ADAC Workshop 2017

Jan. 25, 2017

A Study of Parallel Computation for Deep Learning on TSUBAME

Ikuro Sato, Denso IT Laboratory, Inc.
isato@d-itlab.co.jp

Major collaborators: Yosuke Oyama, Akihiro Nomura, Satoshi Matsuoka
(Tokyo Institute of Technology)

DENSO Group

IT LAB

Outline

- **Introduction**
- Learning algorithm
- Performance modeling

Brief introduction of myself

2005 Received ph.D. in nuclear physics

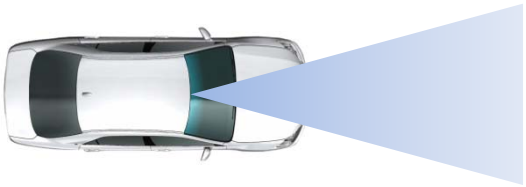
Simulation of quarks and gluons

2005-07 Postdoc at Lawrence Berkeley NL

Thanks to DOE

2008 Employed at Denso IT Lab.

now Engages primarily
on computer vision and pattern recognition
for Advanced Driver Assistance Systems (ADAS).

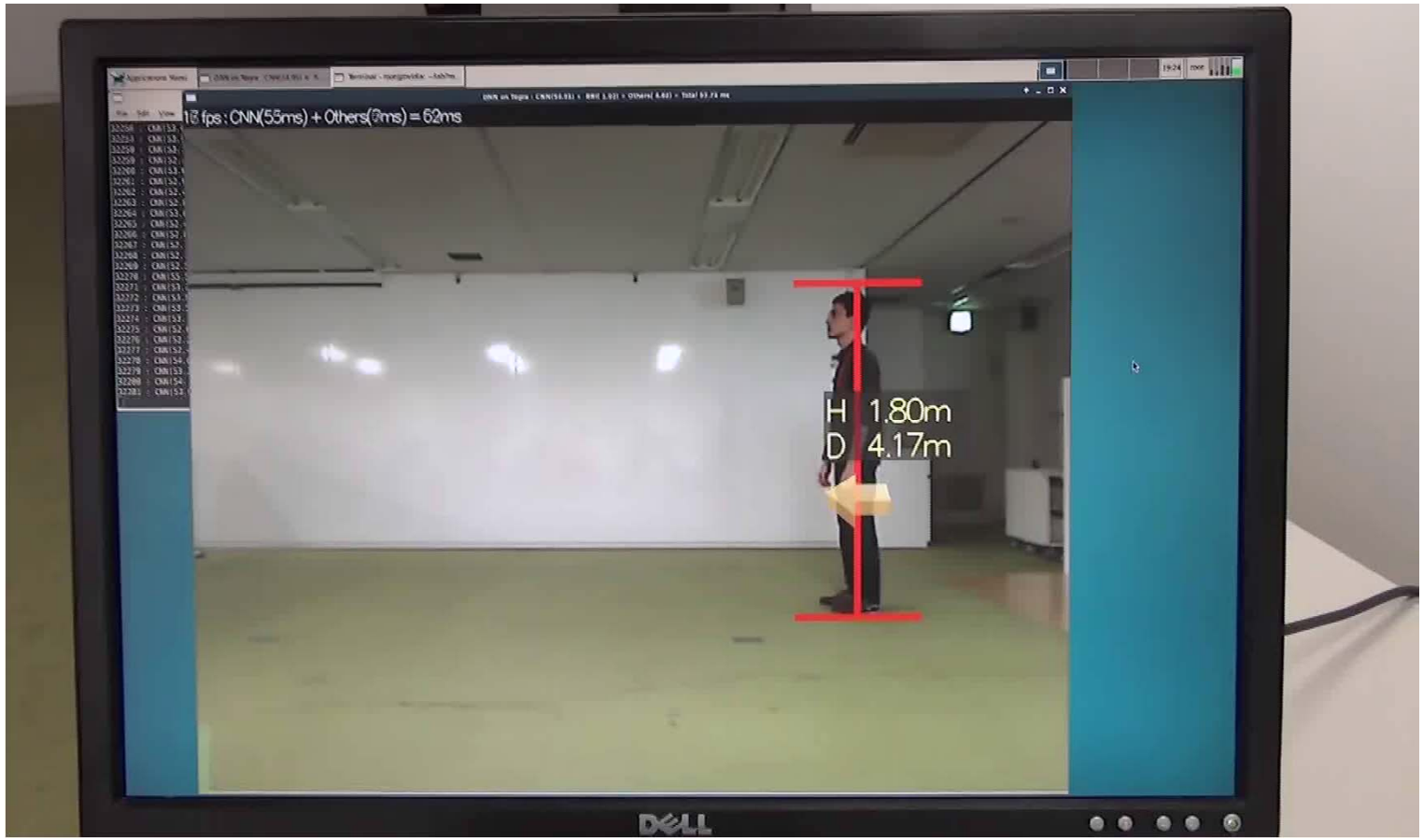


DENSO Group

IT LAB

Pedestrian recognition demo

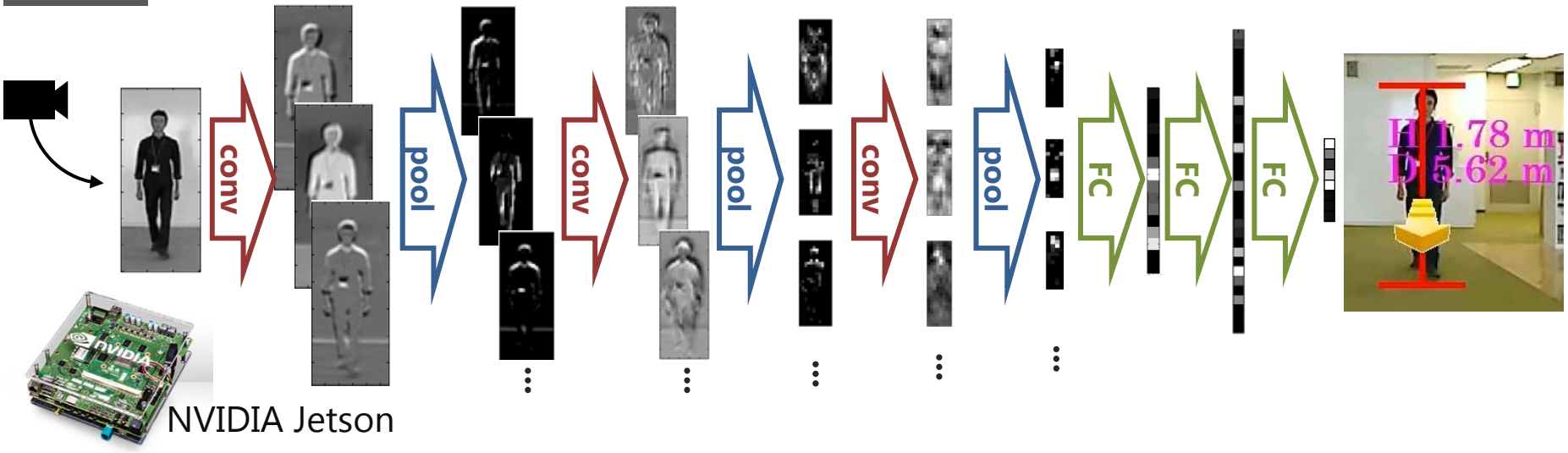
Pedestrian detection with distance, height, & body orientation estimation by a Deep Convolutional Neural Network (CNN).



