F-012

複数データセットを用いたマシンアンラーニング手法の性能比較に関する実験的検討
An experimental study to compare the performance of machine unlearning methods with multiple datasets

柴田 大真† 村上 泰斗† 山内 悠嗣 †
Hiromasa Shibata Taito Murakami Yuji Yamauchi

1 はじめに

近年，機械学習分野は飛躍的な発展を遂げており，さまざまな分野において応用が進められている．一方で，モデルの学習に用いるデータセットに有害なデータやプライバシーに関わる情報が含まれていた場合，ユーザーの権利を侵害する深刻な問題を引き起こす可能性がある．この問題への対処法として，問題となる特定のデータを除外し，残りのデータのみを用いてモデルを再学習する方法が考えられるが，再学習には多大な計算コストが必要であり，現実的な運用が難しいケースも多い．

効率的に特定のデータの影響をモデルから取り除くことを目的としたマシンアンラーニングが注目されている．マシンアンラーニングは，有害または削除要求のあったデータの情報を既存のモデルから部分的に忘却させる技術である．マシンアンラーニング分野では，実用化に向けた課題として確立された手法がないことが挙げられる．そこで，コンペティションプラットフォームである Kaggle にて，新たな手法の作成の促進を目的としたコンペティションが開催され，コンペティションの上位手法は既存の研究の精度を上回る結果が報告された [4]．しかし，このコンペティションでは，異なるデータセットや忘却割合を変更した際の手法の汎用性，実行時間について評価されていない．

そこで本研究では，コンペティションの上位手法を対象に，その性能を比較・分析するための実験的な評価を行う．これにより，各手法の特徴や汎用性を明らかにする．

2 関連研究

マシンアンラーニングが提唱されて以降，様々な手法が提案されている．Cao et al.[1] はマシンアンラーニングを提唱した論文にて，忘却データに対する影響を学習済みモデルの重みから引くことでアンラーニングする手法を提案した．この手法は精度を維持したまま，再学習に要する時間を大幅に短縮したが，アンラーニングできるモデルが限定されるという問題点もある．Trippa et al.[2] は，忘却機構を追加した損失関数を用いてアンラーニングする手法を提案した．この手法は再学習よりも早く実行ができるが，忘却データの量が増えることにより忘却性能が悪化していく問題点を持つ．また，提案されている手法の多くは，保持データで再学習したモデルに近似するように設計されているが，このような手法では，論理的にアンラーニングが確実に実行されたと保証することができない．Chourasia et al.[3] は，モデルのパラメータや学習過程に削除対象データの痕跡が残存している点から，再学習との識別不能性で削除保証を定義する方法は根本的に誤っていると主張している．このような背景を踏まえると，評価指標を統一すると同時に，汎用性と削除保証の信頼性を両立した新たな手法の提案が求められている．

3 NeurIPS 2023 - Machine Unlearning

優れたアンラーニング手法の作成を促進すること，論理的な評価手法を提案することを目的としたコンペティション NeurIPS 2023 - Machine Unlearning*が 2023 年に Kaggle にて開催された．コンペティションには 1,121 チームから合計 1,923 件の投稿があった．独自の評価指標に基づいて高得点を獲得した上位 7 チームが表彰された．

3.1 コンペティションの問題背景

このコンペティションでは，顔画像から年齢を予測するモデルを構築した後，学習データに用いたユーザーの一部が「自分のデータを削除してほしい」と要求するという問題設定が与えられた．この設定は，現実世界でユーザーが個人情報の削除を要求する状況を模している．年齢予測モデルには ResNet-18[5] が使用され，データセットには CASIA-SURF-live[6] が採用された．CASIA-SURF-live は，16 歳から 85 歳までの 1,000 人分の顔画像を含むデータセットであり，個人 ID や年齢のラベルが付与されている．このコンペティションでは，年齢順にラベルを 0 から 9 の 10 クラスに分割し，年齢予測を多クラス分類問題として扱った．また，忘却対象データとしてランダムに 15 人分のデータを選定し，選定されたユーザーに関する情報をモデルからアンラーニングすることが求められた．

