A Lightweight On-device CNN Finetuning using Skip2-LoRA and Quantized Cache

Hiroki Matsutani, Keisuke Sugiura, Masaaki Kondo (Keio Univ),
Radu Marculescu (UT Austin)

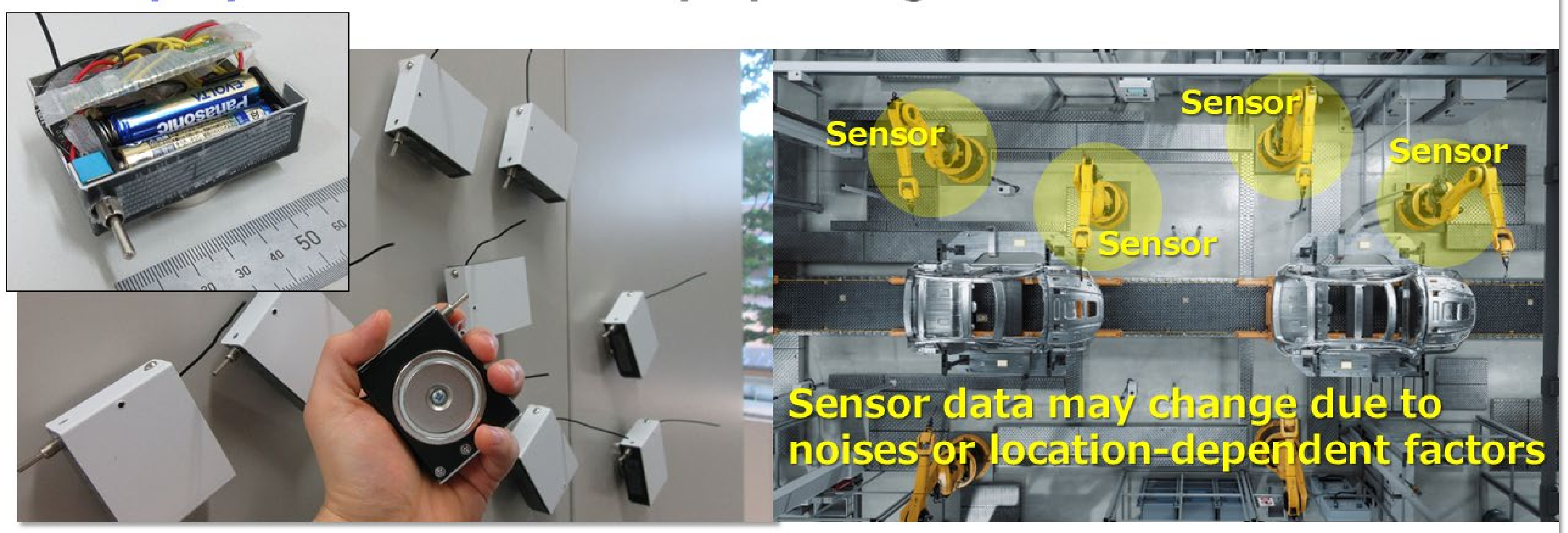
慶應義塾 Keio University

TEXAS The University of Texas at Austin

On-device finetuning for IoT devices

- Motivation for neural network training at edge side

Addressing the gap between pretrained model and deployed environment by updating the model on-device [1,2]

