



適応性と信頼性を両立するオンデバイス 学習技術の確立

松谷宏紀（慶應大学）、渡邊竜司（パナソニック）、吉田康太（立命館大学）



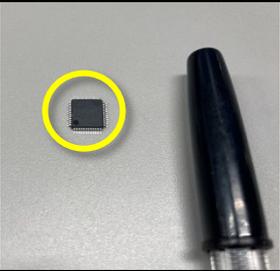
AIの応用ドメイン：計算機の視点

IoT

Mobile

Edge server

Cloud server



20-30mW



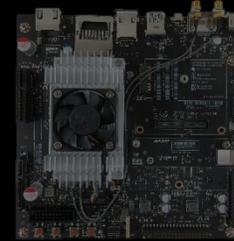
0.2-1W



2-10W



15W+



200W+



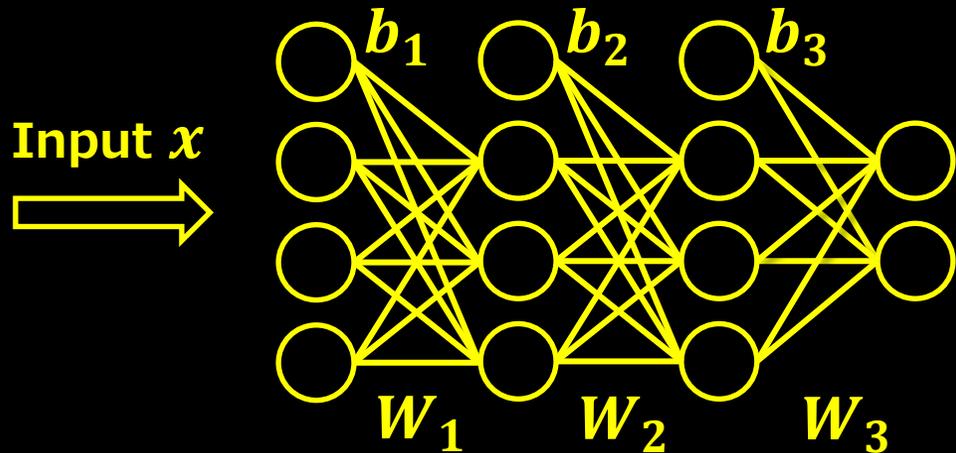
コントローラ

組み込みCPU モバイルCPU

組み込みGPU

高性能GPU

エッジAI：現実空間で動作する組み込みAI



推論 (Prediction)

入力 x と重み係数 W , b をもとに出力を計算

学習 (Training)

教師データをもとに重み係数 W , b を計算

AIの応用ドメイン：計算機の視点

IoT

Mobile

Edge server

Cloud server



20-30mW



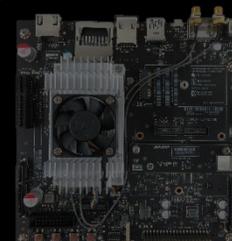
0.2-1W



2-10W



15W+



200W+



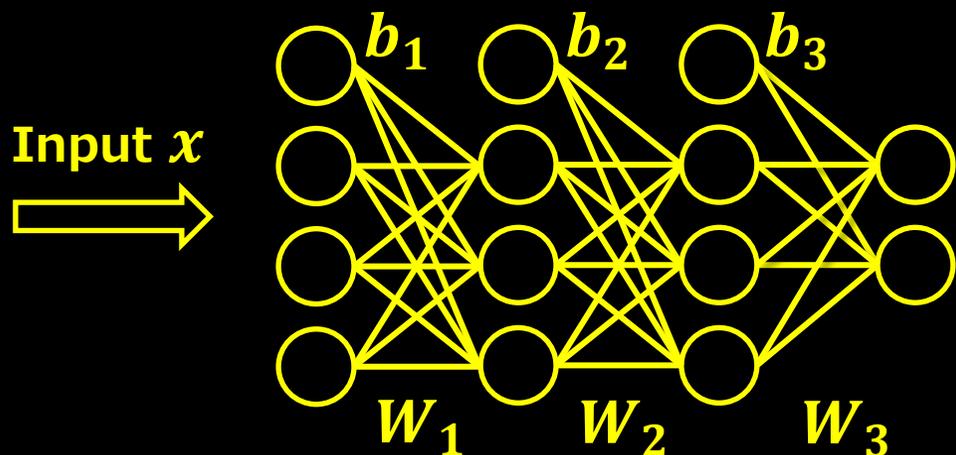
コントローラ

組み込みCPU モバイルCPU

組み込みGPU

高性能GPU

エッジAI：現実空間で動作する組み込みAI



従来のエッジAI

現場では推論のみ（モデルはサーバで学習しておく）



オンデバイス学習

現場で学習もできる（モデルを現場に合わせて更新）

さきがけ (2013~) 、CREST (2017~) 、AIP加速 (2023~)

IoT

Mobile

Edge server

Cloud server

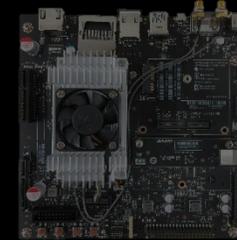
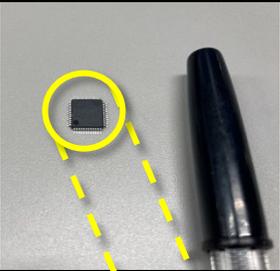
20-30mW

0.2-1W

2-10W

15W+

200W+



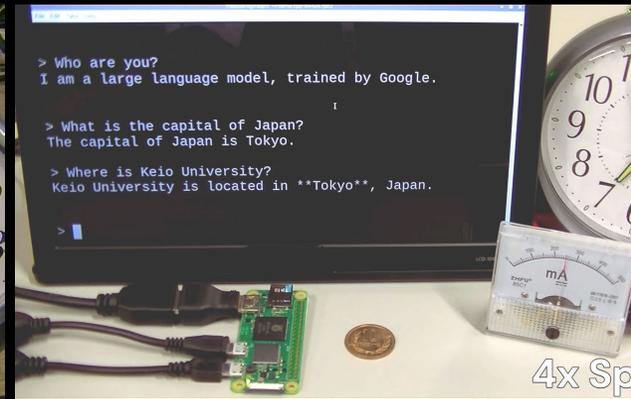
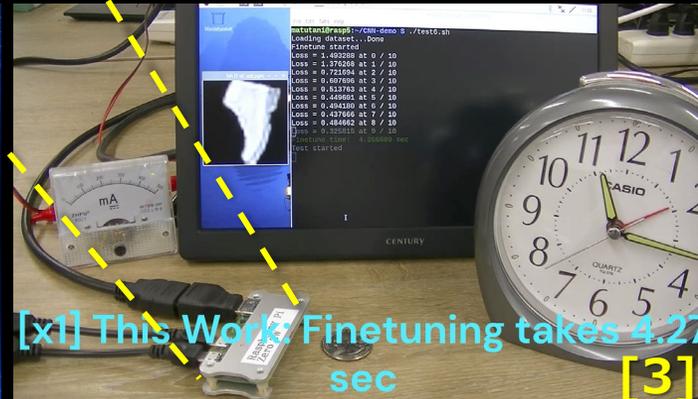
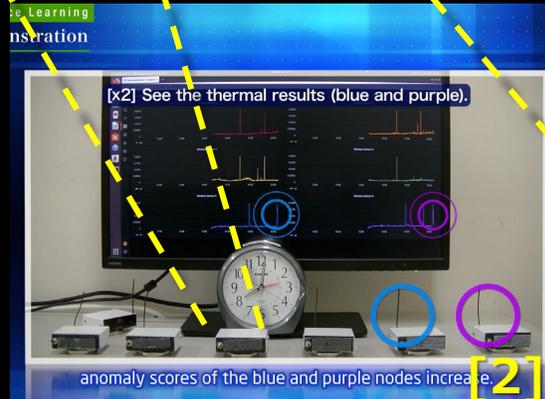
コントローラ

組み込みCPU

モバイルCPU

組み込みGPU

高性能GPU



信号処理用NN

画像認識用CNN

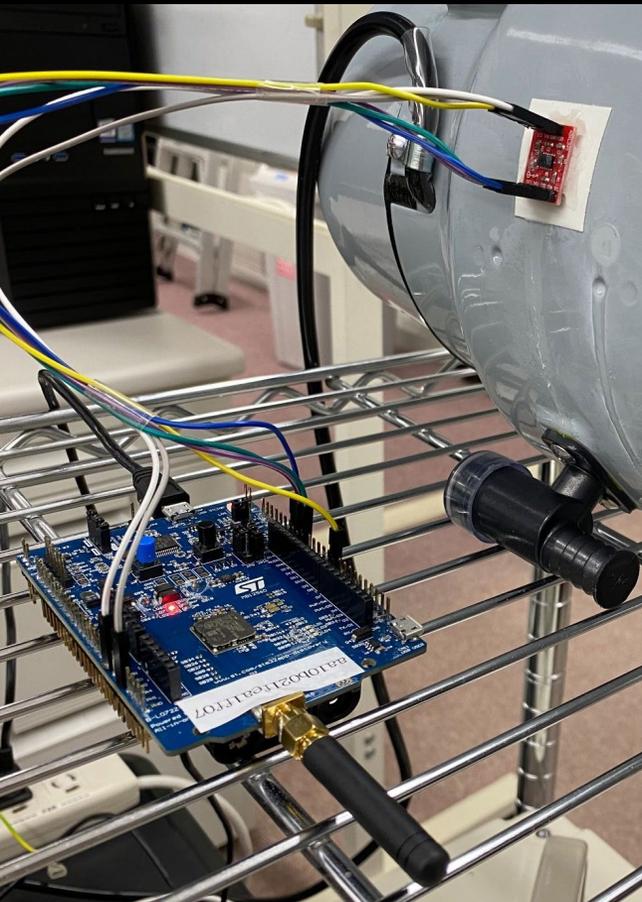
エッジ用LLM

[1] ローム株式会社, "エッジコンピューティング向け完全自立型AIソリューション Solist-AI™", <https://www.rohm.co.jp/support/solist-ai>.
[2] Kazuki Sunaga et al., "Addressing Gap between Training Data and Deployed Environment by On-Device Learning", IEEE Micro (2023).
[3] Keisuke Sugiura et al., "InstantFT: An FPGA-Based Runtime Subsecond Fine-tuning of CNN Models", IEEE Micro (2026).

