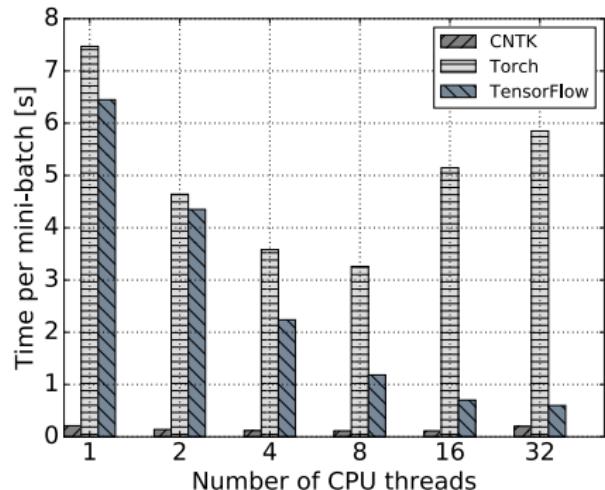
**Figure:** AlexNet-R performance on CPU platform with a mini-batch size of 1024

- Good scalability of TensorFlow kicks in
- Caffe does not scale well on multicore CPUs

# CPU Scalability: RNN LSTM



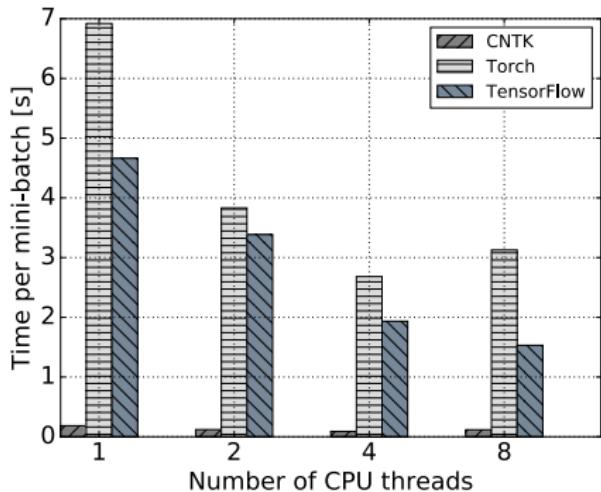
(a) Results on i7-3820.



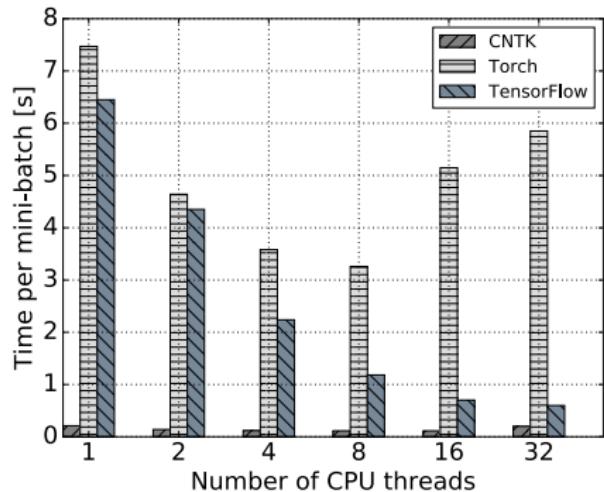
(b) Results on E5-2630.

Figure: LSTM performance on CPU platform with a mini-batch size of 256

# CPU Scalability: RNN LSTM



(a) Results on i7-3820.

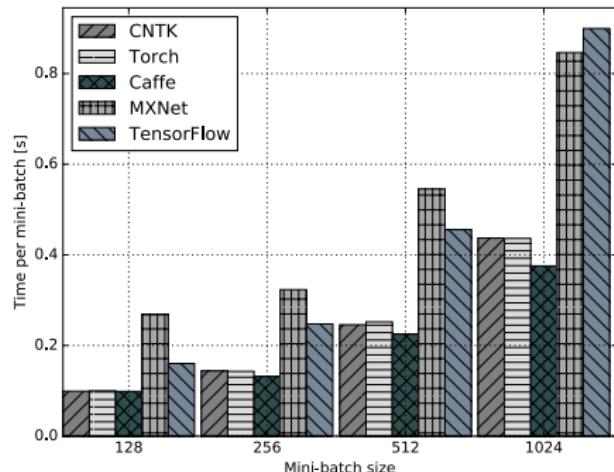


(b) Results on E5-2630.

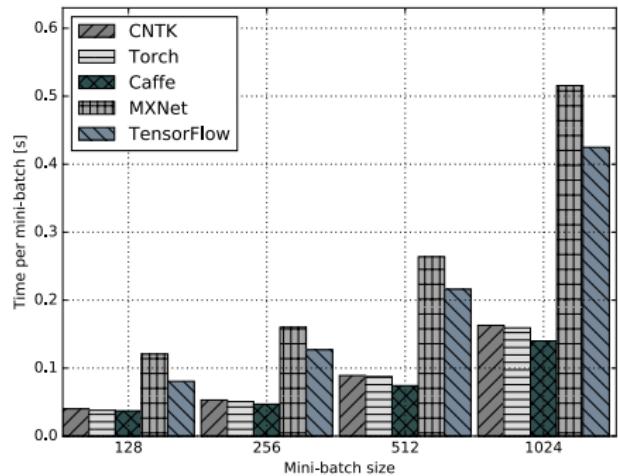
Figure: LSTM performance on CPU platform with a mini-batch size of 256

- Pay more attention to CNTK in the future
- Caffe/MXNet does not support LSTM on CPUs

# GPU Speed: FCN Synthetic



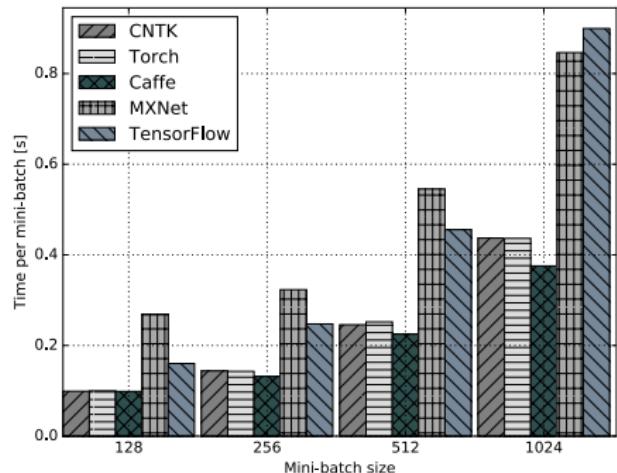
(a) Results on Tesla K80.



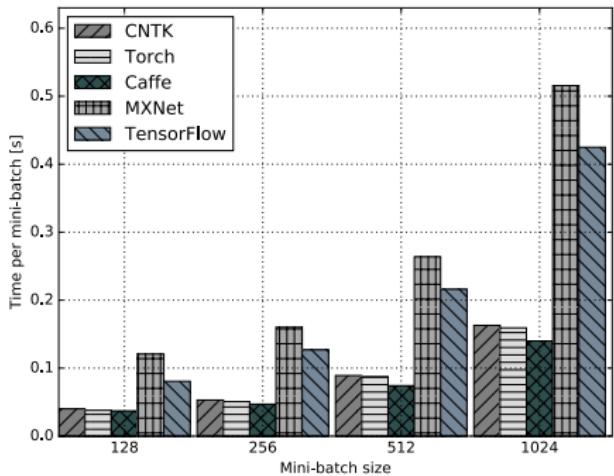
(b) Results on GTX1080.