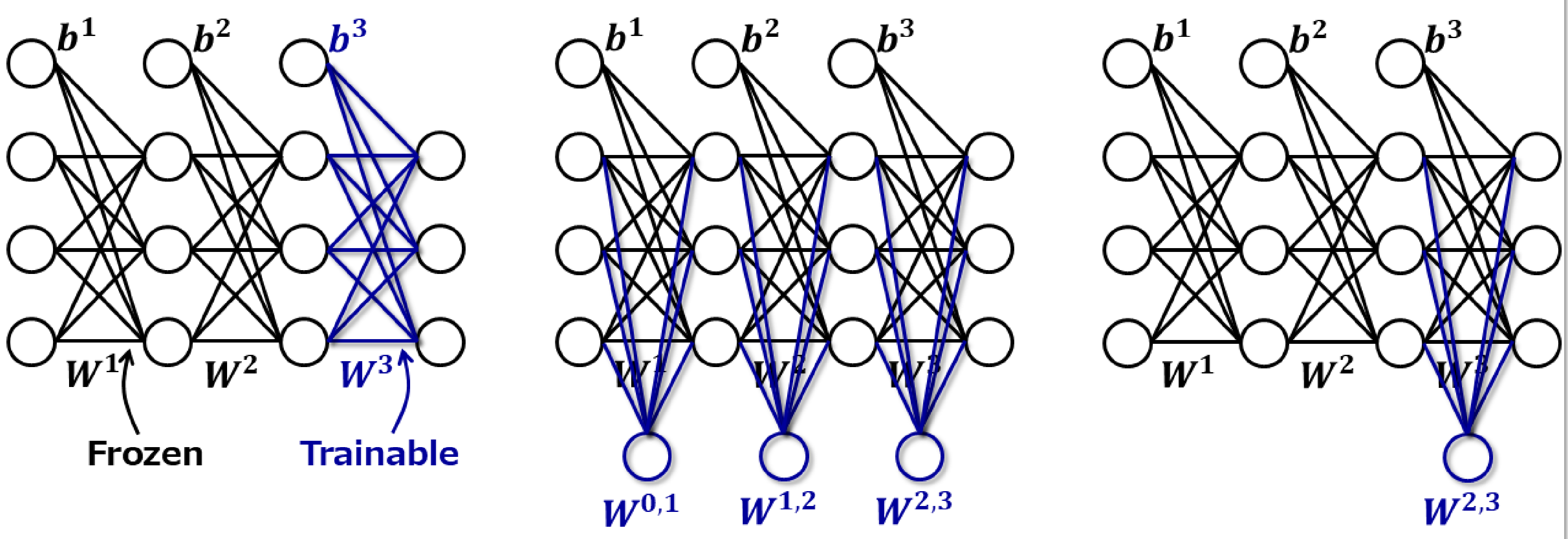


Sensor data may change due to noises or location-dependent factors

[1] Mineto Tsukada et al., "A Neural Network-Based On-device Learning Anomaly Detector for Edge Devices", IEEE Trans. on Computers (2020).
[2] Kazuki Sunaga et al., "Addressing Gap between Training Data and Deployed Environment by On-Device Learning", IEEE Micro (2023).

Baseline finetuning methods

FT-Last [1] | LoRA-All [2] | LoRA-Last



Weights of the last layer (W^3, b^3) are updated

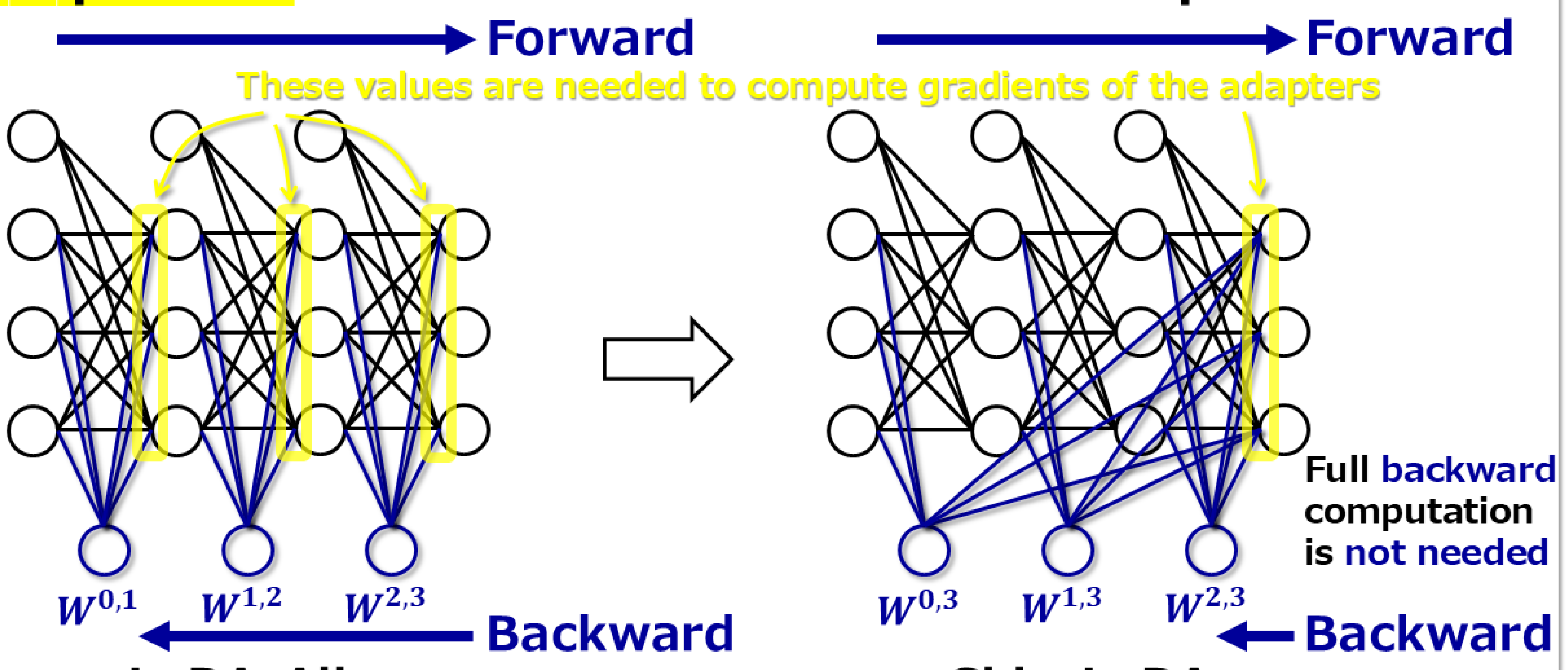
Trainable adapters are attached to all layers

