RIKEN AIP-Toshiba Collaboration Center
理研AIP-東芝連携センター

TOSHIBA

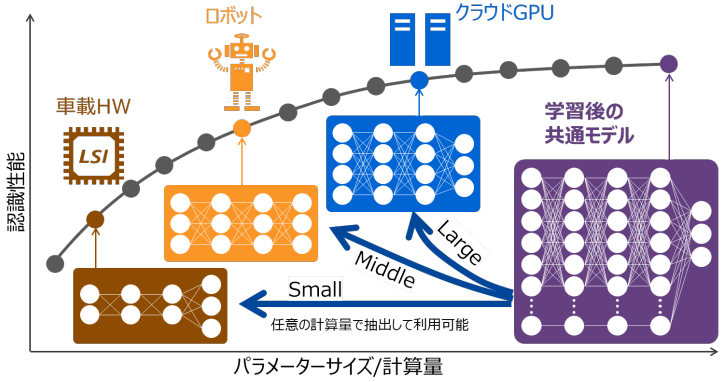
RIKEN

AIP Center for Advanced Intelligence Project

AI開発／運用の困りごとを解決する深層学習基盤技術

データ収集 → データ教示 → モデル設計 → モデル学習 → モデル提供 → モデル監視

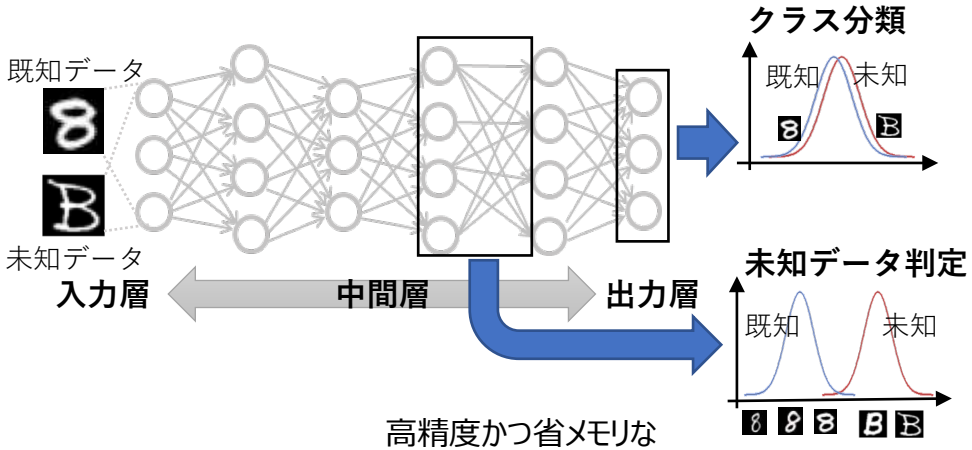
- モデル規模が大きくて組込機器に入らない → **モデル圧縮**
- 学習に時間が掛かりモデルを最適にできない → **並列分散学習**
- デバイス毎にモデルが増えると個別の高速化が困難 → **スケーラブルDNN**
- 運用するモデルが増えると性能監視や保守のコストがかさむ → **モデルメンテナンス**



学習後にモデルサイズや計算量を変更可能な

スケーラブルDNN技術

Atsushi Yaguchi et al., "Decomposable-Net: Scalable Low-Rank Compression for Neural Networks", in IJCAI2021