**Figure:** The performance comparison of FCN-S on GPU platforms

# GPU Speed: FCN Synthetic



(a) Results on Tesla K80.

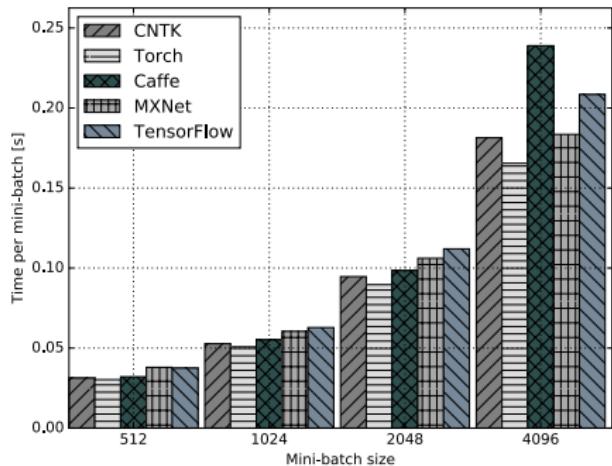


(b) Results on GTX1080.

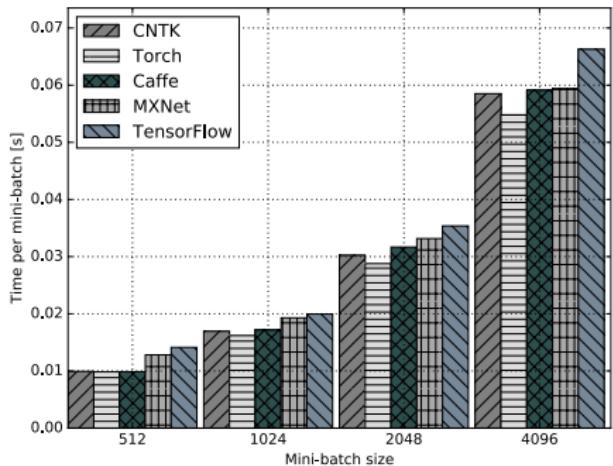
Figure: The performance comparison of FCN-S on GPU platforms

- CNTK/Torch/Caffe out-perform the others

# GPU Speed: FCN Real



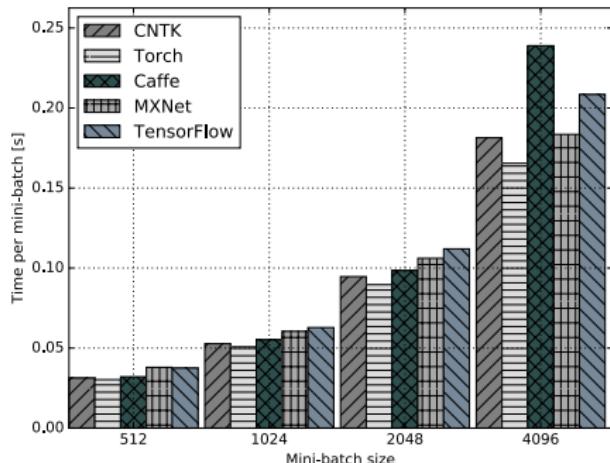
(a) Results on Tesla K80.



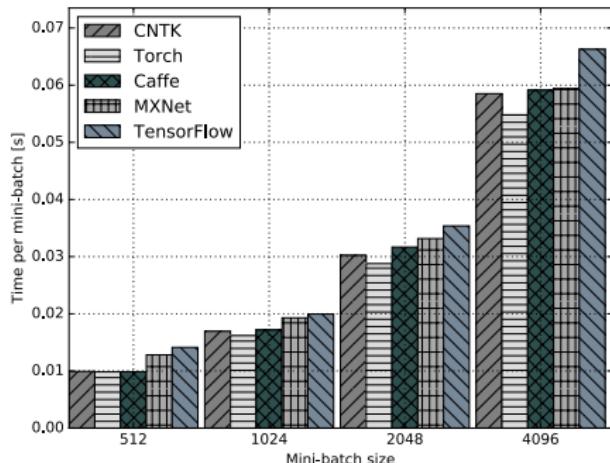
(b) Results on GTX1080.

**Figure:** The performance comparison of FCN-R on GPU platforms

# GPU Speed: FCN Real



(a) Results on Tesla K80.

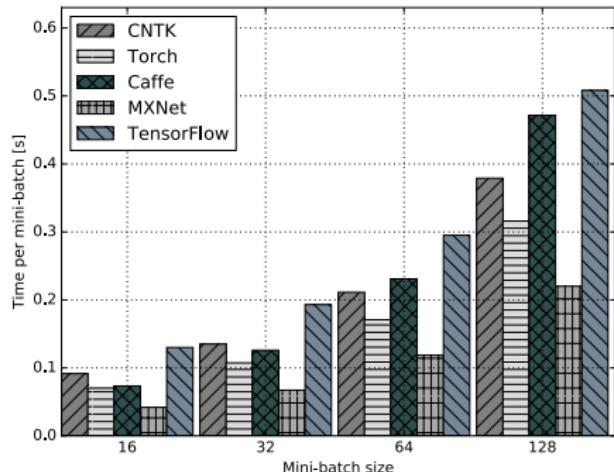


(b) Results on GTX1080.

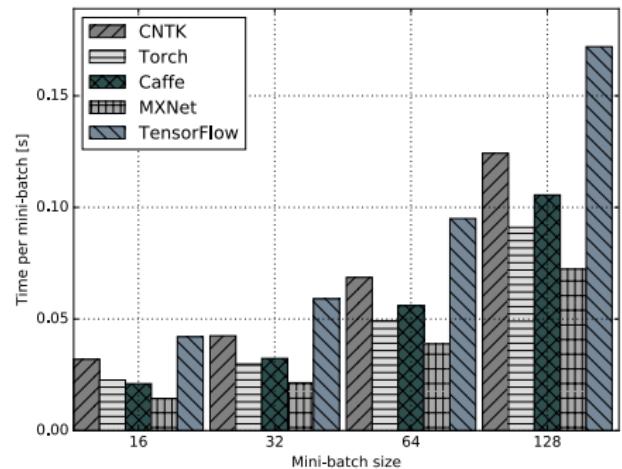
Figure: The performance comparison of FCN-R on GPU platforms

- All packages have similar performance

# GPU Speed: CNN Synthetic



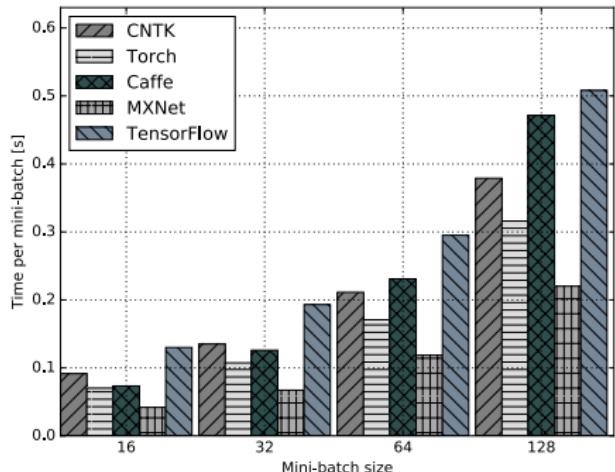
(a) Results on Tesla K80.