Trainable adapter is attached to the last layer

[1] Haoyu Ren et al., "TinyOL: TinyML with Online-Learning onMicrocontrollers", IJCNN'21.
[2] Edward J. Hu et al., "LoRA: Low-Rank Adaptation of Large LanguageModels", arXiv:2106.09685 (2021).

Our proposed approach: Skip-LoRA

- Skip-LoRA can reduce the backward computation



Full backward computation is not needed

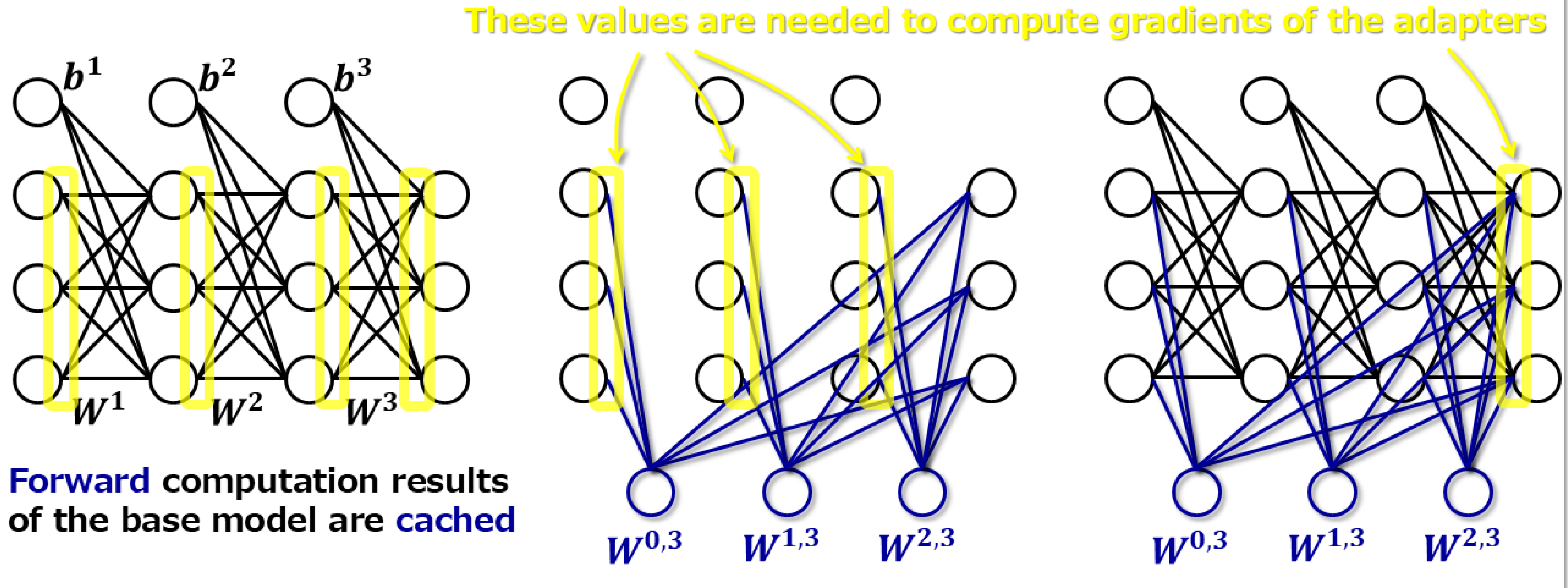
LoRA-All

Skip-LoRA [1]

[1] Hiroki Matsutani et al., "Skip2-LoRA: A Lightweight On-device DNN Fine-tuning Method for Low-cost Edge Devices", ASP-DAC'25.

