Machine Unlearning

Ramin Akbari
Michigan State University
akbarigh@msu.edu

Sachit Gaudi
Michigan State University
gaudisac@msu.edu

Mashrur Morshed
Michigan State University
morshedm@msu.edu

Abstract

Our research tackles the important topic of data privacy and following new rules, especially the European Union's General Data Protection Regulation (GDPR). We're concerned about big AI models making mistakes or remembering things they shouldn't during training, which can harm user privacy. With GDPR's "right to be forgotten," it's crucial to completely remove any sensitive user information. The impracticality of retraining models from scratch for each individual data removal is due to the considerable time and computational resources involved. Consequently, this study aims to devise an efficient unlearning method that optimally addresses both time and memory constraints. Beyond GDPR compliance, these unlearning methods hold promise for removing noisy data points, mitigating instances of hate speech, and helping solve data oriented concern. This study introduces three innovative approaches : i) Soft relabeling, ii) Gradient surgery, and iii) Forget Pruning that particularly target aspects of loss function, optimization, and constraints within the training pipeline. These techniques can be applied individually or in hybrid, providing a comprehensive framework for addressing the challenges associated with data privacy and model compliance.

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 [13, 16]. Large AI models have shown tendencies to either hallucinate or inadvertently memorize training data [1–3, 8, 17, 19], 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 remove noisy data and mitigate hate speech. Code is available here[1]

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 $A_M : D \to \mathcal{R}^l$ is the training regime maps dataset to the weights space $\mathcal{R}^l$ of model M.

The prevalent idea suggests a model is considered "unlearnt" if its weights match the initial training, yet this notion doesn't seamlessly apply to neural networks due to their intricate, non-convex nature. Moreover, the training process's stochasticity introduces variability, resulting in divergent convergence points. Despite these challenges, we leverage two widely-used proxy metrics (F-score, MIA) for unlearning. In the ensuing discussion, we delve into each metric and outline their limitations. Our goal remains to scrutinize the results judiciously and select a metric that aligns with the unique demands of the problem. For instance, in the context of unlearning in the Celeb-A dataset, we opt for F-score due to distinct forget and test distributions. Meanwhile, for CIFAR-10, where the forget set is uniformly sampled from the train set and train/test distributions generally align, we anticipate using MIA for its computational efficiency.

2.1. F-score

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 [18]. The forget quality of unlearning $U(\cdot)$ is said to be

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

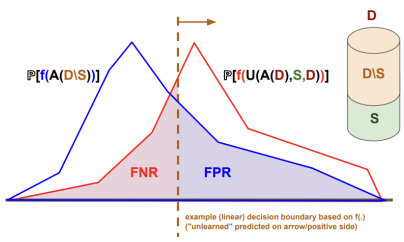

Figure 1. [18] Evaluation metric for unlearning. Any distribution either weights or output space of a sample quantifying unlearning algorithm and training from scratch.

(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 $\mathcal{F} = \sum_{\mathcal{S}} f(\epsilon)$.

. Equation 1 can be further modified [11] as

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

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.

2.2. MIA score

In contrast to F-score, MIA is advantageous as an alternative metric since it doesn't require training multiple models. Another drawback of F-score is the need to compute around 512 models for both training from scratch and under the unlearning paradigm. The paradox here lies in performing unlearning to circumvent training from scratch, yet the method itself relies on training from scratch, rendering it unsuitable for production settings where computational efficiency is crucial.

The concept of the MIA score closely resembles that of the F-score, but with a distinction in the distributions being compared—they are a function of the logits of the test set and forget set. This function is versatile and can take various forms aimed at maximally separating these two distributions. Examples of such functions encompass the logit of the correct class, cross entropy, and loss. To achieve this, a classifier is trained to distinguish between the train and test samples, and the accuracy of the MIA classifier in predicting forget samples as the test set is evaluated. This process establishes that forget set is similar to the test set.