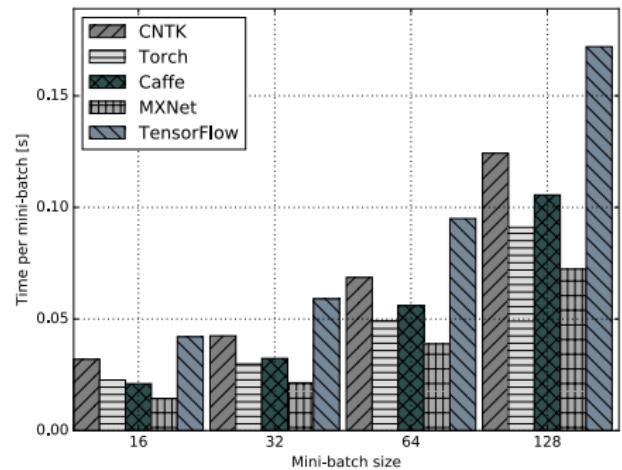
(b) Results on GTX1080.

Figure: The performance comparison of AlexNet-S on GPU platforms

# GPU Speed: CNN Synthetic



(a) Results on Tesla K80.

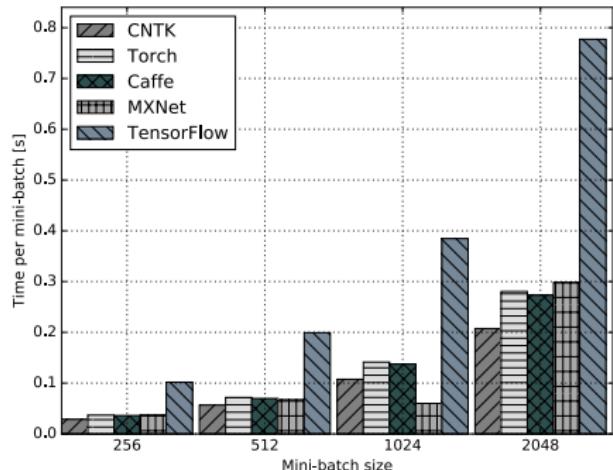


(b) Results on GTX1080.

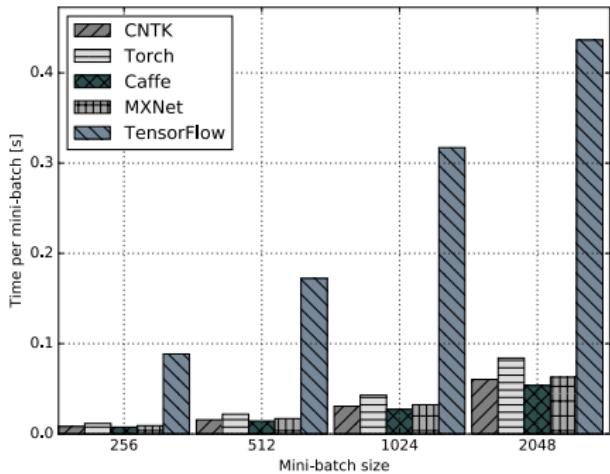
Figure: The performance comparison of AlexNet-S on GPU platforms

- MXNet out-perform the others for CNN on GPUs

# GPU Speed: CNN Real



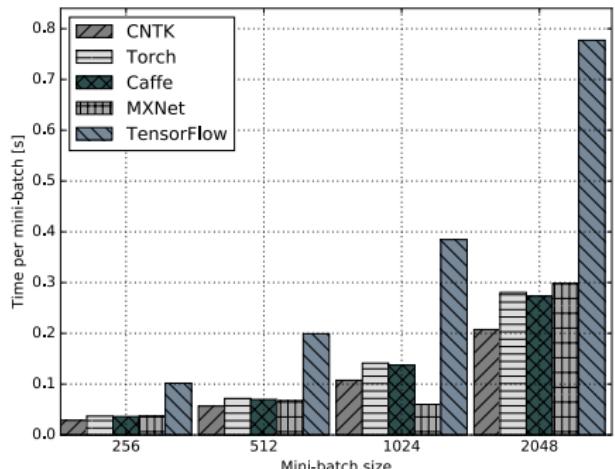
(a) Results on Tesla K80.



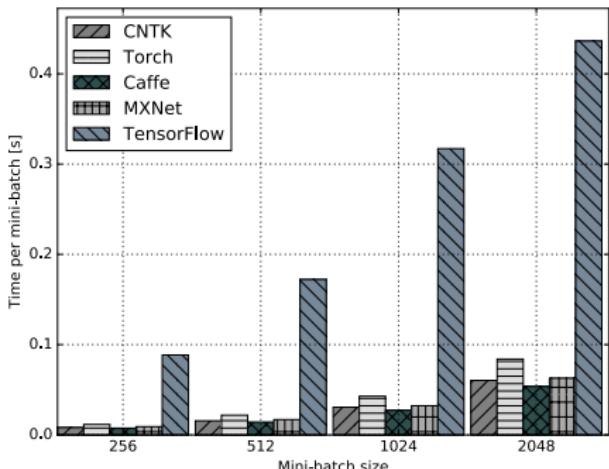
(b) Results on GTX1080.

Figure: The performance comparison of AlexNet-R on GPU platforms

# GPU Speed: CNN Real



(a) Results on Tesla K80.

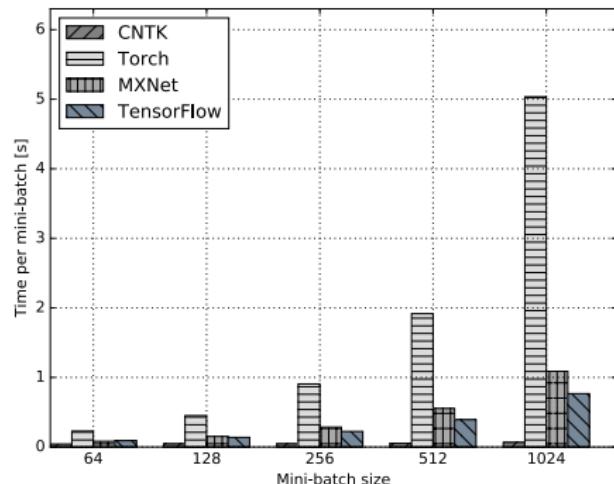


(b) Results on GTX1080.

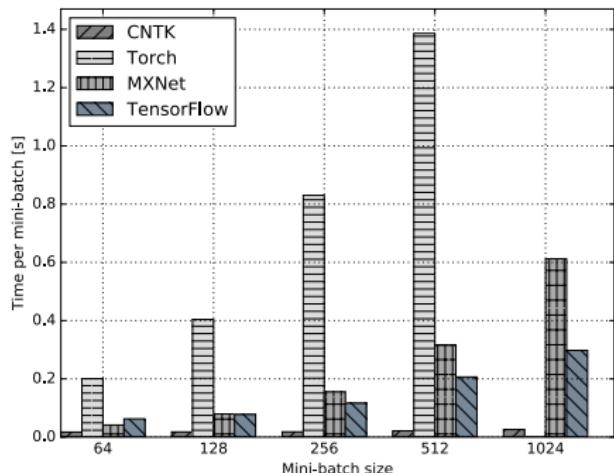
Figure: The performance comparison of AlexNet-R on GPU platforms

- TensorFlow does not have good GPU performance in general

# GPU Speed: RNN LSTM



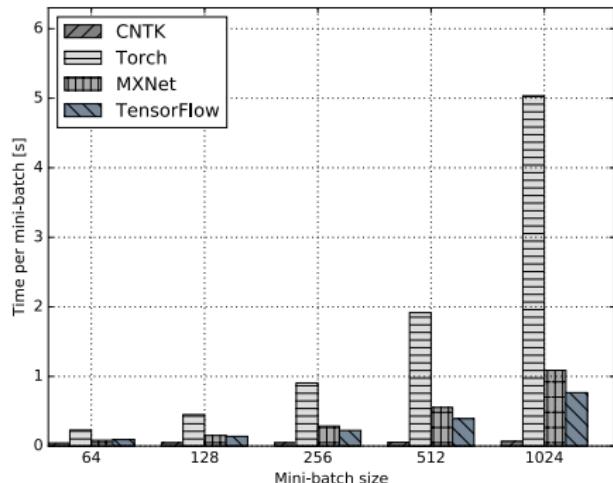
(a) Results on Tesla K80.



