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[1,a)]　柴田 大真[1,b)]　山内 悠嗣[1,c)]

概要

学習済みモデルから問題のある情報を削除するマシン・アンラーニングの重要性が増している．既存手法は高い性能を発揮する一方，破壊的忘却によるモデルの破壊が大きく，性能を回復するためのファインチューニングに多くの時間を要する．そこで，本稿では Piggyback アルゴリズムに基づく選択的な破壊的忘却による高速なマシン・アンラーニング手法を提案する．提案手法は，破壊的忘却フェーズにて忘れたいデータを意味する忘却データを使用し，Piggyback アルゴリズムを応用して重みの使用の有無を選択するバイナリマスクを学習する．これにより，忘却データに寄与する重みを自動かつ高精度に選択できる．選択された重みだけを操作することで，ファインチューニングに要する時間を削減し，高速なマシン・アンラーニングを実現できる．

1. はじめに

機械学習分野の技術が飛躍的な進歩を遂げている一方で，モデルの学習に個人データや有害なデータが混入している場合にはユーザーのプライバシーを侵害する問題を引き起こす．一般データ保護規則 [9] に則り忘れられる権利 [15] などの法律を遵守するため，問題のあるデータを削除した後に問題のないデータ (保持データ) のみで再学習を行う必要がある．しかし，一般的に再学習は膨大な計算量を伴う．そこで，特定の忘れたいデータ (忘却データ) を忘れさせるマシン・アンラーニングと呼ばれる再学習を近似する研究 [21][24] が盛んに取り組まれている．

マシン・アンラーニングの既存の手法では，深層学習により作成したモデルの重みを変更する破壊的忘却を行なった後，保持データを用いてファインチューニングするアプローチが主流である．破壊的忘却時に，モデルを構成する多数の重みを更新すると性能回復のためのファインチューニングに膨大な計算量が必要となる．

そこで，本稿では忘却データを用いた選択的な破壊的忘却によるマシン・アンラーニングの高速化手法を提案する．提案手法は，破壊的忘却時に忘却データのみを用いてモデルにおける重みの使用の有無を選択するバイナリマスクを学習する．これにより，忘却データの推論に寄与する重みを自動かつ高精度に選択できる．バイナリマスク学習により選択された重みだけを操作することで，ファインチューニングの計算コストを削減でき，効率的かつ高精度なアンラーニングを実現することができる．

[1] 中部大学
[a)] er21083-6439@sti.chubu.ac.jp
[b)] er21035-7824@sti.chubu.ac.jp
[c)] yuu@fsc.chubu.ac.jp

2. 関連研究

マシン・アンラーニング及び本研究で採用する枝刈りに関する先行研究について述べる．

2.1 マシン・アンラーニングに関する先行研究

2015 年に提案された Cao *et al.* の研究 [1] を発端としてマシン・アンラーニングの研究 [21][24] が取り組まれている．学習データ $\mathcal{D}$ を用いて学習したモデル $M(\mathcal{D})$ があり，モデル $M(\mathcal{D})$ から忘れたいデータ (忘却データ)$\mathcal{D}_f$ を削除する要求に応え，学習データ $\mathcal{D}$ から忘却データ $\mathcal{D}_f$ を取り除いたデータ (保持データ)$\mathcal{D}_r = \mathcal{D} \setminus \mathcal{D}_f$ で再学習を行うことがマシン・アンラーニングの理想である．しかしながら，再学習には膨大なコストが発生するため，忘却データの削除要求がある度に再学習を行うことは難しい．そこで，少ない計算コストで再学習したモデルに近似するマシン・アンラーニングの研究が盛んに行われており，本研究もこの研究に属する．これまでに，様々なアプローチによるマシン・アンラーニング手法 [21][24] が提案されてきたが，現在では深層学習に基づくアプローチ [4][6][10] が主流となっている．また，2023 年の NeurIPS においてマシン・アンラーニングのコンペティションが Kaggle にて開催され，コンペティション上位の手法は既存の研究を上回る結果であったことが報告されている [18]．コンペティション参加者の上位の手法は深層学習をベースとしており，破壊的忘却を行なった後に保持データを用いてファインチューニングを実施するアプローチを採用している．マシン・アンラーニング手法に求められることは，テストデータに対する精度を確保しながら，再学習したモデルの忘却データの精度と同等の精度に，少ない計算コストで実行することである．しかしながら，この 3 つはトリレンマ (3 種のトレードオフ) に陥りやすい．モデルの重みに対してノイズ