3.2 コンペティションの評価手法

コンペティションにおけるマシンアンラーニングはデータセット D を用いて学習したモデル f_D から特定の画像 $D_f (D_f \subseteq D)$ の影響を取り除くものであり，データセット D から特定の画像 D_f を取り除いたデータ $D_{r=(D-D_f)}$ を用いて再学習したモデル f_r に近似させることを目的とする．そのため，アンラーニング後のモデルの理想は f_r となる．本論文では f_D をターゲットモデル，D_f を忘却データ，D_r を保持データ，f_r を再学習モデルと呼ぶ．この定義に則り，コンペティションでは差分プライバシー [7] の数理的考えを応用した指標を用いて，マシンアンラーニングの性能を定量的に評価する手法を導入した．

3.2.1 アンラーニングの保証

コンペティションでは，アンラーニングの保証について式 (1) のように定式化した．

$$\Pr[f_r \in R] \leq e^{\varepsilon} \cdot \Pr\left[U(f_D, D_f, D) \in R\right] + \delta \quad (1)$$

ここで，R は任意のタスク，ε は出力分布の違いの大きさを制御するパラメータ，δ はこの保証が破られる例外的な確率の上限を表す．式 (1) は，図 1 に示すように

† 中部大学 Chubu University

* https://www.kaggle.com/competitions/neurips-2023-machine-unlearning/overview

f_r と $U(f_D, D_f, D)$ の異なるアルゴリズムの出力分布を比較し，両手法の出力分布が十分に類似していれば，差分プライバシーの数理的考えに則り忘却データ D_f を忘れられたとみなす．また，ε が小さいほどアンラーニングが行われていることが保証され，δ は通常は非常に小さな値に設定される．

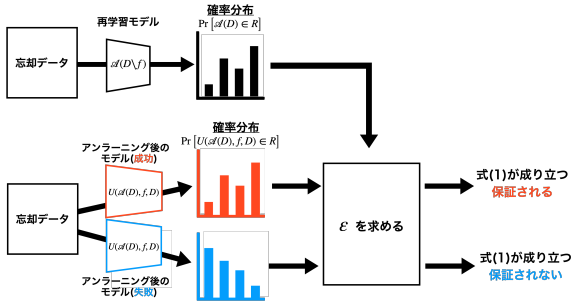

図 1: 差分プライバシーの数理的考えに沿ったアンラーニング保証の流れ

3.2.2 仮説検定

Kairouz et al.[8] は，アンラーニングの評価を仮説検定の枠組みで定式化する手法を提案している．この枠組みでは，データセット D を用いて学習されたモデル f_D が，忘却データセット D_f に対してアンラーニングが適用されたモデルを帰無仮説とし，保持データ D_r で再学習されたモデル f_r を対立仮説とする．識別器が，サンプル X の出力分布の生成元が $U(f_D, D_f, D)$ か f_r かを判別する性能は，第一種過誤（FPR : False Positive Rate）および第二種過誤（FNR : False Negative Rate）によって，式 (2) のように表すことができる．

$$\mathrm{FPR} + e^{\varepsilon} \cdot \mathrm{FNR} \geq 1 - \delta, \quad \mathrm{FNR} + e^{\varepsilon} \cdot \mathrm{FPR} \geq 1 - \delta \quad (2)$$

式 (2) は，FPR と FNR は統計的にトレードオフの関係にあることを示している．この性質を利用すると，任意の固定された δ において式 (3) により，$\hat{\varepsilon}$ を推定することが可能となる．

$$\hat{\varepsilon} = \frac{1}{S} \sum_{n=1}^{S} \max \left\{ \log \left(\frac{1-\delta-\hat{\mathrm{FNR}}}{\hat{\mathrm{FPR}}} \right), \log \left(\frac{1-\delta-\hat{\mathrm{FPR}}}{\hat{\mathrm{FNR}}} \right) \right\} \quad (3)$$

