Computation-in-Memoryにおける不揮発性メモリ書き込み誤差による推論精度劣化の補償

東京大学大学院 工学系研究科

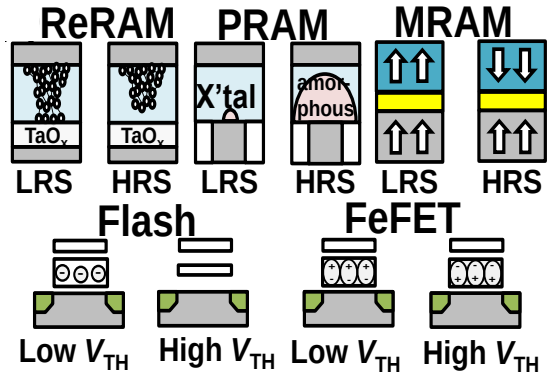
電気系工学専攻 竹内研究室

修士2年　吉清 秦生

目次

背景: Computation-in-Memory(CiM)とは

不揮発性
メモリ・・・　電源を切ってもデータが失われないメモリ
抵抗値としてデータを保存する



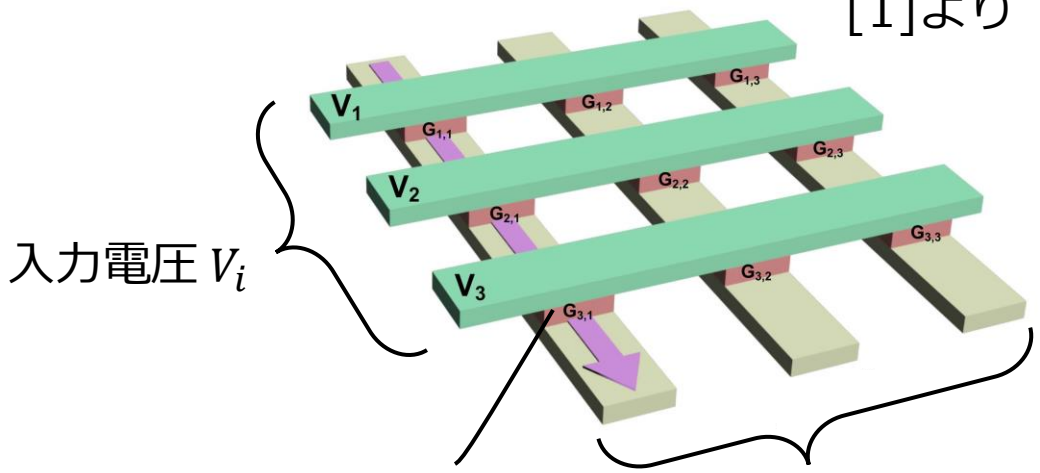
CiM・・・　不揮発性メモリを使った格子状の回路
「積和演算」を低電力、高速に処理できる