エッジAI：設備監視への応用例

無線センサノードによる空調システムの監視 [1]

エッジ側でセンシング、前処理、推論（異常検知）、学習などの知的処理



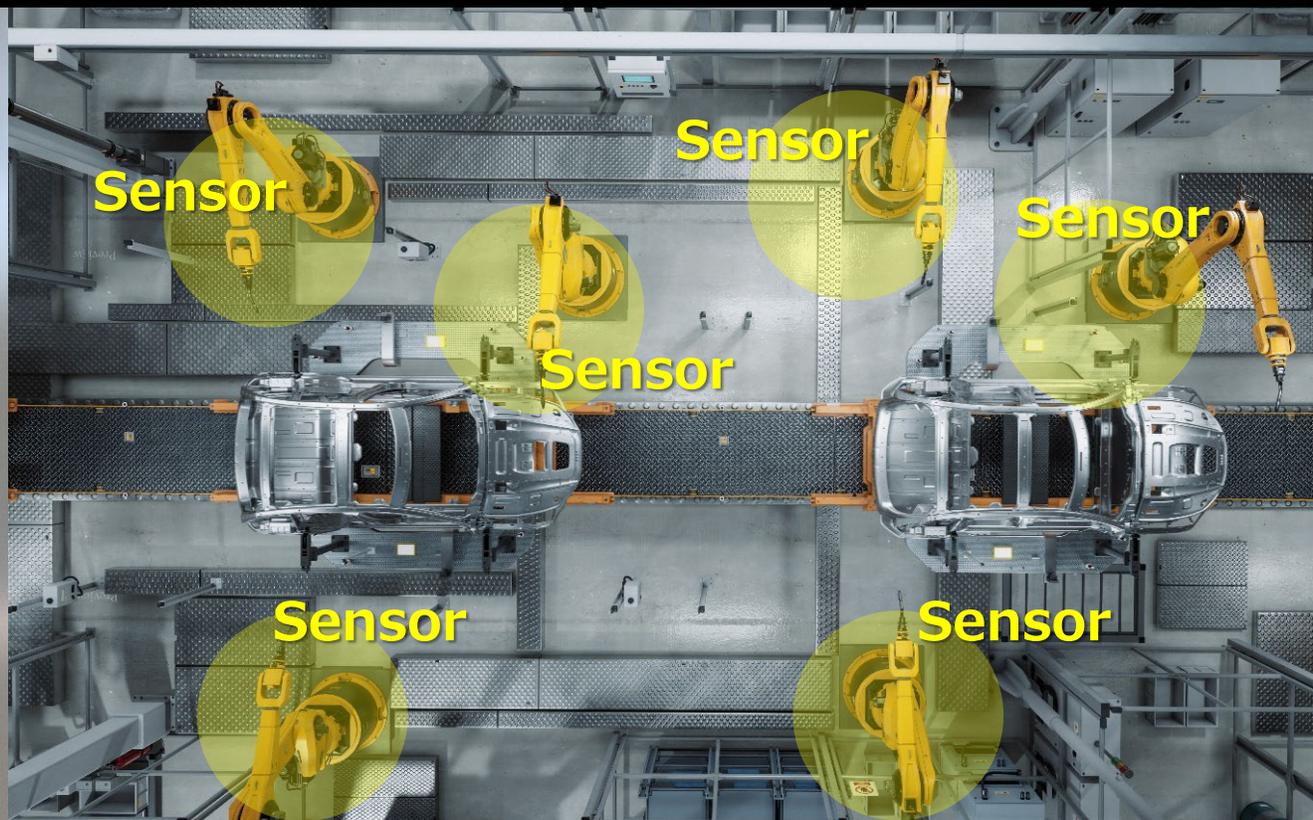
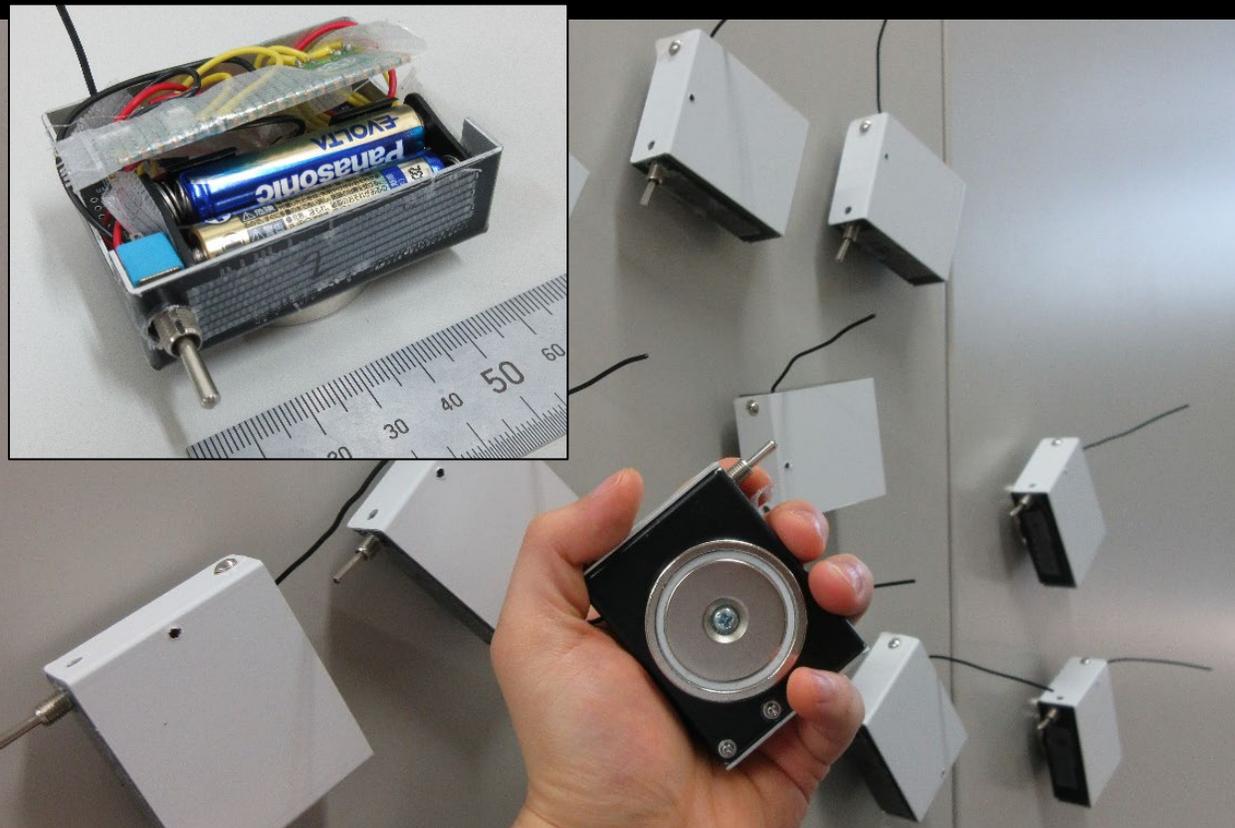
[1] 塚田 峰登, "回転機械の異常振動を検出する無線デバイスの開発", Keio Techno-Mall 2021.

オンデバイス学習：必要性

- 現場ごとにセンサの値の見え方が違う

訓練データ（事前に集めたセンサの値）と現場で実際に得られる値がズれる

原因：センサの角度（加速度センサのXYZ）、周囲のノイズ、個体差、…

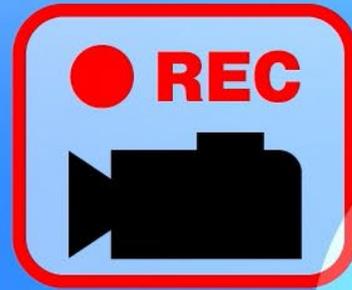


オンデバイス学習：必要性

- 現場ごとにセンサの値の見え方が違う → 現場でのオンデバイス学習

Training phase

Prediction phase



Sequential training mode



Training and anomaly detection
on edge devices

Sequential learning is performed when "train button" is pushed.

オンデバイス学習：基本的なアプローチ

- **解析的重み計算 [1]**

ニューラルネットワークの逐次重み計算アルゴリズムを軽量化

- **パラメータ効率の良いファインチューニング [2]**

誤差逆伝播 & 勾配降下法をベースに順伝播と逆伝播を軽量化

- **同時摂動近似 [3]**

順伝播のみで重みパラメータを最適化

- **専用ハードウェア化**

FPGA (Field-Programmable Gate Array) 等で高速化

[1] N. Y. Liang et al., "A Fast and Accurate Online Sequential Learning Algorithm for Feedforward Networks", IEEE Trans. on NN (2006).