この検定では，多数 ([4] では 512 個) の再学習モデルとアンラーニング後のモデルを用いて，仮説検定を行う．検定の結果から $\hat{\mathrm{FNR}}$，$\hat{\mathrm{FPR}}$ の値が最も大きくなるように閾値を設定し，その $\hat{\mathrm{FNR}}$ と $\hat{\mathrm{FPR}}$ を式 (3) に入力し，サンプル数 S の平均 $\hat{\varepsilon}$ を算出する．前節で述べたように，ε が小さいほどアンラーニングの保証がより確定的なものになる．そのため，式 (3) で求める $\hat{\varepsilon}$ が小さいほど忘却性能が高いと言える．

3.2.3 評価方法

コンペティションでは $\hat{\varepsilon}$ に加え，再学習モデルとアンラーニング後のモデルの保持データにおける精度，テストデータにおける精度を用いた式 (4) を用いて評価した．アンラーニングの保証度合いに加え，保持データやテストデータでの精度も考慮することで，忘却性能とモデルの実用性能のバランスを総合的に評価する指標となっている．

$$\text{Final score} = F \times \frac{\mathrm{Acc}(D \setminus f, \theta_u)}{\mathrm{Acc}(D \setminus f, \theta_r)} \times \frac{\mathrm{Acc}(D_t, \theta_u)}{\mathrm{Acc}(D_t, \theta_r)} \quad (4)$$

ここで，F は式 (3) で求めた $\hat{\varepsilon}$ の値に応じて段階的にスコアリングした値であり，$\hat{\varepsilon}$ が小さいほど高い得点を与える．階層化されたスコアリング基準を設定することで，アンラーニング手法の相対的な性能評価が可能となっている．

3.3 上位手法

本節では，優れた成績を収めた上位 5 手法の概要と特徴を解説する．

3.3.1 1 位手法

本手法は，2 段階の学習で構成されており，忘却データの出力分布を一様化することで情報を消去し，さらに対照学習を用いて特徴空間における分離を強化する二段階構成が特徴である．

1 段階目では，忘却データ $x \in \mathcal{D}_f$ に対し，式 (5) で表す KL ダイバージェンス損失関数を用いて，モデル出力分布 $p(x)$ を一様分布 u に近づけるように学習する．

$$\mathcal{L}_{\mathrm{KL}} = \mathrm{KL}(p(x) \,\|\, u) \quad (5)$$

これにより，ロジット空間における情報消去を促進する．2 段階目では，対照学習に基づくファインチューニングを行い，忘却データと保持データの特徴空間における乖離を強化する．具体的には，忘却データ x とその拡張 x' を正例とし，保持データ $r \in \mathcal{D}_r$ を負例とした式 (6) を用いて学習する．また，τ は温度系数である．

$$\mathcal{L}_{\mathrm{CL}} = -\log \frac{\exp\left(f(x) \cdot f(x')/\tau\right)}{\sum_{r \in \mathcal{D}_r} \exp\left(f(x) \cdot f(r)/\tau\right)} \quad (6)$$

最後に，保持データに対して通常のクロスエントロピー損失によりファインチューニングを行い，性能を補完することで忘却データの忘却とモデルの精度保持の両立を図る．

3.3.2 2 位手法

本手法は，勾配に基づく重みの初期化によって忘却を促進する手法である．保持データと忘却データに対する勾配が類似する場合，保持データでの再学習時に忘却が困難になるという仮定に基づき，両データの勾配が近いモデルの重みを初期化することで忘却を促進する．忘却データにはクロスエントロピー損失による勾配上昇，保持データには通常の勾配降下を適用し，それぞれから勾配を収集する．その際，忘却データと保持データでサンプル数が異なるため，保持データからは忘却データと同数のサンプルをランダムに抽出して勾配を比較する．勾配の絶対値が小さい 30 %の畳み込み層の重みを初期化し，保持データでファインチューニングすることでモデルの精度を回復させる．

3.3.3 3 位手法

本手法は，ターゲットモデルの畳み込み層の重みをランダムに摂動させることで，初期化を行う手法である．まず，重みの平均と，分散を 0.6 とした正規乱数を用いて，入力層に最も近い畳み込み層の重みを初期化する．その後，保持データでファインチューニングする．最終エポック前には局所最適解回避のための軽微な摂動も加える．またクラスの不均衡に対応するため，重み付きクロスエントロピー損失を用いる．