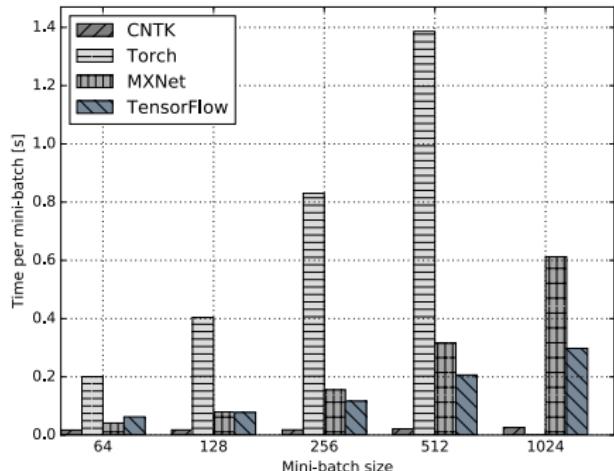
(b) Results on GTX1080.

**Figure:** The performance comparison of LSTM on GPU platforms

# GPU Speed: RNN LSTM



(a) Results on Tesla K80.

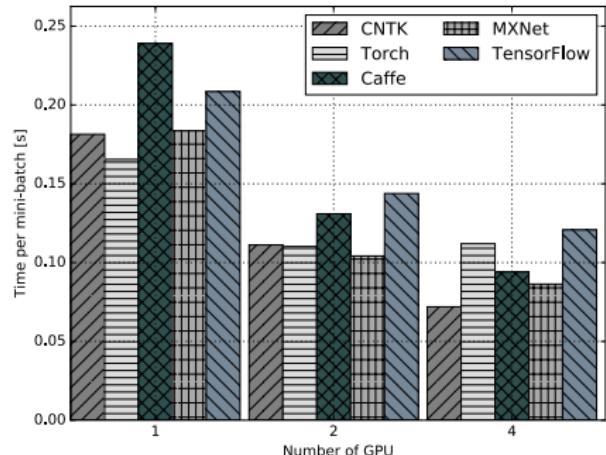


(b) Results on GTX1080.

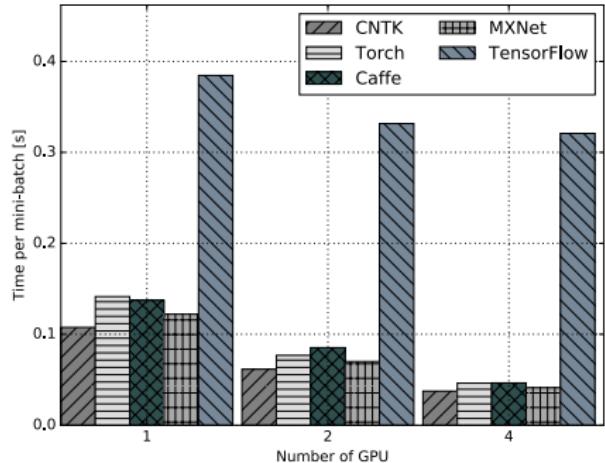
Figure: The performance comparison of LSTM on GPU platforms

- CNTK has excellent RNN performance both on CPU and GPU

# Multi-GPU Scalability



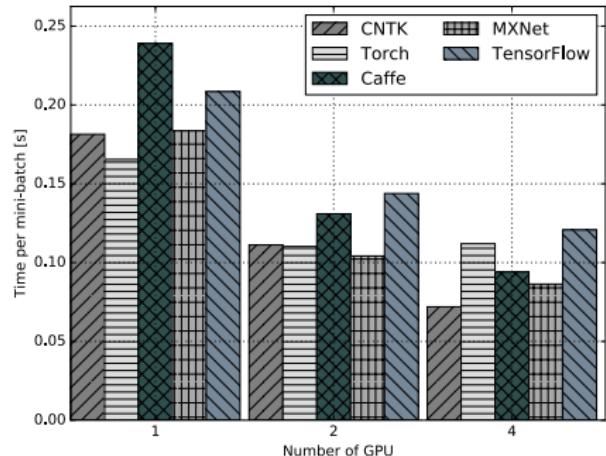
(a) FCN-R



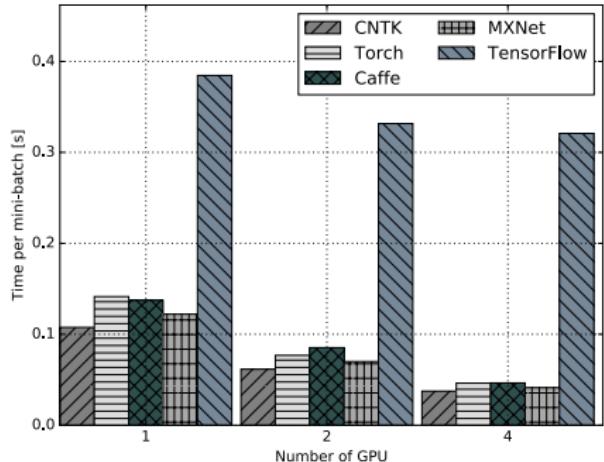
(b) CNN-R

Figure: The scalability on a multi-GPU platform ( $2 \times$  K80)

# Multi-GPU Scalability



(a) FCN-R



(b) CNN-R

Figure: The scalability on a multi-GPU platform ( $2 \times$  K80)

- Multi-GPU can greatly boost the training process of network
- MXNet shows overwhelming advantage over the others
- TensorFlow does not scale well on multi-GPU platform

## §3. Python

20 Python Language

21 Python for Scientific Computing

22 Built-in Data Structures

23 Control Flow

24 Modules

25 Additional Comments

26 Visualization



# Python Language

- Created by Guido van Rossum in 1989 and first released in 1991
- Named after “the Monty Python” (British comedy group)
- An interpreted language—simple, clear, and readable
- Python has many excellent packages for machine learning
- The language of choice in introductory programming courses



# Python for Scientific Computing

## Why Python for scientific computing?

- Dynamic data types and automatic memory management
- Full modularity, supporting hierarchical packages
- Strong introspection capabilities<sup>1</sup>
- Exception-based error handling

## Why consider such a slow language for simulation?

- Code readability and maintenance — **short code, fewer bugs**
- Good for proof-of-concept prototyping
- Implementation time versus execution time
- Well-written Python code is “fast enough” for most computational tasks
- Time critical parts executed through compiled language or **available packages**

---

<sup>1</sup>Code introspection is the ability to examine objects to know what they are, what they do and what they know. Python provides several utilities for code introspection, e.g. `dir()`, `help()`, `type()`.