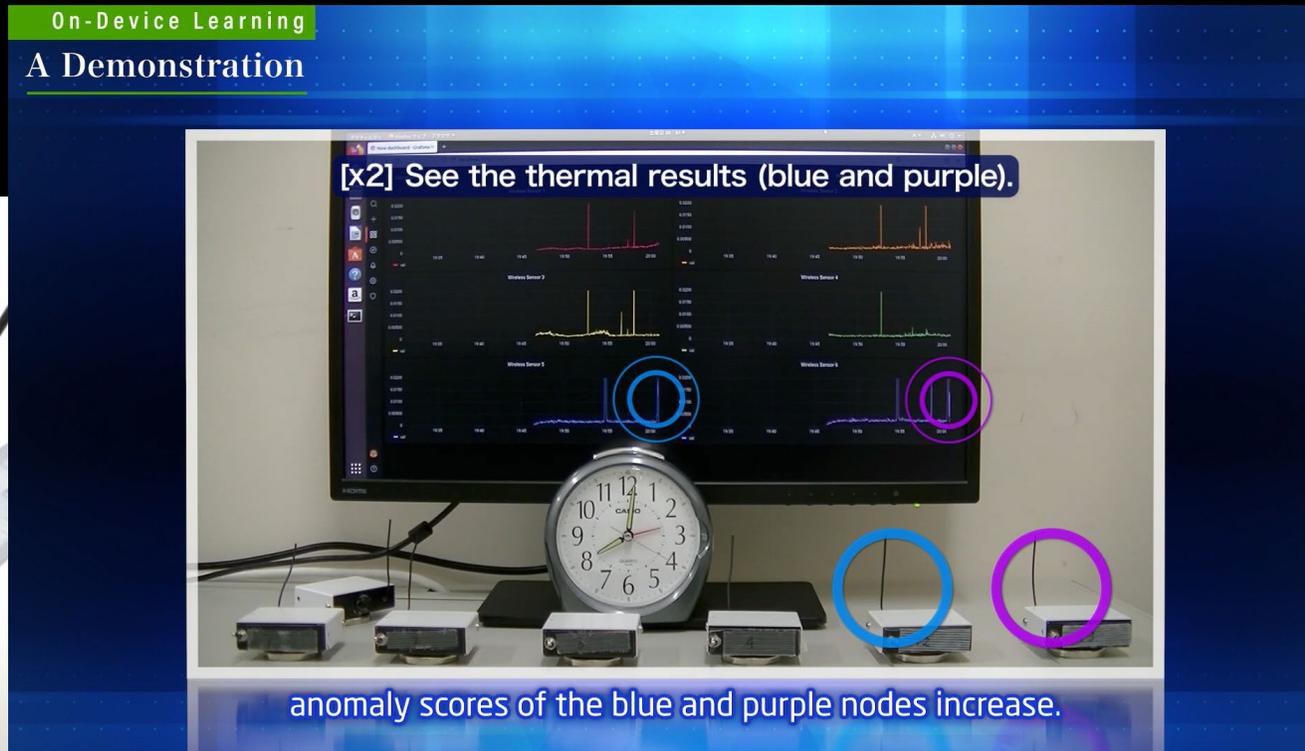
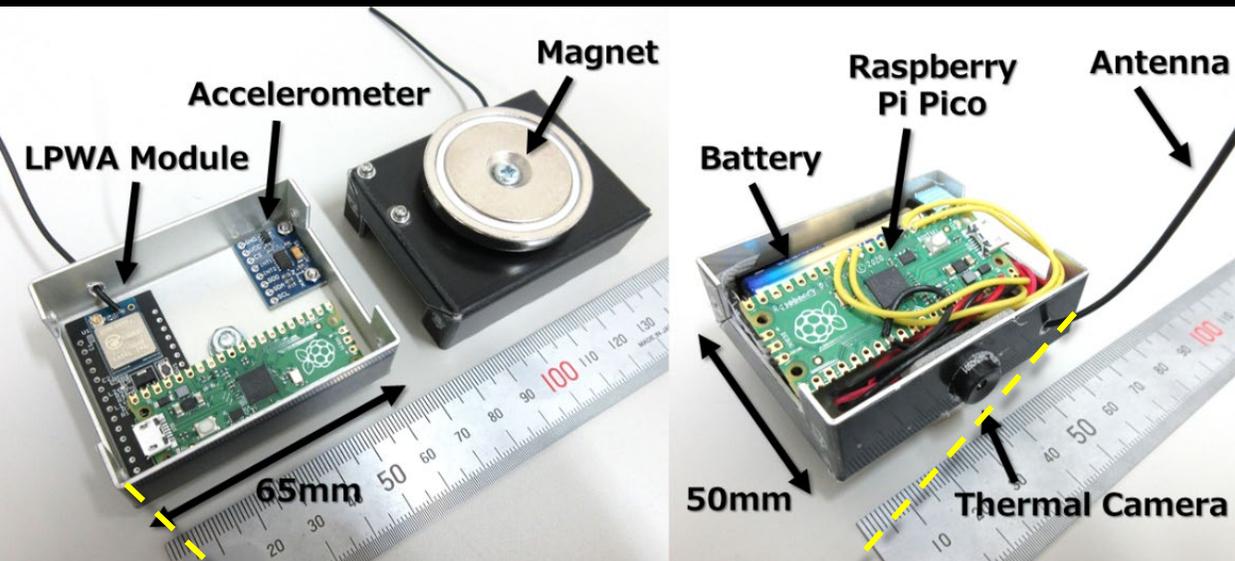
[2] Edward J. Hu et al., "LoRA: Low-Rank Adaptation of Large Language Models", arXiv:2106.09685 (2021).

[3] Sadhika Malladi et al., "Fine-Tuning Language Models with Just Forward Passes", NeurIPS'23.

応用例：貼り付けるだけ異常検知器

- 置かれた現場でモデルを学習

例：貼り付けるだけ異常検知器 [1]



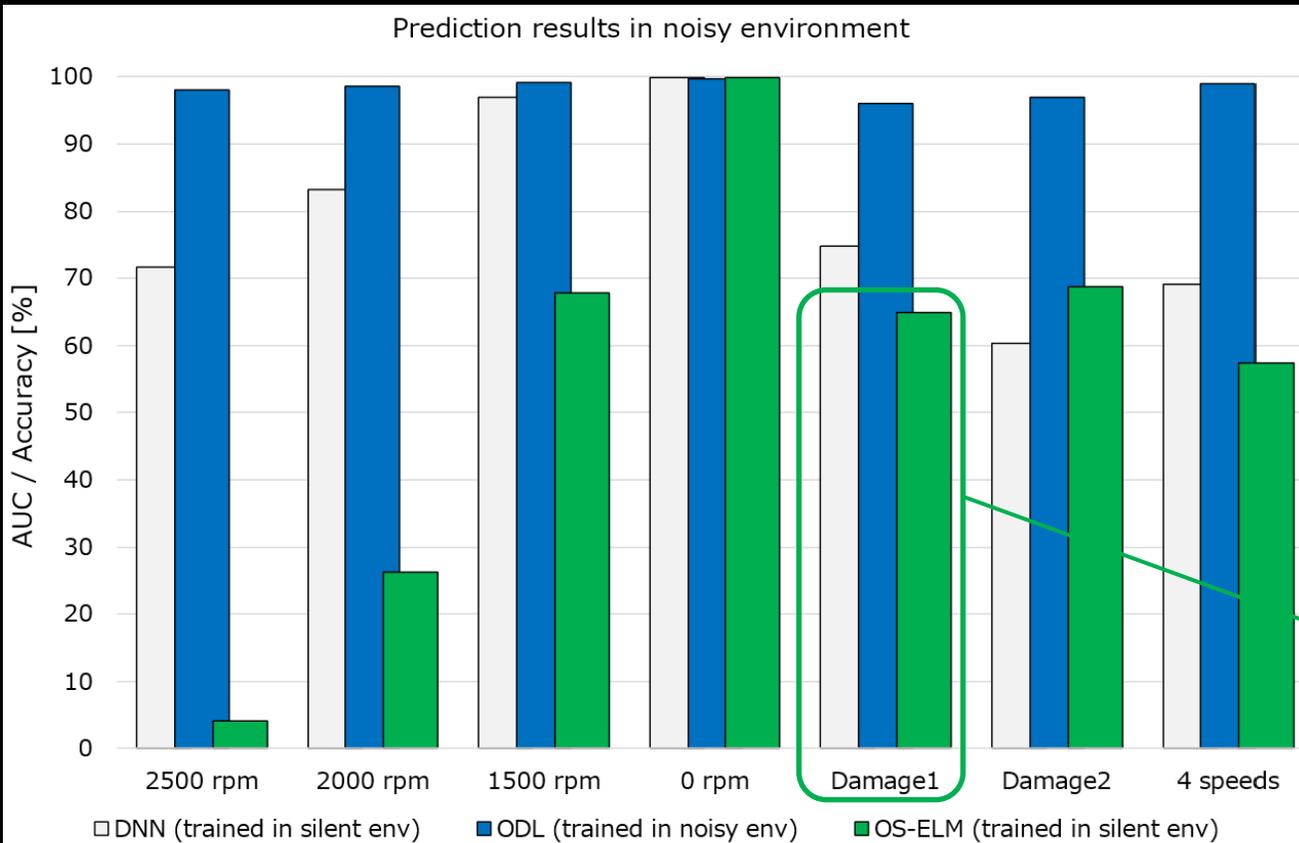
[1] Kazuki Sunaga et al., "Addressing Gap between Training Data and Deployed Environment by On-Device Learning", IEEE Micro (2023).