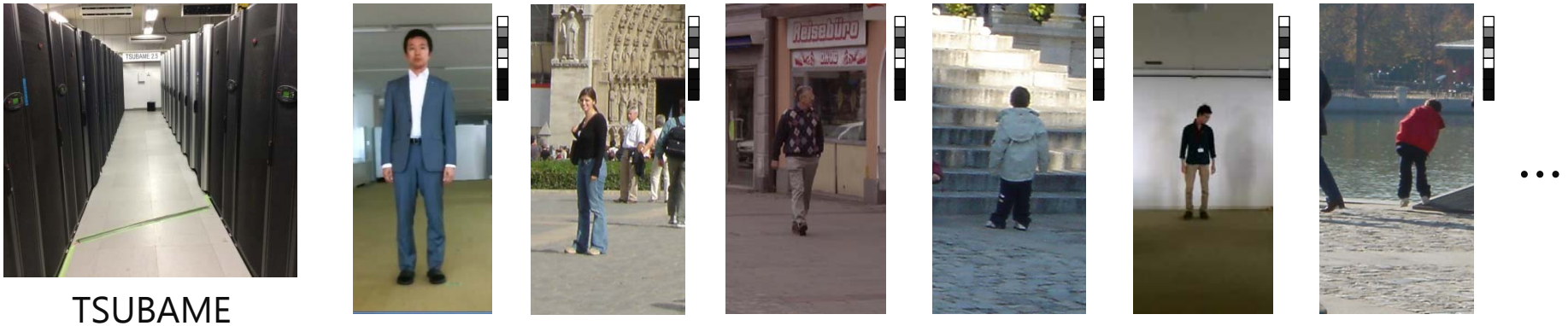
I. Sato and H. Niihara, “Beyond Pedestrian Detection: Deep Neural Networks Level-Up Automotive Safety”, GTC 2014.
I. Sato, et al., “Visual Recognition of Pedestrians with Deep Neural Networks”, ITSWC 2014.

How does this work?

online Feedforward an image one at a time to the parameter-fixed network.



offline Feed many labeled images to the network to tune the parameters.



How much training samples do we need?

As much as possible.

CNN performance does not comparably saturate for large datasets.

dataset	# images	CNN vs non-CNN
ILSVRC	**1,000K**	**CNN** wins
PASCAL	**10K**	**non-CNN** wins

http://image-net.org/challenges/LSVRC/

M. Oquab+, CVPR2014.

How much training samples do we need?

As much as possible.

CNN performance does not comparably saturate for large datasets.

dataset	# images	CNN vs non-CNN
ILSVRC	**1,000K**	**CNN** wins
PASCAL	**10K**	**non-CNN** wins

http://image-net.org/challenges/LSVRC/

M. Oquab+, CVPR2014.

An easy way to get free samples

Data Augmentation.

- Images processed in many ways.
- **Virtually infinite samples generated.**

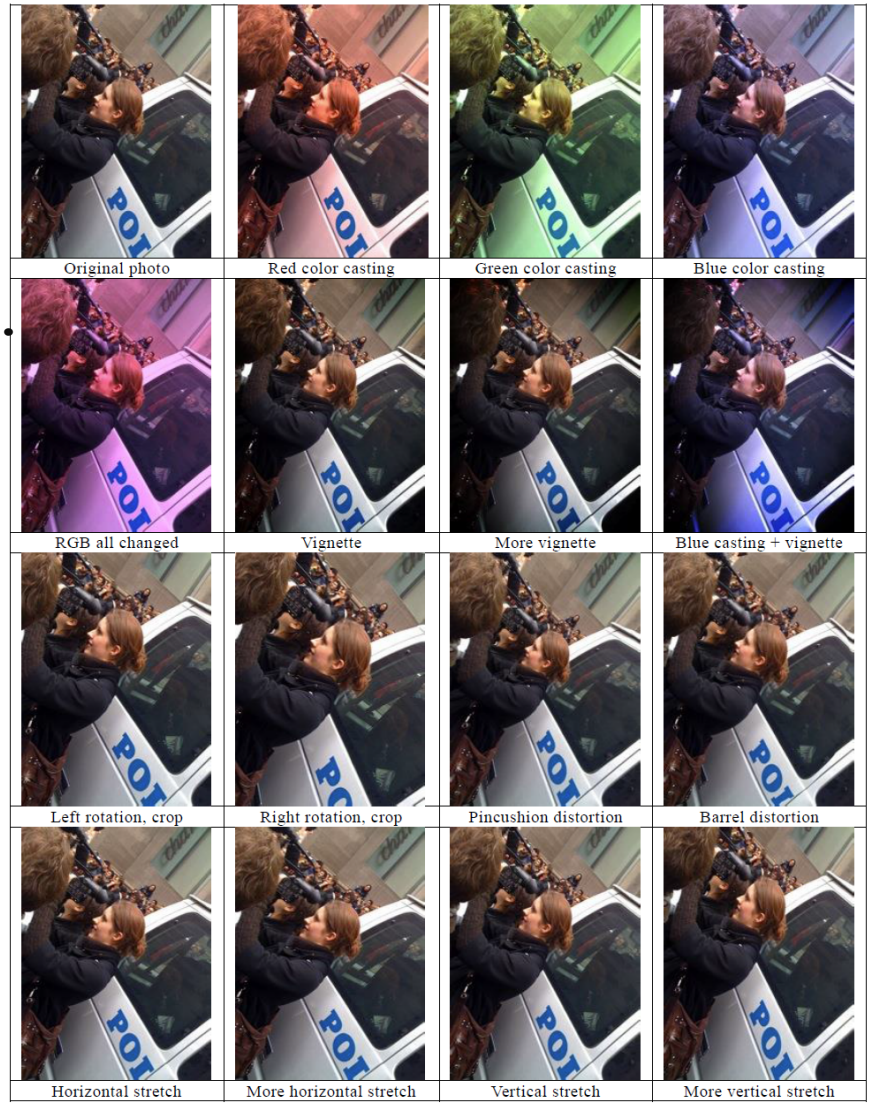


Table 1: The number of possible changes for different augmentation ways.

Augmentation	The number of possible changes
Color casting	68920
Vignetting	1960
Lens distortion	260
Rotation	20
Flipping	2
Cropping	82944(crop size is 224x224, input image size is 512x512)

R. Wu+, arxiv:1501.02876.

An easy way to get free samples

Data Augmentation.

- Images processed in many ways.
- **Virtually infinite samples generated.**

Drove us to use powerful computers!!