[1]より

入出力の関係　$I_j = \sum_i G_{i,j} \cdot V_i$　（積和演算）

↓

積和演算を多く利用する、
ニューラルネットワークの演算を高速化する

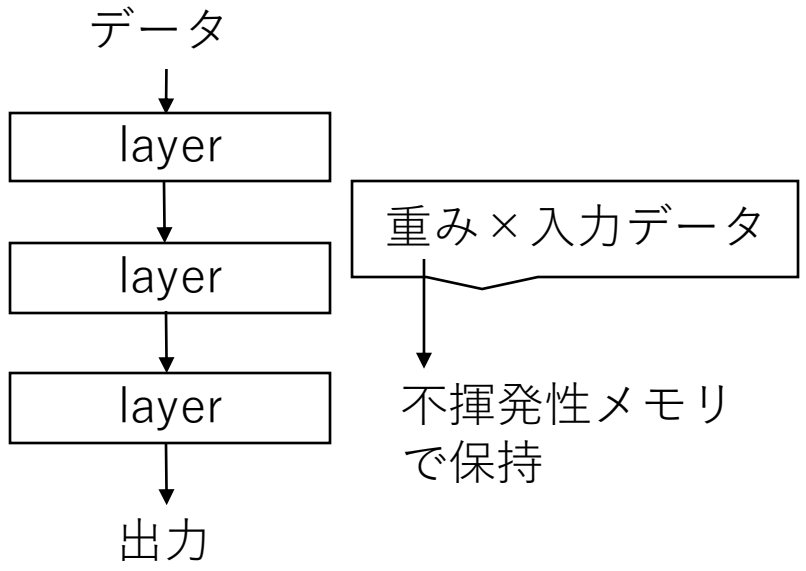


[1] Xi, Y., Gao, B., Tang, J., Chen, A., Chang, M. F., Hu, X. S., Spiegel, J. Van Der, Qian, H., & Wu, H. (2021). In-memory Learning with Analog Resistive Switching Memory: A Review and Perspective. Proceedings of the IEEE,109(1),14–42.

背景: ニューラルネットワーク

- 画像や音声などの情報を入力し、層構造の演算を繰り返し、出力を出す機構
- 各層には重みというパラメータがあり、基本的に入力データと重みの積和演算を行うことで出力を計算する
- 重みの値は、大量のデータを入力して学習する。