In datasets with notable variations in factors like illumination and lighting between train and test inputs, the resulting larger separability leads to higher MIA scores. However, it's important to emphasize that a higher MIA score doesn't automatically indicate superior unlearning. To account for these differences and maintain a consistent comparison, we introduce a metric based on the absolute difference of MIA concerning the model trained from scratch.

Despite this refinement, a significant limitation persists when the input train and test distributions differ. This discrepancy is expected to result in distinct test and forget distributions, challenging the assumption that the logit distributions of forget and test should align.

3. Methods

We present three innovative approaches for unlearning in neural networks. The first, Soft-Relabeling, involves assigning labels to the forget-set based on their distance to samples in the retain-set. The second, Gradient Surgery, accelerates fine-tuning by projecting gradients into the null space of the forget-set. Lastly, Activation Pruning selectively removes activations that significantly contribute to the forget-set but not to the retain-set. These approaches are systematically compared to established methods discussed in Section 6 to assess their effectiveness in addressing the challenges associated with unlearning.

3.1. Soft-Relabeling

This insight stems from our observation of a model trained in the absence of a specific class. Remarkably, when encountering previously unseen class data, the model does not exhibit maximum uncertainty, treating all classes as equiprobable. Rather, it makes confident errors. Subsequent examination reveals that these misclassifications are

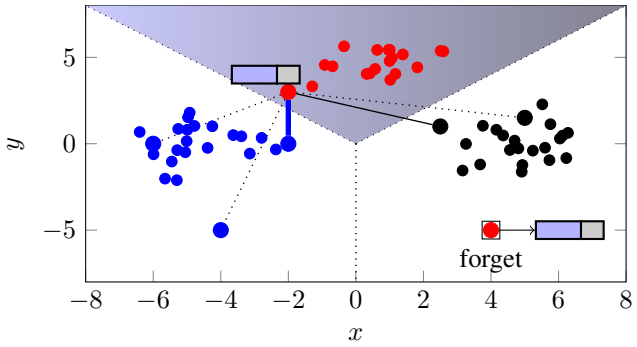

Figure 2. Soft-Relabeling: Task is to forget ● in 3 class classification, we replace the label of ● with similarity weighted combination from the retain data points (●, ●)

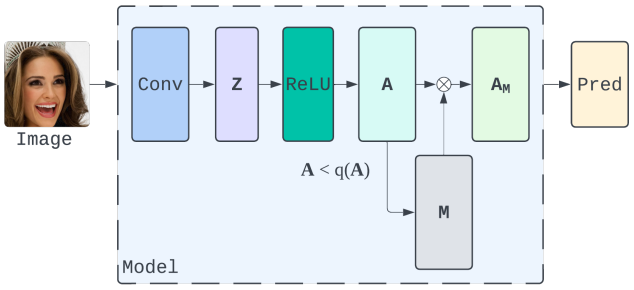

Figure 3. A brief overview of activation pruning. Given $\mathbf{A}$, the output of an activation function on forget images, we compute a pruning mask that zeros out any activation higher than the q^{th} percentile of the batch-wise average of $\mathbf{A}$. The masks are not further recomputed for retain images.

often linked to the closest classes, exhibiting shared properties. For instance, the model may mistakenly classify planes as ships or birds due to the common blue regions in sky and water. Such erroneous judgments appear to arise from the model's reliance on spurious correlations. Based on these insights, we hypothesize that we can unlearn a class, the label of the forget samples is obtained from the masked attention of the retained samples in a batch.

3.2. Gradient Surgery

The process of unlearning may be conceptualized as the intentional removal of particular segments within the model's memory that are entangled with the forget set. Subsequently, there is an attempt to implicitly acquire and integrate this knowledge through the retain set. We begin by pondering: How can we focus exclusively on the elements of a model connected to the forget set without affecting those acquired through the retain set? From a mathematical point of view, one way to erase model's memory is to perform gradient ascend. In this context, we can hypothesize that gradient's direction is equivalent to writing to the different segments in the model's memory. Following this, we can use the idea of gradient surgery to tackle this problem. Defining $\mathcal{F}$ as the forget set and $\mathcal{R}$ as the retain set, this process is analogous to implementing a Gram-Schmidt procedure on the gradients obtained from each set. We can use stochastic gradients to alleviate the computational cost of calculating the gradients. This procedure can be written as follows :

$$\nabla\mathcal{L}_{R_\perp} = \nabla\mathcal{L}_F - \frac{\nabla\mathcal{L}_F^T \nabla\mathcal{L}_R}{\nabla\mathcal{L}_R^T \nabla\mathcal{L}_R} \nabla\mathcal{L}_R \tag{3}$$