# Built-in Data Structures

Numeric types: int, float, complex

```

1 a=1      # int
2 b=1L     # long int
3 c=0xf    # int (hex format)
4 d=010    # int (octal format)
5 e=1.0    # float
6 f=1+2j   # complex
  
```

Sequence types: list, tuple, str, dict

```

1 t=(3.14, True, 'Yes', [1], (0xf,))
2 l=[3.14, True, 'Yes', [1], (1L, 0xf)] + [None]*3      # tuple example
3 s='Hello' + " " + 'world!'                            # list example
4 s=("Hello", " " "world!")
5 d={1: 'int', 'pi': 3.14}                             # str example 1
6 s="Python"; s.find('thon')                           # str example 2
7
8 # dict example
9 # find substring
  
```

Formatted output

```

1 print('%(lang)s has %(num)02d quote types.' %{"lang":"Python", "num":3})
  
```

User defined functions<sup>23</sup>

```

1 def square(x):
2     return x*x
  
```

<sup>2</sup>Function overhead is high. Not to call a function repeatedly; Using aggregation instead.

<sup>3</sup>Python always passes (the value of) the reference to the object to the function.

# Control Flow

## If-then-else

```
1 a = 1
2 if a > 0:
3     print "a is positive"
4 elif a==0:
5     print "a is zero"
6 else:
7     print "a is negative"
```

## For loop

```
1 # loop from 0 to 9
2 for i in range(10):
3     print i
4
5 # loop over the list named by oldlist
6 newlist = [s.upper() for s in oldlist]
7
8 a = range(5)           # create a new list a = [0,1,2,3,4]
9 b = a                 # b points to the list a
10 c = [item for item in a] # copy list a to a new list
```

## While loop

```
1 sum = 0; i = 0
2 while i < 10:
3     sum += i
4     i += 1
```

# Modules

- This way will only introduce the module name into the name space in which the import command was issued. The names within the module will not appear in the enclosing namespace: they must be accessed through the module name.

```
import math  
math.sin(3.14)
```

- This way does not introduce the name math into the current namespace. It does introduce all public names of the math module into the current namespace.

```
from math import *  
sin(3.14)
```

- This will only import the sin function from math module and introduce the name sin into the current namespace, but it will not introduce the name math into the current namespace, directly use

```
from math import sin  
sin(3.14)
```

- Make it as local as possible to avoid import overhead; But avoid calling it repeatedly; **If possible, avoid it!**

# Additional Comments

- ① In Python, everything (including functions, modules, and files) are objects. A variable is created through assignment:

```
x = y = z = 0.1
```

- ② help() is a function which gives information about the object. For example,

```
help('modules')      # generate a list of all modules that can be imported  
help('modules time') # generate a list of modules with 'time' in description
```

- ③ Use a profiler to find optimization possibilities

```
import profile        # cProfile is now recommended  
profile.run('main()')
```

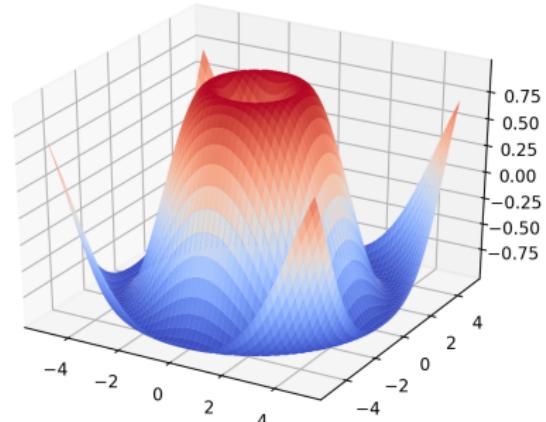
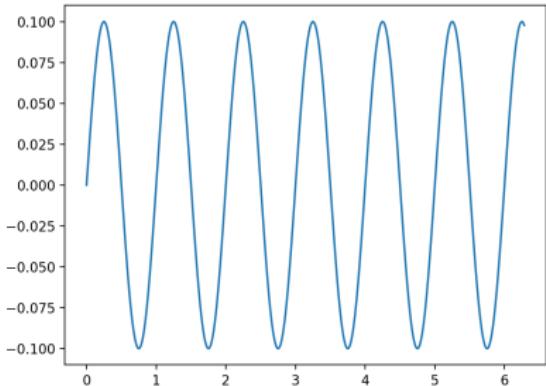
- ④ Some useful and important packages

- NumPy: for scientific computing
- Matplotlib: for visualising data
- SciPy: providing lots of numerical algorithms
- SymPy: for symbolic mathematics

# Visualization

```
1 # import pyplot module for visualization
2 import matplotlib.pyplot as plt
3 import numpy as np
4
5 # generate points (x,y)
6 dx = 0.001
7 x = np.arange(0, 2*np.pi, dx)
8 y = 0.1*np.sin(2*np.pi*x)
9
10 # plot (x,y) curve
11 plt.plot(x, y)
12 plt.show()
```

```
1 from mpl_toolkits.mplot3d import Axes3D
2 import matplotlib.pyplot as plt
3 from matplotlib import cm
4 import numpy as np
5
6 X = np.arange(-5, 5, 0.01)
7 Y = np.arange(-5, 5, 0.01)
8 X, Y = np.meshgrid(X, Y)
9 R = np.sqrt(X**2 + Y**2)
10 Z = np.sin(R)
11
12 # Plot the surface
13 fig = plt.figure()
14 ax = fig.gca(projection='3d')
15 surf = ax.plot_surface(X, Y, Z,
16                         cmap=cm.coolwarm)
17 plt.show()
```



## §4. TensorFlow

- 27 Basic Concepts
- 28 TensorFlow Operations
- 29 Computational Graph
- 29 Computational Graph
- 30 Example: SoftMax in TF
- 31 Example: SoftMax in TF (Cont)
- 32 Visualization: TensorBoard
- 33 TensorBoard: Scalars
- 34 TensorBoard: Graphs
- 35 TensorBoard: Distributions
- 36 TensorBoard: Histograms



# Basic Concepts

- ☞ **Graph:** In TensorFlow, ML algorithms are represented as computational graph. A computational graph (data flow graph) is a form of directed graph where vertices (nodes) describe operations, while edges represent data flowing between these operations.
- **Operation:** An operation may represent a variable or constant, a control flow directive, a mathematical function, a file I/O, or a network communication port.
- **Tensor:** A tensor is a multi-dimensional collection of homogeneous values with a fixed static type.
- **Variable:** Variables can be described as persistent, mutable handles to in-memory buffers storing tensors.
- ☞ **Session:** In TensorFlow the execution of operations and evaluation of tensors may only be performed in a special environment called session.

# TensorFlow Operations

Operation groups	Operations Examples
Maths	Add, Sub, Mul, Dic, Exp, Log, Greater than, Less than, Equal
Array	Concat, Slice, Constant, Rank, Shape
Matrix	MatMul, MatrixInverse, MatrixDeterminant
Neuronal Network	SoftMax, Sigmoid ReLU, Convolution2D, MaxPool
Checkpointing	Save, Restore
Queues and sincronizations	Enqueue, Dequeue, MutexAcquire, MutexRelease
Flow control	Merge, Switch, Enter, Leave, NextIteration

# Computational Graph

```
1 import tensorflow as tf
2
3 # Define a computational graph
4 graph = tf.Graph()
5
6 # Define operation nodes in the above graph
7 with graph.as_default():
8     # Define two constants(two nodes)
9     a = tf.constant(1.0)
10    b = tf.constant(2.0)
11    # Define an operation node: c = a * b
12    c = tf.multiply(a, b)
13
14    # Add scalar summary to operation node
15    tf.summary.scalar('c', c)
16
17    # Merge all the summaries and write to ./board
18    merged = tf.summary.merge_all()
19    writer = tf.summary.FileWriter('./board', graph)
20
21 with tf.Session(graph=graph) as sess:
22     # Output the result(c=a*b)
23     result = sess.run(c)
24     print(result)
25     # Write data to ./board
26     writer.add_summary(merged.eval())
```