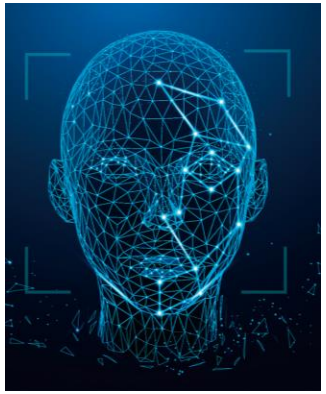
- CiMでは、学習済み重みを不揮発性メモリに記憶する。



- エッジデバイス(IoTデバイス, スマホ, 車など)上での使用
- エッジデバイスは計算能力、電力が限られている

→CiMデバイスを使用することで、
推論演算を高速化、低消費電力化

物体検知



顔認識

音声認識

課題: 不揮発性メモリの書き込みばらつきエラー

- 物性として値を保持するため、完全には制御できない

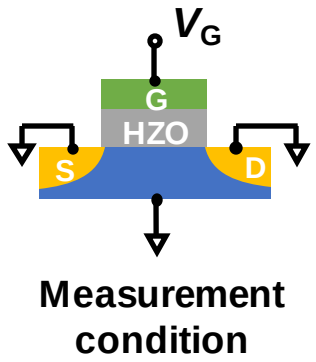
さらに

- 不揮発性メモリは新興の素子であり制御精度が低い
- アナログ値として値を保存するため、補正も難しい

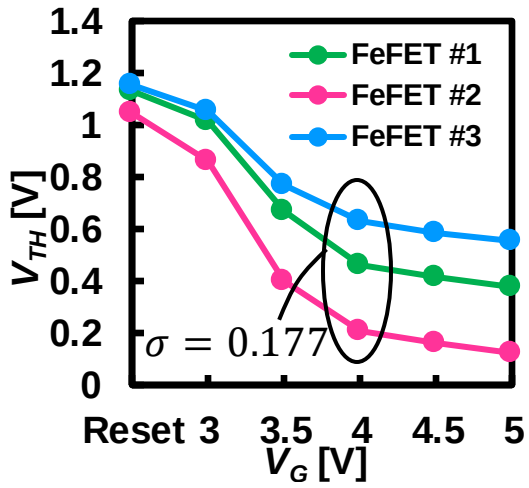
- 結果として、値の書き込み時に、保存した値がばらついてしまう

➡ ニューラルネットワークの推論精度が低下

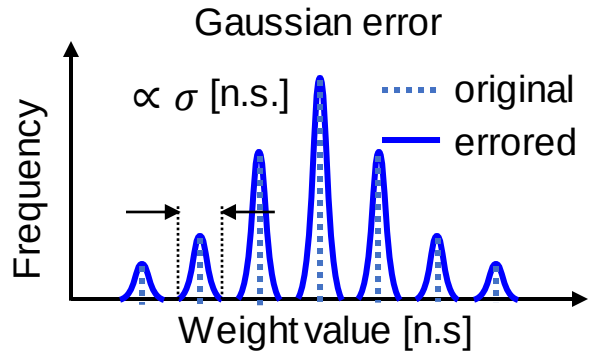
FeFETでの測定例



Measurement condition

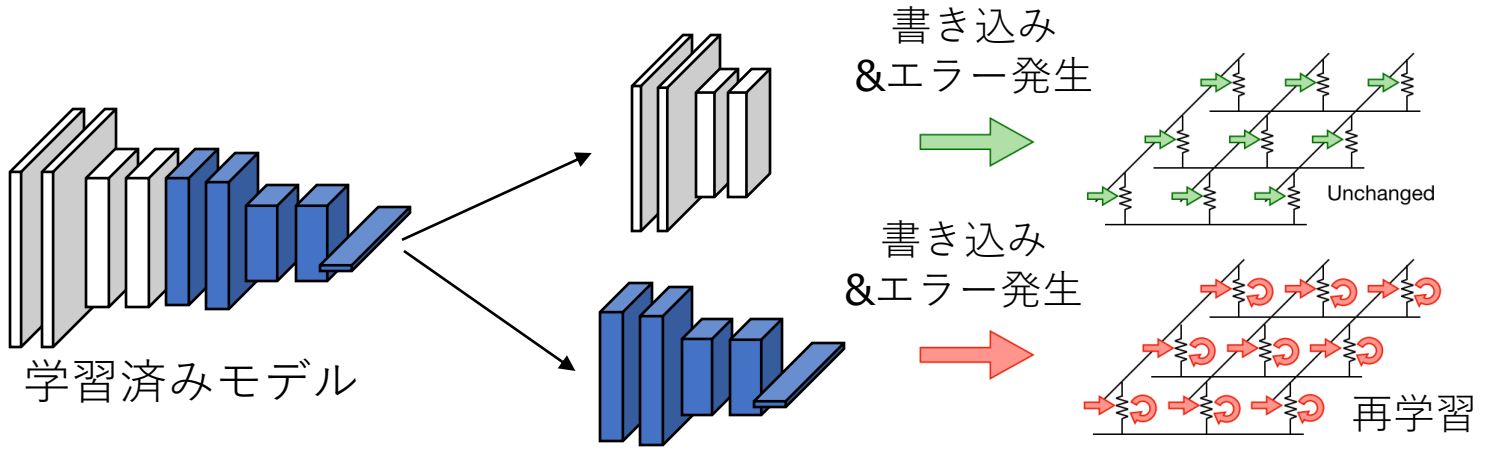


書き込みばらつき(エラー)による重み変化の様子



解決方針と先行研究: 再学習によるばらつきの補償