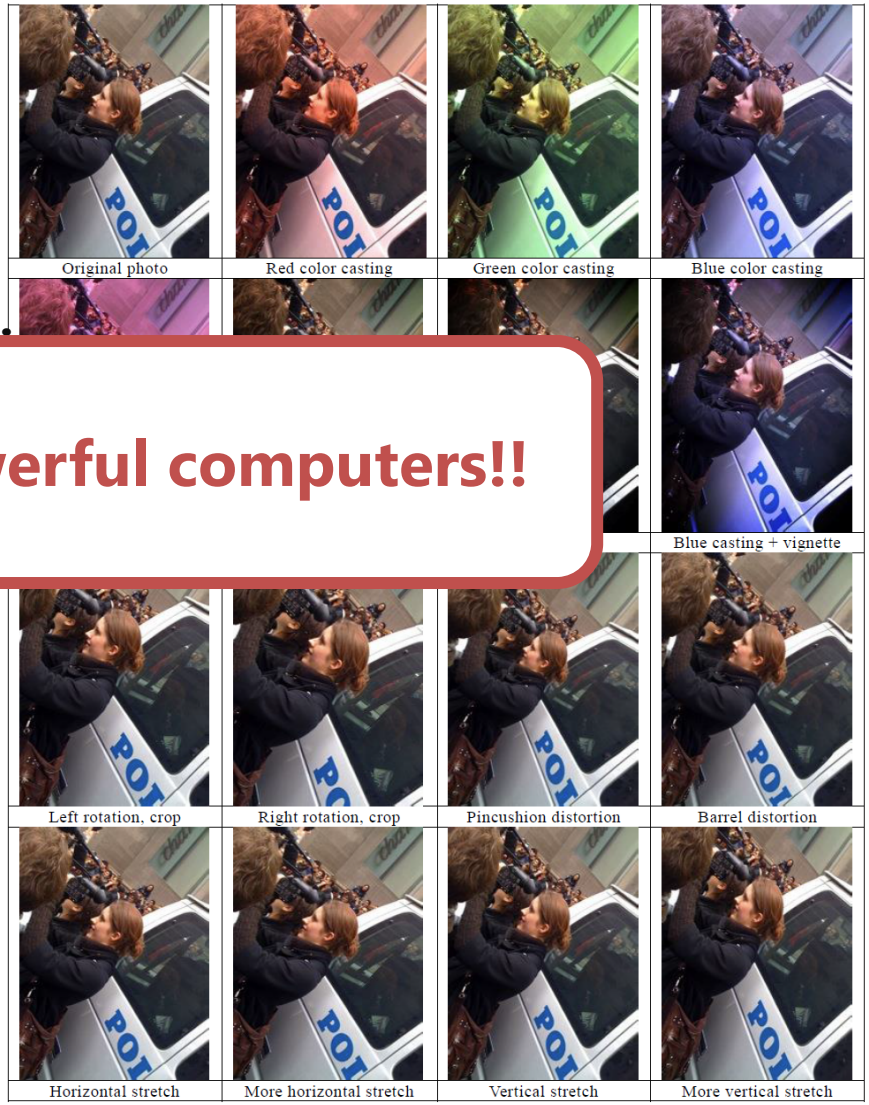


Table 1: The number of possible changes for different augmentation ways.

Augmentation	The number of possible changes
Color casting	68920
Vignetting	1960
Lens distortion	260
Rotation	20
Flipping	2
Cropping	82944(crop size is 224x224, input image size is 512x512)

R. Wu+, arxiv:1501.02876.

Outline

- Introduction
- **Learning algorithm**
- Performance modeling

Our algorithm, a brief overview

SPRINT, SuPeRcomputation In Neural Training

	Methods	
Who communicate	Parameter Server	Collective Comm. (MPI-AllReduce)
When to communicate	Synchronous	Asynchronous
What to communicate	Data-parallel	Model-parallel

Y. Oyama, A. Nomura, I. Sato, H. Nishimura, Y. Tamatsu, S. Matsuoka, BigData2016.

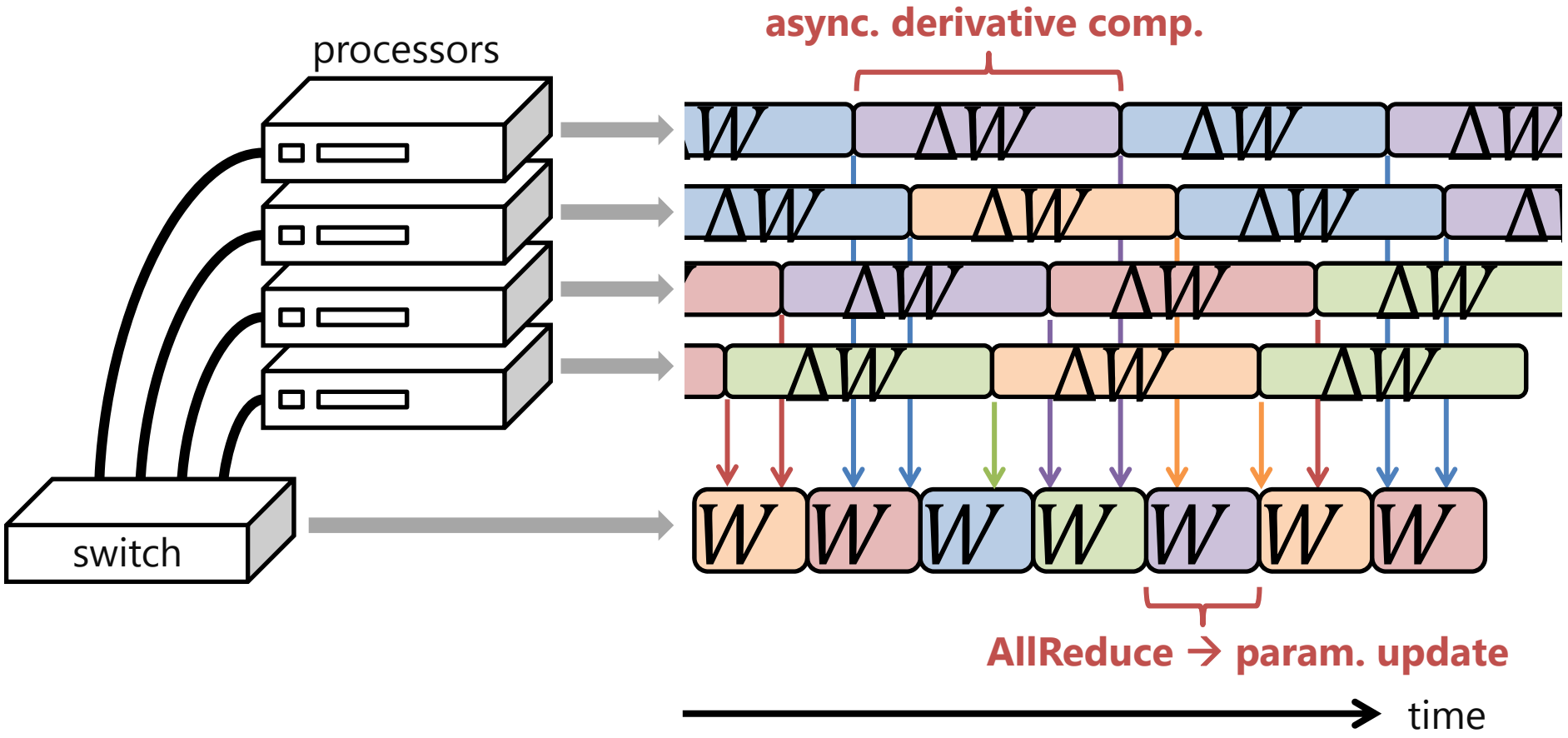
Our algorithm, a brief overview

SPRINT, SuPeRcomputation In Neural Training

	Methods	
Who communicate	Parameter Server	**Collective Comm.** **(MPI-AllReduce)**
		➡ **Exploit T2's all-to-all comm. capability.**
When to communicate	Synchronous	**Asynchronous**
		➡ **Gain fast update frequency.**
What to communicate	**Data-parallel**	Model-parallel
		➡ **Final usage: embedding system**