3.3.4 4 位手法

本手法では，まずターゲットモデルを L1 ノルムに基づいて 99 %初期化する．これにより忘却データの情報を除去する．その後，保持データのみを用いてファインチューニングを行う．その際にエントロピーの類似性を保つ正則化項を追加した式 (7) で学習する．

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CE}} + \mathcal{L}_{\text{RegEntropy}} \quad (7)$$

ここで，$\mathcal{L}_{\text{CE}}$ は保持データに対するクロスエントロピー損失，$\mathcal{L}_{\text{RegEntropy}} = \text{MSE}(H_{\text{orig}}, H_{\text{ft}})$ はターゲットモデルとファインチューニング後のモデルの予測エントロピーの平均二乗誤差である．これにより，忘却と性能保持のバランスをとっている．

3.3.5 5 位手法

本手法では，2 つのアプローチが提案された．1 つ目のアプローチでは，ターゲットモデルの畳み込み層に対し空間的次元の転置を適用し，その後，保持データを用いてファインチューニングする．これは，ターゲットモデルの特徴表現を部分的に維持しつつ，空間的情報を意図的に破壊することで忘却データへの依存性を低下させ，学習を安定させる狙いがある．

2 つ目のアプローチでは，ターゲットモデルと簡易的にアンラーニングされたモデルの予測結果を比較することで，忘却データ上の予測に顕著な変化が生じるサンプルを選別し，顕著に変化したサンプルに対して擬似ラベルを作成し，ファインチューニングを行う．これにより，忘却データに特化した情報を構造的に破棄しつつ，保持データに対する性能を維持する狙いがある．

4 評価実験

本章では，コンペティションの上位手法を用いた比較実験について述べる．コンペティション終了後に主催者が結果や評価手法についてまとめたレポート [4] では，手法の汎用性についての実験結果を踏まえながら議論している．そこでは，評価指標の有効性とコンペティション上位手法と既存のアンラーニング手法の性能比較していたが，忘却データの割合変更を伴う実験を行なっていない．しかし実用環境においては，削除要求が発生する頻度や規模は状況によって大きく異なる．そのため，忘却対象データの割合が変動した場合にも安定して高い忘却性能を維持できるかどうか検証することは，手法の汎用性および実用性を評価する上で極めて重要である．そこで本実験では，忘却データの割合変更を伴う汎用性と異なるデータセットを用いた際の汎用性を調査する．

4.1 実験環境

本実験では，コンペティションの問題背景に則り実験を行う．モデルには ResNet-18 を使用し，データセットは CASIA-SURF-live を使用する．これに加え，データセットの違いによる汎用性を評価するために，一般的な画像分類で使用される CIFAR-10 を用いて同様の実験を行う．忘却データの割合は，コンペティションと同様の 2 %を使用すると同時に，0.5 %でも実験を行う．これにより忘却データの割合が変化した際の各手法の性能を確認する．評価指標は，コンペティションと同じ評価指標に加え，実行時間も比較する．比較する手法はコンペティションの上位 5 手法とする．なお，5 位の手法は重みを転置したものを代表として使用する．ハイパーパラメータ調整により忘却性能に大きな差を生み，手法の能力を正確に評価できなくなることを避けるため，本実験ではファインチューニング時のハイパーパラメータのみ最適化している．また本実験では，ファインチューニング時に保持データとテストデータの精度が収束したタイミングをアンラーニングの終了とみなす．実験環境は，CPU が Intel Corei7-14700KF，メモリが 64GB，GPU が Nvidia GeForce RTX 4090 の PC を用いる．