ばらつきによる推論精度の低下の対策（先行研究）

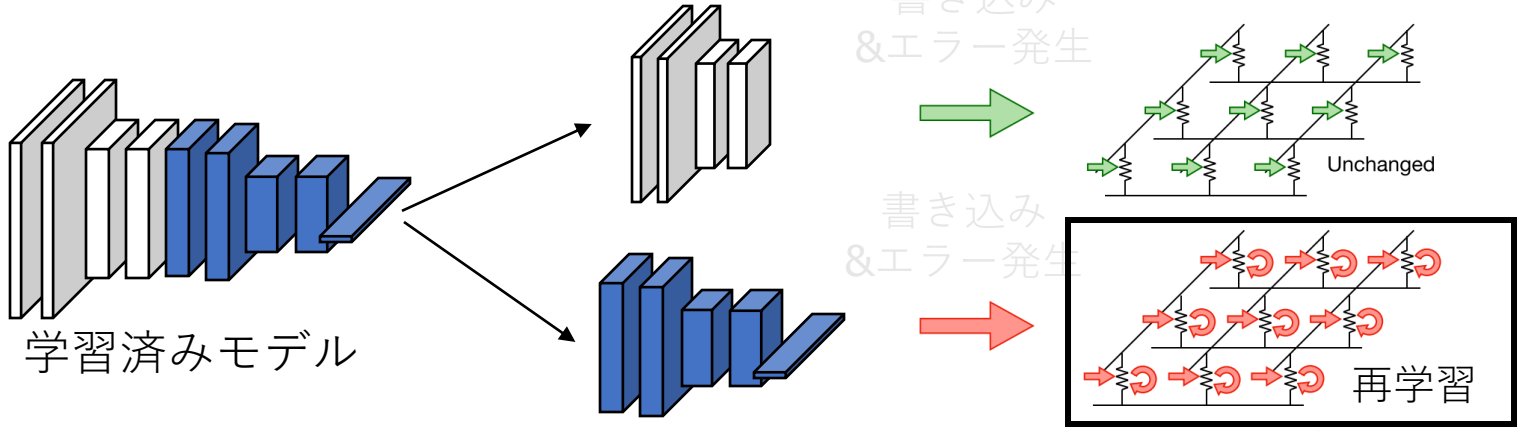


[1] P. Yao, H. Wu, B. Gao, J. Tang, Q. Zhang, W. Zhang, J. J. Yang, and H. Qian, “Fully hardware-implemented memristor convolutional neural network,” Nature, vol.577, no. 7792, pp.641–646, 2020.

書き込まれたモデルの後段の層を部分的に再学習し、ばらつきに対応する

解決方針と先行研究: 再学習によるばらつきの補償

ばらつきによる推論精度の低下の対策（先行研究）

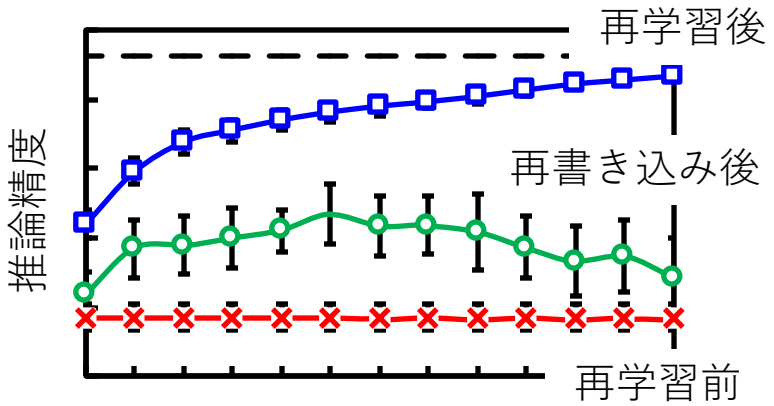


[1] P. Yao, H. Wu, B. Gao, J. Tang, Q. Zhang, W. Zhang, J. J. Yang, and H. Qian, “Fully hardware-implemented memristor convolutional neural network,” Nature, vol.577, no. 7792, pp.641–646, 2020.