Our proposed approach: Skip2-LoRA

- Skip2-LoRA can reuse forward computation results



Forward computation results of the base model are cached

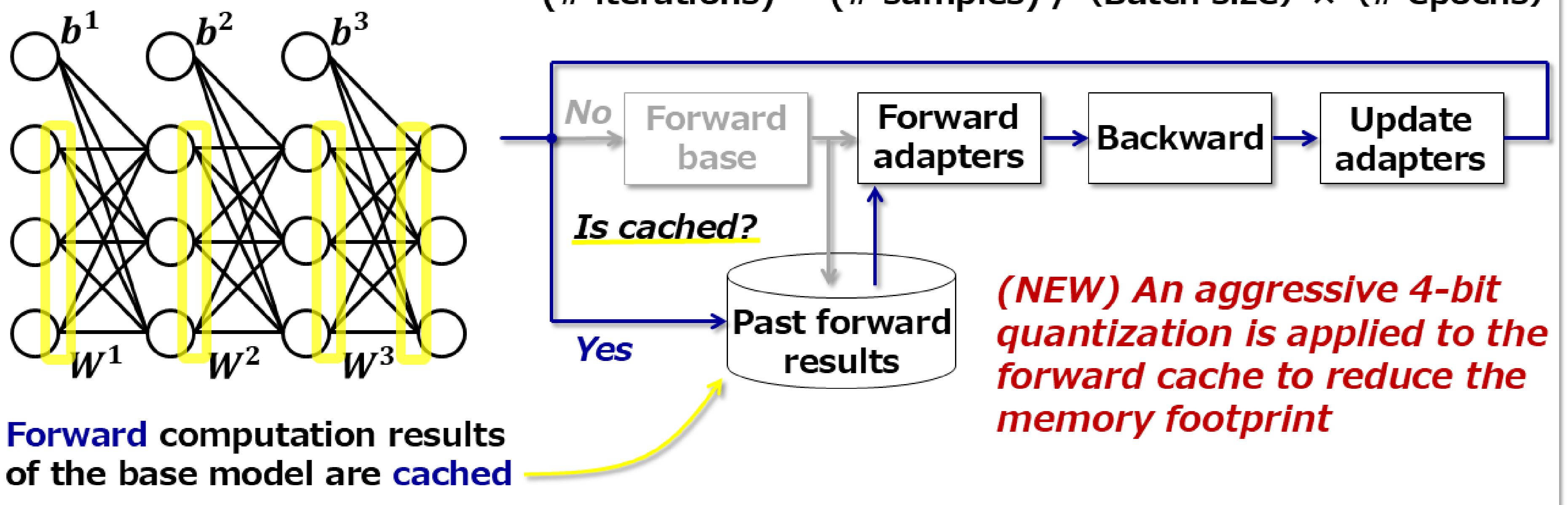
Base model + Adapters = Skip-LoRA [1]

[1] Hiroki Matsutani et al., "Skip2-LoRA: A Lightweight On-device DNN Fine-tuning Method for Low-cost Edge Devices", ASP-DAC'25.

Our proposed approach: Skip2-LoRA

- Skip2-LoRA can reuse forward computation results

(# iterations) = (# samples) / (Batch size) × (# epochs)



Forward computation results of the base model are cached

(NEW) An aggressive 4-bit quantization is applied to the forward cache to reduce the memory footprint