を加える手法や初期化する手法も提案されてはいるものの，操作対象のパラメータが，忘却データの認識精度への寄与を考慮していないため，不必要な性能低下を招きやすい他，ファインチューニングによる性能回復に時間を要する．従って，トリレンマを解決するためには，忘却データの影響が強い重みを操作するような効率的な破壊的忘却が必要である．

2.2 枝刈りとモデルのスパース化

モデルを軽量化するためのアプローチの 1 つに枝刈り [2] がある．枝刈りでは，モデルの重み $\mathbf{W}$ と 2 値で構成されるバイナリマスク $\mathbf{Y}$ のアダマール積を計算する．バイナリマスクは 0 と 1 で構成されており，マスクの値が 0 の場合にはモデルの重みが削除される (重みへの接続がなくなる) ため，モデルを圧縮することが可能となる．枝刈りは，重み単位で 0 と 1 の 2 値を決定する非構造的枝刈りと，行や列単位で 2 値を決定する構造的枝刈りの 2 つに分類できる．本節では，提案手法で採用する非構造的枝刈りについて述べる．

非構造的枝刈りでは，モデルの個々の重みに 0 または 1 のバイナリ値を割り当て，不要な重みを削除することでスパース化を実現する．この手法の利点は，モデルのパラメータ数を削減し，計算コストを大幅に削減できる点にある．Han *et al.* は，閾値以下の小さな重みは学習済みのモデルへの貢献が小さい重みとし，その重みを削減して，その後のファインチューニングにより性能を回復させることで，高い圧縮率を実現する手法 [7] を提案した．また，Gale *et al.* は，Transformer[19] や ResNet-50[8] などの大規模モデルに対して非構造的枝刈り手法 [23][14][12] がどの程度有効かを検証し，一般的な閾値設定による枝刈りだけではモデルの精度が低下しやすいことを指摘している [5]．そのため，枝刈り後のファインチューニングや適応的な閾値選択が重要となることが示されている．Liu *et al.* は，モデルの各層単位で閾値ベクトルを用意し，誤差を逆伝播させる際に閾値を最適化する手法 [11] を提案した．

枝刈りの研究は，複数の異なるタスクを順番に学習する継続学習 [20] にも応用されている．継続学習では，モデルが新しいタスクを学習する過程で，過去に学習したタスクに対する性能が大きく低下する破壊的忘却を防ぐことが目的である．本研究で応用する Piggyback[13] は，継続学習を実現するために枝刈りの研究にインスパイアされて提案された．Piggyback では，モデルの重みを直接的に変更する操作は行わず，0 と 1 で構成されるバイナリマスクを学習することで異なるタスクに適応する．学習済みモデルに対してバイナリマスクを適用することで，学習済みモデルの性能を保持しつつ，各タスクに適したモデルを実現する．

3. 提案手法

本稿では，選択的な破壊的忘却による高速なマシン・アンラーニング手法を提案する．提案手法の特徴は以下の通りである．

- 忘却データのみを用いたマスク学習
  提案手法は，忘却データの推論に寄与する重みを特定するために，忘却データのみを用いて Piggyback アルゴリズムによりバイナリマスクを学習する．End-to-End で微分可能であるため，自動的に重みを特定できる．保持データを使用しないため高速なマスク学習が実現できる．

- 選択的な破壊的忘却
  マスク学習により選択された重みに減衰係数を掛け合わせることによりノルムを減衰させる．特定の重みのみを操作することにより，不必要な認識精度の低下を防ぐことができる．また，特定した重みを初期化するのではなく，ノルムを小さくすることで，モデルの過剰な破壊を防ぐ．

- 操作する重みの限定
  本研究ではベースとなるモデルとして ResNet-18[8] を用いる．Piggyback アルゴリズムによるバイナリマスク学習時には ResNet-18 の全ての重みを対象とするが，破壊的忘却をする対象は，マスク学習により選択された重みの中でも，ResNet-18 の最後の layer(2 つの残差ブロック) に限定して重みのノルムを減衰させる．データセットに共通して重要なパラメータを保持することで，不必要なモデルの破壊を防ぎ，ファインチューニングに要する学習回数を削減できる．