高精度かつ省メモリな

未知データ検知技術

上松 和樹 他, “中間層出力の射影に基づいた分布外検知による性能と軽量性の両立”, in IBIS 2021

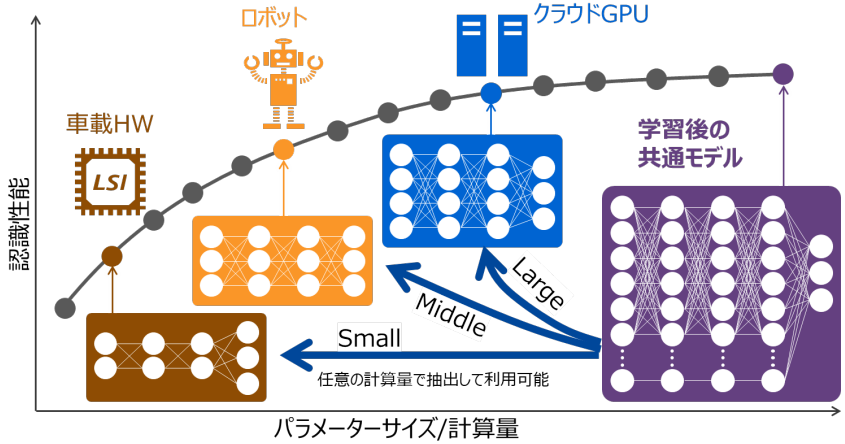
スケーラブルDNN（Decomposable-Net）

提案手法のポイント

TOSHIBA　RIKEN　AIP Center for Advanced Intelligence Project

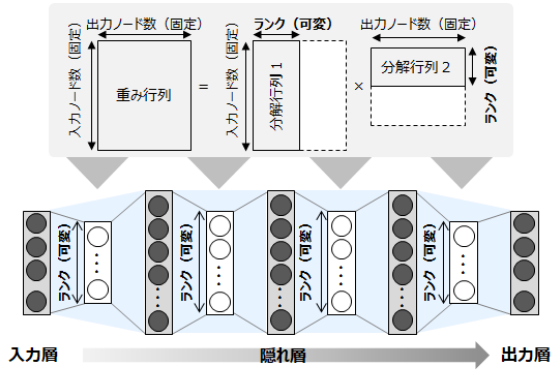
1 学習後の重み行列を分解することで計算量を学習後に変更可能

エッジ組み込みの再学習・試行錯誤が不要に

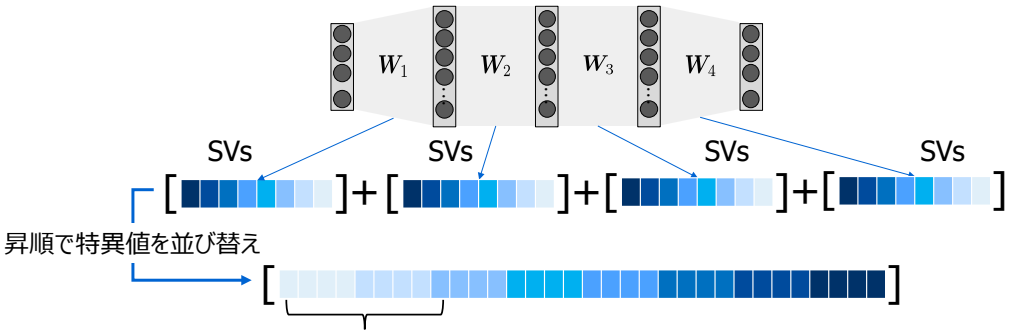


2 フルサイズとコンパクトモデルの誤差を低減する同時学習手法

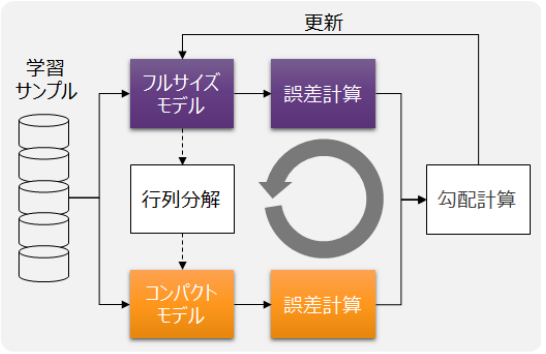
学習の工夫でコンパクトモデルでも性能低下を抑制



行列分解によるサイズの変更方法



ネットワーク全体の特異値に基づいて、各層のランク（コンパクト化度合い）を決定
→ 冗長な層を重点的にコンパクト化できるため、性能低下を抑制できる