書き込まれたモデルの後段の層を部分的に再学習し、ばらつきに対応する

先行研究の課題

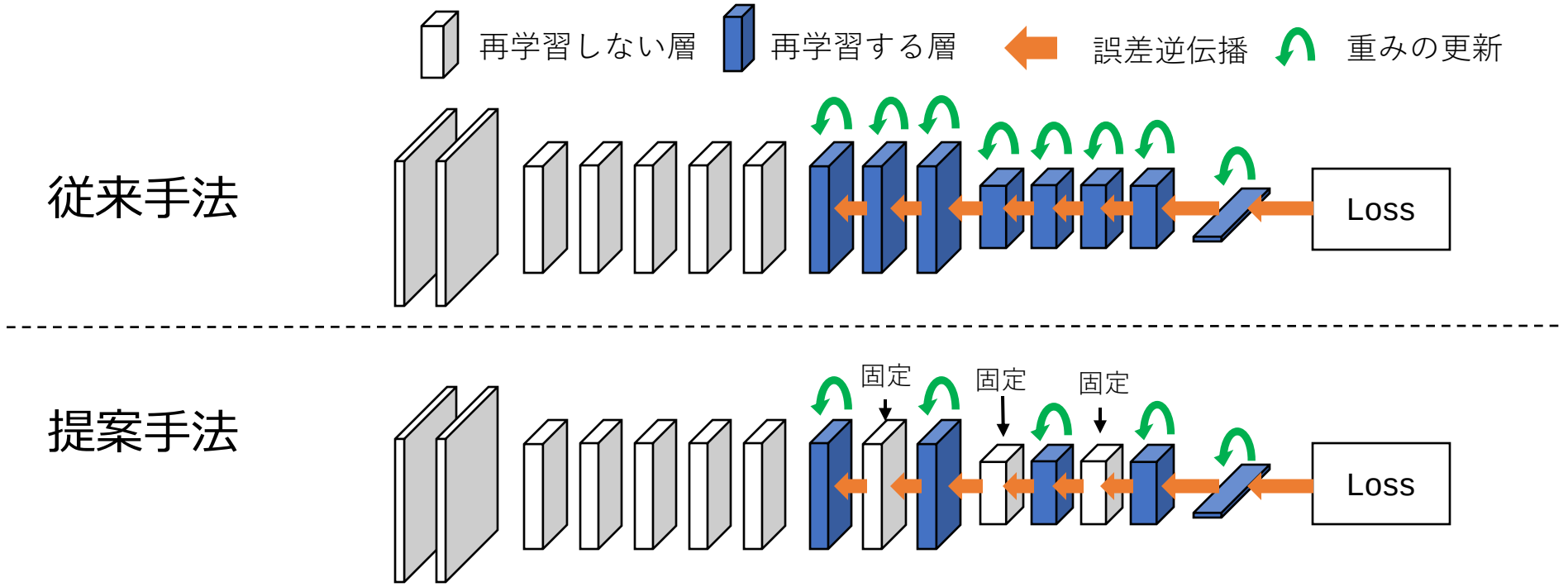
再書き込み
&エラー再発生



再書き込み時に、再び書き込みばらつきが生じる

提案手法：従来手法との比較

アイデア: 再書き込みする重みを減らして、再書き込み時の精度低下を抑制



- 従来手法: モデル後段の数層を再学習
- 提案手法: モデル後段の数層の中で、エラー耐性が高い層のみを再学習
  →再書き込み時のばらつきによる精度低下が減少