# Computational Graph

```
1 import tensorflow as tf
2
3 # Define a computational graph
4 graph = tf.Graph()
5
6 # Define operation nodes in the above graph
7 with graph.as_default():
8     # Define two constants(two nodes)
9     a = tf.constant(1.0)
10    b = tf.constant(2.0)
11    # Define an operation node: c = a * b
12    c = tf.multiply(a, b)
13
14    # Add scalar summary to operation node
15    tf.summary.scalar('c', c)
16
17    # Merge all the summaries and write to ./board
18    merged = tf.summary.merge_all()
19    writer = tf.summary.FileWriter('./board', graph)
20
21 with tf.Session(graph=graph) as sess:
22     # Output the result(c=a*b)
23     result = sess.run(c)
24     print(result)
25     # Write data to ./board
26     writer.add_summary(merged.eval())
```





# Example: SoftMax in TF

```
1 import tensorflow as tf
2
3 # Import training and test data from MNIST
4 import tensorflow.examples.tutorials.mnist.input_data as input_data
5 mnist = input_data.read_data_sets("MNIST_data/", one_hot=True)
6
7 # Create a new TensorFlow graph
8 graph = tf.Graph()
9
10 with graph.as_default():
11
12     # Nodes can be grouped into visual blocks for TensorBoard
13     with tf.name_scope('input_features'):
14         # Placeholder for input variables (None = variable dimension)
15         x = tf.placeholder(tf.float32, shape=[None, 784], name='input_x')
16
17     with tf.name_scope('input_labels'):
18         # Placeholder for labels
19         y_ = tf.placeholder(tf.float32, shape=[None, 10], name='labels')
20
21     with tf.name_scope('parameters'):
22         W = tf.Variable(tf.zeros([784, 10]), name='weights')
23         b = tf.Variable(tf.zeros([10]), name='biases')
24
25         # Track tensor distributions over time for TensorBoard
26         tf.summary.histogram('WEIGHTS', W)
27         tf.summary.histogram('BIASES', b)
28
29     with tf.name_scope('use_softmax'):
30         # Apply softmax to the input data and get prediction y
31         y = tf.nn.softmax(tf.matmul(x, W) + b)
```

# Example: SoftMax in TF (Cont)

```
32     with tf.name_scope('train'):
33         # Compute the cross entropy of real label y_ and prediction label y
34         cross_entropy = -tf.reduce_sum(y_*tf.log(y))
35         # Create a gradient-descent optimizer with learning rate = 0.01
36         train_step = tf.train.GradientDescentOptimizer(0.01).minimize(cross_entropy)
37
38     with tf.name_scope('test'):
39         correct_prediction = tf.equal(tf.argmax(y,1), tf.argmax(y_,1))
40         accuracy = tf.reduce_mean(tf.cast(correct_prediction, "float"))
41         # Track tensor values over time for TensorBoard
42         tf.summary.scalar('Accuracy', accuracy)
43
44     merged = tf.summary.merge_all()    # Merge all the summaries
45     writer = tf.summary.FileWriter('./board', graph) # Write summary file to ./board
46
47 with tf.Session(graph=graph) as sess:
48     # Initialize all variables
49     tf.global_variables_initializer().run()
50
51     for step in range(1000):
52         if (step%10) == 0:
53             # Feed test data to compute accuracy
54             feed = {x: mnist.test.images, y_: mnist.test.labels}
55             _, acc = sess.run([merged, accuracy], feed_dict=feed)
56             print('Accuracy at %s step: %s' %(step, acc))
57         else:
58             # Feed training data to train the model
59             batch_x, batch_y = mnist.train.next_batch(100)
60             sess.run(train_step, feed_dict={x: batch_x, y_: batch_y})
61             writer.add_summary(merged.eval(feed_dict={x:batch_x, y_:batch_y}),
62                               global_step=step)
```

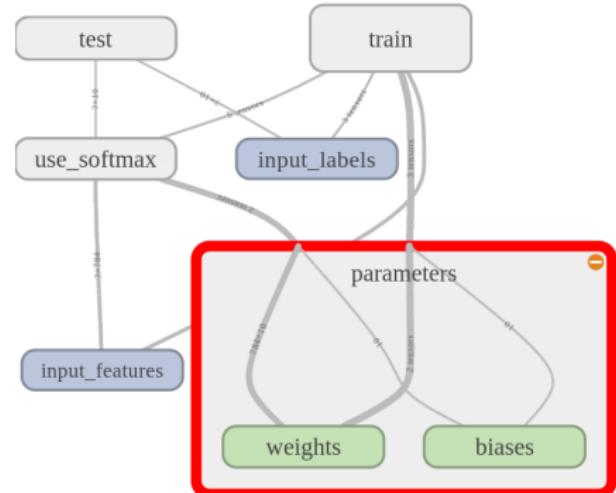
# Visualization: TensorBoard

## Declarative style

- Same style as Teano
- Easy to understand
- Possible for graph optimization
- More difficult for debugging

Computation graphs are powerful but complicated

- Thousands of nodes or more
- Network is deep
- Visualization is helpful for debugging



To run TensorBoard, use the following command:

```
$ tensorboard --logdir=path/to/log_directory
```

Then navigate your web browser to **localhost:6006** to view the TensorBoard.



# TensorBoard: Scalars

TensorBoard

SCALARS IMAGES AUDIO GRAPHS DISTRIBUTIONS HISTOGRAMS EMBEDDINGS

test 1

Split on underscores

Data download links

Tooltip sorting method: default

Smoothing:

Horizontal Axis: STEP (selected), RELATIVE, WALL

Runs

.  Toggle All Runs

./board

Figure: test/Accuracy

Step	Accuracy
0.000	0.500
200	0.900
400	0.900
600	0.900
800	0.900
900	0.900

# TensorBoard: Graphs

TensorBoard      SCALARS      IMAGES      AUDIO      GRAPHS      DISTRIBUTIONS      HISTOGRAMS      EMBEDDINGS

Fit to screen      Download PNG

Run (1)

Session runs (0)

Upload Choose File

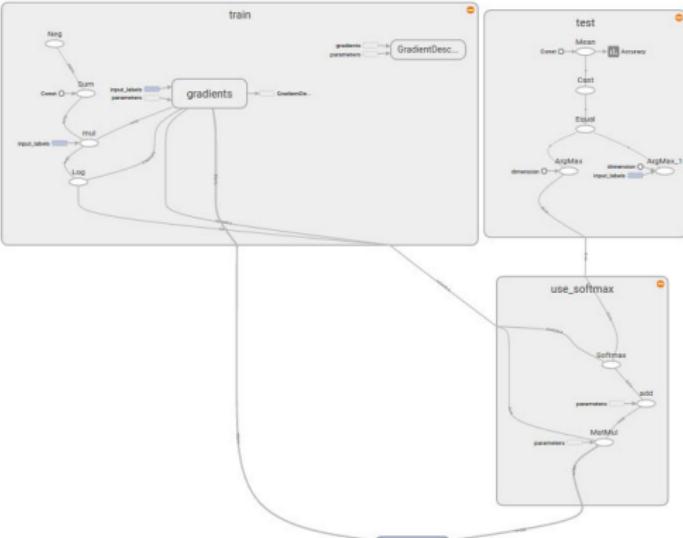
Trace inputs

Color  Structure  Device

colors same substructure  
unique substructure (\* = expandable)

Graph Namespace\*  
OpNode  
Unconnected series\*  
Connected series\*  
Constant  
Summary  
Dataflow edge  
Control dependency edge  
Reference edge