Y. Oyama, A. Nomura, I. Sato, H. Nishimura, Y. Tamatsu, S. Matsuoka, BigData2016.

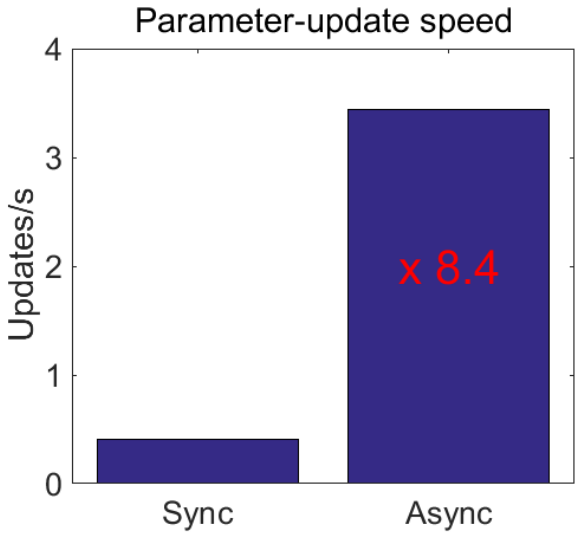
Asynchronous & AllReduce approach



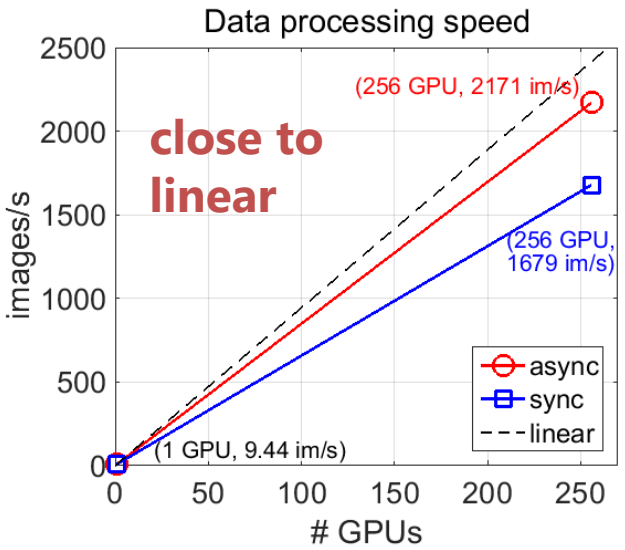
Asynchronous strategy

Pros.

- Parameter-update speed



- Data-processing speed



Ex) a fully-convolutional model

Cons.

- **Non-steepest descent** due to the **staled** parameters

Async. update rule:

$$W^{t+1} = W^t - \eta \sum_n \left. \frac{\partial J(x_n; W)}{\partial W} \right|_{W^{t-s}}$$

Derivatives computed by old param.

TSUBAME Ground Challenge 2015

Resources 420 TSUBAME2.5-nodes (1260 Tesla K20Xs), 7 days

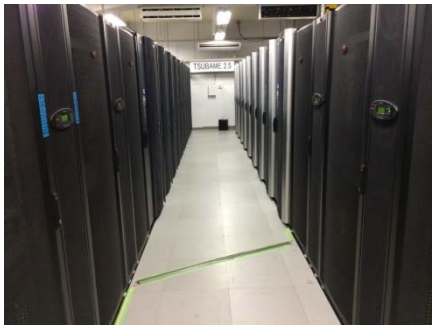
Task To train multiple deep CNN models
to participate the ImageNet challenge

Some facts

Data-parallel scale: 96 GPUs at largest

Performance: ~1 TFOPS/GPU in cost derivative comp.

Resource usage: 1146 GPUs used simultaneously at peak



TSUBAME 2.5

Quantitative evaluation

Not record breaking, but not bad…

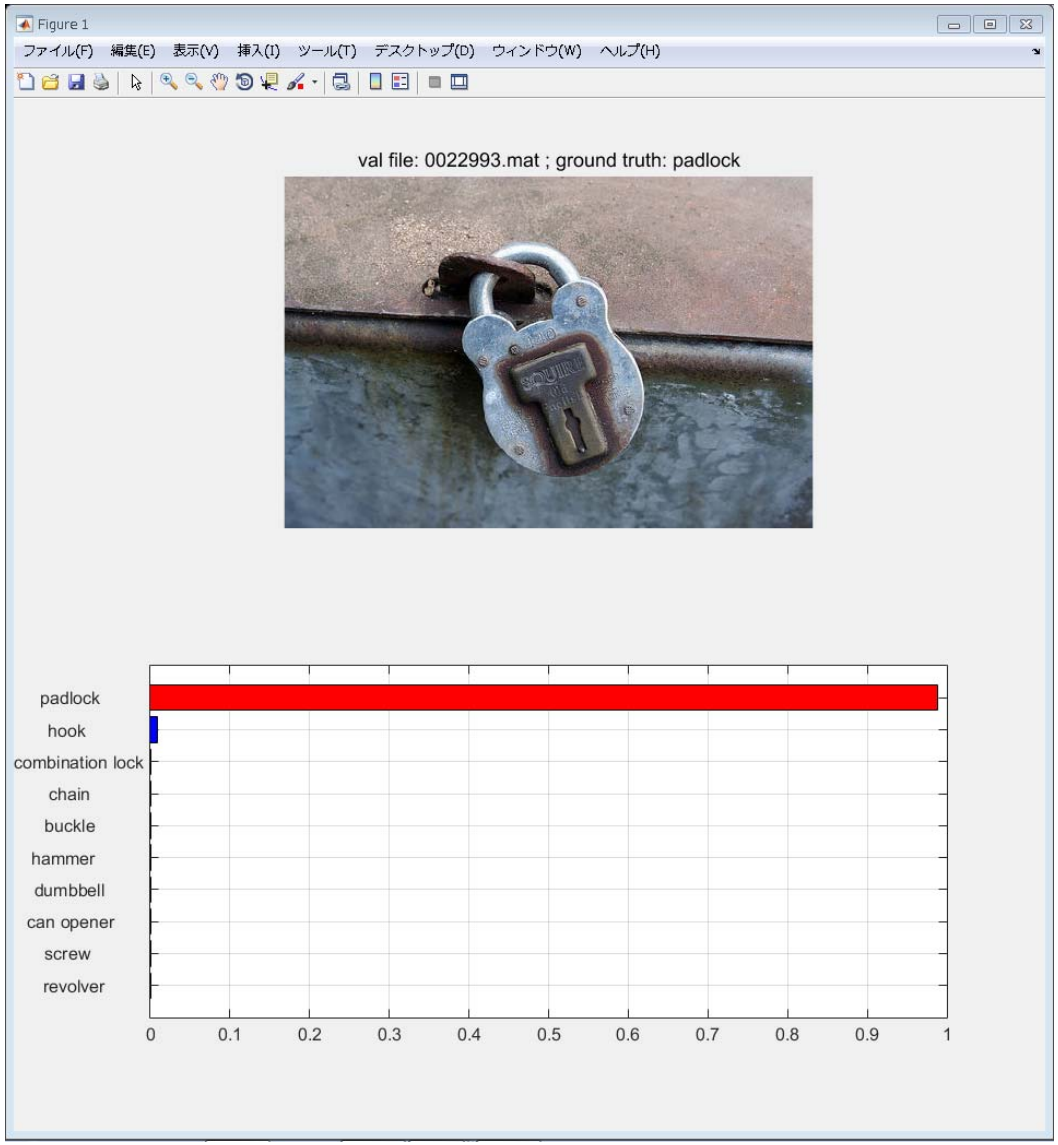
Table 4. Evaluation of our ensemble model with representative ones.

