Machine Unlearning

Akbari, Ramin
Michigan State University
akbarigh@msu.edu

Morshed, Mashrur Mahmud
Michigan State University
morshedm@msu.edu

Sachit Gaudi
Michigan State University
gaudisac@msu.edu

1. Introduction

Our research delves into the critical challenge of data privacy and compliance with emerging regulations, specifically the EU's General Data Protection Regulation (GDPR) as outlined in [12, 15]. Large AI models have shown tendencies to either hallucinate or inadvertently memorize training data [1–3, 7, 16, 17], posing a significant threat to user privacy. In light of GDPR's "right to be forgotten" imperative, the necessity to eradicate any traces of sensitive user information is evident. Retraining models from scratch for each individual removal is impractical due to the substantial time and computational resources involved. This research centers on developing an efficient unlearning method, both in terms of time and memory, to effectively eliminate sensitive user data. These unlearning methods can extend their utility to the removal of noisy data points and the mitigation of hate speech.

2. Problem Statement

This section formulates the problem and the metrics to determine the effectiveness of the algorithm. The unlearning $U(\cdot)$ is defined as to "forget" samples $S \subset D$, from the trained model $A_M(D)$, where $A_M : D \rightarrow \mathcal{R}^l$ is the training regime maps dataset to the weights space $\mathcal{R}^l$ of model M.

The notion of forgetting is measured relative to training the model from scratch without the samples S, i.e $A_M(D \backslash S)$. We cannot compare exact weights due to the randomness from the process. Therefore, to measure the forget quality, We recall the definition of unlearning metric, which draws inspiration from Differential privacy(DP). For a reference, we refer the reader to neurips machine unlearning competition [6]. The forget quality of unlearning $U(\cdot)$ is said to be (ϵ, δ) if

$$Pr[A_M(D \backslash S) \in \mathcal{R}^l] \leq e^{\epsilon} Pr[U(A_M(D), S, D) \in \mathcal{R}^l] + \delta \tag{1}$$

This metric is employed to assess the distribution of weights between training from scratch and the unlearning process. As the weights form a distribution rather than a unique point, owing to randomness in the initial seed of weights and the order of training samples. As the weight space is very high dimensional (11M for ResNet-18) the output space can be considered as a suitable proxy ($d << l$). The metric's computation involves processing each sample from $\mathcal{S}$ through K different seeds of the model, generating output distributions for both the unlearned method and training from scratch. The distance between these distributions using measures like KL-divergence, Bayesian decision boundaries, or any Model Inference Attack (MIA) forms the metric. The cumulative distance for all samples in the forget set $\mathcal{S}$ contributes to the forget quality, which is expressed as $\mathcal{F} = \sum_{\mathcal{S}} f(\epsilon)$.

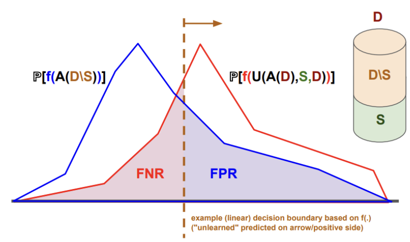

Figure 1. [6] Evaluation metric for unlearning. Any distribution either weights or output space of a sample quantifying unlearning algorithm and training from scratch.

. Equation 1 can be further modified [10] as

$$\begin{aligned}\epsilon = \sup_{i \in \mathit{MIA}} \big[&\max(\log(1 - \delta - \mathit{FPR}[i]) - \log(\mathit{FNR}[i]), \\ &\log(1 - \delta - \mathit{FNR}[i]) - \log(\mathit{FPR}[i]))\big]\end{aligned} \tag{2}$$

One noteworthy aspect to consider is the trade-off between utility, as represented by retain-set accuracy, and forget-quality. While it's possible to completely 'forget' by initializing the model, such a model would offer no utility. On the other hand, an existing model containing information about the forgotten samples might compromise privacy. Therefore, the task for unlearning methods, as previously explored in the literature, is to find the balance between accuracy and privacy. To account for utility, accuracy can be incorporated into the metric. Finally, $\mathcal{F} = g(Acc(R), Acc(T)) \times \sum_{\mathcal{S}} f(\epsilon)$, with R and T representing the retain and test sets.

3. Experimental Setup

The dataset comprises natural images of individuals' faces (X_i) along with associated identity (I_i) and age (a_i) information. We represent this dataset as $\mathcal{D} = (X_i, a_i) \quad \forall \quad i$, as specified in [6]. The 'forget set,' denoted as $\mathcal{S}$, is constructed to include 2% of the training dataset's identities. Importantly, these identities are selected in a non-I.I.D manner from the training data, with a notable emphasis on individuals with smaller ages within the forget set.

Our training procedure adheres to the $A_M(D)$ framework, where M corresponds to a ResNet-18 model. This model is trained for 30 epochs, with the inclusion of class weights to address class imbalance effectively

We aim to selectively 'forget' samples from $\mathcal{S}$. To assess the quality of this 'forgetting' process, we employ the $\mathcal{F}$ metric with $K = 512$ random seeds, as defined in Section 2.

As the dataset is hidden, we work with CelebA dataset which is similar to the nature of the problem.

