Networks with Sub-Networks

Kyushu Institute of Technology
Graduate School of Life Science and Systems Engineering

ICAROB presentation, Jan 14, 2020

Ninnart Fuengfusin, Hakaru Tamukoh

Kyutech
Kyushu Institute of Technology

Tamukoh Lab.
A Brain-Like Computer System Lab.

Content

Background

- **Mobile devices have varied specifications:**
  - ✓ From high to low-end devices.

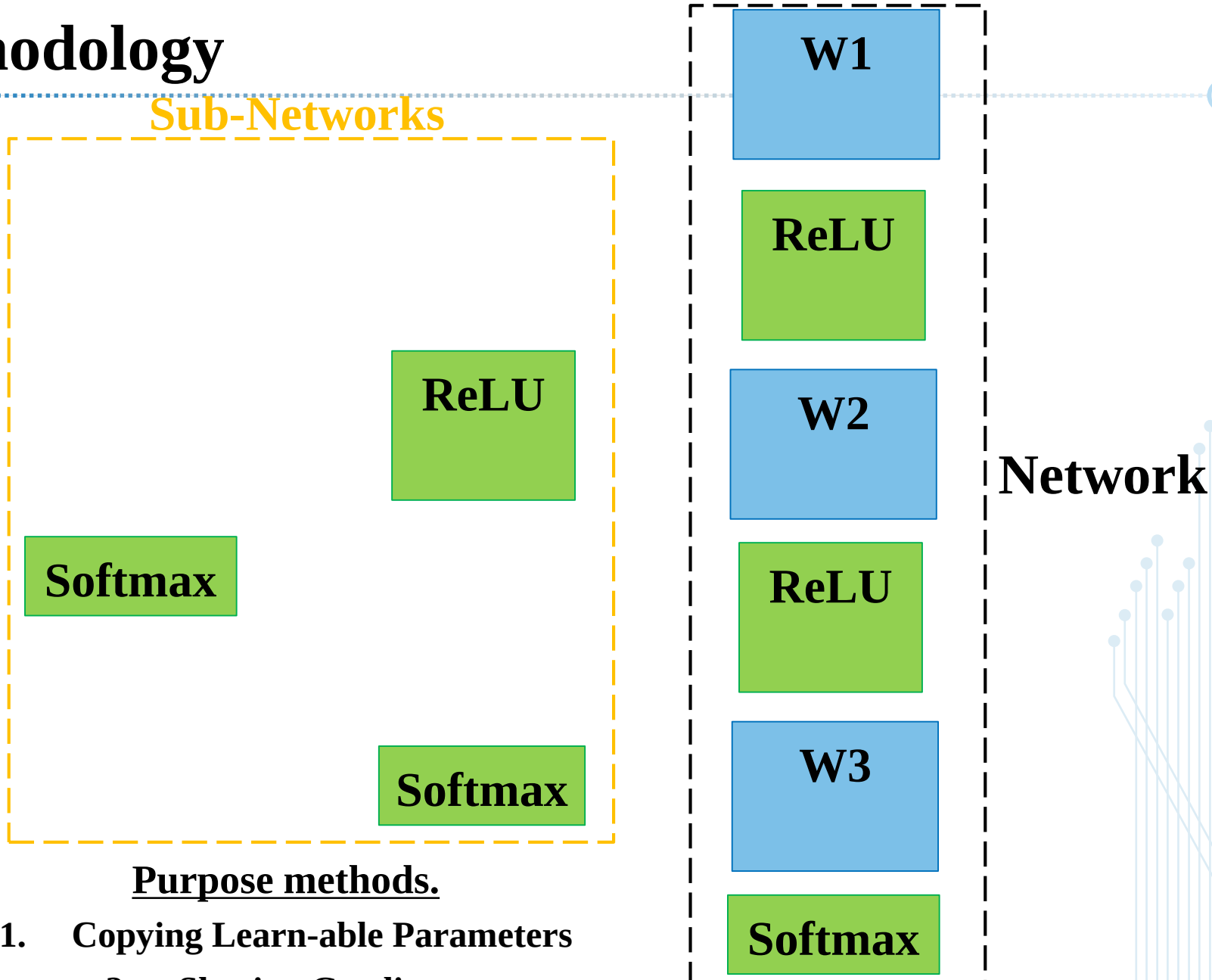
- **To put a neural network (NN) into mobile devices:**
  - ✓ Balance trade-off between latency and performance of a NN given the specification of devices. (Middle point)
- **A NN that is suitable to both of devices.**

Objective

- **To explore NN model that can:**
  - ✓ Adjust the performance and latency on demand.
    - ➢ More depth, better performance, worse the latency.
  - ✓ Without consuming the high amount memory footprint.