応用例：貼り付けるだけ異常検知器

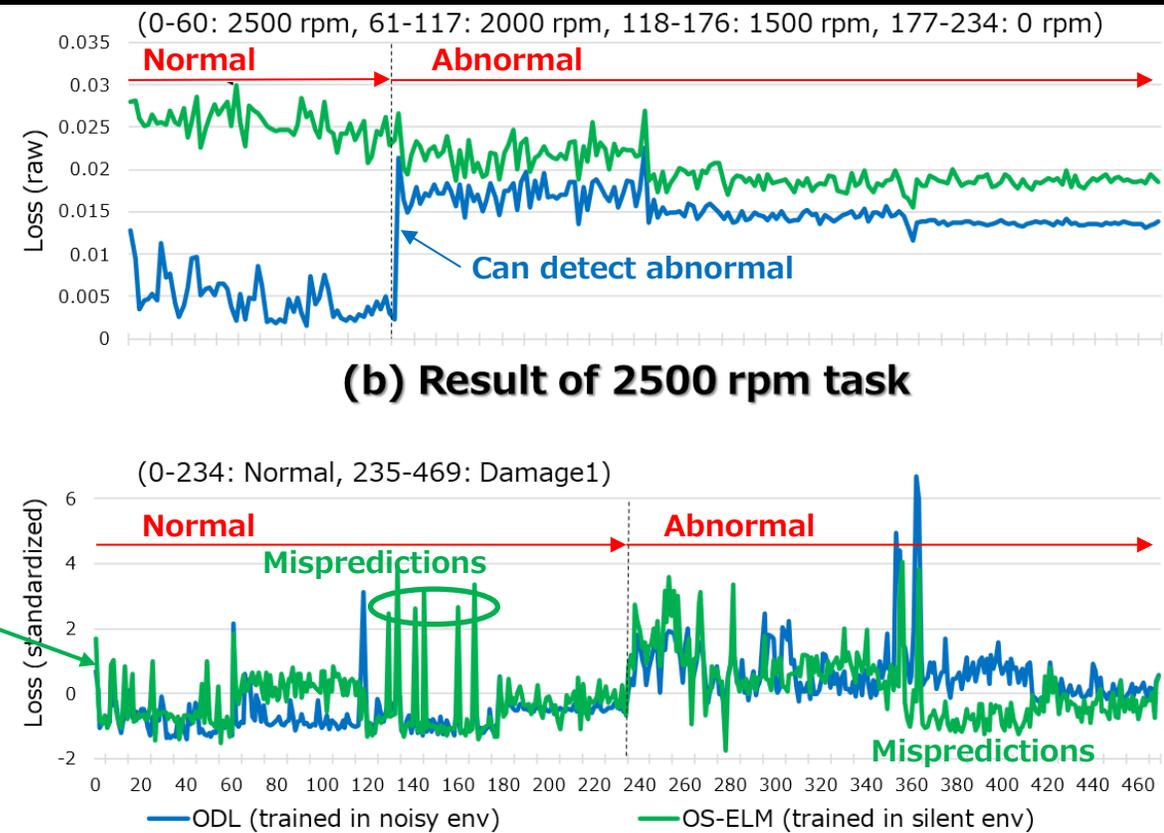
- 静音環境にて学習後、騒音環境に置かれて異常検知

■ 推論のみAI：騒音環境に適応できない → 精度が大幅に低下 ☹️

■ オンデバイス学習：騒音環境に適応できる → 高い精度をキープ 😊



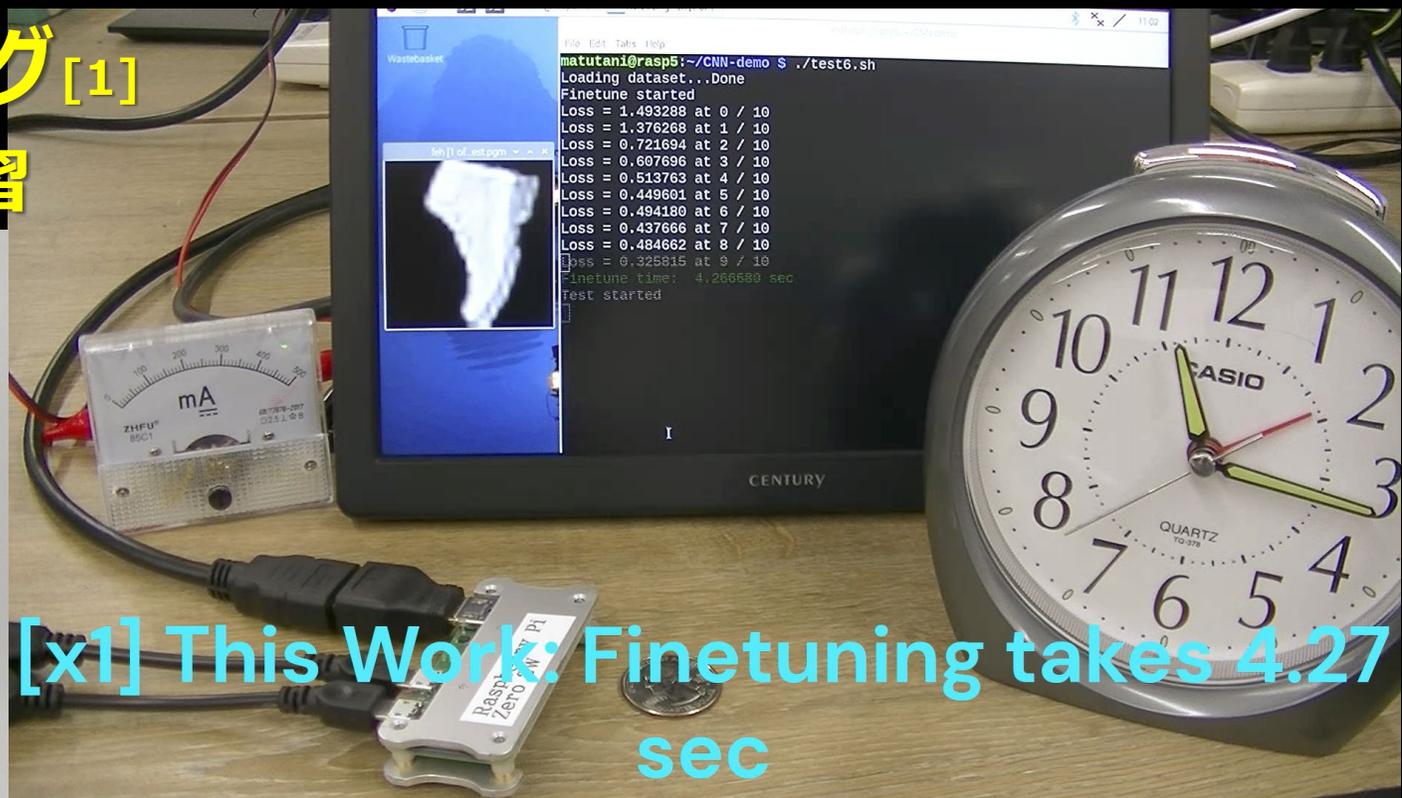
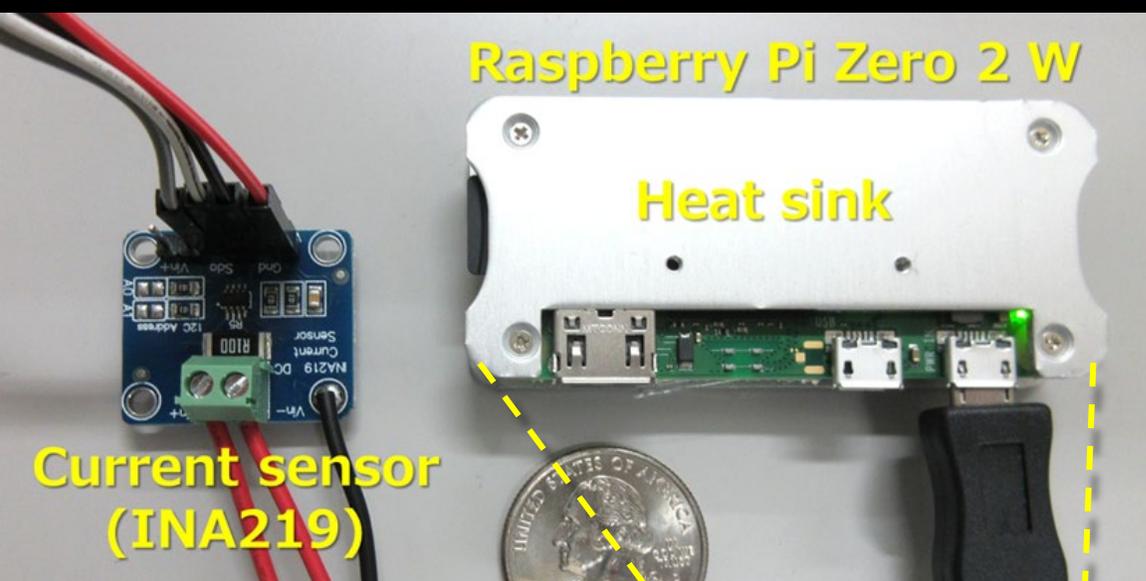
(a) Summary of seven tasks



(c) Result of Damage1 task

応用例：画像認識AIのファインチューニング

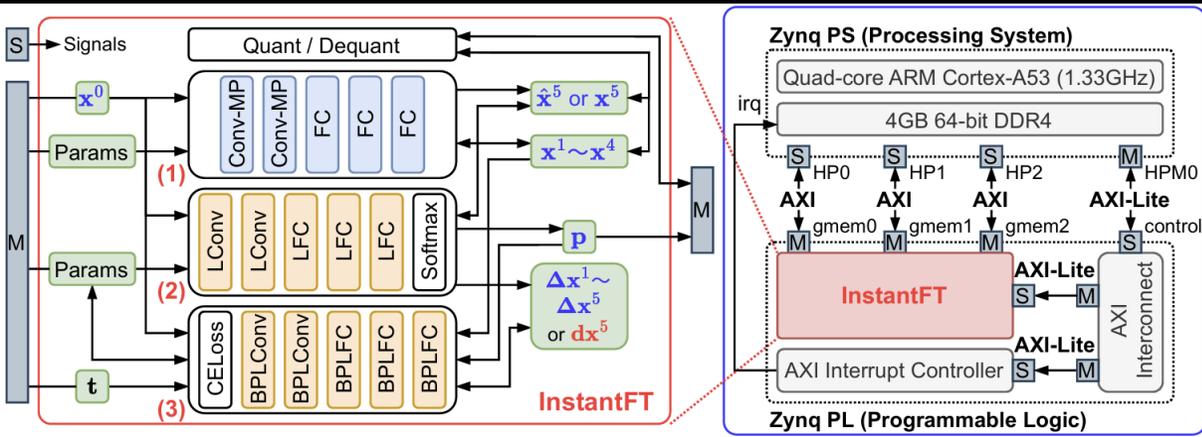
- ・ オンデバイスファインチューニング [1]
エッジ向けDNNモデルの現場学習