Here $\nabla\mathcal{L}_F$ and $\nabla\mathcal{L}_R$ correspond to the gradient of forget and retain set respectively. $\nabla\mathcal{L}_{R_\perp}$ is the component of gradient of the forget set which is orthogonal to the retain set. Following this, in the descent phase, we use a rather large initial learning rate and restrict the operation exclusively to the retain set. This choice is motivated by the objective of circumventing the basin where the model previously converged, aiming for a more generic solution in the new loss landscape. Moreover, the intention is to confine this training process which can resemble a soft pruning approach. This restriction ensures that only a specific portion of the model undergoes training. To achieve this, we introduce elastic-net regularization, which incorporates both $\ell_1$norm and $\ell_2$norm as penalty terms.

3.3. Activation Pruning

One of the motivations behind the development of unlearning techniques is that retraining a model from scratch on the retain set R is prohibitively expensive in numerous scenarios. The previous two approaches, detailed in Sec. 3.1 and Sec. 3.2, still involve performing gradient computations on the forget set $\mathcal{F}$. A question that may naturally arise: can we develop an unlearning algorithm that exploits the information in $\mathcal{F}$ without any backward gradient computations?

Activation pruning is motivated by the above question. Let $\mathsf{A}_n = r_n(\mathsf{Z}_n)$ be an intermediate activation map, where $r_n(.)$ denotes an arbitrary activation function inside the model (such as ReLU). The pruning algorithm works in two stages: (i) First, we make a forward pass through the model with a batch of forget data, $\mathcal{F}$, thus getting access to the set of intermediate activations $\{\mathsf{A}_n\}$. (ii) Second, based on a pre-determined threshold q, we prune (set to zero) any activation values above the q^{th} percentile. The idea is that, if the tensor A_n has high values in certain elements, those elements are propagating important information about the forget set through the model. Fig. 3 shows a visual overview of this process. If we zero out or *prune* some of the high-valued elements in A_n, then naturally the model performance on the forget set would drop—and so would its performance on the retain set $\mathcal{R}$, as the activation map elements important to $\mathcal{F}$ can also be important to $\mathcal{R}$.

Despite the initial performance drop on retain, further

fine-tuning solely on the retain set can reinforce new *pathways* for the retain data; it will encourage retain images to forge new connections through the model which will be different from that of the forget set. In practice, this can be implemented by creating a wrapper on the activation functions, that first computes the average activation $\mathsf{A} \in \mathbb{R}^{C \quad H \quad W}$ corresponding to a forget batch, and then computes the bool mask based on q. We set q to very high values for the activation functions in the early layers of the model (for instance, $q = 0.99$), while setting it to lower values, like $q = 0.90$, for deeper layers. This fuelled by the rationale that shallower layers in convolutional models learn more generic and low-level features, while deeper layers learn high-level (and thus sample-specific features). Further, if we prune too many activations in the early stages of the model, we are destroying information too early, which will cause serious detriments in model performance.

4. Experimental setup

We want to study the effectiveness of the unlearning algorithm under 2 settings. The selection of the two datasets aims to explore the diverse capacity of unlearning. In CelebA, the forget set and test set exhibit distinct distributions, while CIFAR represents a scenario with a uniform distribution. Notably, the initial model in CelebA tends to be overly fitted due to the substantial gap between train and test sets (11%). Furthermore, the forget set in CelebA features a different distribution compared to both the retain set and test set. In contrast, the initial model in CIFAR is better generalized, with a smaller performance gap (2%), and the forget set shares the same distribution as the retain set and test set.