フルサイズとコンパクトモデルの同時学習手法

スケーラブルDNN（Decomposable-Net）

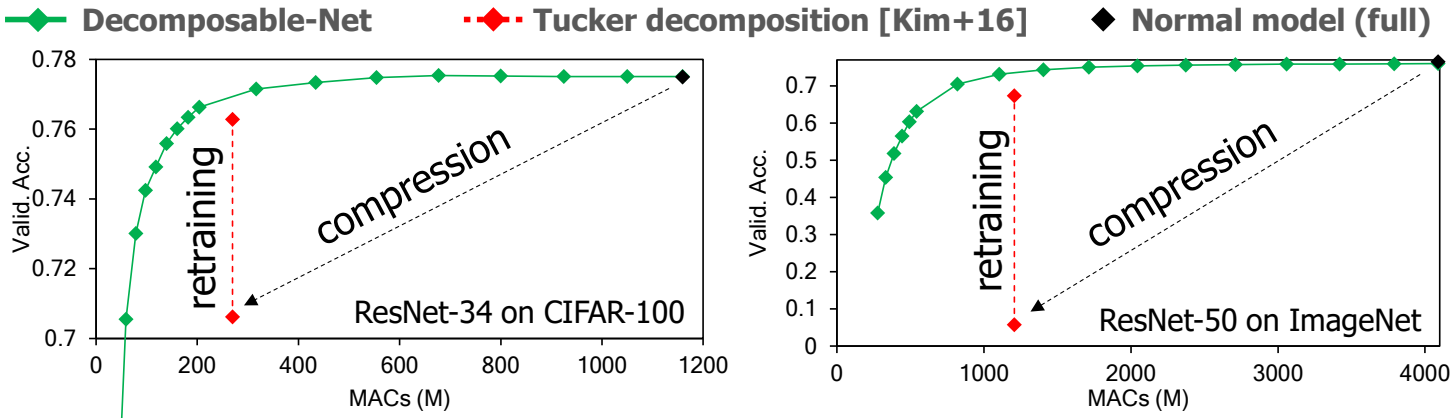
既存の低ランク圧縮との比較

TOSHIBA　RIKEN　AIP Center for Advanced Intelligence Project

Tucker decomposition：低ランク近似による圧縮＋再学習

Yong-Deok Kim et al., "Compression of Deep Convolutional Neural Networks for Fast and Low Power Mobile Applications", in ICLR2016.

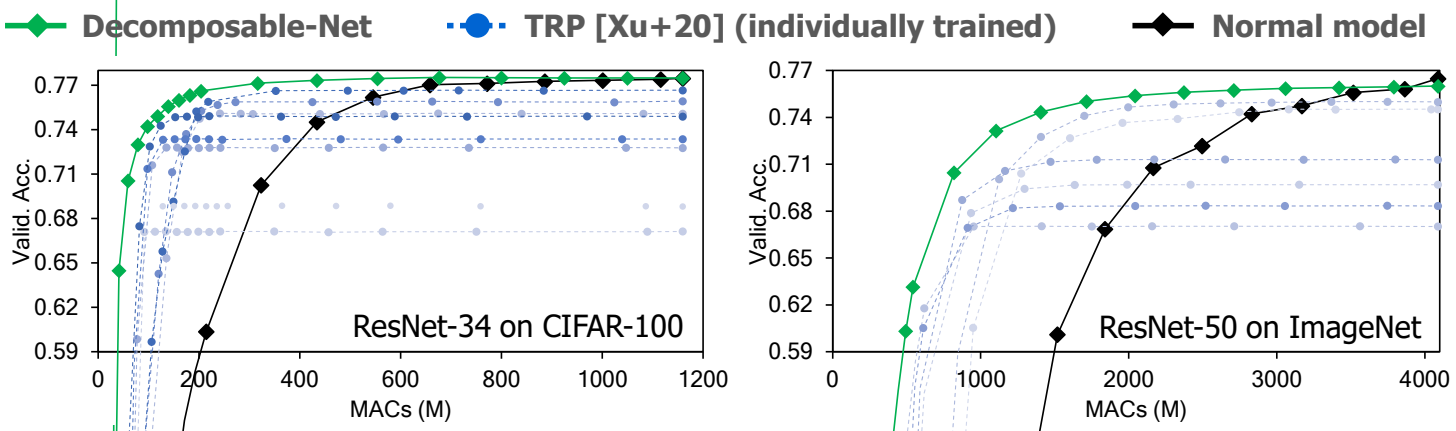
→ トレードオフのポイントで再学習が必要



TRP：低ランク近似の正則化付き学習

Yuhui Xu et al., "TRP: Trained Rank Pruning for Efficient Deep Neural Networks", in IJCAI2020.