- **We propose Network with Sub-Networks (NSN).**
  - ✓ A NN that can rearrange its weights into the smaller NN.

Methodology



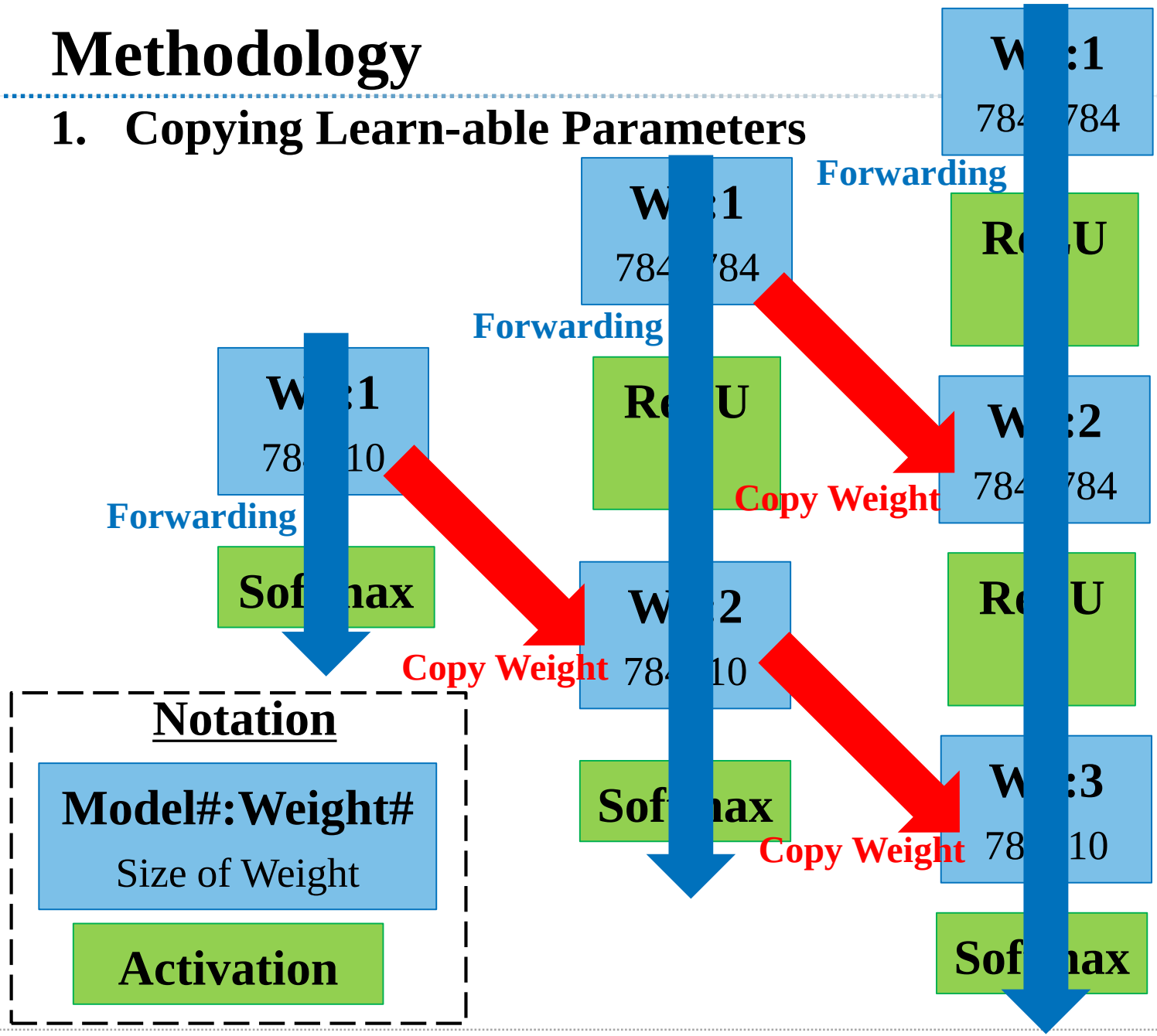
Purpose methods.