図 1 に提案手法の流れを示す．まず，忘却データのみを用いた Piggyback アルゴリズムによりバイナリマスクを学習し，忘却データの情報を含む重みを特定する．そして，マスク学習で特定された重みのうち，ResNet-18 の最後の layer に限定して重みのノルムを減衰させる．その後，保持データを用いたファインチューニングによりテストデータに対する性能を回復させる．

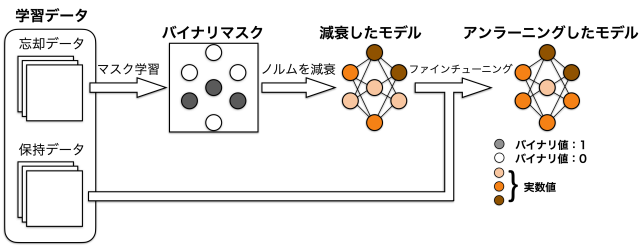

図 1: 提案手法の流れ．

3.1 Piggyback アルゴリズムによるバイナリマスク学習

提案手法では，忘却データの情報を含む重みを選択するために Piggyback アルゴリズムを使用する．Piggyback[13] は，すでに学習したタスクのパフォーマンスに影響を与えることなく，複数のタスクに適応させることを目的とした継続学習 [3] で用いられる．

本研究では忘却データを追加タスクのデータとみなし，Piggyback でタスク固有のバイナリマスクを学習する．モデルの事前学習に用いた忘却データでバイナリマスク学習をすることで，事前学習済みモデルから，忘却データの情報を含む重みを選択することが可能となる．

図 2 に提案手法のバイナリマスク学習の流れを示す．本稿では ResNet-18 を用いるが，式の簡略のため全結合層の例で表す．まず，学習データを用いて学習された学習済みモデルの全パラメータ Θ_{pre} を固定する．

$$\Theta_{\text{pre}} = \{\mathbf{W}_{\text{pre}}, \mathbf{b}_{\text{pre}}\} \tag{1}$$

ここで，$\mathbf{W}_{\text{pre}}$ は学習済みモデルの重み，$\mathbf{b}_{\text{pre}}$ はバイアスを表す．次に，学習済みモデルの各重みに対応した実数マスク x_i を作成し，全ての実数マスクの値を閾値 t 以上の同じ値で初期化する．

$$x_i = x_{\text{init}}, \quad \forall i \tag{2}$$

その後，実数マスク x_i の値が閾値 t 未満の値を 0 に，閾値以上の値を 1 にバイナリ化し，バイナリマスク y_i を作成する．

$$y_i = \begin{cases} 1, & \text{if } x_i \geq t \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

そして，作成したバイナリマスク y_i を学習済みモデルの重み $\mathbf{W}_{\text{pre}}$ に掛け合わせた重み $\mathbf{W}_{\text{masked}}$ で構成された，マスク処理モデルを作成する．

$$\mathbf{W}_{\text{masked}} = \mathbf{W}_{\text{pre}} \odot Y \tag{4}$$

ここで，$Y = [y_1, y_2, \ldots, y_d]$ はバイナリマスクのベクトルであり，$\odot$ はアダマール積を表す．

このマスク処理モデルに忘却データ $\mathcal{D}_f$ を入力し，順伝播，逆伝播を行い，実数マスク x_i の値を更新する．この一連の操作は忘却データのみを用いるため高速な処理が可能である．

3.2 損失関数

Piggyback アルゴリズムを用いたバイナリマスクの学習では式 (5) に示す損失を最小化する．

$$\mathcal{L}_{cross} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c} \log \hat{y}_{i,c} \tag{5}$$

ここで，N はバッチサイズ，C はクラス数，$y_{i,c}$ は正解ラベル，$\hat{y}_{i,c}$ はモデルの予測確率である．バイナリマスク学習により実現したいことは，忘却データに寄与する重みを特定することである．すなわち，忘却データに寄与する重みに対応する実数マスクの値を閾値以上にし，忘却データに寄与しない重みに対応する実数マスクの値を閾値未満にすることである．しかし，式 (5) に示す損失関数によりバイナリマスク学習をすると，学習が進行しない問題が発生する．すべての実数マスクは閾値以上の同じ値で初期化される．従って，すべてのバイナリマスクの値が 1 となり，マスク処理モデルは学習済みモデルそのものである．学習済みモデルは，忘却データを含んだ学習データで学習されたモデルであるため，忘却データに対する損失が小さく，バイナリマスク学習が進行しない．