4. Relevant Works

Unlearning is an emerging field marked by a lack of standardized definitions and evaluation criteria. This evolving landscape has given rise to diverse perspectives, resulting in multiple definitions and assessment measures. Notably, certain evaluation metrics center around the concept that effective unlearning algorithms should align the logit distributions of samples from the 'forget set' ($\mathcal{S}$) with those of a test dataset. This perspective leads to the direct optimization of GAN loss between the test and forget sets, as proposed by [4]. Alternatively, other approaches, such as [14], leverage a challenge inherent in deep learning, catastrophic forgetting, to their advantage. Additionally, [5] demonstrates that fine-tuning on the retained set ($\mathcal{D} \setminus \mathcal{S}$) leads to effective unlearning. Some works, including [11], address the more stringent case of unlearning, class unlearning, by maximizing KL-divergence on the forget set labels. Furthermore, works like [19] and [8] employ Fisher's discriminant, originally designed for unlearning in classical machine learning, though challenges arise when adapting it to large models due to its $O(W^2)$ time complexity.

However, the aforementioned approaches exhibit instability in optimization, lack theoretical guarantees or the ability to balance accuracy and privacy, as mentioned in Section 2. These methods do not provide clear explanations for the emergence of unlearning properties. Specifically, the GAN approach may falter when faced with a non-I.I.D forget set, while the KL-divergence approach may prove less effective for an I.I.D forget set. Our problem statement, which involves forgetting specific identities within a dataset characterized by class imbalance, does not neatly fit into either the strong I.I.D or non-I.I.D category.

For a more comprehensive understanding of the evolving field of unlearning, we recommend that interested readers refer to recent survey papers on the topic [9, 13, 18] available at [1].

5. Acknowledgements

The problem definition and the evaluation criteria is part of Neurips competition [6]. We will compete with all the participants and report the leader board and our approaches.

The code base [2] is organised as follows, We have a base template which contains, the benchmark and the training code. In the folder $Unlearning$, we implement different algorithms $U(A_M(\mathcal{D}), S, \mathcal{D}) \rightarrow \hat{M}$. We report the results in the project.

It is essential to note that this collaborative effort has been jointly undertaken by the three authors, each making equal and significant contributions. The order of authorship is organized alphabetically.

References

[1] Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *28th USENIX Security Symposium (USENIX Security 19)*, pages 267–284, 2019. 1

[2] Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650. USENIX Association, 2021.

[3] Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. Quantifying memorization across neural language models. In *The Eleventh International Conference on Learning Representations*, 2023. 1

[4] Kongyang Chen, Yao Huang, and Yiwen Wang. Machine unlearning via GAN. *CoRR*, abs/2111.11869, 2021. 2

[5] Min Du, Zhi Chen, Chang Liu, Rajvardhan Oak, and Dawn Song. Lifelong anomaly detection through unlearning. In *Proceedings of the 2019 ACM SIGSAC conference on computer and communications security*, pages 1283–1297, 2019. 2

[6] Jamie Hayes Peter Kairouz Isabelle Guyon Meghdad Kurmanji Gintare Karolina Dziugaite Peter Triantafillou Kairan Zhao Lisheng Sun Hosoya Julio C. S. Jacques Junior Vincent Dumoulin Ioannis Mitliagkas Sergio Escalera Jun Wan Sohier Dane Maggie Demkin Walter Reade Eleni Triantafillou, Fabian Pedregosa. Neurips 2023 - machine unlearning, 2023. 1, 2

[1] https://github.com/jjbrophy47/machine_unlearning

[2] https://github.com/sachit3022/unlearning

[7] Vitaly Feldman. Does learning require memorization? a short tale about a long tail. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, pages 954–959, 2020. 1

[8] Aditya Golatkar, Alessandro Achille, and Stefano Soatto. Eternal sunshine of the spotless net: Selective forgetting in deep networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[9] Yiwen Jiang, Shenglong Liu, Tao Zhao, Wei Li, and Xianzhou Gao. Machine unlearning survey. In *Fifth International Conference on Mechatronics and Computer Technology Engineering (MCTE 2022)*, page 125006J. International Society for Optics and Photonics, SPIE, 2022. 2

[10] Peter Kairouz, Sewoong Oh, and Pramod Viswanath. The composition theorem for differential privacy. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 1376–1385, Lille, France, 2015. PMLR. 1

[11] Meghdad Kurmanji, Peter Triantafillou, Jamie Hayes, and Eleni Triantafillou. Towards unbounded machine unlearning, 2023. 2

[12] Alessandro Mantelero. The eu proposal for a general data protection regulation and the roots of the 'right to be forgotten'. *Computer Law Security Review*, 29(3):229–235, 2013. 1

[13] Thanh Tam Nguyen, Thanh Trung Huynh, Phi Le Nguyen, Alan Wee-Chung Liew, Hongzhi Yin, and Quoc Viet Hung Nguyen. A survey of machine unlearning, 2022. 2

[14] Vinay Venkatesh Ramasesh, Aitor Lewkowycz, and Ethan Dyer. Effect of scale on catastrophic forgetting in neural networks. In *International Conference on Learning Representations*, 2022. 2

[15] Christopher Rees and Debbie Heywood. The 'right to be forgotten' or the 'principle that has been remembered'. *Computer Law Security Review*, 30(5):574–578, 2014. 1

[16] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pages 3–18. IEEE, 2017. 1

[17] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021. 1

[18] Haibo Zhang, Toru Nakamura, Takamasa Isohara, and Kouichi Sakurai. A review on machine unlearning. *SN Computer Science*, 4(4):337, 2023. 2

[19] Yongjing Zhang, Zhaobo Lu, Feng Zhang, Hao Wang, and Shaojing Li. Machine unlearning by reversing the continual learning. *Applied Sciences*, 13(16), 2023. 2