For a fair evaluation of the unlearning paradigm, we refrain from tuning hyperparameters for both datasets. Consistently, we apply the same paradigm to assess how each approach performs across various levels of unlearning.

4.1. Celeb-A: Train-Forget-Test are sampled non-uniformly

The dataset consists of natural images featuring individuals' faces (X_i), along with associated identity (I_i) and attribute (a_i) information. The attributes are a combination of three identity-related binary attributes. We represent this dataset as $\mathcal{D} = (X_i, a_i) \quad \forall \quad i$, as detailed in [18]. The 'forget set,' denoted as $\mathcal{S}$, is meticulously curated to encompass 2% of the training dataset's identities. Crucially, these identities are chosen in a non-I.I.D manner from the training data, wherein only 2 classes out of 8 are selected, as per the methodology outlined in [18].

4.2. CIFAR-10: Train-Forget-Test are sampled uniformly

For the CIFAR-10 dataset, we forget 5000 samples, sampled uniformly out of the 50000 samples present in the training set. The examples to forget are obtained from the CIFAR example provided in the starter code of the Unlearning Challenge [18].

4.3. Model

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.

While ResNet-18 serves as an excellent initial framework, we anticipate that the unlearning methods applied to ResNet-18 will generalize to other model architectures. To account for diverse levels of memorization in different architectures and varying model capacities, we extend our evaluation to include ResNet-50 and Vision Transformers [6]. Detailed results are presented in Section 5.

4.4. Hyper parameters of metrics

Our objective is to selectively 'forget' samples from $\mathcal{S}$. To evaluate the effectiveness of this 'forgetting' process, we utilize the $\mathcal{F}$ metric with $K = 32$ random seeds, as outlined in Section 2. In this evaluation, cross-entropy is employed as a function, and a max-separable Bayesian classifier serves as a discriminator to estimate the F-score. For MIA, we employ four different functions, namely correctness, confidence, entropy, and probability.

5. Results

.

Attacks	correctness	confidence	entropy	prob	best attack MIA ↓
Retrain	20.92	52.06	40.68	52.14	0.00
Finetune [7]	14.12(6.8)	53.91 (1.85)	45.07 (4.39)	56.73 (4.59)	4.41
Random Labels	10.68 (10.24)	54.79 (2.73)	48.39 (7.71)	57.39 (5.25)	6.48
SCRUBS [12]	17.04 (3.88)	62.57 10.51	54.96 (14.28)	58.72 (6.58)	8.81
Boundary Unlearning [5]	15.680 (5.24)	53.65 (0.26)	44.56 (12.17)	59.00 (2.27)	4.98
Soft-Relabeling	13.88 (3.2)	42.97 (10.94)	38.96 (1.72)	53.13 (0.99)	4.21
Gradient Surgery	21.68 (0.76)	55.56 (3.5)	38.30 (2.38)	52.31 (0.17)	(1.70)
Forget Pruning	50.62 (29.7)	49.90 (2.16)	51.74 (11.06)	49.40 (2.74)	11.41
No Unlearning	9.96 (10.96)	54.78 (0.89)	50.61 (9.93)	60.89 (8.75)	7.63

Table 1. MIA scores for CIFAR10 dataset, ( ) indicates the gap to the retrained from scratch. lower the better.

From Tab. 2 we can observe that all our methods show competitive results compared to existing approaches. We analyze the results from the perspective of each approach:
Soft-Relabeling: Soft-Relabeling demonstrates consistent reliability across various levels of unlearning, as indicated by the MIA scores for both CIFAR-10 and Celeb-A in Tables 1 and 3. In contrast, some other approaches excel in specific settings but show a decline in performance in other

Metric	RA "	TA "	UA #	F-score#	total score"
Retrain	87.34	81.94	77.96	1	1
Finetune [7]	88.71	82.50	85.36	0.524	0.488
Random Labels	89.31	82.66	89.40	0.595	0.605
SCRUBS [12]	82.26	82.63	83.33	0.488	0.480
Boundary Unlearning [5]	86.31	82.75	83.18	0.504	0.502
Soft-Relabeling	87.91	82.18	84.72	0.521	0.525
Gradient Surgery	84.06	81.49	82.71	0.637	0.604
Forget Pruning	82.73	81.25	82.85	0.640	0.547
No Unlearning	89.22	83.38	89.39	-	-