Base model

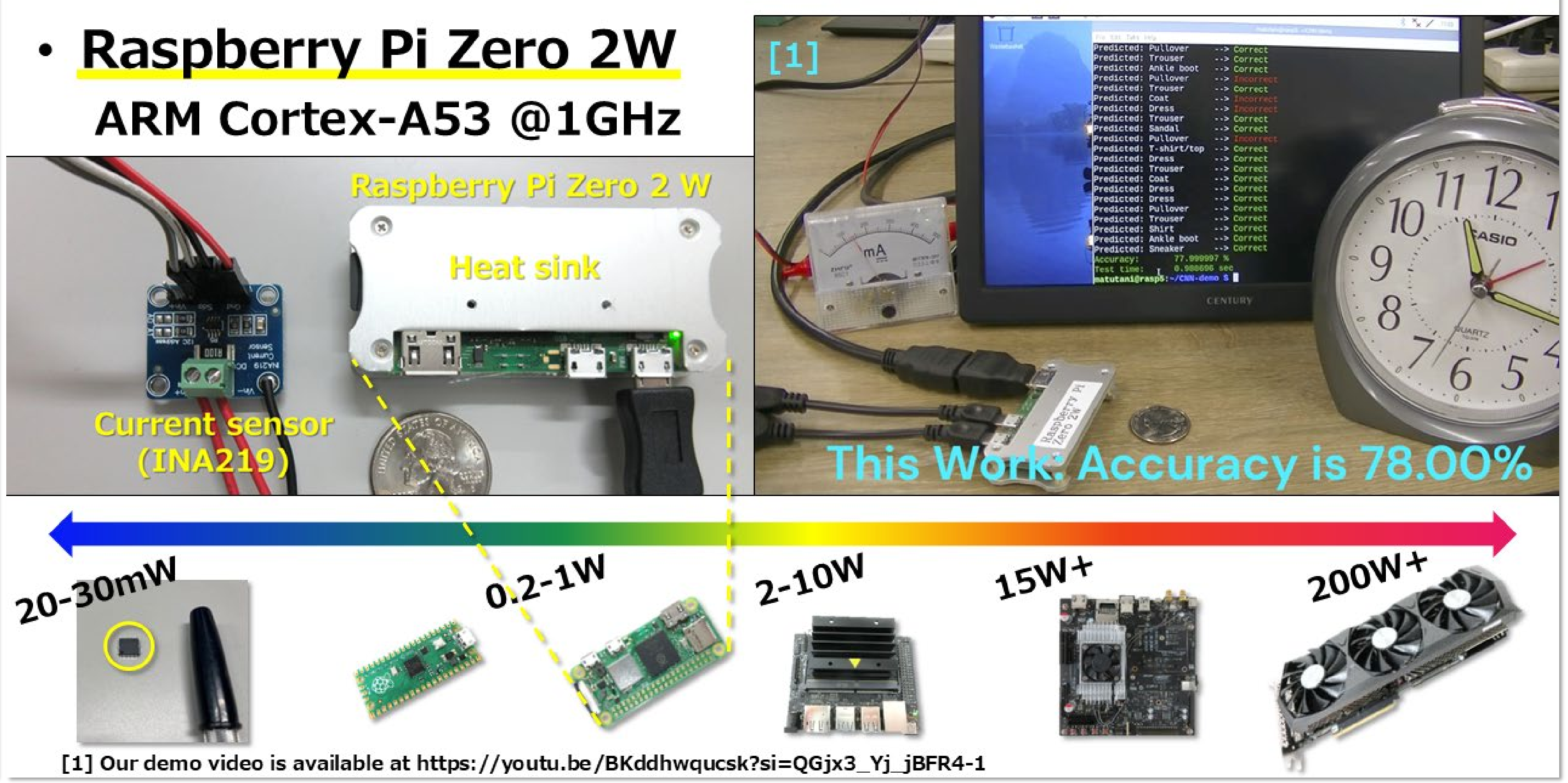
iterations is typically larger than 1;
so, we can skip most of the forward computation [1]

[1] Hiroki Matsutani et al., "Skip2-LoRA: A Lightweight On-device DNN Fine-tuning Method for Low-cost Edge Devices", ASP-DAC'25.