実験：Resnet-32の構造

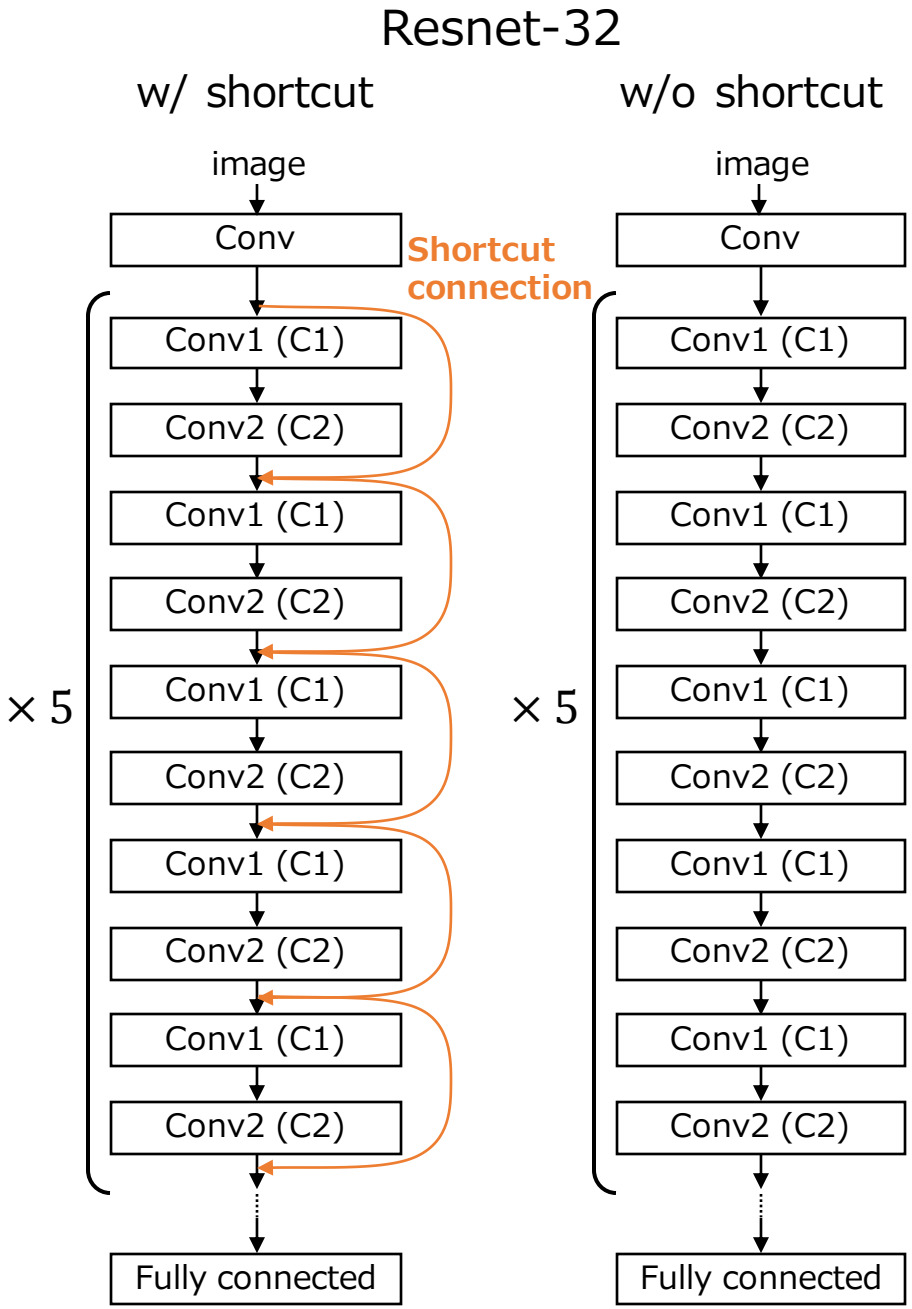
どの層を再学習するか → 実験的に求める

モデル: Resnet-32

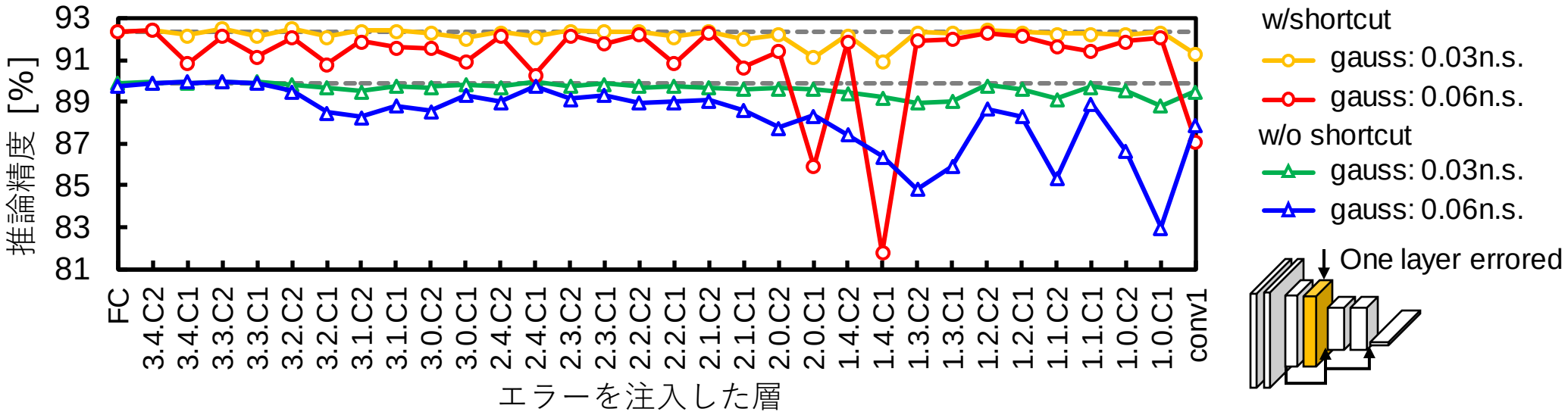
- 畳み込み層(convolution)を持つ
- shortcut connectionを持つ
- shortcut connectionの間の畳み込み層をC1, C2と呼ぶ