Model	#constituent models	#layers with product	#parameters	Top-5 error rate (%) on test set
AlexNet [1]	5	8	60M	16.64
ours	6	16	39M	13.89
OxfordNet [2]	7	19	144M	7.33
GoogLeNet [3]	7	22	5M	6.66
MSRA [4]	6	152	57M*	3.57

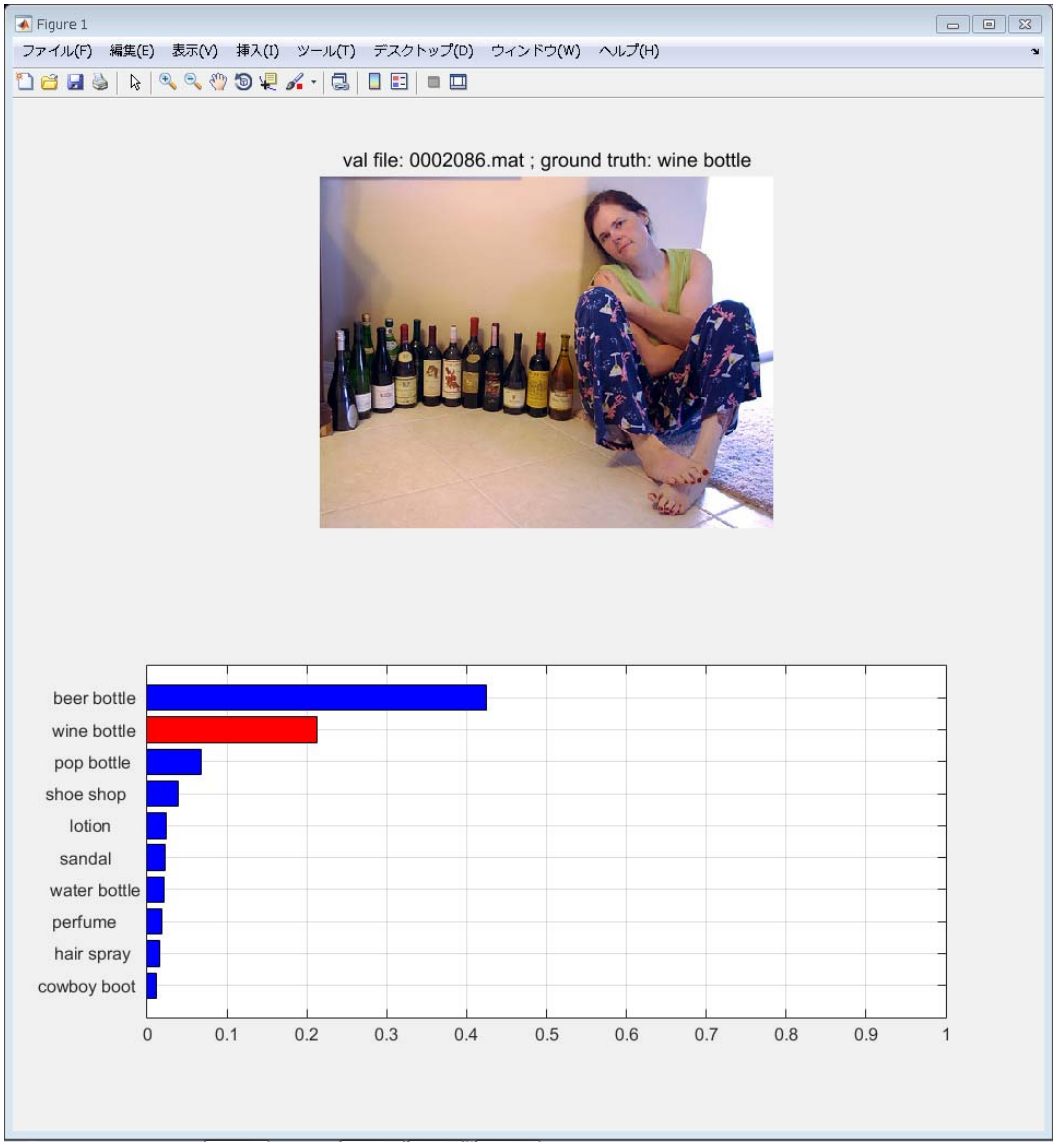
I. Sato, R. Watanabe, H. Nishimura, A. Nomura, S. Matsuoka, TSUBAME ESJ Vol.14, 2016.

[1] A. Krizhevsky+, NIPS 2012.
[2] K. Simonyan and A. Zisserman, ICLR 2015.
[3] C. Szegedy+, CVPR 2015.
[4] K. He+, CVPR 2016.

