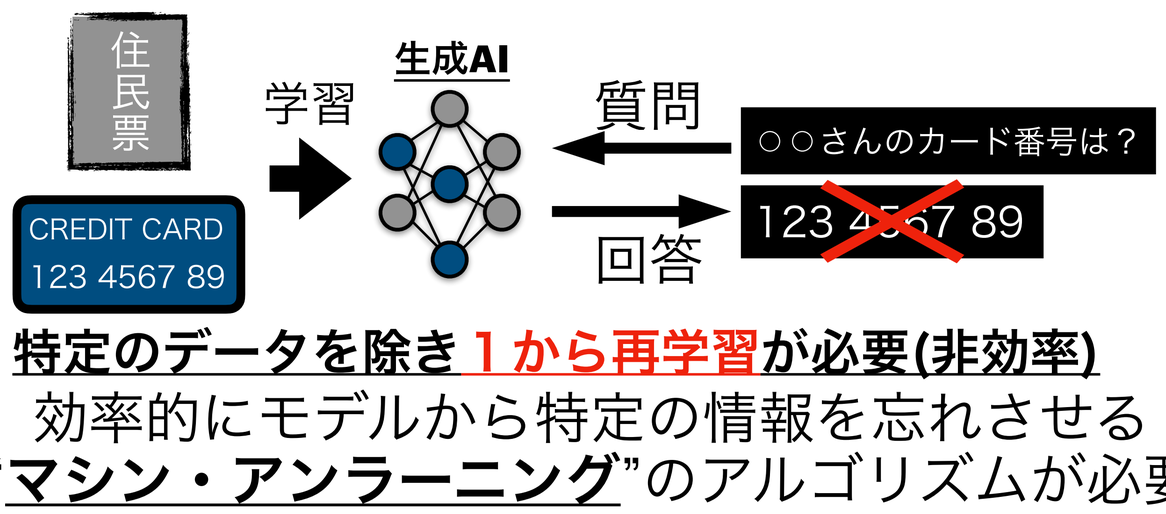
選択的破壊的忘却に基づく マシン・アンラーニングの高速化

村上泰斗，柴田大真，山内悠嗣(中部大学)

研究背景・目的

対話型AIや画像生成AIが広く普及
学習に個人情報が混入するとプライバシー侵害の問題



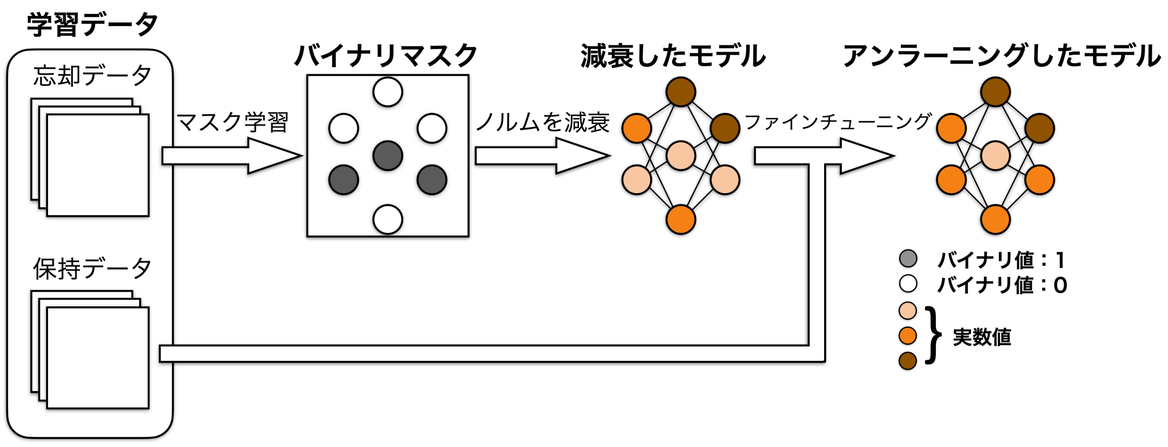
特定のデータを除き１から再学習が必要(非効率)
効率的にモデルから特定の情報を忘れさせる
“**マシン・アンラーニング**”のアルゴリズムが必要