1. Copying Learn-able Parameters
2. Sharing Gradient.

Low Latency Model ⟷ High Latency Model

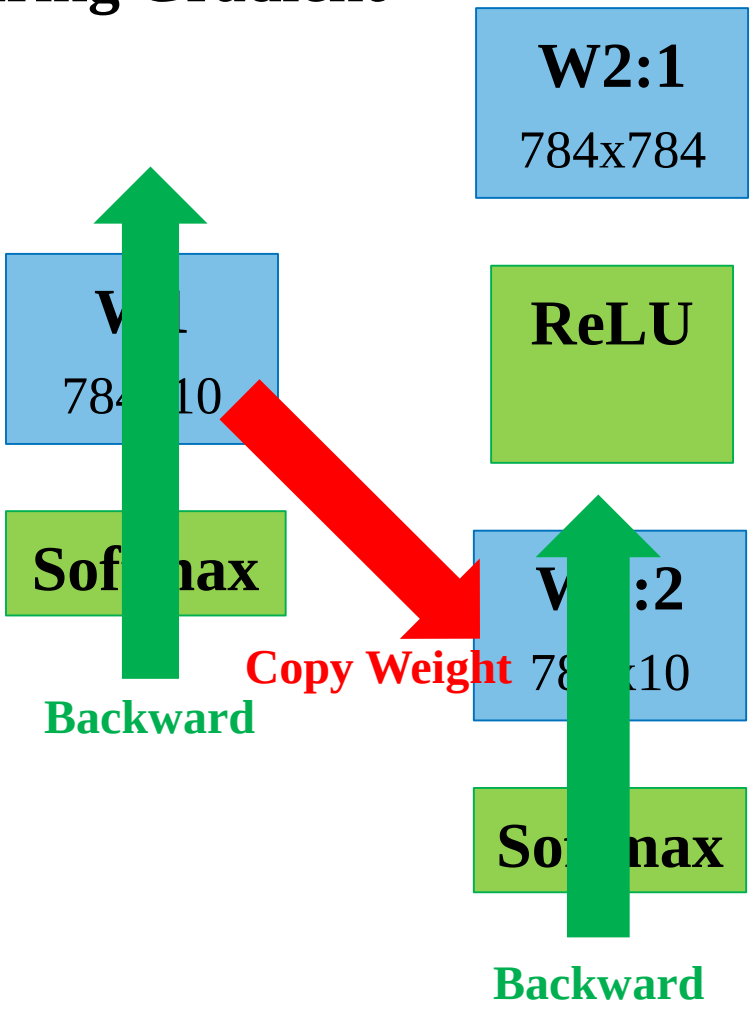
Methodology

1. Copying Learn-able Parameters



Methodology

2. Sharing Gradient



General Updating:

$$w_{1:1} = w_{1:1} - lr\frac{\partial L}{\partial w_{1:1}}$$

$$w_{2:2} = w_{2:2} - lr\frac{\partial L}{\partial w_{2:2}}$$

Since, $w_{2:2}$ is a copy of $w_{1:1}$.

$$w_{2:2} = w_{2:2} - \frac{lr}{2}\left(\frac{\partial L}{\partial w_{1:1}} + \frac{\partial L}{\partial w_{2:2}}\right)$$

$$w_{1:1} = w_{2:2}$$

Notation

Model#:Weight#
Size of Weight

Activation

Experimental Results and Discussion

Experimental Setting:

- **Two settings:** dense NN with a hidden layer and **two hidden layers as Network.**
- Dataset: MNIST[1]
- Training Epoch: 600
- Training batch: 128
- Learning rate (lr): 0.3, step down every 200 epoch.
- Optimization: Gradient Descent with Momentum
- Regularization: Dropout [2] and L2 regularization.