データセット: CIFAR-10

- 車, 鳥など10種の画像を分類するベンチマークデータセット



実験：各層のエラー耐性



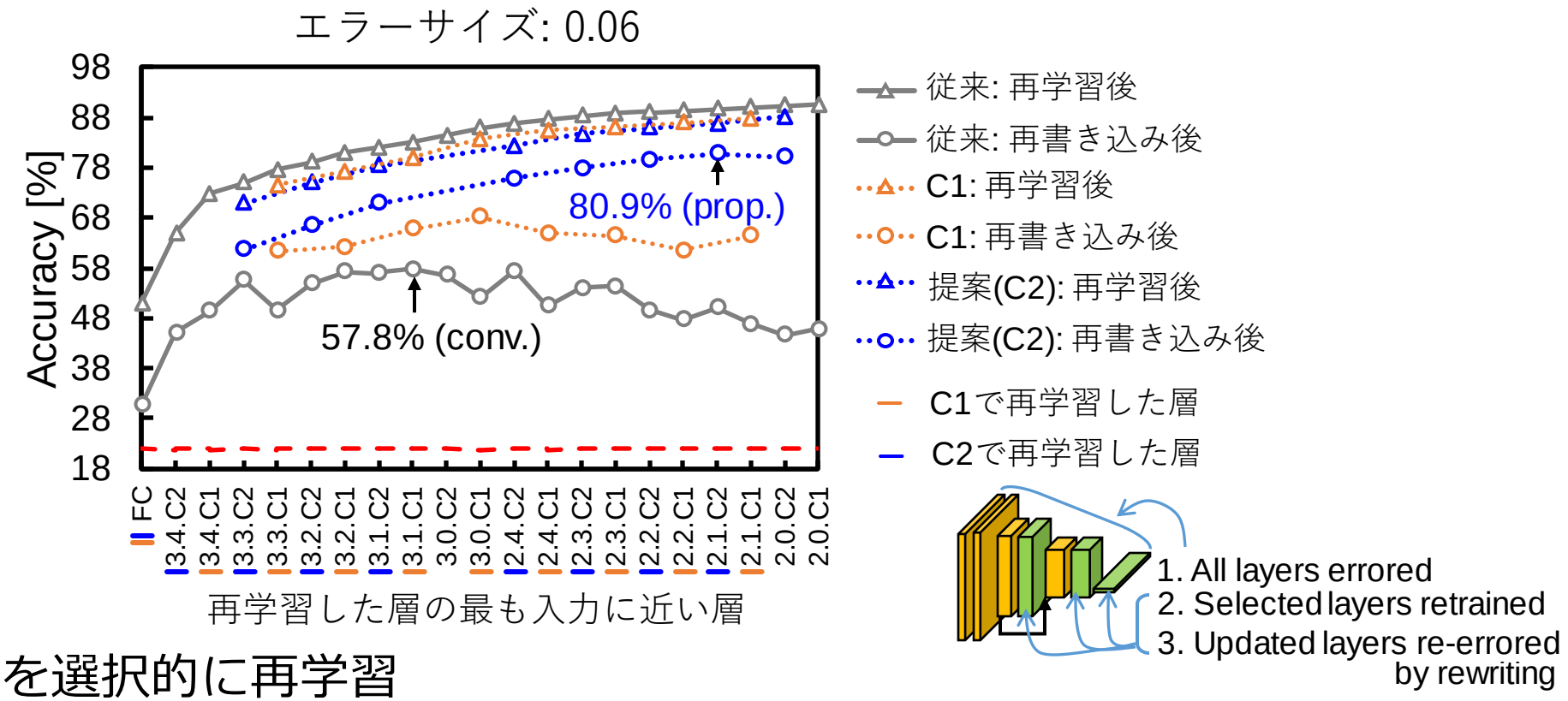
実験：1層のみにエラーを注入する実験を実施(シミュレーション)