提案手法

変更が必要な重みだけを選択し，ノルムを減衰

提案手法の流れ

1. 忘却データに寄与する重みを特定(バイナリマスク学習)
2. 特定した重みのノルムを減衰
3. ファインチューニング

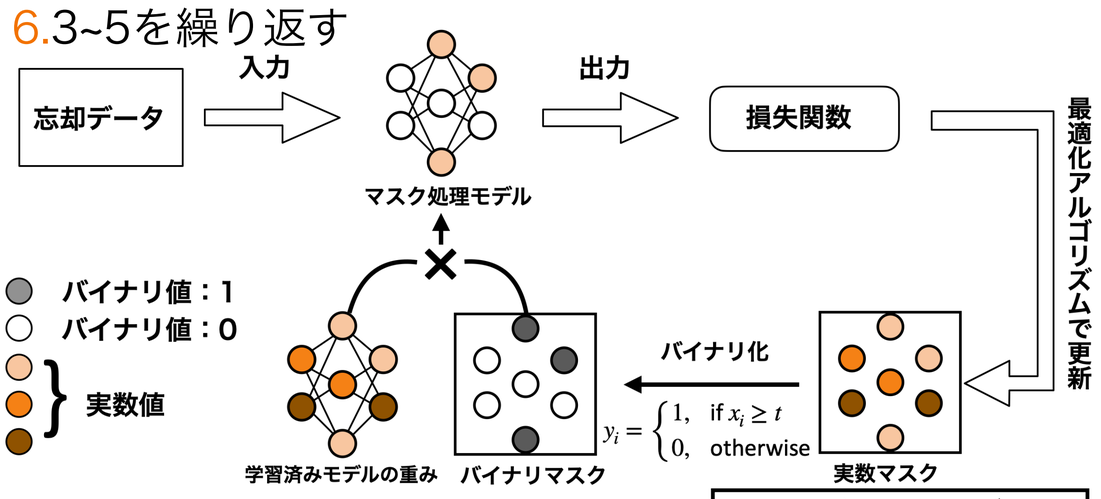


忘却データに寄与する重みの特定方法

Piggyback[1]アルゴリズムに基づくマスク学習

バイナリマスク学習の流れ

1.学習済みモデルの重みを固定
2.1つ1つの重みに対応した実数マスクを初期化
3.実数マスクの値を閾値でバイナリ化(0 or 1)
4.バイナリマスクを学習済みモデルに掛け合わせる
5.マスク処理モデルで順，逆伝播(実数マスクを更新)
6.3~5を繰り返す



L1正則化の導入

$$\mathscr{L}_{\text{cross}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c}\log \hat{y}_{i,c}$$

$$\mathscr{L}_{\text{mask}} = \mathscr{L}_{\text{cross}} + \lambda\|\boldsymbol{\theta}\|_1$$

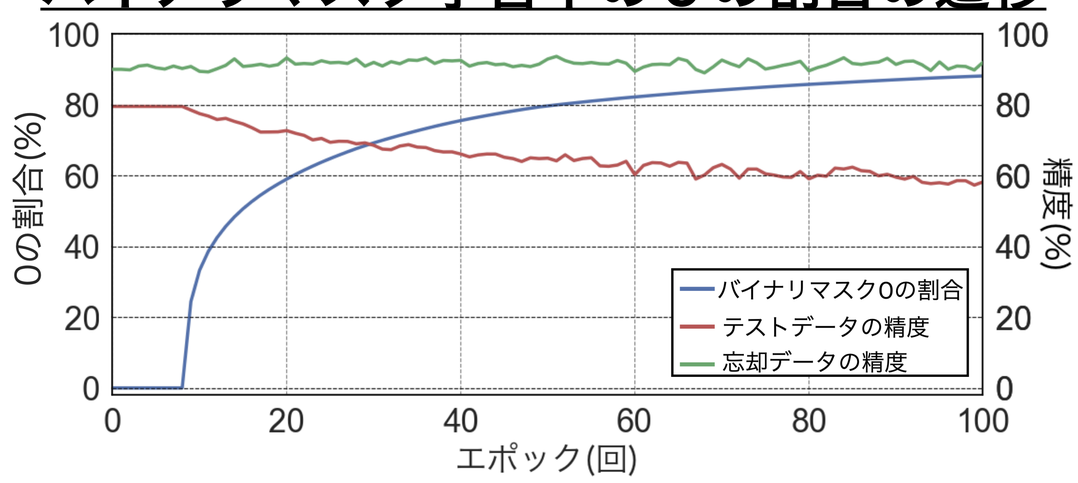
N：バッチサイズ
C：クラス数
$y_{i,c}$：正解ラベル
$\hat{y}_{i,c}$：モデルの予測確率
$\|\boldsymbol{\theta}\|_1$：L1ノルム
λ：正則化の強さ

忘却データの認識精度を向上 ＋ 有効な重みを減少

忘却データに寄与する重みだけが有効に

[1] A.Mallya et al.: “Piggyback: Adapting a single network to multiple tasks by learning to mask weights”, ECCV, 2018.