Experimental Results and Discussion

Base-Line Neural Network:

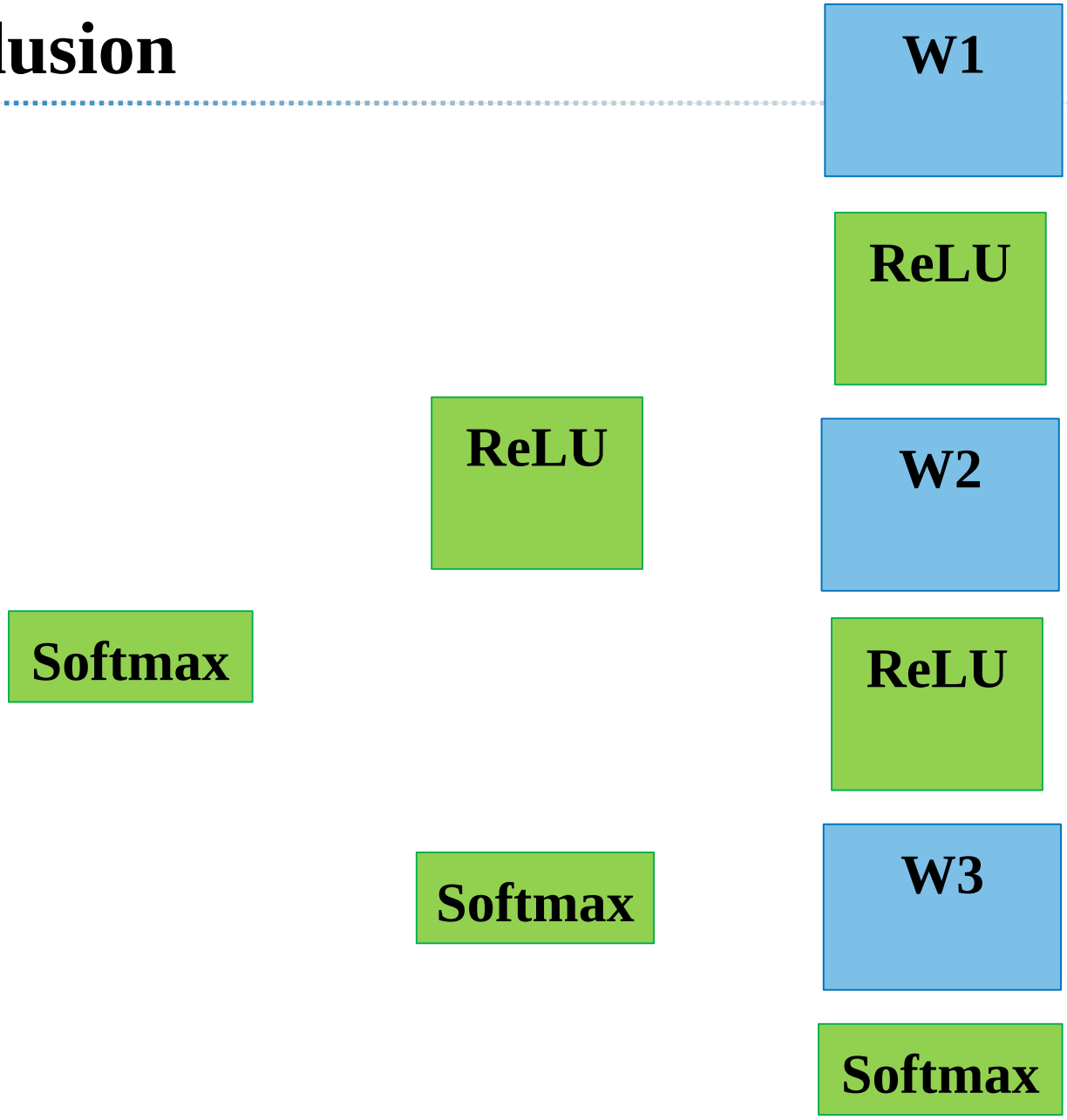
	Test Accuracy	Number Parameters
NN with 2 hidden layers	**0.9886**	1.24M
NN with a hidden layers	0.9882	0.62M
Softmax-Regression	0.9241	7.85k

Proposed Method NSN:

	Test Accuracy	Number Parameters
Network (NN with 2 hidden layers)	**0.989**	1.24M
Sub-network#1 (NN with a hidden layers)	0.9843	0.62M
Sub-network#2 (Softmax-Regression)	0.926	7.85k

- Bias toward **Network**. Sub-Networks have less test accuracy than base-line.
- Better result from regularization effect?

Conclusion



The code will be published on: *https://github.com/ninnart-fuengfusin/nsn*

Reference

- [1] Y. LeCun, C. Cortes, and C. Burges, “Mnist hand-written digit database,” *AT&T Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, vol. 2, p. 18, 2010.

- [2] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, “Dropout: a simple way to prevent neural networks from overfitting”, *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

Thank you!

ご清聴ありがとうございました。

http://www.brain.kyutech.ac.jp/~tamukoh/