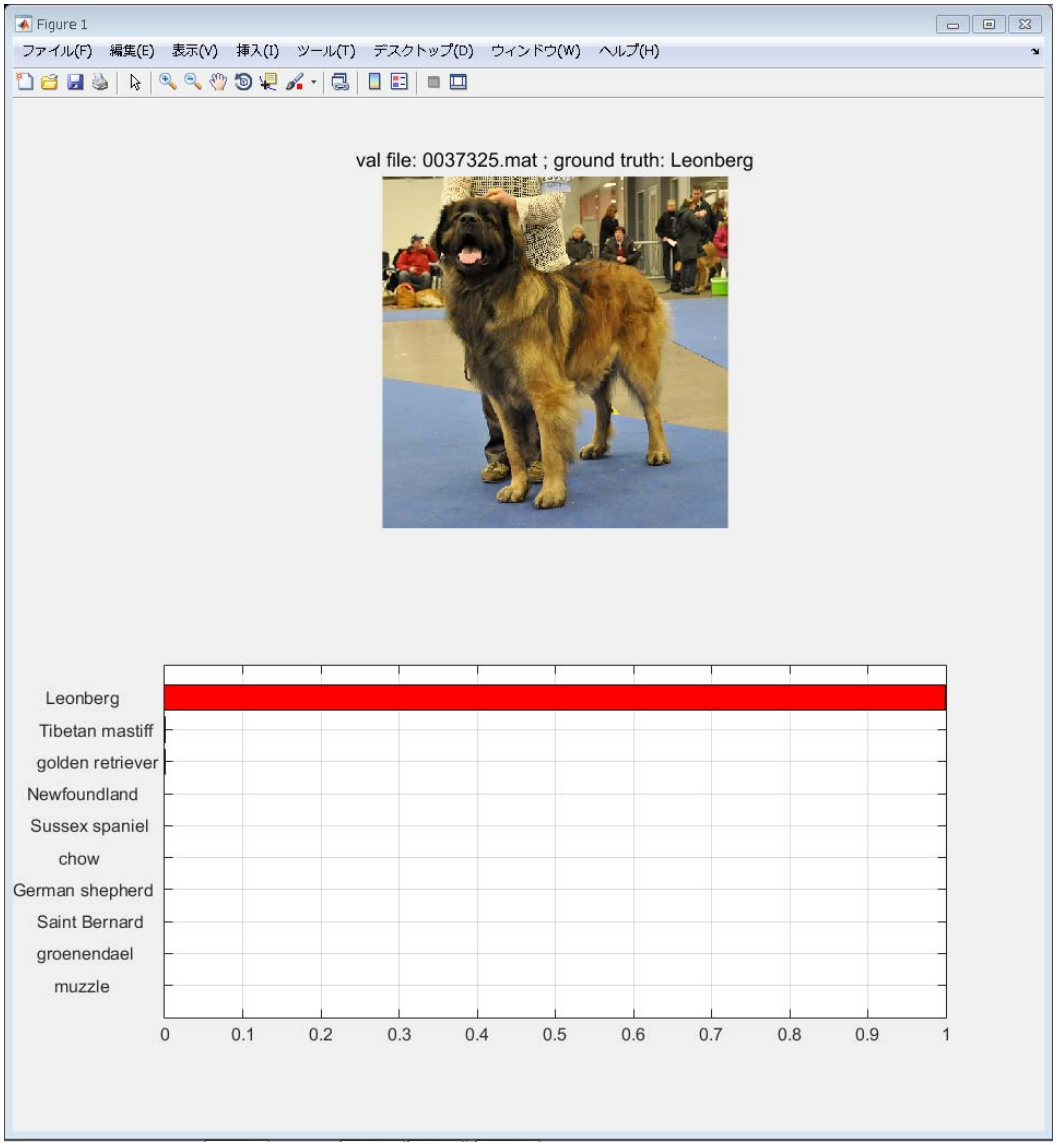
Qualitative evaluation



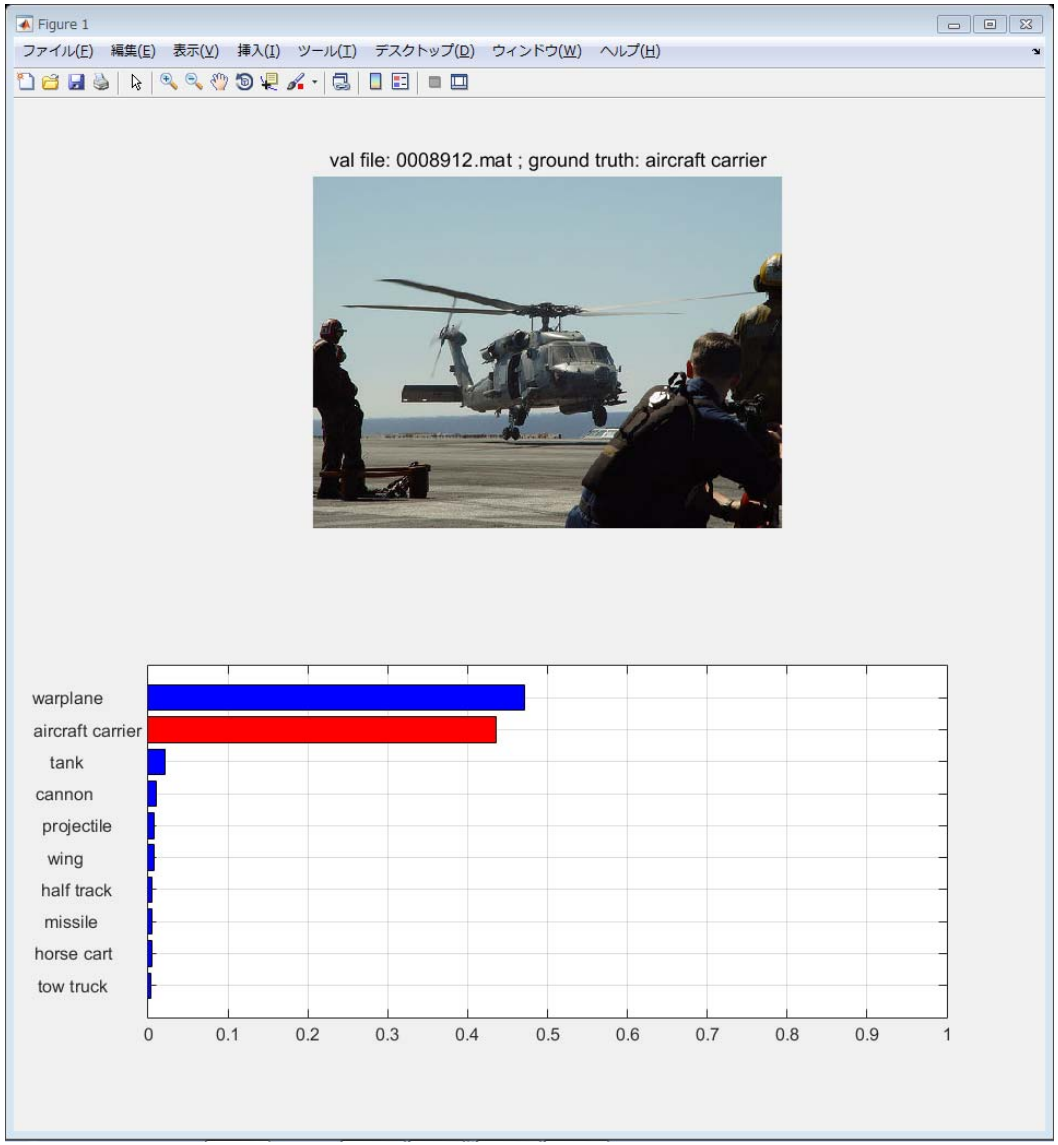
Qualitative evaluation



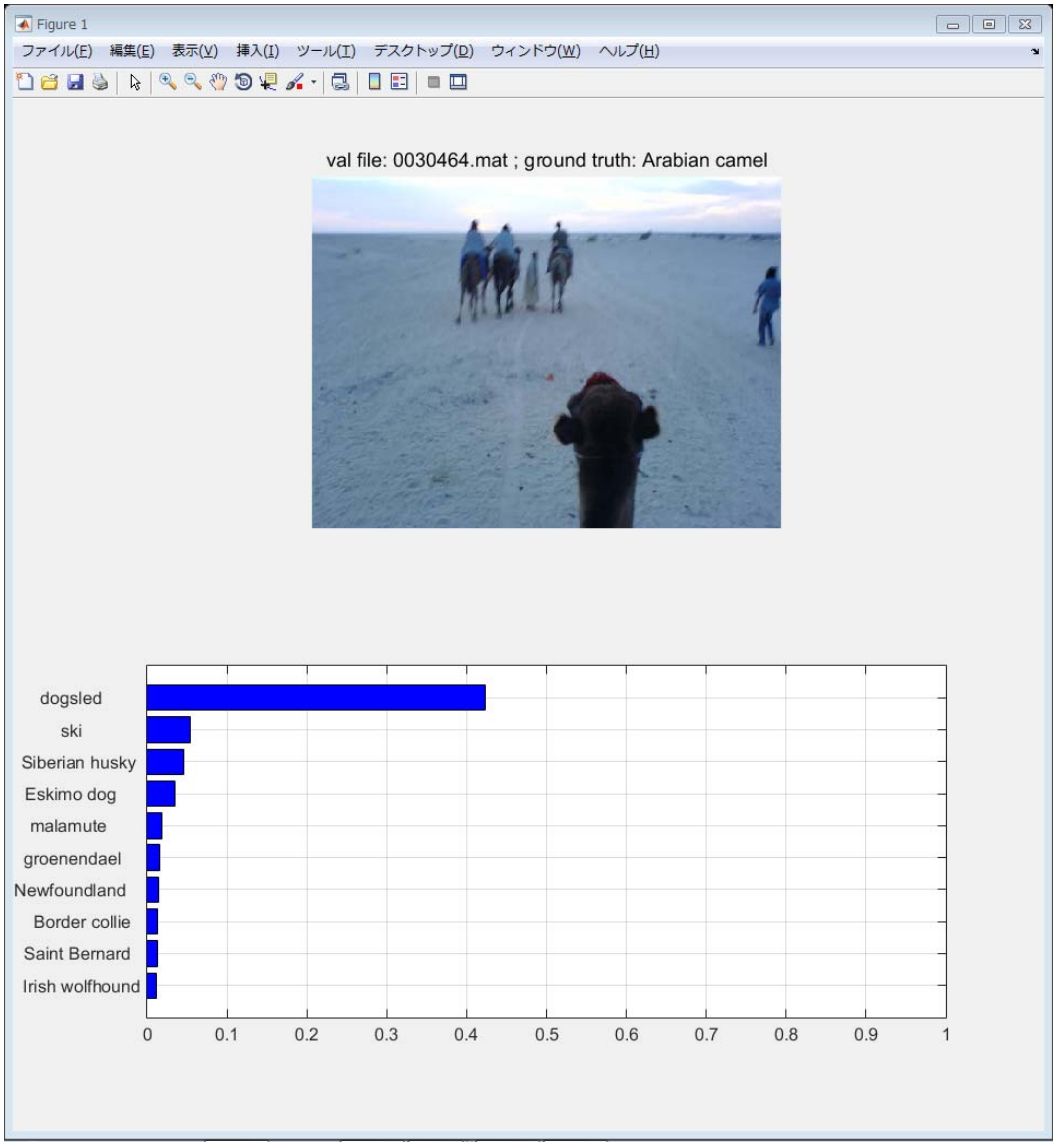
Qualitative evaluation



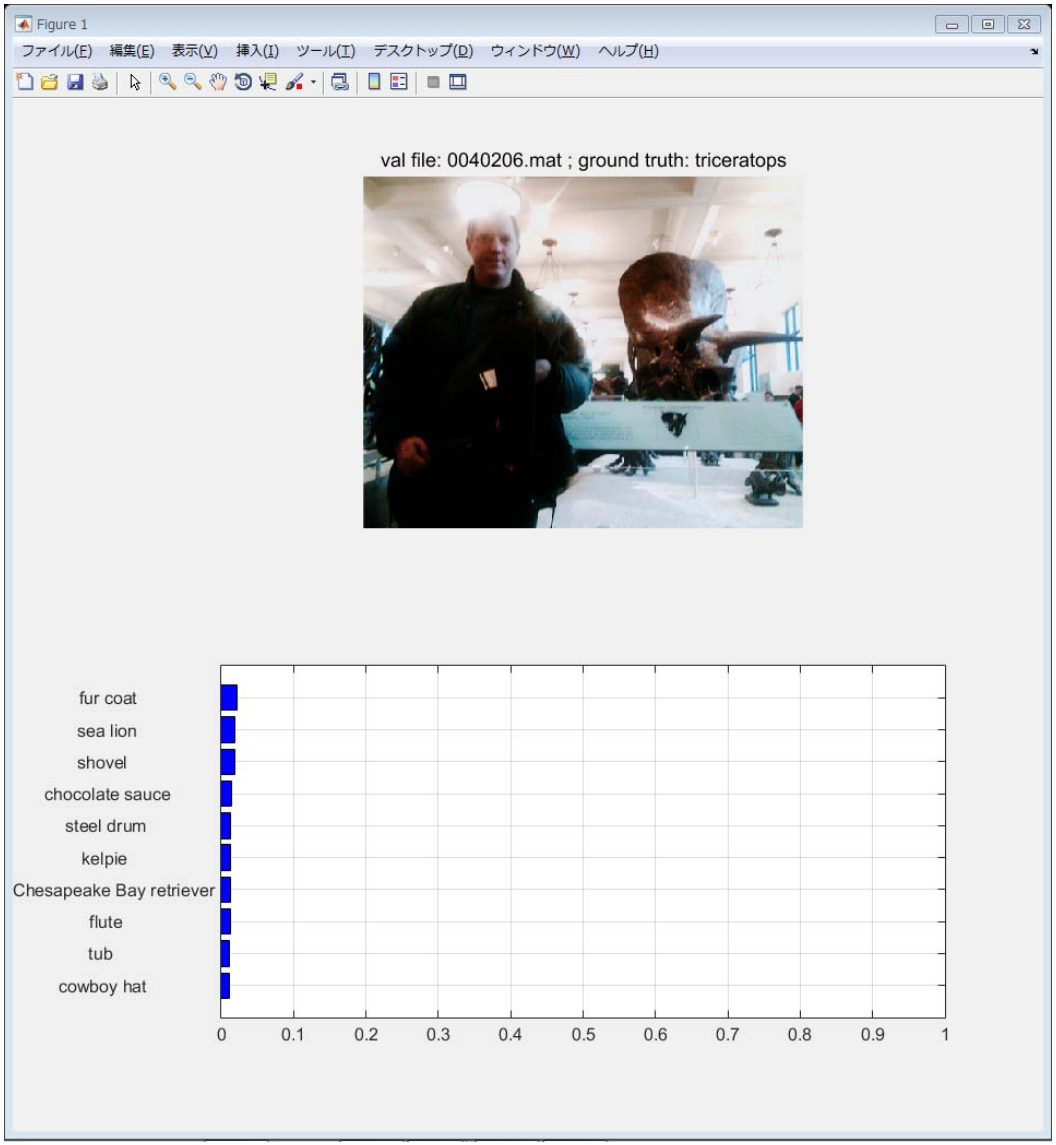
Qualitative evaluation



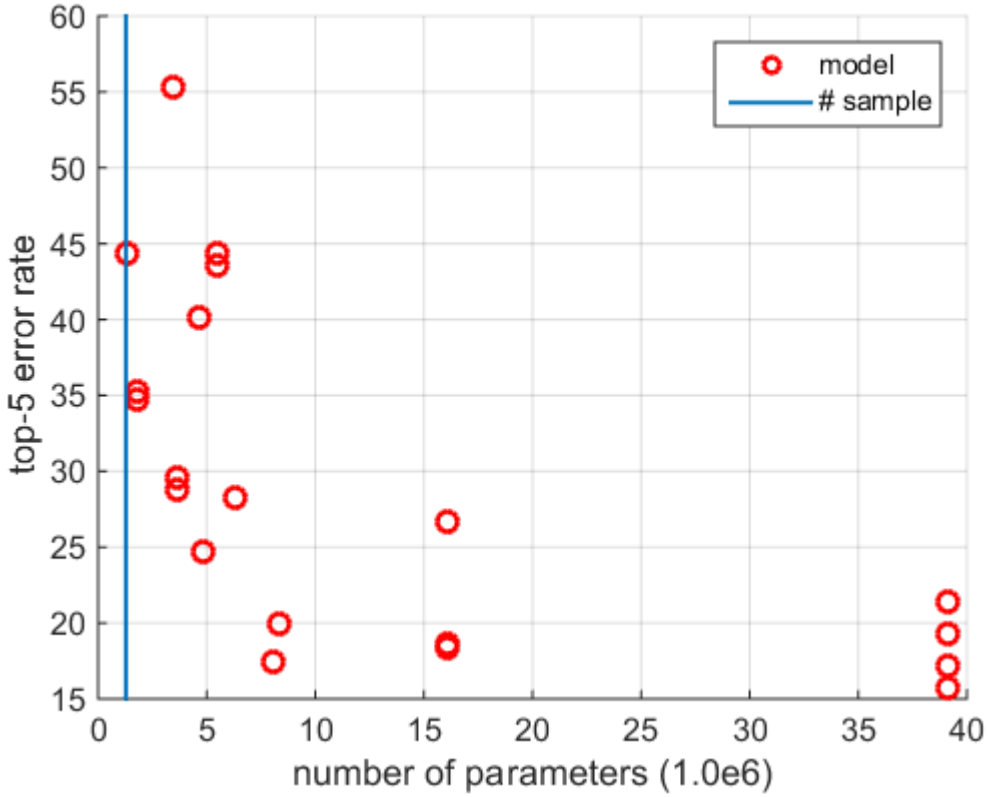
Qualitative evaluation



Qualitative evaluation



An interesting observation