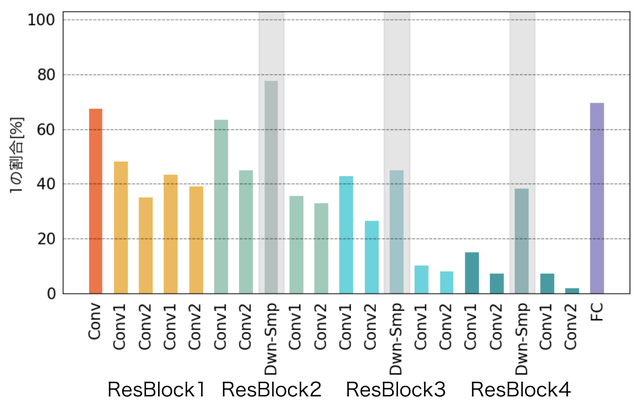
バイナリマスク学習中の０の割合の遷移



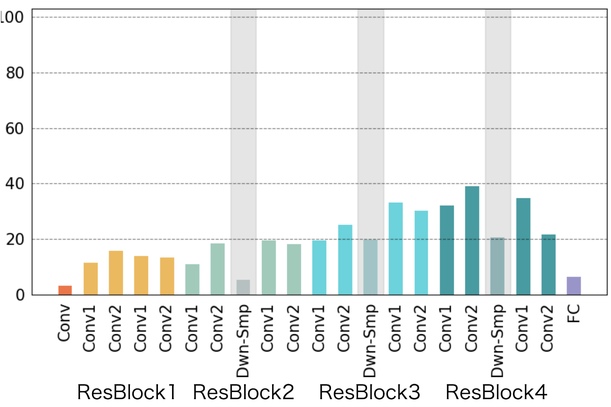
操作する重みを限定

最後のlayer(２つの残差ブロック)のみを減衰の対象
入力層に近い層の重みはデータに共通して必要

忘却マスクの１の割合



忘却マスクのみが１の割合



特定した重みのノルムを減衰

$$\mathscr{W}_{\text{new}} = \mathscr{W}_{\text{pre}} \odot \eta$$

$\mathscr{W}_{\text{new}}$：更新後の重み
$\mathscr{W}_{\text{pre}}$：モデルの重み
η：減衰係数$(0<\eta<1)$

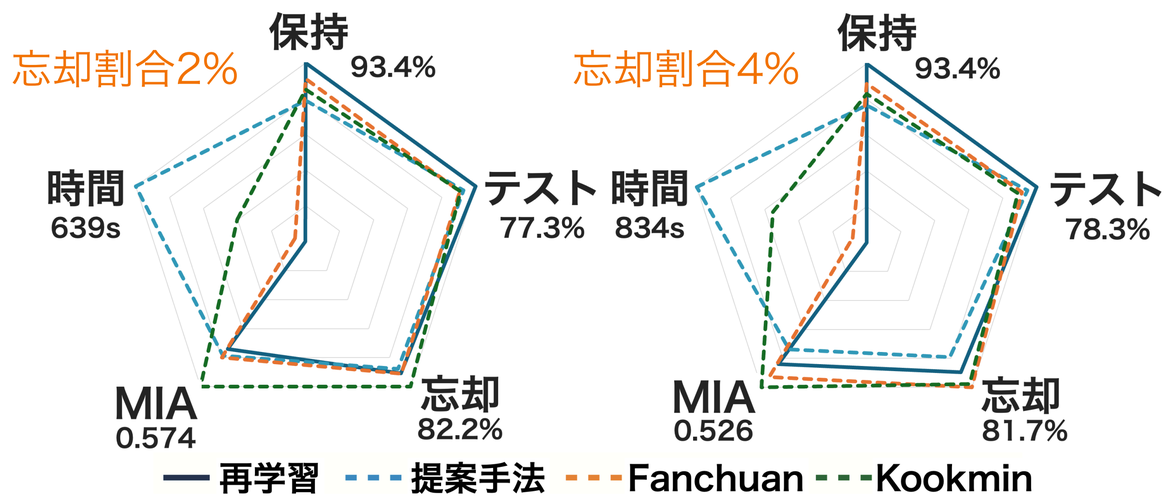
初期化ではなく減衰することで過剰なモデル破壊を防ぐ

実験概要

- **データセット(Imagenet-100)**
  - 訓練データの2%，4%を忘却データに(ランダムに選択)
  - 残りのデータを保持データに(98%，96%)
- **評価指標**
  - テストデータ，忘却データの認識精度
  - Membership Inference Attack(MIA)
  - 実行時間
- **比較手法**
  - Fanchuan(kaggle1位)
  - Kookmin(kaggle2位)
- **実験環境**
  - CPU：Intel Corei7-14700KF
  - GPU：Nvidia GeForce RTX4090

実験結果

- **テストデータ，忘却データの認識精度とMIA**
  - 既存の手法と**同等か僅かに上回る**精度
- **実行時間**
  - 他の手法より**遥かに高速**なアンラーニングが可能



- **今後の展望：大規模モデルに対しての性能について検証**