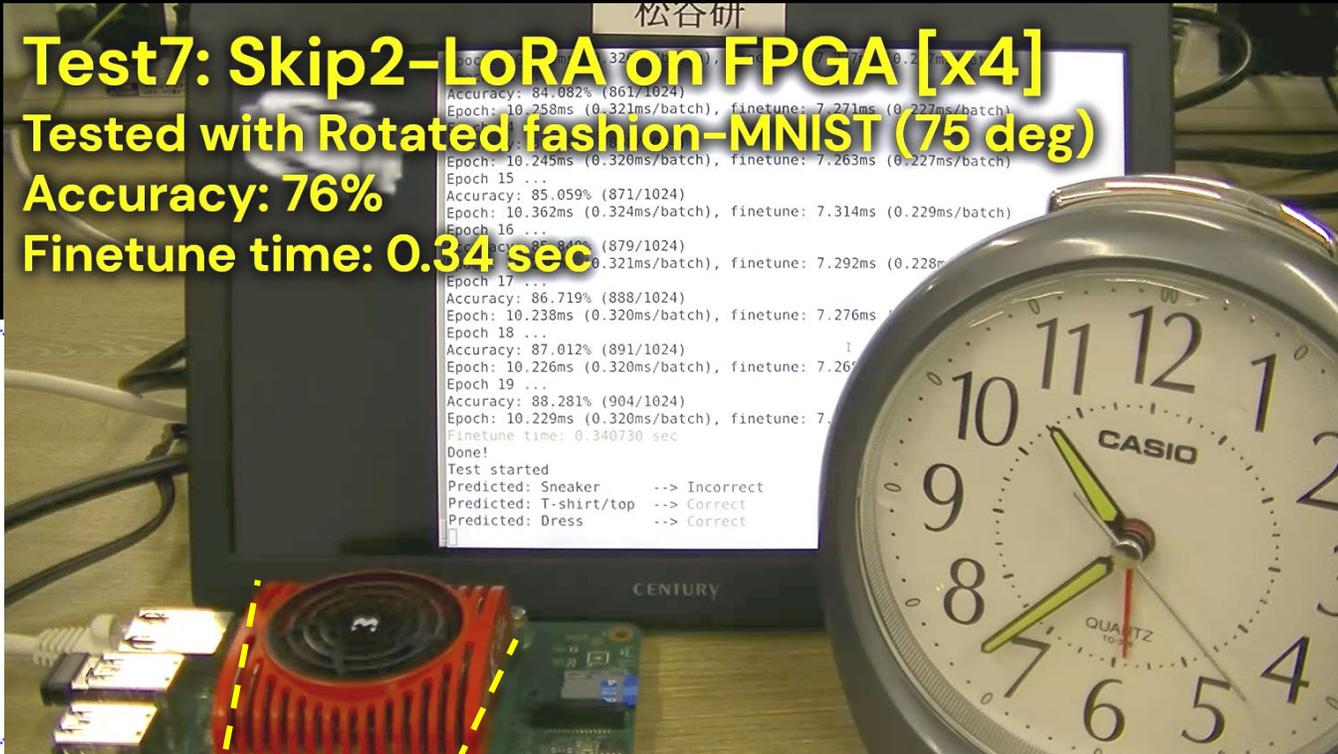
[1] Hiroki Matsutani et al., "Skip2-LoRA: A Lightweight On-device DNN Fine-tuning Method for Low-cost Edge Devices", ASP-DAC'25.

応用例：画像認識AIのファインチューニング

- 超高速ファインチューニング [1]
FPGAによる専用回路化によって
0.34秒でファインチューニング



Test7: Skip2-LoRA on FPGA [x4]
Tested with Rotated fashion-MNIST (75 deg)
Accuracy: 76%
Finetune time: 0.34 sec



```

Accuracy: 84.082% (861/1024)
Epoch: 10.258ms (0.321ms/batch), finetune: 7.271ms (0.227ms/batch)
Epoch 15 ...
Accuracy: 85.059% (871/1024)
Epoch: 10.362ms (0.324ms/batch), finetune: 7.314ms (0.229ms/batch)
Epoch 16 ...
Accuracy: 85.884% (879/1024)
Epoch: 10.245ms (0.320ms/batch), finetune: 7.263ms (0.227ms/batch)
Epoch 17 ...
Accuracy: 86.719% (888/1024)
Epoch: 10.238ms (0.320ms/batch), finetune: 7.276ms (0.228ms/batch)
Epoch 18 ...
Accuracy: 87.012% (891/1024)
Epoch: 10.226ms (0.320ms/batch), finetune: 7.265ms (0.227ms/batch)
Epoch 19 ...
Accuracy: 88.281% (904/1024)
Epoch: 10.229ms (0.320ms/batch), finetune: 7.265ms (0.227ms/batch)
Finetune time: 0.340730 sec
Done!
Test started
Predicted: Sneaker ---> Incorrect
Predicted: T-shirt/top ---> Correct
Predicted: Dress ---> Correct
    
```

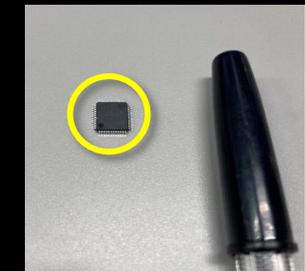
20-30mW

0.2-1W

2-10W

15W+

200W+

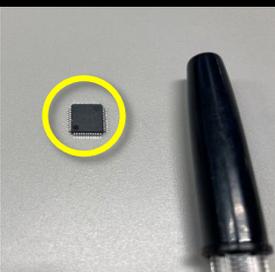


応用例：画像認識AIのファインチューニング

- 超高速ファインチューニング [1]
FPGAによる専用回路化によって
0.34秒でファインチューニング

Test7: Skip2-LoRA on FPGA [x4]
Tested with Rotated fashion-MNIST (75 deg)
Accuracy: 76%
Finetune time: 0.34 sec

```
Accuracy: 84.082% (861/1024)  
Epoch: 10.258ms (0.321ms/batch), finetune: 7.271ms (0.227ms/batch)  
Epoch: 10.245ms (0.320ms/batch), finetune: 7.263ms (0.227ms/batch)  
Epoch 15 ...  
Accuracy: 85.059% (871/1024)  
Epoch: 10.362ms (0.324ms/batch), finetune: 7.314ms (0.229ms/batch)  
Epoch 16 ...  
Accuracy: 86.884% (879/1024)  
Epoch: 10.245ms (0.321ms/batch), finetune: 7.292ms (0.228ms/batch)  
Epoch 17 ...  
Accuracy: 86.719% (888/1024)  
Epoch: 10.238ms (0.320ms/batch), finetune: 7.276ms (0.227ms/batch)  
Epoch 18 ...  
Accuracy: 87.012% (891/1024)  
Epoch: 10.226ms (0.320ms/batch), finetune: 7.261ms (0.227ms/batch)  
Epoch 19 ...  
Accuracy: 88.281% (904/1024)  
Epoch: 10.229ms (0.320ms/batch), finetune: 7.261ms (0.227ms/batch)  
Finetune time: 0.340730 sec  
Done!  
Test started  
Predicted: Sneaker ---> Incorrect  
Predicted: T-shirt/top ---> Correct  
Predicted: Dress ---> Correct
```



20-30mW



0.2-1W



2-10W



15W+



200W+



[1] Keisuke Sugiura et al., "InstantFT: An FPGA-Based Runtime Subsecond Fine-tuning of CNN Models", IEEE Micro (2026).

応用例：オンデバイス学習AIチップ[®]



※「Solist-AITM」はローム株式会社の商標または登録商標です。



モータ軸受損傷の検知



バッテリーの劣化・残量推定



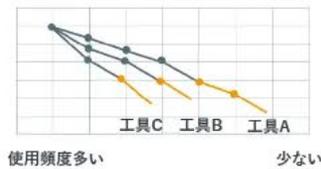
サーバーの異常発熱の検知



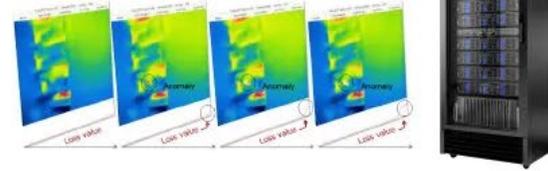
その他



定速運転での正常加速度波形を学習し、軸受損傷をパターン検知



個々のバッテリー状態を追加学習し、残量や劣化を高精度に推定



AIによって特異点での異常発熱を検知し、異常度の上昇を検出

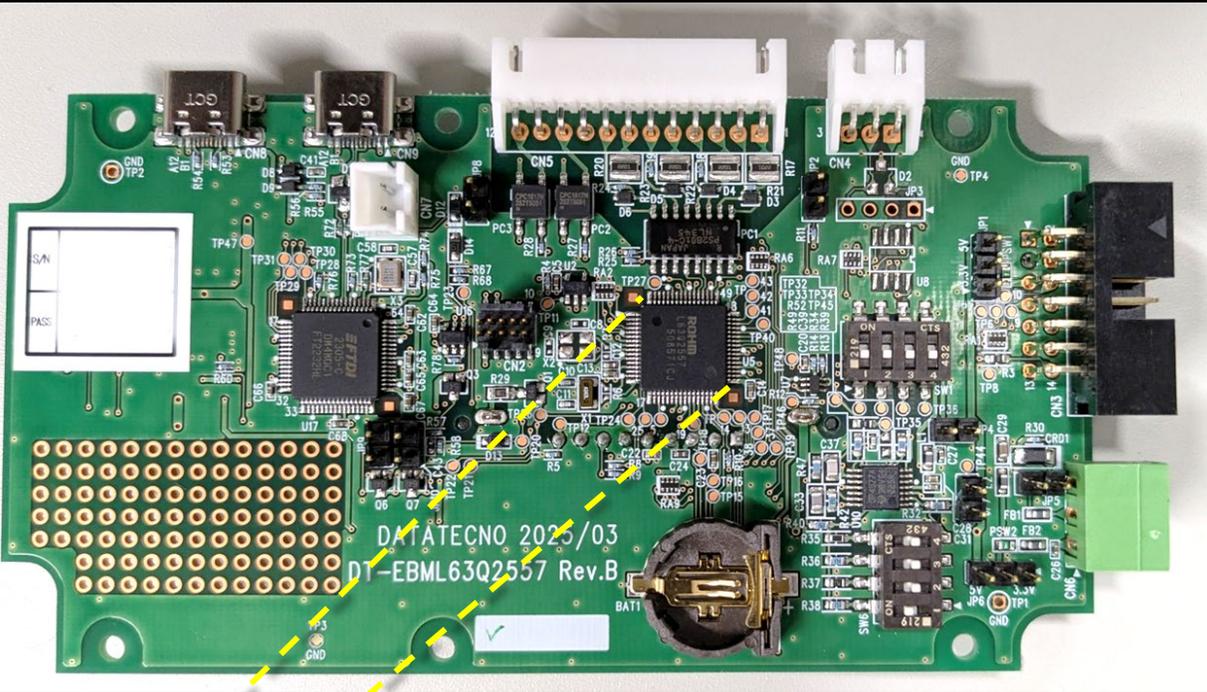
- ・不良品選別
人間の感覚による処理を自動化
- ・測定環境へのマッチング
最適なパラメータを現場で更新
- ・劣化予知
耐用年数の長いインフラ設備の保全
- ・介護補助
独居老人、寝たきりの人の見守り

[1] ローム株式会社, "エッジコンピューティング向け完全自立型AIソリューション Solist-AITM", <https://www.rohm.co.jp/support/solist-ai>.

[2] 日本経済新聞, "ローム、AI搭載マイコンを開発 ネット使わず自動学習", 日本経済新聞電子版 (2025年3月18日).