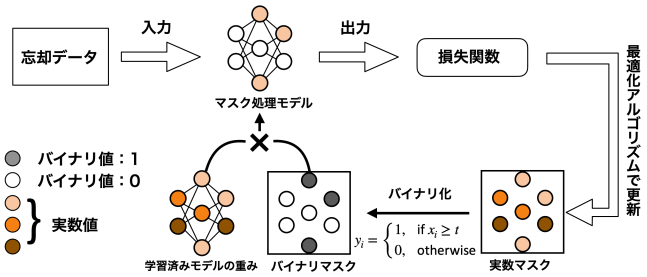

図 2: バイナリマスク学習の流れ．

そこで，本研究では式 (5) に示す損失関数に L1 正則化の項を加えた式 (6) により最適化する．

$$\mathcal{L}_{mask} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c} \log \hat{y}_{i,c} + \lambda \|\boldsymbol{\theta}\|_1 \tag{6}$$

ここで，$\|\boldsymbol{\theta}\|_1$ は実数マスク行列の L1 ノルム，λ は正則化の強さを制御するハイパーパラメータである．実数マスク行列に対しての L1 正則化項を加えることにより，忘却データに寄与しない重みに対応する実数マスクの値を閾値未満にし，忘却データに寄与する重みの特定が可能となる．提案手法によるバイナリマスク学習中のバイナリマスクが 0 となる割合の変化を図 3 に示す．学習データには ImageNet-100 を使用し，学習データの 2% を忘却データとした．なお，忘却データはランダムに選択した．一定の学習回数を繰り返すことでバイナリマスク 0 の割合が急激に増加し，後半には一定の割合に落ち着くことがわかる．先に述べた条件の場合には，100 回程度のエポックで一定の割合に収束し，バイナリマスク学習が完了する．この回数は多いように見えるが，バイナリマスク学習時には忘却データのみを用いるため，非常に少ない計算量で処理することが可能である．また，図 3 より，バイナリマスクの 0 の割合が約 88% で収束していることがわかる．このことは，L1 正則化項を加えたことにより，約 88% の実数マスクの値を閾値未満に出来るということである．

3.3 ノルムの減衰とファインチューニング

Piggyback アルゴリズムに基づいて作成されるバイナリマスクが 1 の値をとる場合には，忘却データが重みに対し

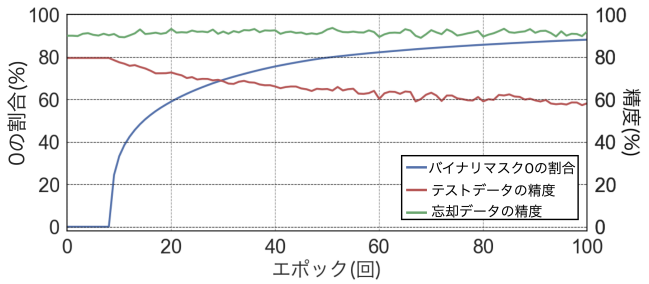

図 3: バイナリマスク 0 の割合の変化.

て大きな影響を与えていることを表す．選択された重みに減衰係数を掛け合わせることにより，重みのノルムを小さくする．減衰係数は 1 より小さい値で調整可能なハイパーパラメータである．提案手法は，バイナリマスク学習により選択された重みのうち，ResNet-18 の最後の layer に限定して重みのノルムを減衰させる．文献 [22] より，CNN は入力層に近い層では，入力画像のコーナーやエッジなどの局所的な特徴を抽出することが示されている．データセットに共通して重要なパラメータを保持することで，不必要なモデルの破壊を防ぎ，ファインチューニングに要する学習回数を削減できる．

4. 評価実験

4.1 既存の手法との比較実験の概要

提案手法の有効性を確認するために既存の手法と画像分類問題の結果を比較する．2023 年に開催された NeurIPS のコンペティションの報告 [18] にて，最新の研究 [4][6][10] よりもコンペティション参加者の手法の方が良い結果が得られたことについて言及されている．従って，コンペティションで優勝した手法 (Fanchuan) 及び準優勝した手法 (Kookmin) と比較する．コンペティションにて優勝，準優勝したチームの手法の詳細については [18] を参照されたい．なお，全ての手法においてベースとなるモデルとして ResNet-18 を用いる．