結果：

- w/ shortcut (黄, 赤):
  - ギザギザした形状になっている
  - C1にエラーを注入した際に、精度が大きく低下
- w/o shortcut (緑, 青):
  - 一貫した特徴は見られない

→
- C1より、C2の方がエラー耐性が高い
- C1,C2のエラー耐性は、shortcut connectionに起因している

実験：提案手法による精度回復



提案手法: C2層を選択的に再学習

実験: 全層にエラーを注入し、後段の[全層, C1層, C2層]を数層再学習

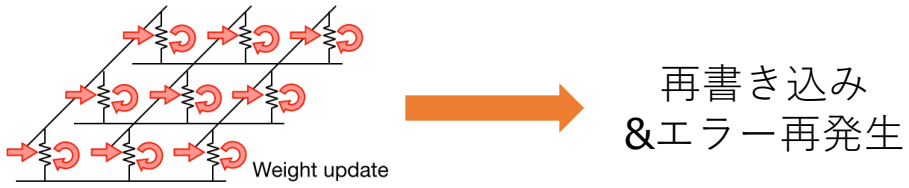
- 再学習直後の精度は従来手法(灰)の方が高い
- 再書き込み後の推論精度は提案手法(青)の方が高い
- 比較のための、C1のみの再学習(橙)は、再書き込みによる低下が大きい

実験：提案手法の実験結果

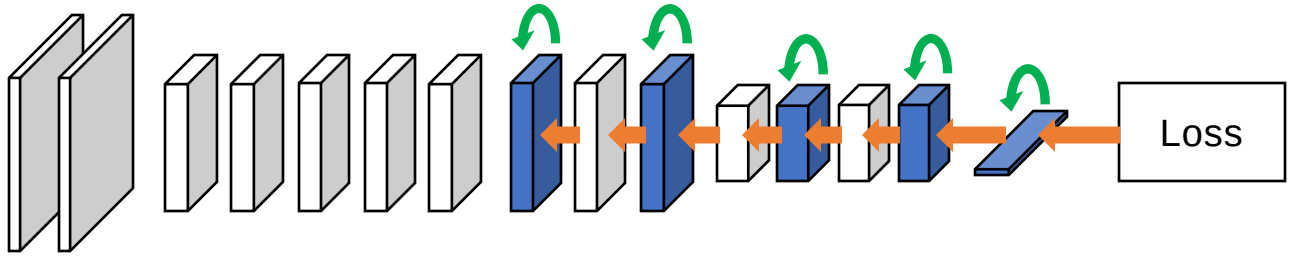
	Conv.	Prop.	Conv.	Prop.
エラーサイズ [n.s.]	0.03		0.06	
エラー注入後の推論精度 [%]	79.8		21.9	
再書き込み後の推論精度 [%]	88.7 ⇨	90.4	57.5 ⇨	80.9
誤差逆伝播する層数	14	18	9	20
再書き込みする重みの個数	380k ⇨	185k	297k ⇨	194k

- 再書き込みする重みの個数を削減　(380k→185k, 297k→194k)
  - 再書き込み後の推論精度を向上　(88.7→90.4, 57.5→80.9)

結論



- 書き込みばらつきの補正後、再書き込みの際に再び書き込みばらつきが発生



- エラー耐性の高い層のみを再学習することで、再書き込み時の精度低下を抑制

- shortcut connectionに起因するエラー耐性の差を利用