Table 2. Forget Metrics for CIFAR10 dataset

Attacks	correctness	con dence	entropy	prob	best attack MIA
Retrain	7.15	12.31	12.39	67.09	0.00
Finetune [7]	6.8 (0.35)	11.2 (1.11)	12.17 (0.22)	66.64 (0.45)	0.53
Random Labels	7.6 (0.45)	11.73 (0.38)	10.72 (1.67)	67.36 (0.27)	0.66
SCRUBS [12]	6.6 (0.55)	10.5(1.81)	10.18 (2.21)	66.45 (0.64)	1.30
Boundary Unlearning [5]	7.22 (0.07)	12.36 (0.05)	12.59 (0.2)	68.62 (0.46)	(0.195)
Soft-Relabeling	7.12 (0.03)	11.89 (0.69)	11.68 (0.49)	67.39 (0.3)	0.37
Gradient Surgery	6.7 (0.45)	11.80 (0.51)	11.39 (1)	66.80 (0.29)	0.56
Forget Pruning	37.64 (30.49)	25.42 (13.11)	42.26 (29.87)	75.81 (8.72)	20.55
No Unlearning	6.3 (0.85)	11.59 (0.7)	12.23 (0.72)	67.65 (0.56)	0.70

Table 3. MIA scores for CelebA dataset.( ) indicates the gap to the retrained from scratch. lower the better.

Metric	RA "	TA "	UA #	F-score#	total score"
Retrain	99.99	86.33	90.20	1	1
Finetune [7]	99.99	86.35	90.33	0.058	0.058
Random Labels	99.99	86.35	85.41	0.093	0.093
SCRUBS [12]	90.74	86.16	90.62	0.012	0.011
Boundary Unlearning [5]	87.00	85.69	90.78	0.004	0.003
Soft-Relabeling	89.23	82.83	88.87	0.110	0.104
Gradient Surgery	87.40	86.44	91.01	0.008	0.009
Forget Pruning	99.99	86.36	90.31	0.488	0.406
No Unlearning	99.15	86.48	98.92	-	-

Table 4. Forget Metrics for CelebA dataset

scenarios. The forget metric score in Table 2 aligns with the MIA scores and competes well with existing solutions. While Soft-Relabeling may not be the best in terms of utility for CIFAR-10, with only a1:31%deviation from the true utility of retraining from scratch, it is within the1 2% range of the best method.

Soft-Relabeling performs well on the CelebA dataset, particularly when the model is overly tted. The method effectively erases the decision boundary of over tted samples, emulating the behavior of retraining from scratch. However, occasional drops in performance may occur due to stochastic updates. Increasing the batch size of the retain set leads to a broader exploration of the space, resulting in superior performance on the score.

Gradient Surgery: Gradient Surgery performs well in terms of both MIA scores and forget metrics for CIFAR, as observed in Tab. 1 and Tab. 2. It achieves the best forget metric score among our proposed methods. This success can be attributed to the uniformity of the forget-set and retain-set. By disallowing gradients in the direction of the forget set, the model accelerates the ne-tuning approach. It can be demonstrated that the slight modi cation of performing one iteration of gradient ascent on the forget set helps prevent the model from converging into saddle points (a possibility for netuning) ensuring the model reaches the global minima. In the case of CelebA, the forget metric score, as shown in Tab. 4, is lower compared to that of Soft-Relabeling but higher than other existing approaches. The utility is on the lower end compared to other existing method.