評価実験では，比較的規模が大きなデータセット ImageNet-100 を用いる．ImageNet-100 は，ImageNet-1k[16] のサブセットである．ImageNet-1k からランダムに選択された 100 クラスの画像で構成されている．忘却データの割合を変化させた際の傾向を確認するために，学習用画像 $\mathcal{D}$ の 2% と 4% を忘却データ $\mathcal{D}_f$ として，各割合での評価をする．また，学習用画像 $\mathcal{D}$ から忘却データ $\mathcal{D}_f$ を取り除いたデータを保持データ $\mathcal{D}_r = \mathcal{D} \setminus \mathcal{D}_f$ とする．評価では下記に示す 5 つの指標により評価する．

- 保持データに対する精度
- 忘却データに対する精度
- テストデータに対する精度
- Membership Inference Attack(MIA)
- 実行時間

Membership Inference Attack(MIA)[17] は，モデルが特定のデータを学習に使用したかどうかを推測する攻撃である．MIA の値が再学習と同じであればアンラーニング性能が高いといえる．実験環境は，CPU が Intel Corei7-14700KF, メモリが 64GB, GPU が Nvidia GeForce RTX4090 の PC を用いる．

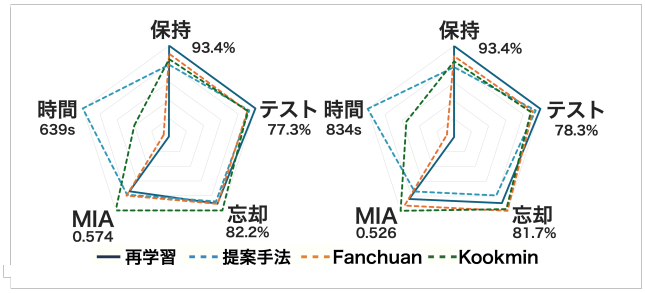

図 4: 各手法の性能を 5 つの指標により可視化した．なお，各指標は全てのアンラーニング手法の中で最大値 (時間は最小値) を基準に [0,1] の範囲に正規化した．また，各指標における数値は，全てのマシン・アンラーニング手法における最大値 (時間は最小値) を表す．

4.2 実験結果

比較結果を図 4 に示す．両実験において，テストデータに対する精度は，若干の性能差が見られるもののほぼ同等であった．忘却データに対する精度を比較すると，Kookmin, Fanchuan は再学習よりも高く，この結果が MIA にも影響を及ぼしている．次に実行時間を比較する．再学習は，膨大な計算量を要するのに対し，提案手法や他の手法は再学習よりも少ない実行時間でマシン・アンラーニングを可能としている．特に提案手法は他の手法に比べ，大幅に実行時間を削減することができている．

マシン・アンラーニングの分野において，一般的に実行時間とテストデータに対する精度，忘却データに対する精度がトリレンマ (3 種のトレードオフ) となる．しかし，提案手法では，3 つの指標のバランスがよく，トリレンマを緩和できているといえる．

5. おわりに

本稿では，Piggyback アルゴリズムに基づく選択的な破壊的忘却による高速なマシン・アンラーニング手法を提案した．提案手法は，破壊的忘却フェーズにて忘却データを使用し，Piggyback アルゴリズムを応用して重みの使用の有無を選択するバイナリマスクを学習する．これにより忘却データに寄与する重みを自動的に，かつ高精度に選択可能となる．選択された重みのみを操作することにより，ファインチューニング時間を削減し，高速なマシン・アンラーニングを実現した．今後は，大規模モデルに対してのマシン・アンラーニング性能について検証する予定である．

参考文献

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

[2] Cheng, H., Zhang, M. and Shi, J. Q.: A Survey on Deep Neural Network Pruning: Taxonomy, Comparison, Analysis, and Recommendations, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 46, No. 12, pp. 10558–10578 (2024).

[3] De Lange, M., Aljundi, R., Masana, M., Parisot, S., Jia, X., Leonardis, A., Slabaugh, G. and Tuytelaars, T.: A continual learning survey: Defying forgetting in classification tasks, *IEEE transactions on pattern analysis and machine intelligence*, Vol. 44, No. 7, pp. 3366–3385 (2021).