→ トレードオフのポイントが正則化強度で異なるため複数のモデル学習が必要



スケーラブルDNN（Decomposable-Net）

既存のスケーラブル技術との比較

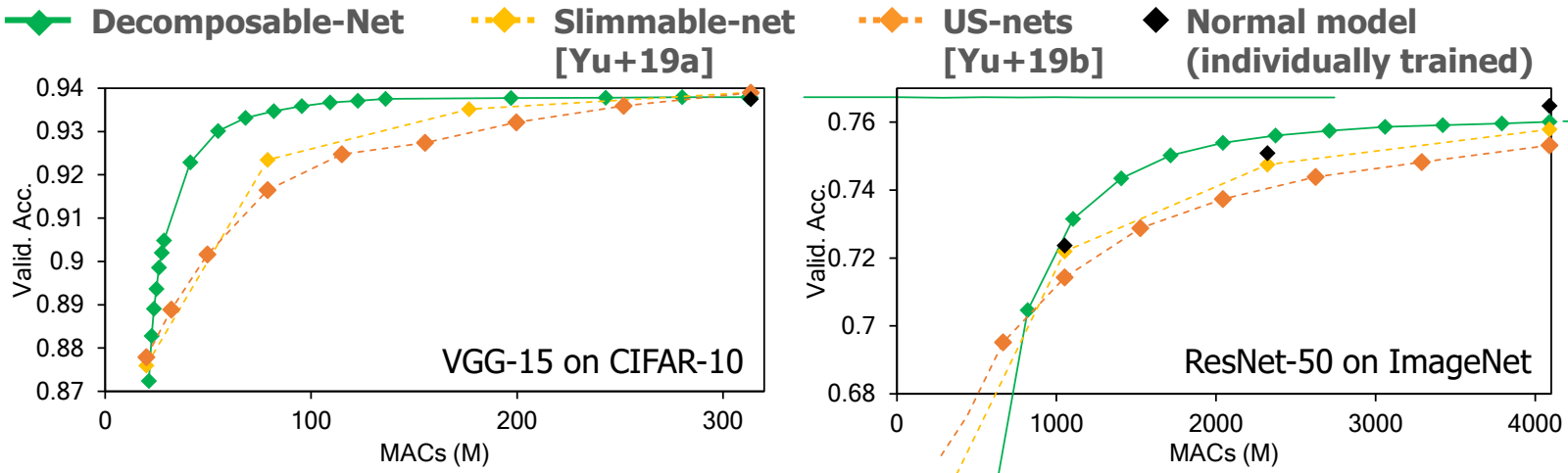
TOSHIBA　RIKEN　AIP Center for Advanced Intelligence Project

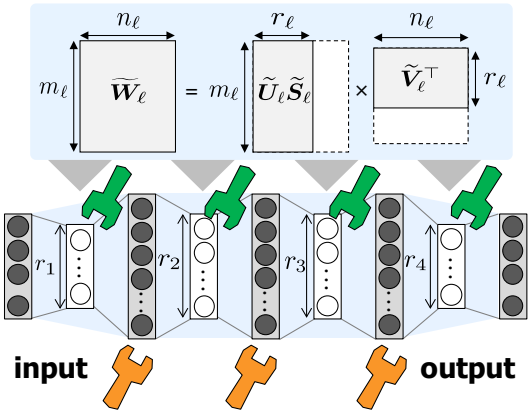
Slimmable-net, US-nets：各層のチャネル数割合を一律で削減して計算量を変更

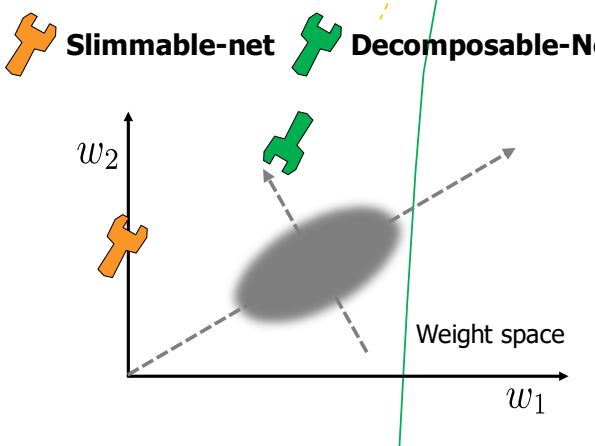
Jiahui Yu et al., “Slimmable Neural Networks,”, in ICLR 2019. Jiahui Yu et al., “University Slimmable Networks,” in ICCV2019.