Activation Pruning : In the case of CIFAR, activation pruning seems to have a problem with membership inference attack and the forget metrics for CIFAR, as observed in Tab. 1; this effect is less pronounced for CelebA, but still present to some extent, as seen in Tab. 3. For CIFAR, the forget metric score is comparable to existing methods; it is higher than Soft-Relabeling but less than that of Gradient Surgery, as seen from Tab. 2. However, for CelebA, in Tab. 4, the forget metric score is signi cantly higher than approaches. We believe the strange difference between CIFAR and CelebA can be explained by the fact that the base model we used for unlearning on CelebA was a very con dent, over tted model with a retrain accuracy of almost 1. But the model for CIFAR was less accurate; this suggests pruning works better for models which are tted strongly on train, but has problems with under tted models; perhaps a less aggressive pruning would work for CIFAR. When the initial model has high amount of over tting, Activation Pruning outperforms any other method of a huge margin. The model also preserves the utility of the model in the CelebA dataset.

Discussion on Metrics Activation pruning exhibits a notable disparity in MIA and F -score metrics for the CelebA dataset. According to MIA score, Activation Pruning performs poorly, while in terms of F -score, it outperforms every method by a signi cant margin. Given the substantial difference between the test and forget distributions in Celeb-A, the MIA metric proves less reliable, as the initial distribution differs, and the expectation of the logit distribution similarity between the forget set and test set is not met. In contrast, F -score demonstrates greater robustness to such assumptions, making it the preferred metric for datasets with distinct distributions, like Celeb-A.

For CIFAR-10 datasets, achieving a robust estimate of F -score demands a higher number of models to be trained. Unfortunately, due to limited computational resources, we had to settle for 32 models, which may not provide the most reliable estimate. In these circumstances, MIA proves to be a more dependable metric. In the case of the CIFAR-10 dataset, where the test and forget distributions are uniform, and the retain distribution differs from forget and test, as illustrated in Fig. 4, MIA emerges as a reliable metric.

6. Related Work

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' (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 [15], leverage a challenge inherent in deep learning, catastrophic forgetting, to their advantage. Additionally, [7] demonstrates that fine-tuning on the retained set ($D(n S)$) leads to effective unlearning. Some works, including [12], address the more stringent case of unlearning, class unlearning, by maximizing KL-divergence on the forget set labels. Furthermore, works like [21] and [9] 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 [10, 14, 20] available at[2].

7. Conclusion

In this study, we approach the machine unlearning problem from three different perspectives. We restrict ourselves to just Computer Vision, specifically image classification, though the problem can be considered a general paradigm applicable over most sub-fields in machine learning. We found that all three approaches we proposed perform competitively with existing methods in the field, and are promising directions of future study.

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 Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar 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. 6

[5] Min Chen, Weizhuo Gao, Gaoyang Liu, Kai Peng, and Chen Wang. Boundary unlearning, 2023. 4, 5

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 4

[7] 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. 4, 5, 6

[8] 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

[9] 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. 6

[10] 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. 6

[11] 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. 2

[12] Meghdad Kurmanji, Peter Triantafillou, Jamie Hayes, and Eleni Triantafillou. Towards unbounded machine unlearning, 2023. 4, 5, 6

[13] 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

[14] 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. 6

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

[15] Vinay Venkatesh Ramasesh, Aitor Lewkowycz, and Ethan Dyer. Effect of scale on catastrophic forgetting in neural networks. In International Conference on Learning Representations, 2022. 6

[16] 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

[17] 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

[18] Eleni Triantafillou, Fabian Pedregosa, 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, and Walter Reade. Neurips 2023 - machine unlearning, 2023. 1, 2, 4

[19] 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

[20] Haibo Zhang, Toru Nakamura, Takamasa Isohara, and Kouichi Sakurai. A review on machine unlearning. SN Computer Science, 4(4):337, 2023. 6

[21] Yongjing Zhang, Zhaobo Lu, Feng Zhang, Hao Wang, and Shaojing Li. Machine unlearning by reversing the continual learning. Applied Sciences, 13(16), 2023. 6

(a) Loss distribution of forget-test before unlearning

(b) Loss distribution of forget-test after unlearning

Figure 4. Effect of unlearning Algorithm, showing the test and forget distributions are similar compared to the not apply unlearning.

8. Appendix

To fully understand the results and fully understand the relative performace, we plot the results.