Skip2-LoRA for CNNs: Platform

- Raspberry Pi Zero 2W

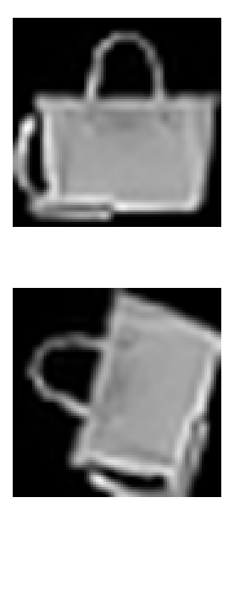
ARM Cortex-A53 @1GHz

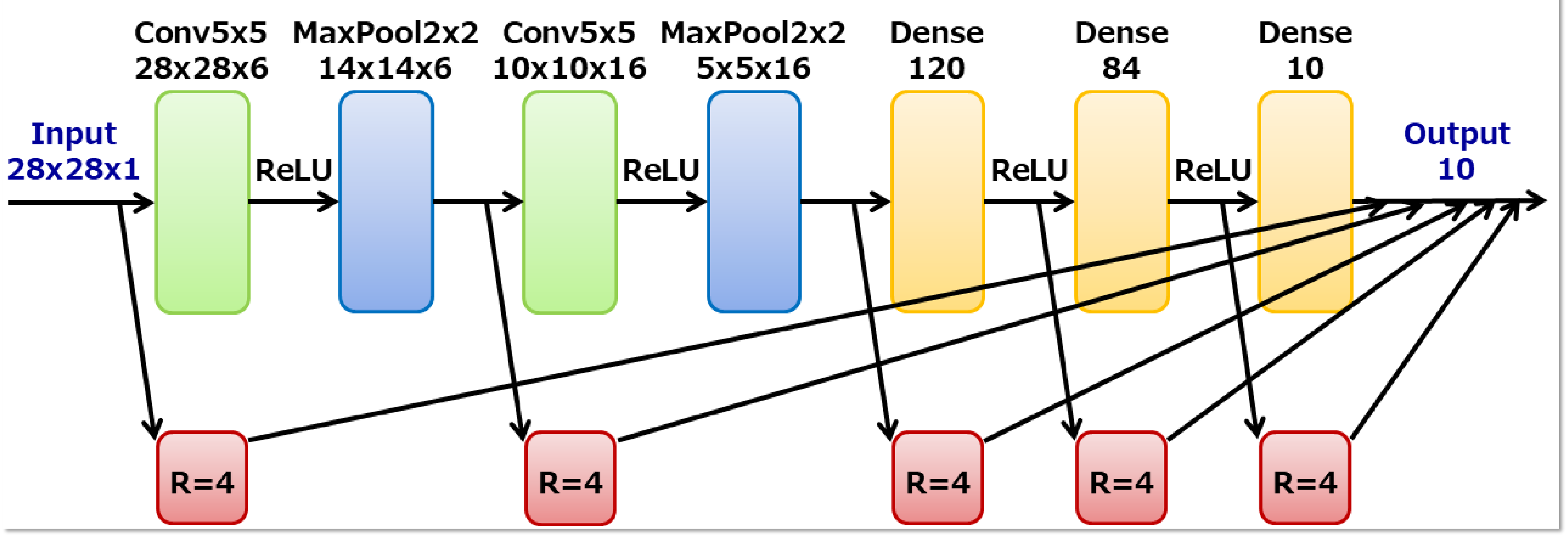


[1] Our demo video is available at https://youtu.be/BKddhwqucsk?si=QGjx3_Yj_jBFR4-1

Skip2-LoRA for CNNs: Model

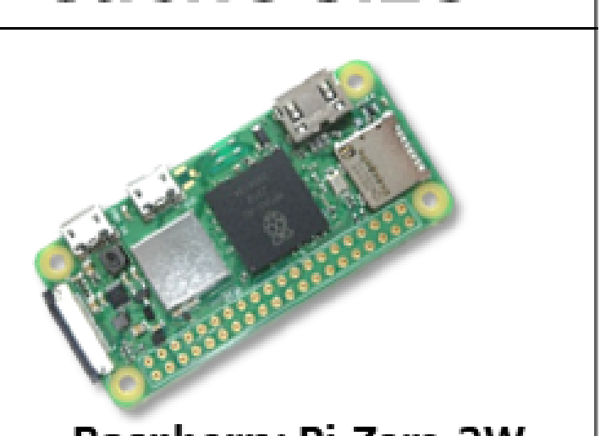
- Pretrained with fashion-MNIST ← Data drift
- Finetuned with Rotated fashion-MNIST (75 deg)
- Tested with Rotated fashion-MNIST (75 deg)





Skip2-LoRA for CNNs: Results

- In this work, Skip2-LoRA [1] is applied to CNNs
- An aggressive 4-bit quantization is applied to the forward cache to reduce the memory footprint

Model	Accuracy	FT time @RPZ2	Cache size
No Finetuning (FT)	9.18 %		 Raspberry Pi Zero 2W (aka $15 computer)
FT-Last	60.94 %	18.09 sec	
LoRA-Last	53.81 %	18.09 sec	
LoRA-All	75.59 %	114.15 sec	
Skip-LoRA	73.54 %	19.84 sec	
Skip2-LoRA	73.54 %	3.90 sec	7,336 kB
Quant Skip2-LoRA	74.02 %	4.27 sec	1,036 kB

*Number of FT samples: 1024, Number of epochs for FT: 10

[1] Hiroki Matsutani et al., "Skip2-LoRA: A Lightweight On-device DNN Fine-tuning Method for Low-cost Edge Devices", ASP-DAC'25.

Acknowledgements: H.M. was supported in part by JST AIP Acceleration Research JPMJCR23U3, Japan.