4.2 実験結果

表 1 に実験結果を示す．コンペティションの結果をまとめた文献 [4] において，CASIA-SURF-live，忘却データの割合を 2 %にした際に 1 位手法が最も高い Final スコアを得ていた．しかし，本実験ではコンペティションの結果とは異なり，4 位手法が Final スコア 0.087 と最も高いスコアを出力し，次点が 1 位の手法であった．文献 [4] では，モデルを 512 個使用した評価では 1 位手法が最も高い Final スコアを得ていたが，モデルを 1024 個使用して Final スコアを比較した際には 4 位手法が一番精度が良いという報告がされている．このことから，この評価指標はモデルの数によって性能が変化する指標であると言える．実環境でこの指標を使用する場合に膨大な量のモデルを作成する必要があり，多大なコストがかかっていまう点が，この評価指標の課題である． また，忘却データの割合を 0.5 %にした際には，3 位手法が最も高い Final スコア 0.025 を出力し，次点で 1 位手法が 0.012 を得た．各データに対する精度を比較しても 3 位，1 位は他手法より再学習モデルに近似できている手法であることがわかる．しかし，3 位手法は忘却データに対する精度が著しく低くなっている．これらは忘却ができているという視点で見ると優秀な手法であるが，再学習モデルに近似させるという観点においては良い数値とは言い難い．また 2 位手法に関しても，テストデータに対する精度が再学習モデルより高いことがわかる．3 位手法と同様に，再学習モデルに近似させるという観点で見ると良い数値とは言えない．このような再学習モデルより精度が良くなった場合についての扱いは，今後議論の余地があると考える．

CIFAR-10 を用いた実験では，忘却データの割合を 0.5 %及び 2 %にした際に 2 位手法が最も高い Final スコアを得た．CASIA-SURF-live と CIFAR-10 の順位は類似していることが確認できたが，CIFAR-10 を用いた実験において 1 位手法は非常に低い Final スコアとなった．これは，信頼度分布に大きな違いがあることに起因すると考えられる．図 2 に，1 位手法と 3 位手法に同一の CIFAR-10 の画像を入力した際の信頼度分布を示し，図 3 に，1 位手法と 3 位手法に同一の CASIA-SURF-live の画像を入力した際の信頼度分布を示す．図 2 から，1 位手法の信頼度分布は平均値に近い値を取る頻度が高く，非常に狭い分布となった．一方，3 位手法は信頼度分布の標準偏差が大きく，分布が広がる傾向にあることが確認できた．また，図 3 から，CASIA-SURF-live ではより類似する信頼度分布を出力していることが確認できる．この結果より，1 位手法はデータセットを変えた場合において，再学習モデルとの信頼度分布の差が明確になることを示しており，データセットに対する汎用性が低いと考えられる．また，1 位手法は出力の一様化と特徴空間の乖離を併用する複雑な構成を取っており，対照学習のパラメータや温度パラメータ等の設計が難しい．そのため，実環境で使用するには実装に細心の配慮が求められる手法といえる．

次に実行時間を比較する．表 1 より，どの手法も再学習と比較して大幅に実行時間を削減できていることがわかる．一方で，データセットや忘却データの割合を変更した際には実行時間の順位が変化することも確認でき

表 1: 比較実験の実行結果

-	CASIA-SURF-live								CIFAR-10							
忘却割合	0.5 %				2 %				0.5 %				2 %			
手法	Final	Test	Forget	Time[s]	Final	Test	Forget	Time[s]	Final	Test	Forget	Time[s]	Final	Test	Forget	Time[s]
1 位	0.012	93.6	96.1	8.9	0.081	94.1	89.3	8.9	0.0003	87.4	86.9	23.6	0.0004	86.3	85.9	56.5
2 位	0.006	98.1	94.3	19.7	0.074	91.0	86.6	9.6	0.021	86.1	94.3	34.7	0.017	86.3	92.3	68.2
3 位	0.025	94.5	91.5	8.3	0.071	89.7	85.9	8.2	0.009	83.7	84.9	51.5	0.005	84.4	87.1	70.1
4 位	0.001	96.2	94.9	48.3	0.087	86.1	86.6	7.3	0.006	86.3	87.8	49.2	0.004	84.2	86.8	81.4
5 位	0.002	98.1	95.7	37.6	0.049	93.2	85.8	5.8	0.003	81.8	79.5	42.4	0.006	85.7	85.4	61.5
再学習	-	94.4	94.0	70.6	-	95.8	85.6	57.1	-	87.5	85.8	348.9	-	87.4	87.2	329.9