→ チャネル数の直接削減は、提案の固有空間上の削減よりも近似誤差が増加しやすい

→ 各層一律割合の削減は、提案の冗長な層の重点的な圧縮よりも近似誤差が増加しやすい







← DNNのパラメータ集合（イメージ図）と提案手法（緑）と既存手法（オレンジ）のパラメータ圧縮軸の違い

モデルメンテナンス（未知データ検知）

提案手法のポイント

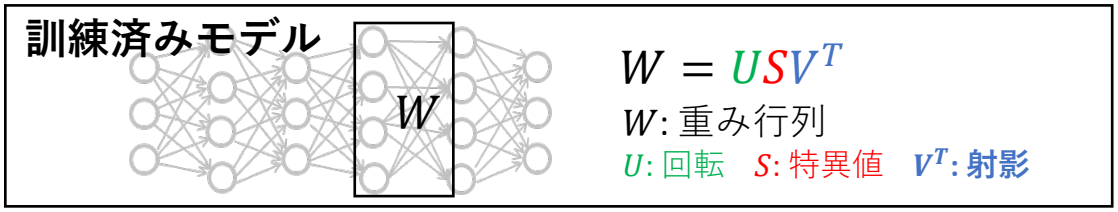
TOSHIBA　RIKEN　AIP Center for Advanced Intelligence Project