[4] Fan, C., Liu, J., Zhang, Y., Wong, E., Wei, D. and Liu, S.: Salun: Empowering machine unlearning via gradient-based weight saliency in both image classification and generation, *arXiv preprint arXiv:2310.12508* (2023).

[5] Gale, T., Elsen, E. and Hooker, S.: The State of Sparsity in Deep Neural Networks, *arXiv preprint arXiv:1902.09574* (2019).

[6] Golatkar, A., Achille, A. and Soatto, S.: Eternal sunshine of the spotless net: Selective forgetting in deep networks, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9304–9312 (2020).

[7] Han, S., Pool, J., Tran, J. and Dally, W. J.: Learning both Weights and Connections for Efficient Neural Networks, *Advances in Neural Information Processing Systems 28 (NIPS 2015)*, pp. 1135–1143 (2015).

[8] He, K., Zhang, X., Ren, S. and Sun, J.: Deep Residual Learning for Image Recognition, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778 (2016).

[9] Hoofnagle, C. J., Van Der Sloot, B. and Borgesius, F. Z.: The European Union general data protection regulation: what it is and what it means, *Information & Communications Technology Law*, Vol. 28, No. 1, pp. 65–98 (2019).

[10] Kurmanji, M., Triantafillou, P., Hayes, J. and Triantafillou, E.: Towards unbounded machine unlearning, *Advances in neural information processing systems*, Vol. 36 (2024).

[11] Liu, J., Xu, Z., Shi, R., Cheung, R. C. C. and So, H. K. H.: Dynamic Sparse Training: Find Efficient Sparse Network From Scratch With Trainable Masked Layers, *International Conference on Learning Representations (ICLR)* (2020).

[12] Louizos, C., Welling, M. and Kingma, D. P.: Learning Sparse Neural Networks through L0 Regularization, *CoRR*, Vol. abs/1712.01312 (2017).

[13] Mallya, A., Davis, D. and Lazebnik, S.: Piggyback: Adapting a Single Network to Multiple Tasks by Learning to Mask Weights, *Proceedings of the European Conference on Computer Vision (ECCV)* (2018).

[14] Molchanov, D., Ashukha, A. and Vetrov, D.: Variational Dropout Sparsifies Deep Neural Networks, *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pp. 2498–2507 (2017).

[15] Rosen, J.: The right to be forgotten, *Stan. L. Rev. Online*, Vol. 64, p. 88 (2011).

[16] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C. and Fei-Fei, L.: ImageNet Large Scale Visual Recognition Challenge, *International Journal of Computer Vision*, Vol. 115, No. 3, pp. 211–252 (2015).

[17] Shokri, R., Stronati, M., Song, C. and Shmatikov, V.: Membership inference attacks against machine learning models, *2017 IEEE symposium on security and privacy (SP)*, IEEE, pp. 3–18 (2017).

[18] Triantafillou, E., Kairouz, P., Pedregosa, F., Hayes, J., Kurmanji, M., Zhao, K., Dumoulin, V., Junior, J. J., Mitliagkas, I., Wan, J. et al.: Are we making progress in unlearning? findings from the first neurips unlearning competition, *arXiv preprint arXiv:2406.09073* (2024).

[19] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Łukasz Kaiser and Polosukhin, I.: Attention Is All You Need, *Advances in Neural Information Processing Systems*, Vol. 30, pp. 5998–6008 (2017).

[20] Wang, L., Zhang, X., Su, H. and Zhu, J.: A Comprehensive Survey of Continual Learning: Theory, Method and Application, *arXiv*, Vol. 2302.00487v3 (2024).

[21] Xu, H., Zhu, T., Zhang, L., Zhou, W. and Yu, P. S.: Machine Unlearning: A Survey, *ACM Computing Surveys*, Vol. 56, No. 1 (2024).

[22] Zeiler, M. D. and Fergus, R.: Visualizing and understanding convolutional networks, *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*, Springer, pp. 818–833 (2014).

[23] Zhu, M. and Gupta, S.: To prune, or not to prune: exploring the efficacy of pruning for model compression, *CoRR*, Vol. abs/1710.01878 (2017).

[24] 張　海波，櫻井幸一：機械アンラーニングの研究に関する現状と課題，人工知能学会誌，Vol. 38, No. 2, pp. 197–205 (2023).