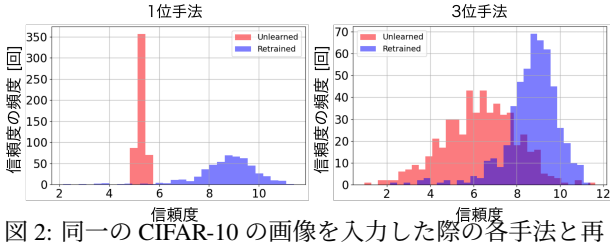

図 2: 同一の CIFAR-10 の画像を入力した際の各手法と再学習モデルの信頼度分布比較

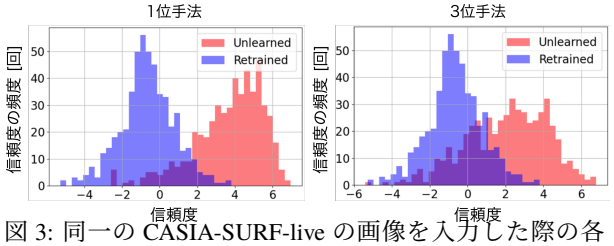

図 3: 同一の CASIA-SURF-live の画像を入力した際の各手法と再学習モデルの信頼度分布比較

た．例えば，1 位手法のような重みを間接的に操作する手法は実行時間が早いことがわかる．これは忘却と精度を補完するためのファインチューニングを効率的に行うことにより，効率的なアンラーニングを行っているためと考えられる．一方で，4 位手法のようなモデルの重みのほとんどを初期化するような手法では，モデルサイズやパラメータ数によって計算量が大きく変動すると考察する．

表 1 を通して，3 位，4 位手法がデータセットの違いによる汎用性が期待できる手法であると考える．しかし，両手法ともにファインチューニングによる精度補完が十分に行えず，再学習モデルよりスコアが低い点が今後の課題である．

5 おわりに

本稿では，マシンアンラーニングコンペティションの上位手法を対象に，その性能を比較・分析するための実験的な評価を行った．忘却後のテストデータや忘却データに対する精度のみならず，再学習モデルとの出力分布の類似度がマシンアンラーニングの品質評価において重要な要素であることを確認した．また，データセットや忘却割誌の変更に伴う，性能の違いについても確認した．今後は，出力分布の整合性を高めるための新たな手法の提案や，忘却の効率性・安定性を両立させる手法の開発が望まれる．

参考文献

[1] Cao, Y. and Yang, J.: “Towards Making Systems Forget with Machine Unlearning” , IEEE Symposium on Security and Privacy, pp. 463–480 (2015).

[2] Trippa, D., Campagnano, C., Bucarelli, S, M. et al.: “∇ τ : Gradient-based and Task-Agnostic Machine Unlearning”, CoRR, abs/2403.14339 (2024).

[3] Chourasia, R. and Shah, N.: “Forget Unlearning: Towards True Data-Deletion in Machine Learning”, International conference on machine learning, pp. 6028–6073 (2023).

[4] Triantafillou, E., Kairouz, P., Pedregosa, F. et al.: “Are we making progress in unlearning? Findings from the first NeurIPS unlearning competition”, CoRR, abs/2406.09073 (2024).

[5] He, K., Zhang, X., Ren, S. et al.: “Deep residual learning for image recognition”, Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778 (2016).

[6] Zhang, S., Liu, A., Wan, J. et al.: “Casia-surf:A large-scale multi-modal benchmark for face anti- spoofing.”, IEEE Transactions on Biometrics, Behavior, and Identity Science, pp. 182–193 (2020).

[7] Dwork, C., McSherry, F., Nissim, K. et al.: “Calibrating noise to sensitivity in private data analysis”, Theory of Cryptography: Third Theory of Cryptography Conference, pp. 265–284 (2006).

[8] Kairouz P., Oh, S., and Viswanath, P.: “The composition theorem for differential privacy. In International conference on machine learning”, International conference on machine learning, pp. 1376–1385 (2015).