### Main Graph



### Auxiliary Nodes

- input\_labels → test
- parameters → use\_softmax
- input\_features → use\_softmax

# TensorBoard: Distributions

TensorBoard

SCALARS IMAGES AUDIO GRAPHS DISTRIBUTIONS HISTOGRAMS EMBEDDINGS

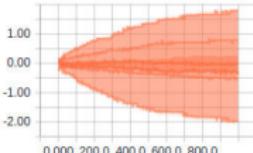
Split on underscores X

Horizontal Axis  
 STEP  RELATIVE  WALL

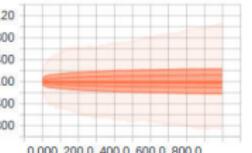
Runs  
Write a regex to filter runs

parameters

parameters/BIASES



parameters/WEIGHTS



TOGGLE ALL RUNS

./board

# TensorBoard: Histograms

TensorBoard

SCALARS    IMAGES    AUDIO    GRAPHS    DISTRIBUTIONS    HISTOGRAMS    EMBEDDINGS

Split on underscores X

Write a regex to create a tag group

Histogram Mode OVERLAY OFFSET

Offset Time Axis STEP RELATIVE WALL

Runs

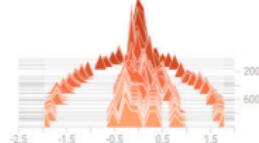
Write a regex to filter runs ✓ ○ .

TOGGLE ALL RUNS

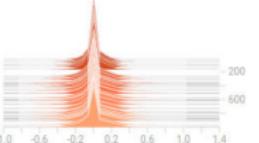
./board

parameters

parameters/BIASES



parameters/WEIGHTS



## §5. MXNet

37 General Comments

38 Example: SoftMax in MXNet

39 Example: SoftMax in MXNet



40 Visualization of the SoftMax Example

# General Comments

- ① Used to power Amazon Web Services (AWS)
- ② Support many different applications (e.g. computer vision, natural language processing, speech recognition, unsupervised machine learning, support embedded APIs, visualization)
- ③ Mixed programming style: imperative and declarative
  - Light-weighted (around 50K lines of core code)
  - Data parallelism with multi-devices: 88% scalability with 256 GPUs
  - Support many front-ends, including JavaScript (run on web browsers)
  - Provide intermediate-level and high-level interface modules
  - Provide abundant IO functions
- ④ Fully compatible with Torch: modules and operators
- ⑤ Visualize neural network graphs
  - Call `mx.viz.plot_network()`
- ⑥ Not well documented, code not easy to read



# Example: SoftMax in MXNet

```
1 import os, urllib, gzip, struct
2 import numpy as np
3
4 # Fetch the MNIST dataset
5 def download_data(url, force_download=True):
6     fname = url.split('/')[-1]
7     if force_download or not os.path.exists(fname):
8         urllib.urlretrieve(url, fname)
9     return fname
10 def read_data(label_url, image_url):
11     with gzip.open(download_data(label_url)) as flbl:
12         magic, num = struct.unpack(>II", flbl.read(8))
13         label = np.fromstring(flbl.read(), dtype=np.int8)
14     with gzip.open(download_data(image_url), 'rb') as fimg:
15         magic, num, row, col = struct.unpack(>IIII", fimg.read(16))
16         image = np.fromstring(fimg.read(), dtype=np.uint8).reshape(len(label),row,col)
17     return (label, image)
18 path='http://yann.lecun.com/exdb/mnist/'
19 train_lbl, train_img = read_data(path+'train-labels-idx1-ubyte.gz',
20                                 path+'train-images-idx3-ubyte.gz')
21 test_lbl, test_img = read_data(path+'t10k-labels-idx1-ubyte.gz',
22                                 path+'t10k-images-idx3-ubyte.gz')
23
24 # Create data iterators for MXNet
25 import mxnet as mx
26 def to4d(img):
27     return img.reshape(img.shape[0], 1, 28, 28).astype(np.float32)/255
28 batch_size = 100
29 train_iter = mx.io.NDArrayIter(to4d(train_img), train_lbl, batch_size, shuffle=True)
30 test_iter = mx.io.NDArrayIter(to4d(test_img), test_lbl, batch_size)
```

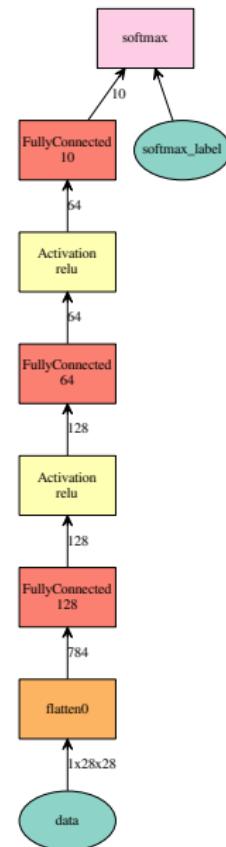


# Example: SoftMax in MXNet

```
32 import logging
33 logging.getLogger().setLevel(logging.INFO)
34
35 # A multilayer perceptron contains several fully-connected layers
36 data = mx.sym.Variable('data') # Create a place holder variable for the input data
37 data = mx.sym.Flatten(data=data) # Flatten the data from 4-D shape into 2-D
38
39 # The first fully-connected layer and according activation function
40 fc1 = mx.sym.FullyConnected(data=data, name='fc1', num_hidden=128)
41 act1 = mx.sym.Activation(data=fc1, name='relu1', act_type="relu")
42
43 # The second fully-connected layer and the according activation function
44 fc2 = mx.sym.FullyConnected(data=act1, name='fc2', num_hidden = 64)
45 act2 = mx.sym.Activation(data=fc2, name='relu2', act_type="relu")
46
47 # The thrid fully-connected layer, note that the hidden size should be 10
48 fc3 = mx.sym.FullyConnected(data=act2, name='fc3', num_hidden=10)
49 # The softmax and loss layer
50 out = mx.sym.SoftmaxOutput(data=fc3, name='softmax')
51 mod = mx.mod.Module(out)
52
53 # Visualize the network structure with output size
54 mx.viz.plot_network(symbol=out, shape={'data': (batch_size, 1, 28, 28)}).view()
55
56 # Train the model with the above network and data iterators
57 mod.fit(train_data=train_iter, eval_data=test_iter, num_epoch=10)
```

[https://github.com/dmlc/mxnet/blob/master/example/svm\\_mnist/svm\\_mnist.py](https://github.com/dmlc/mxnet/blob/master/example/svm_mnist/svm_mnist.py)

# Visualization of the SoftMax Example



## §6. Torch

41 Programming Interface

42 Example: Two-Layer Network

43 Example: Two-Layer Network (Cont)

44 Example: Linear Regression in Lua

45 Example: Linear Regression in Lua (Cont)

46 Example: Linear Regression in Lua (Cont)



# Programming Interface

- ➊ Very fast: Fastest scripting language LuaJIT is used
- ➋ Flexible with wide range of applications
  - Speech, image, and video applications
  - Large-scale machine-learning applications
  - Used by Facebook, Twitter, Deepmind
  - Easy extensibility: integrate any library into Torch
- ➌ Portable to any platform
  - Torch can run on iPhone with no modification to scripts
  - Embeddable, with ports to iOS, Android and FPGA backends
- ➍ No automatic differentiation