応用例：オンデバイス学習AIチップ

- オンデバイス学習機能付きAIチップ [1]



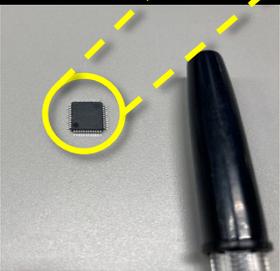
20-30mW

0.2-1W

2-10W

15W+

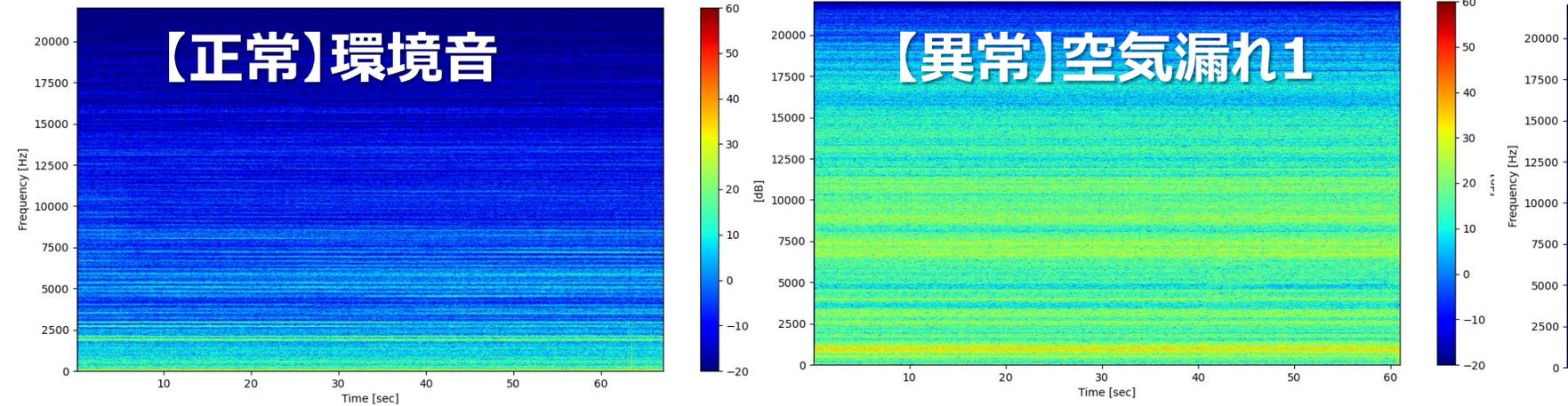
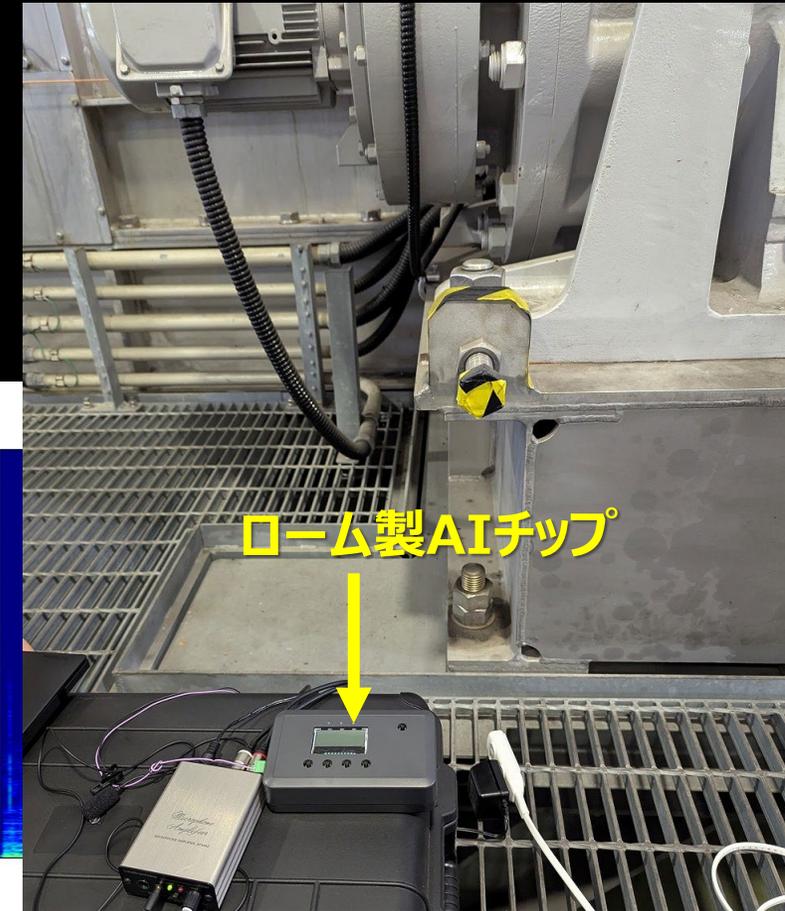
200W+



[1] ローム株式会社, "エッジコンピューティング向け完全自立型AIソリューション Solist-AI™", <https://www.rohm.co.jp/support/solist-ai>.

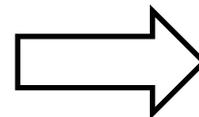
応用例：上下水道施設の異常検知

- オンデバイス学習を用いたポンプ室の異常検知
 雑音環境下において、空気漏れを検出
 正常音を雑音込みでオンデバイス学習 → 異常検知



正常音のみ学習 (ベースライン)

	正常音と判定	異常音と判定
実際の正常音	681	1691 <small>誤検知</small>
実際の異常音	2 <small>見逃し</small>	710

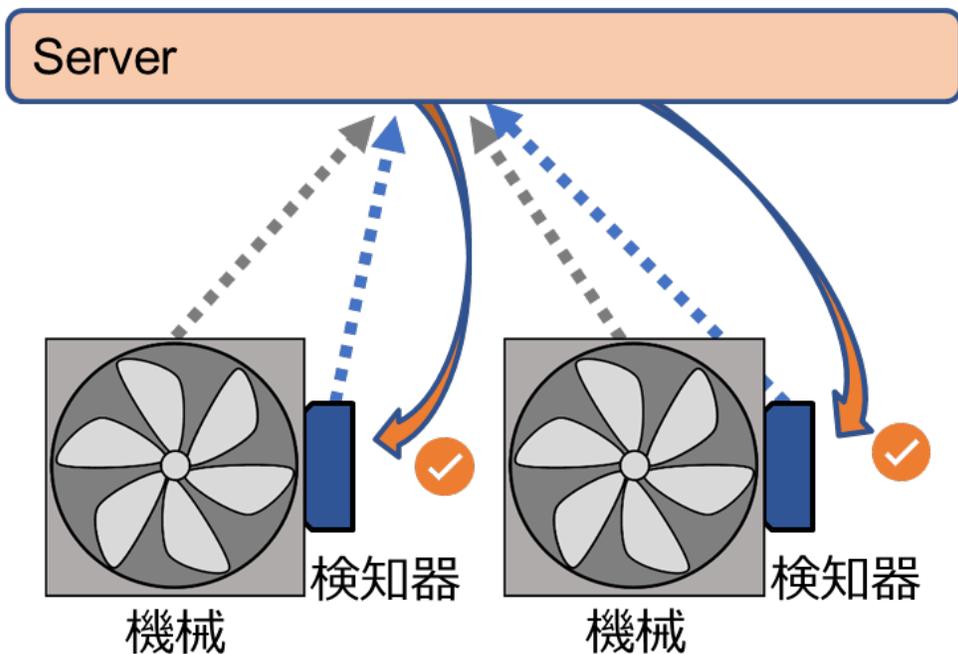


現場の雑音をオンデバイス学習 (提案手法)

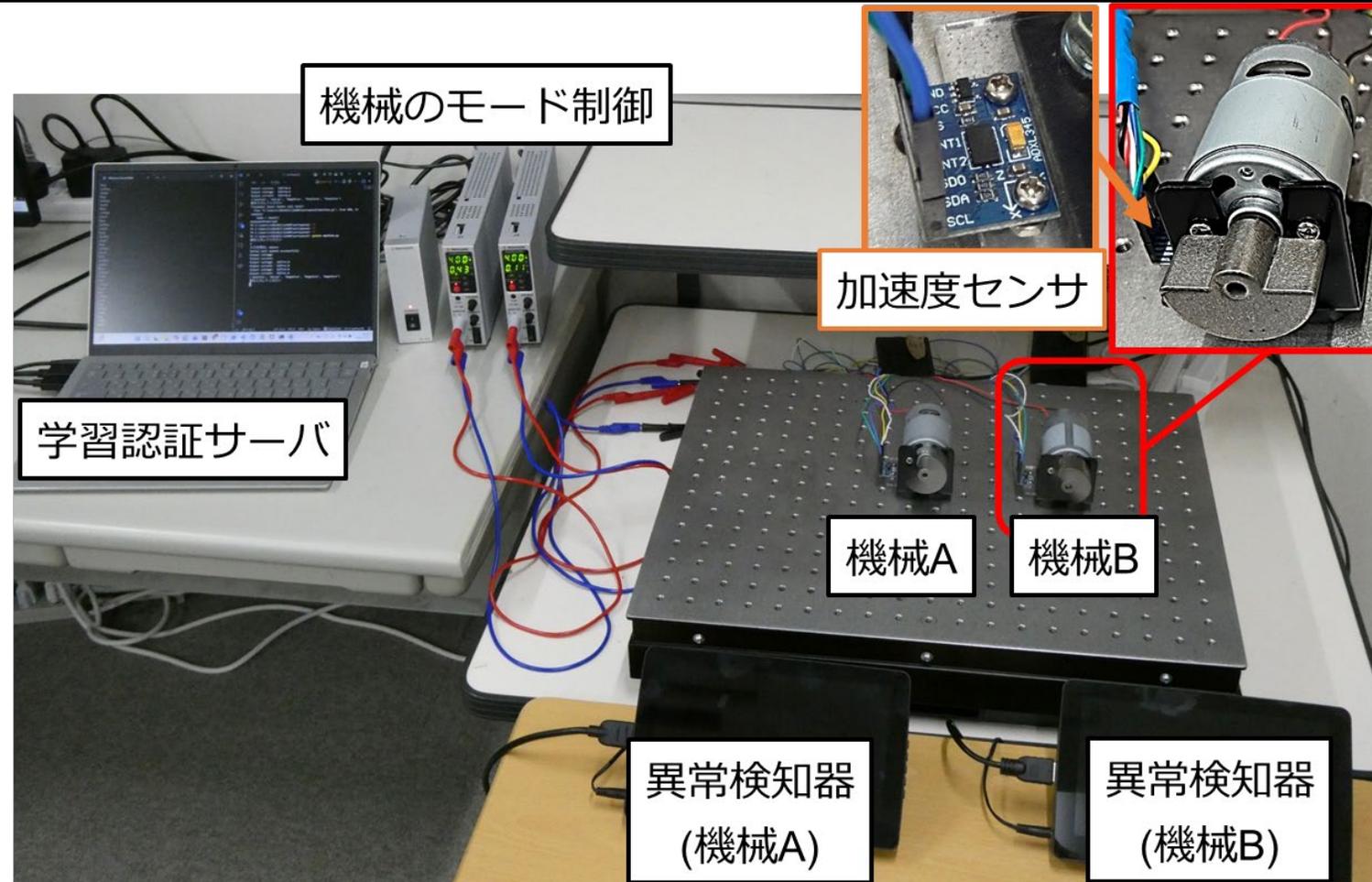
	正常音と判定	異常音と判定
実際の正常音	1016	58 <small>誤検知</small>
実際の異常音	5 <small>見逃し</small>	716