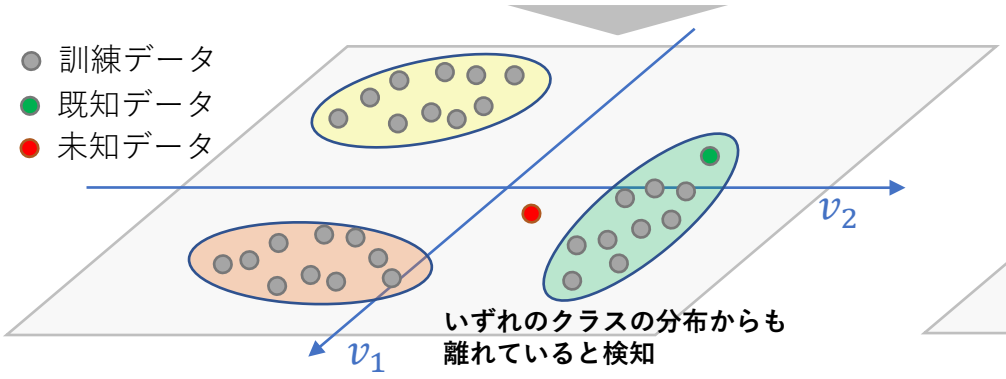
1 学習済みモデルの中間特徴量を使う再学習が不要な検知手法

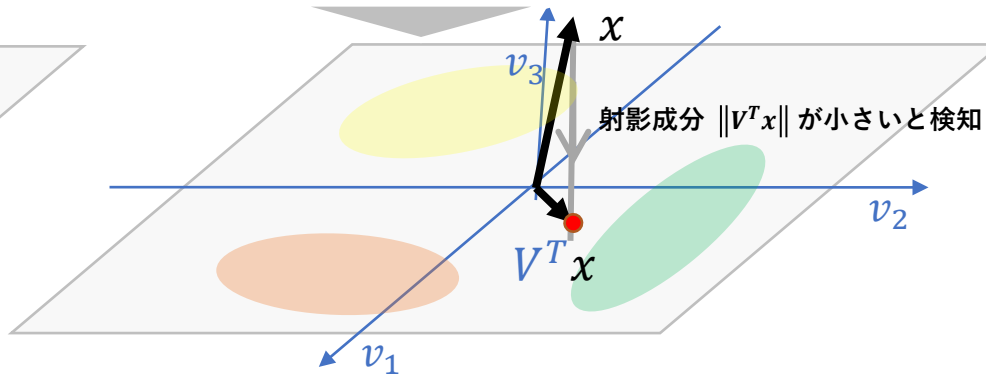
未知データの検知を想定したモデルの再調整が不要

2 既知データの分布の保存が不要で省メモリな検知手法

計算資源の限られるエッジデバイスなどにも有効







マハラノビス距離（既存手法）

- 訓練データとの距離から未知データを検出
- マハラノビス距離でデータ分布の形状を反映
- 分布の形状の保存で必要な追加メモリが多い

射影成分（開発手法）

- 中間層射影成分の大小で未知データを検出
- 特徴量が伝搬しないデータを未知と判定
- 分布の形状の保存が不要で省メモリ

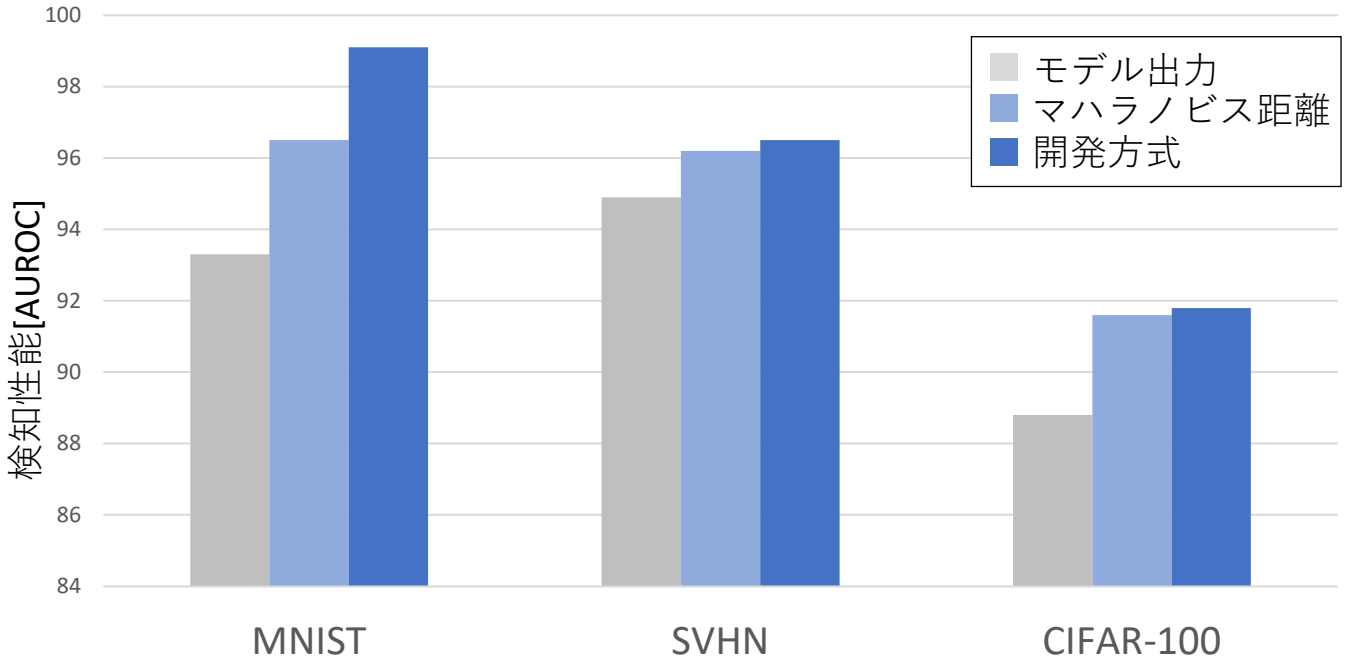
モデルメンテナンス（未知データ検知）

公開データセットでの評価結果

TOSHIBA　RIKEN　AIP Center for Advanced Intelligence Project