# Example: Two-Layer Network

```
1 import torch
2 from torch.autograd import Variable
3
4 # N is batch size; D_in is input dimension;
5 # H is hidden dimension; D_out is output dimension.
6 N, D_in, H, D_out = 64, 1000, 100, 10
7
8 # Create random Tensors to hold inputs and outputs, and wrap them in Variables.
9 x = Variable(torch.randn(N, D_in))
10 y = Variable(torch.randn(N, D_out), requires_grad=False)
11
12 # Use the nn package to define our model as a sequence of layers.
13 model = torch.nn.Sequential( torch.nn.Linear(D_in, H),
14                             torch.nn.ReLU(),
15                             torch.nn.Linear(H, D_out) )
16
17 # The nn package also contains definitions of popular loss functions;
18 loss_fn = torch.nn.MSELoss(size_average=False)
19
20 learning_rate = 1e-4
21
22 for t in range(500):
23     # Forward pass: compute predicted y by passing x to the model.
24     y_pred = model(x)
25     # Compute and print loss.
26     loss = loss_fn(y_pred, y)
27     print(t, loss.data[0])
28     # Zero the gradients before running the backward pass
29     model.zero_grad()
```



## Example: Two-Layer Network (Cont)

```
30 # Backward pass: compute gradient of the loss
31 loss.backward()
32
33 # Update the weights using gradient descent
34 for param in model.parameters():
35     param.data -= learning_rate * param.grad.data
36     end
37 end
```

<https://github.com/jcjohnson/pytorch-examples>

# Example: Linear Regression in Lua

```
1 require 'torch'
2 require 'optim'
3 require 'nn'
4
5 # write the loss to a text file and read from there to plot it as training proceeds
6 logger = optim.Logger('loss_log.txt')
7
8 # input data
9 data = torch.Tensor{{40, 6, 4},{44, 10, 4},{46, 12, 5},
10 {48, 14, 7},{52, 16, 9},{58, 18, 12},{60, 22, 14},
11 {68, 24, 20},{74, 26, 21},{80, 32, 24}}
12
13 # define the container
14 model = nn.Sequential()
15 ninputs = 2; noutputs = 1
16
17 # define the only module
18 model:add(nn.Linear(ninputs, noutputs))
19
20 # Define a loss function
21 criterion = nn.MSECriterion()
22
23 # retrieve its trainable parameters
24 x, dl_dx = model:getParameters()
25
26 # compute loss function and its gradient
27 feval = function(x_new)
28     # set x to x_new, if differnt
29     if x ~= x_new then
30         x:copy(x_new)
31     end
32 end
```

# Example: Linear Regression in Lua (Cont)

```
32 # select a new training sample
33 _nidx_ = (_nidx_ or 0) + 1
34 if _nidx_ > (#data)[1] then _nidx_ = 1 end
35
36 local sample = data[_nidx_]
37 local target = sample[{ {1} }]
38 local inputs = sample[{ {2,3} }]
39
40 # reset gradients
41 dl_dx:zero()
42
43 # evaluate the loss function and its derivative wrt x
44 local loss_x = criterion:forward(model:forward(inputs), target)
45 model:backward(inputs, criterion:backward(model.output, target))
46
47 # return loss(x) and dloss/dx
48 return loss_x, dl_dx
49 end
50
51 # define SGD
52 sgd_params = {
53     learningRate = 1e-3,
54     learningRateDecay = 1e-4,
55     weightDecay = 0,
56     momentum = 0
57 }
58
59 # we cycle 10,000 times over our training data
60 for i = 1,1e4 do
61     #this variable is used to estimate the average loss
62     current_loss = 0
```



# Example: Linear Regression in Lua (Cont)

```
63 #an epoch is a full loop over our training data
64 for i = 1,(#data)[1] do
65     # return new x and value of the loss functions
66     _, fs = optim.sgd(feval,x,sgd_params)
67     # update loss
68     current_loss = current_loss + fs[1]
69 end
70
71 # report average error on epoch
72 current_loss = current_loss / (#data)[1]
73 print('current loss = ' .. current_loss)
74
75 logger:add{['training error'] = current_loss}
76 logger:style{['training error']} = '-'
77 logger:plot()
78
79
80 # Test the trained model
81 text = {40.32, 42.92, 45.33, 48.85, 52.37, 57, 61.82, 69.78, 72.19, 79.42}
82
83 for i = 1,(#data)[1] do
84     local myPrediction = model:forward(data[i][{{2,3}}])
85     print(string.format("%2d %6.2f %6.2f", i, myPrediction[1], text[i]))
86 end
```

## §7. Caffe

47 General Comments

48 Example: Image Classification

49 Example: Extend Layers

50 Example: Extend Layers (Cont)

Caffe

Deep learning framework

by BAIR

# General Comments

- ➊ Mainly focus on (and well suited for) CNN and image recognition
  - Use an expressive architecture
  - Provide simple command line tools for training and prediction
  - Allow defining models and optimization by configuration
  - Use Google protocol tool to define parameters for nets and solvers
- ➋ Not so convenient to extend
  - Write layers in C++ or Python, handwritten CUDA code
  - Not well documented
  - No automatic differentiation
  - Not as flexible as other engines
  - Python interface not transparent
- ➌ Lots of dependencies; can be tricky to install



# Example: Image Classification

```
1 import caffe
2 import matplotlib.pyplot as plt
3
4 # paste your image URL here
5 my_image_url = "https://wikipedia/Orang_Utan/2C_Malaysia.JPG"
6 !wget -O image.jpg $my_image_url
7
8 # transform it and copy it into the net
9 image = caffe.io.load_image('image.jpg')
10 caffe.net.blobs['data'].data[...] = transformer.preprocess('data',image)
11
12 # perform classification
13 caffe.net.forward()
14
15 # obtain the output probabilities
16 output_prob = net.blobs['prob'].data[0]
17
18 # sort top five predictions from softmax output
19 top_inds = output_prob.argsort()[:-1][:-5]
20
21 #
22 plt.imshow(image)
23
24 #
25 print 'probabilities and labels:'
26 zip(output_prob[top_inds], labels[top_inds])
```

# Example: Extend Layers

```
1 import caffe
2 import numpy as np
3
4 class EuclideanLoss(caffe.Layer):
5     def setup(self, bottom, top):
6         if len(bottom) != 2: #check number of input data
7             raise Exception("Need two inputs to compute distance")
8
9     def reshape(self, bottom, top):
10        #check input dimensions match
11        if bottom[0].count != bottom[1].count:
12            raise Exception("Inputs must have the same dimension")
13        #difference in shape of inputs
14        self.diff = np.zeros_like(bottom[0].data, dtype=np.float32)
15        # loss output is scalar
16        top[0].reshape(1)
17
18    def forward(self, bottom, top):
19        self.diff[...] = bottom[0].data - bottom[1].data
20        top[0].data[...] = np.sum(self.diff**2)/bottom[0].num/2.
21
22    def backward(self, top, propagate_down, bottom):
23        for i in range(2):
24            if not propagate_down[i]:
25                continue
26            if i == 0:
27                sign = 1
28            else:
29                sign = -1
30            bottom[i].diff[...] = sign * self.diff / bottom[1].num
```

# Example: Extend Layers (Cont)

Define a class in Python to extend Layer

```
31 layer{
32     type: "Python"
33     python_param {
34         module: "layers"
35         layer: "EuclideanLoss"
36     }
37 }
```

[https://docs.google.com/presentation/d/1UeKXVgRvvxg9OUdh\\_UiC5G71UMscNPlvArsWER41PsU/edit#slide=id.gc2fdcce7\\_216\\_0](https://docs.google.com/presentation/d/1UeKXVgRvvxg9OUdh_UiC5G71UMscNPlvArsWER41PsU/edit#slide=id.gc2fdcce7_216_0)

Thank You!