オンデバイス学習：意図せぬ追加学習への対策

- 分散合意：オンデバイス学習の「安価にバラまける」という特徴を活用
複数オンデバイス学習器の合意によって追加学習をトリガ [1]



仮定：環境起因のドリフトなら、同一空間内の他検知器においてもドリフトが検出されるはず

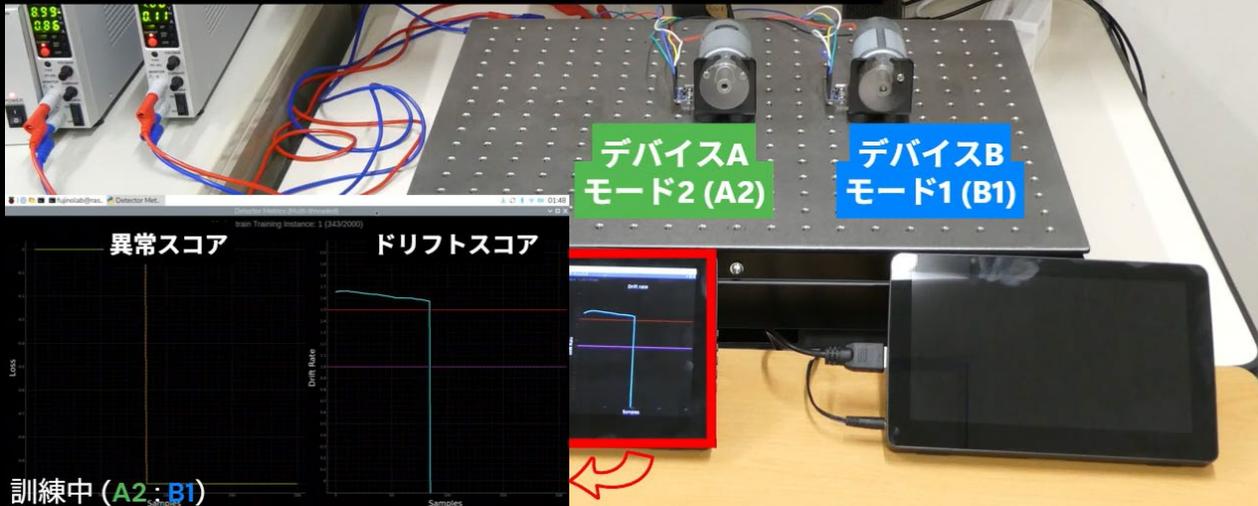


[1] Takahito Ino et al., "Mitigating Data Poisoning Attack in On-Device Learning Anomaly Detectors via Peer Consensus", AICompS'25.

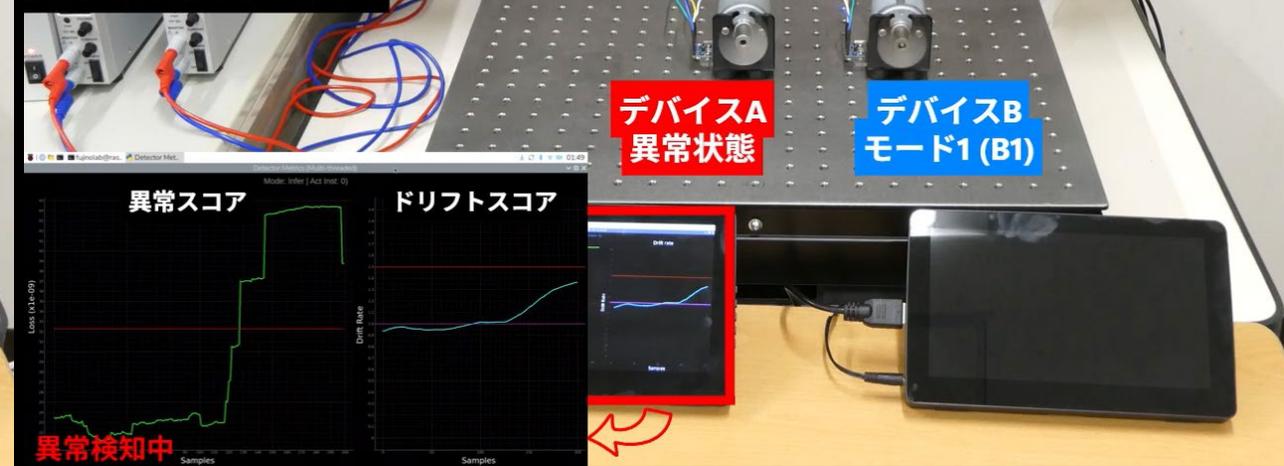
オンデバイス学習：意図せぬ追加学習への対策

- 分散合意：オンデバイス学習の「安価にバラまける」という特徴を活用
複数オンデバイス学習器の合意によって追加学習をトリガ [1]

デバイスAのモードが2に切り替わる。
異常スコアが上昇し、ドリフトが検知される。
モード変更とドリフト検知の通知がそれぞれサーバに送られ、
サーバから学習許可が下り、訓練が開始される。



デバイスAのモードが異常状態に切り替わる。
異常スコアが上昇し、ドリフトが検知される。
モード切替の通知がサーバに送られていないため、サーバは学習許可を
出さない。
異常と判定される。



追加学習が許可されるケース（既知のドリフト）

追加学習が許可されないケース（異常）

オンデバイス学習：保有技術と応用ドメイン

IoT

Mobile

Edge server

Cloud server

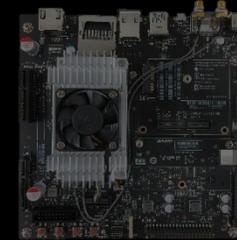
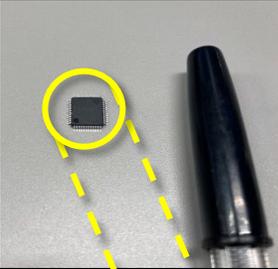
20-30mW

0.2-1W

2-10W

15W+

200W+



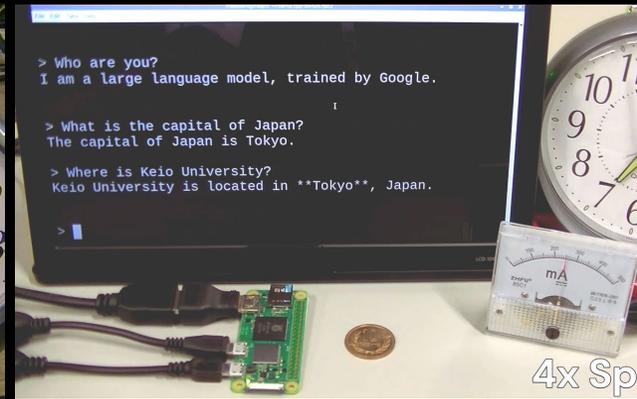
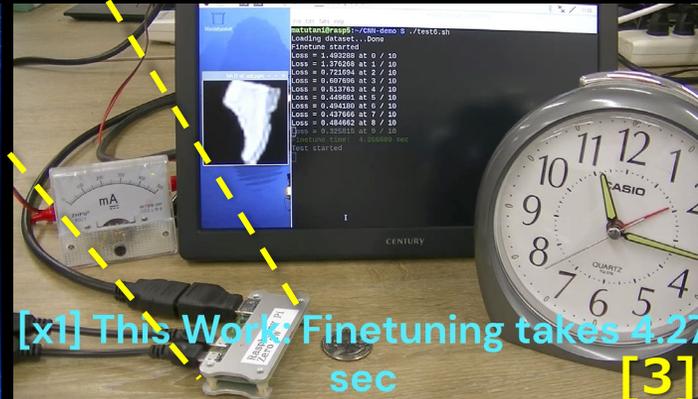
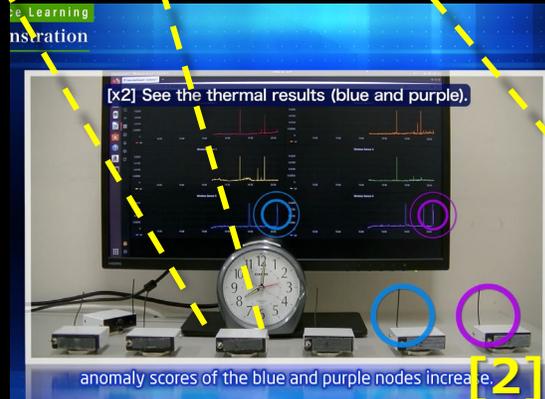
コントローラ

組み込みCPU

モバイルCPU

組み込みGPU

高性能GPU



信号処理用NN

画像認識用CNN

エッジ用LLM

[1] ローム株式会社, "エッジコンピューティング向け完全自立型AIソリューション Solist-AI™", <https://www.rohm.co.jp/support/solist-ai>.

[2] Kazuki Sunaga et al., "Addressing Gap between Training Data and Deployed Environment by On-Device Learning", IEEE Micro (2023).