An interesting observation

Outline

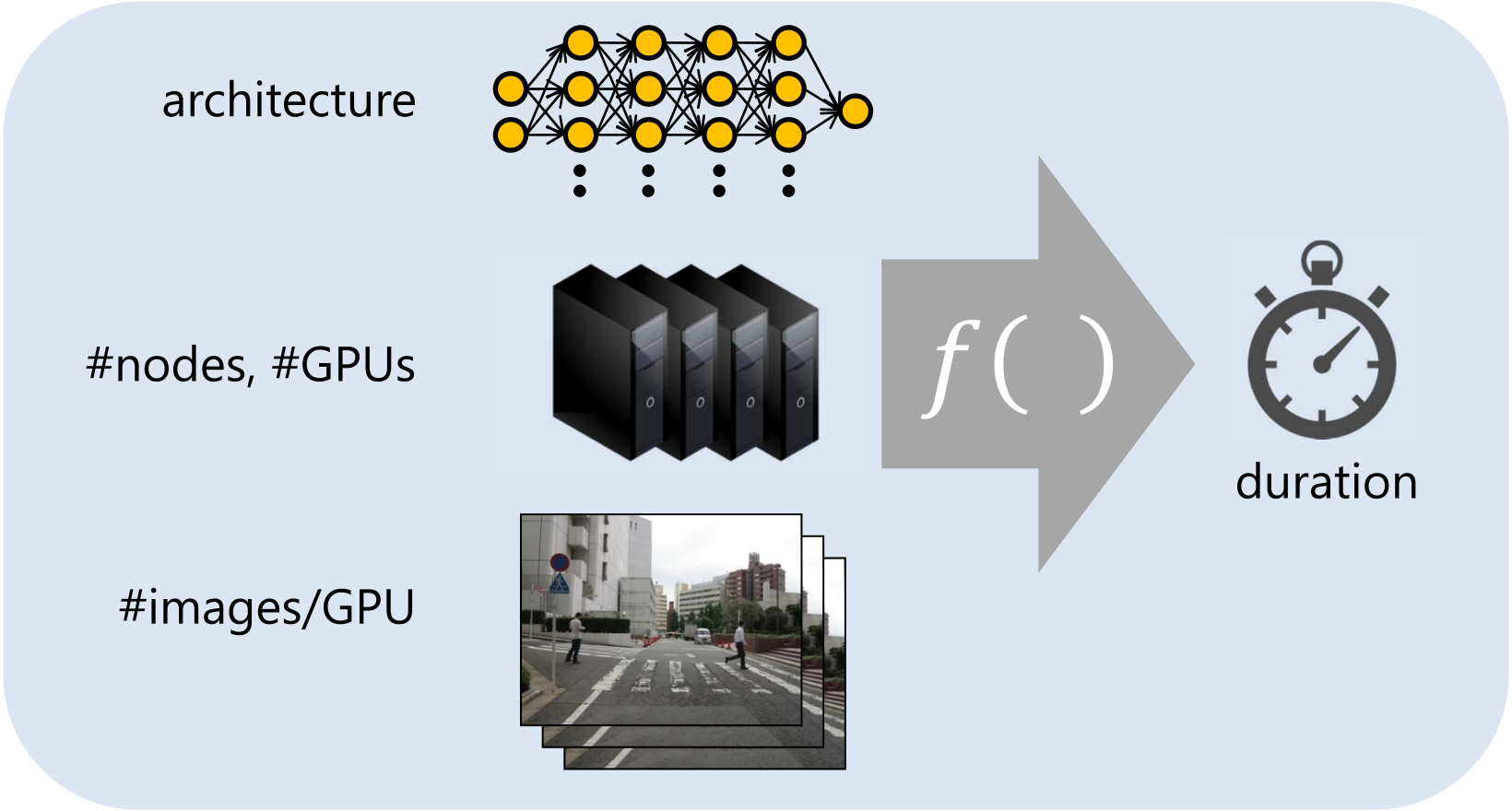
- Introduction
- Learning algorithm
- **Performance modeling**

Credit: Yosuke Oyama, Akihiro Nomura, Satoshi Matsuoka

(Tokyo Institute of Technology)

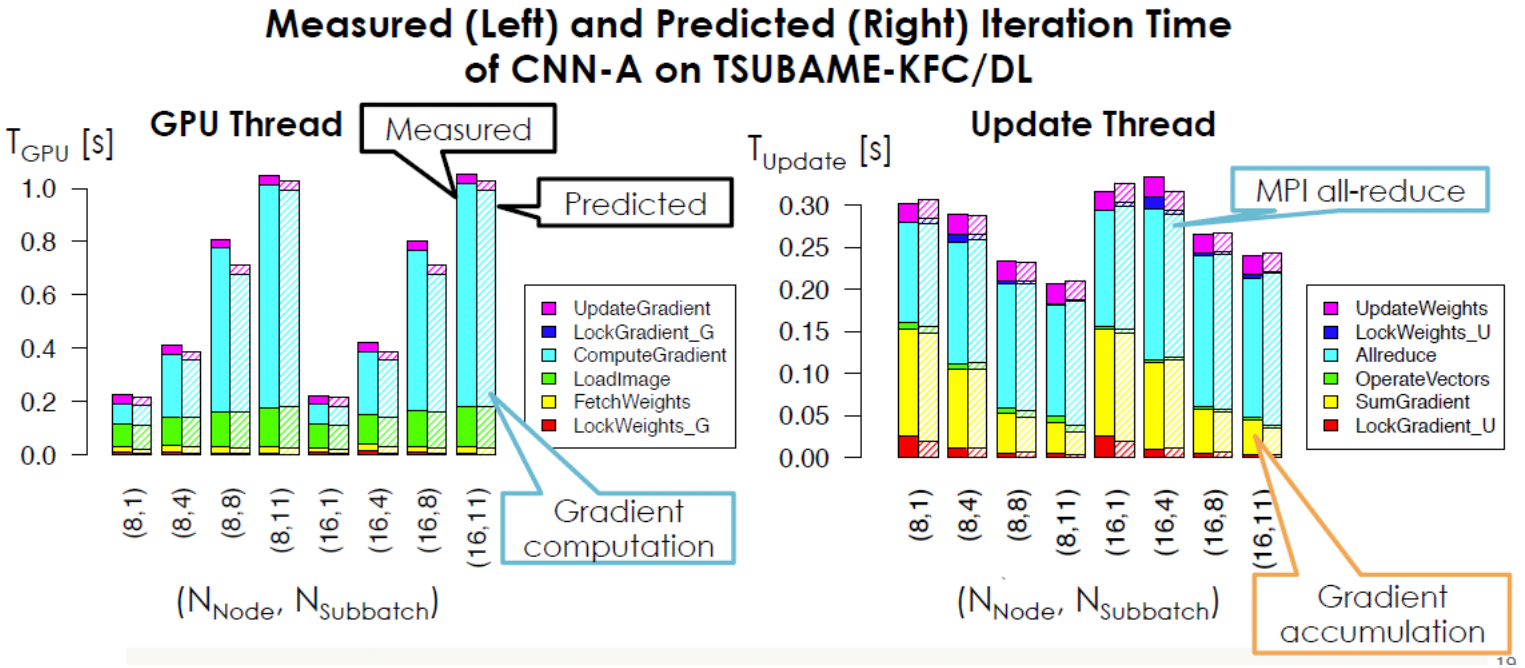
Concept

Design a simple mathematical model that predicts training duration, given…



Y. Oyama, A. Nomura, I. Sato, H. Nishimura, Y. Tamatsu, S. Matsuoka, BigData2016.

Effects



Epoch duration can be accurately predicted.

Tells the best resource usage.

e.g., #nodes, #images/GPU

Summary

- Need to learn a large-scale image dataset for automotive applications.
- Investigate GPU-asynchronous & AllReduce distributed deep learning.
- Propose a training duration prediction model.