評価方法

既知データとして CIFAR-10 で学習を行い
未知データとして MNIST, SVHN, CIFAR-100 の検出性能を比較



追加必要メモリ

モデル出力	マハラノビス距離	開発方式
0.00GB	3.50GB	0.05GB

省メモリで高い検知性能を実現

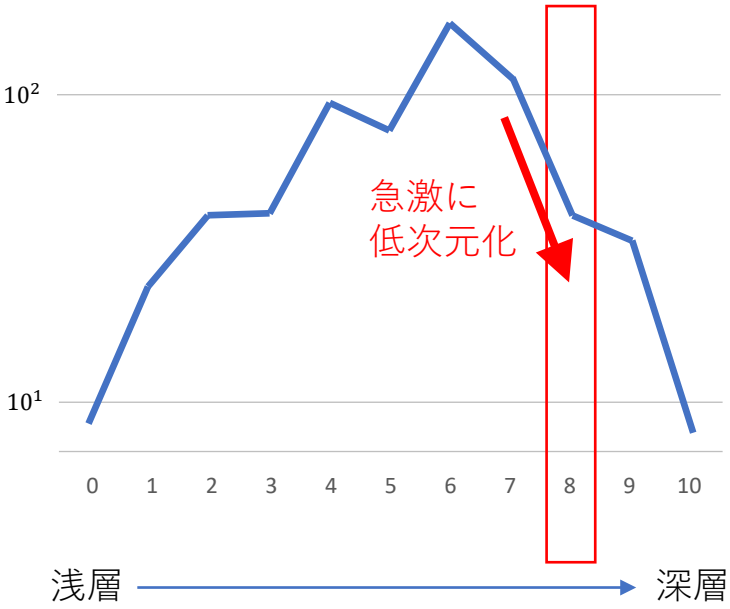
検知に最適な層の選択

TOSHIBA RIKEN AIP Center for Advanced Intelligence Project

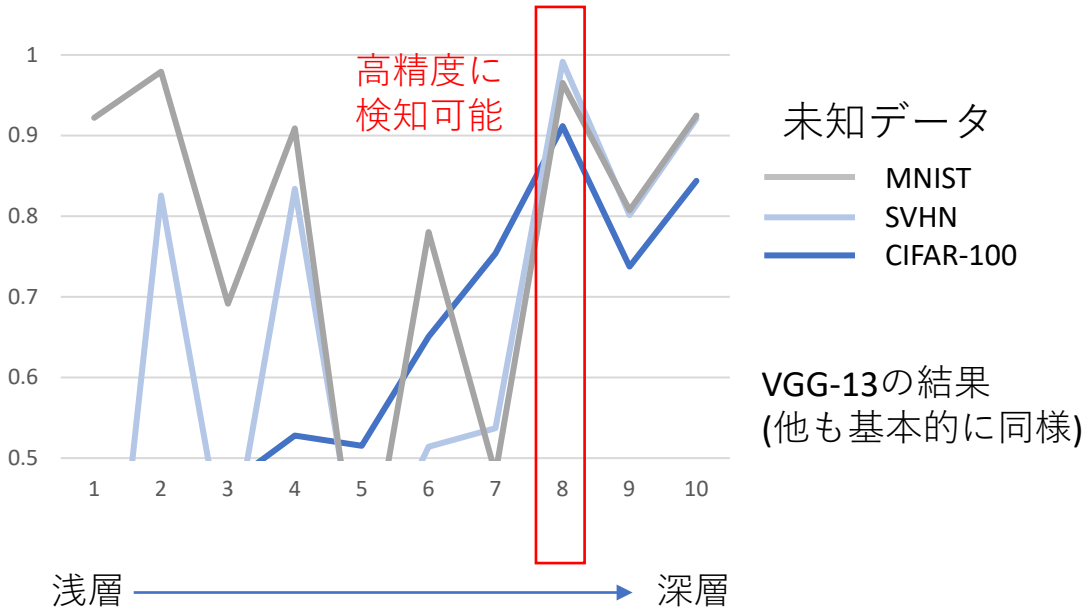
訓練済みモデルの各層で安定ランクと検知性能を比較

→ 安定ランクは層を経るにしたがって増加し、ある層で急激に減少
→ どのデータセットでも、安定ランクが急激に落ちる層の検知性能が最大

安定ランク ($\|W\|_F^2 / \|W\|_2^2$)



検知性能 (AUROC)



VGG-13の結果
(他も基本的に同様)

低次元化が起こる層で未知データ検知性能が最大