[3] Keisuke Sugiura et al., "InstantFT: An FPGA-Based Runtime Subsecond Fine-tuning of CNN Models", IEEE Micro (2026).

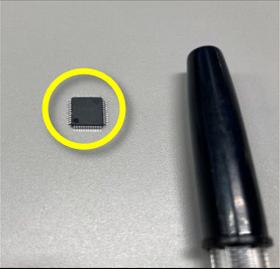
オンデバイス学習：まとめと今後の方向性

IoT

Mobile

Edge server

Cloud server



20-30mW



0.2-1W



2-10W



15W+



200W+



コントローラ

組み込みCPU

モバイルCPU

組み込みGPU

高性能GPU

エッジAI

クラウドAI

クラウドへの一極集中



エッジに付加価値を持たせる

データや計算をデータセンタに集約
強力なGPU（最先端の半導体技術）
エッジはセンシングだけ
エッジの付加価値は小さい

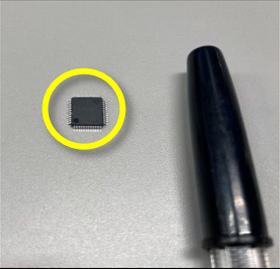
オンデバイス学習：まとめと今後の方向性

IoT

Mobile

Edge server

Cloud server



20-30mW



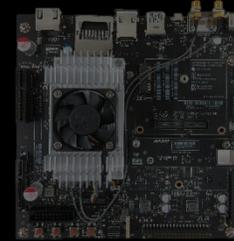
0.2-1W



2-10W



15W+



200W+



コントローラ

組み込みCPU

モバイルCPU

組み込みGPU

高性能GPU

エッジAI

クラウドへの一極集中



エッジに付加価値を持たせる

従来のエッジAI

現場では推論のみ（モデルはサーバで学習しておく）

オンデバイス学習

現場で学習もできる（モデルを現場で更新できる）

例：ユーザさんが自分の環境に合わせて、
ボタン1つでモデルの構築まで出来たら便利では？

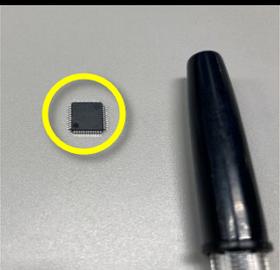
最後に：本研究がもたらす未来イメージ

IoT

Mobile

Edge server

Cloud server



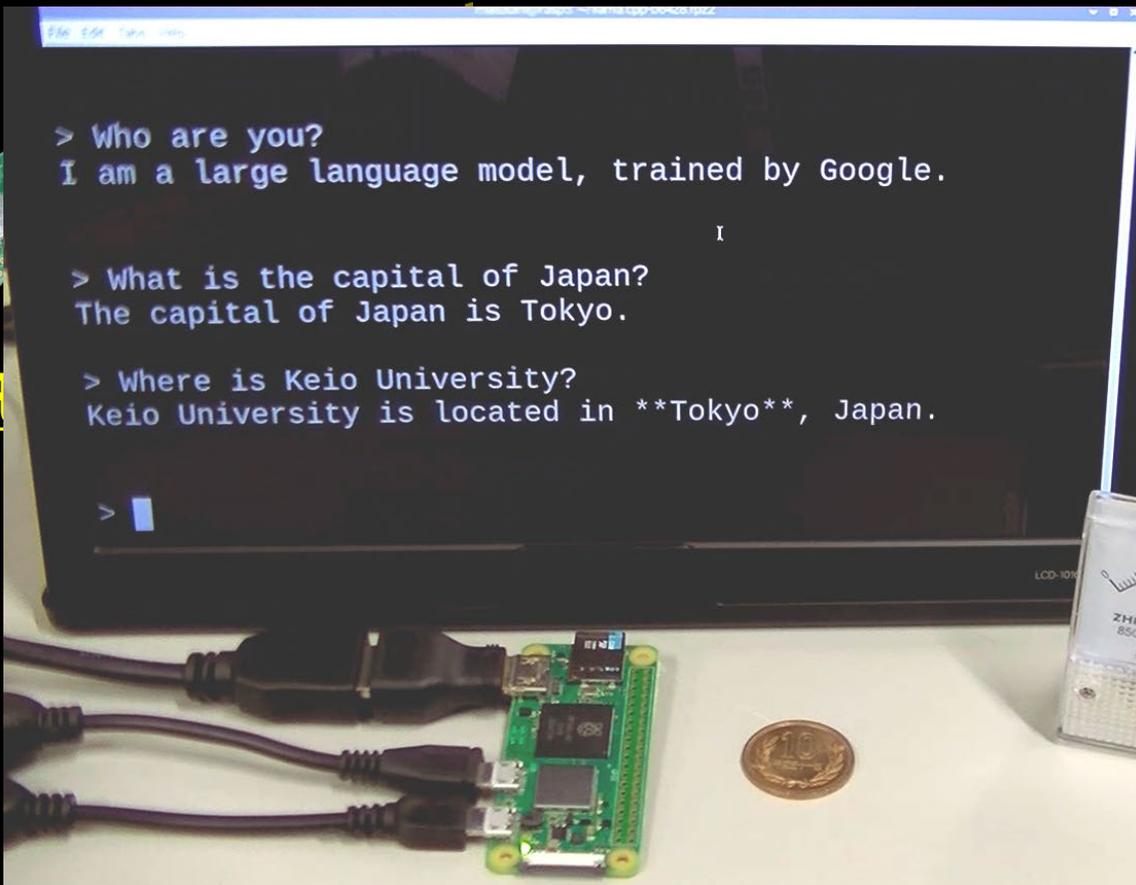
20-30mW

コントローラ



組み込みCPU

0.2-1W



4x Speed

エッジAI

クラウドへの一極集中



エッジに付加価値を持たせる

今後の方向性：

Physical AIに資する高度なAI技術開発

[1] Hiroki Matsutani et al., "Accelerating Local LLMs on Resource-Constrained Edge Devices via Distributed Prompt Caching", arXiv Preprint (2026).

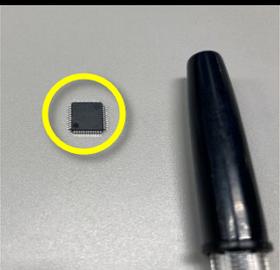
最後に：本研究がもたらす未来イメージ

IoT

Mobile

Edge server

Cloud server



20-30mW



0.2-1W



コントローラ

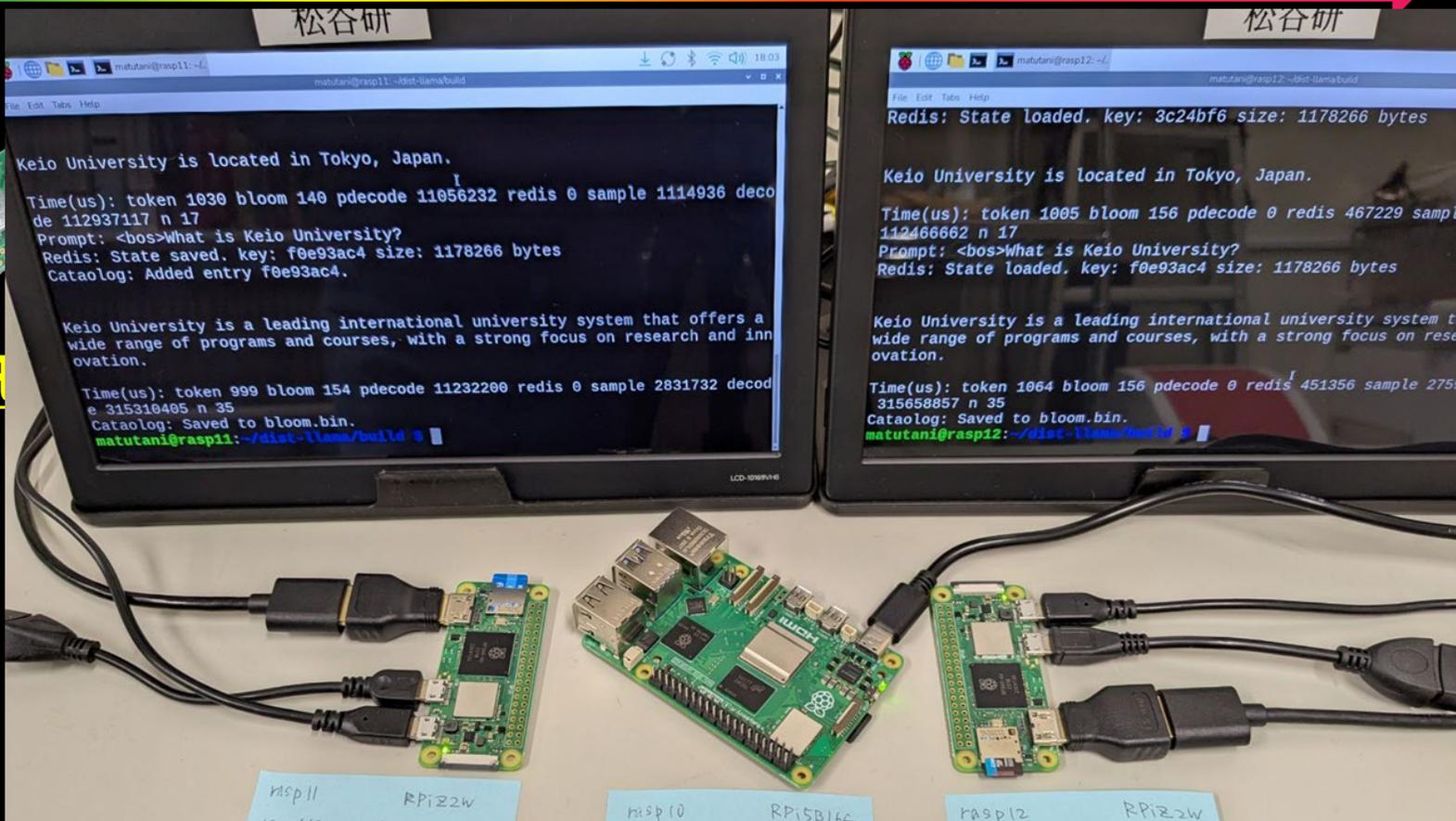
組み込みCPU

エッジAI

クラウドへの一極集中



エッジに付加価値を持たせる



今後の方向性：

Physical AIに資する高度なAI技術開発