Adversarial Federated Unlearning with Representation Decoupling

Yu Jiang[⋆†], Kwok-Yan Lam[⋆†], and Chee Wei Tan[⋆]

[⋆]College of Computing and Data Science, Nanyang Technological University
[†]Digital Trust Centre, Nanyang Technological University, Singapore
{yu012, kwokyan.lam, cheewei.tan}@ntu.edu.sg

Abstract. Federated unlearning (FU) removes the influence of specific clients' data from a collaboratively trained model without requiring complete retraining when the target clients request unlearning, addressing "the right to be forgotten" requirements in distributed learning environments, which has attracted significant interest. However, existing FU methods can achieve certain unlearning effectiveness but rarely address internal representation dependencies. To address this limitation, we propose a novel adversarial framework for FU that achieves representation decoupling through mutual information minimization. In our framework, the unlearned model acts like a generator that learns to generate outputs that appear not to have seen unlearning data, while a discriminator evaluates unlearning effectiveness by distinguishing between original and unlearned model states. Additionally, we design an auxiliary network within the discriminator to estimate the statistical dependence between original and unlearned model representations, guiding the unlearned model to achieve effective representation decoupling by minimizing the estimated mutual information. Experimental results reveal a trade-off between utility preservation and unlearning effectiveness, where our method achieves enhanced unlearning effectiveness compared to existing approaches.

Keywords: Machine unlearning · Federated learning · Adversarial learning · Mutual information

1 Introduction

The widespread adoption of federated learning (FL) systems across various domains has fundamentally transformed how machine learning models are trained on distributed data. While FL enables collaborative model training without requiring direct data sharing among participants, it faces new challenges related to data privacy, security, and regulatory compliance. Privacy regulations such as the General Data Protection Regulation (GDPR) [13] and the California Consumer Privacy Act (CCPA) [3] have established "the right to be forgotten," mandating that individuals have the right to request deletion of their data. The concept of machine unlearning has emerged as a promising solution to these requirements, enabling the removal of influence from specific data points or client contributions from trained models without requiring complete retraining.

In federated settings, unlearning faces distinct challenges compared to centralized scenarios. Unlike centralized unlearning where target data can be directly identified and accessed, federated learning involves multiple training rounds where individual client contributions become deeply permeated within the global model through iterative parameter aggregation. This distributed training process makes it difficult to isolate and remove specific client influences without affecting the overall model performance. Consequently, effective federated unlearning methods need to erase influence from target clients while preserving the utility performance from remaining clients. Existing FU approaches that aim to remove client data influence from FL models include fine-tuning the global model [8], model scrubbing that approximates training without the unlearning data [11], and synthetic data generation for unlearning data to help the model unlearn specific information [10]. While these methods can achieve certain effectiveness, such as reducing attack success rates of backdoor attacks after unlearning, they primarily focus on output-level changes and overlook whether the model has eliminated the underlying feature representations learned from target clients. These methods do not consider addressing representation-level dependencies where the model continues to remember patterns originally learned from target client data, allowing adversaries to potentially infer sensitive information about target clients even when the output appears to have forgotten the data.

To address this limitation, we propose an adversarial framework for FU that achieves representation decoupling through mutual information minimization. The unlearned model acts like a generator that learns to generate outputs appearing to have forgotten the target client data while ensuring its internal feature representations become statistically independent from the original FL model. To enable this dual objective, we design a discriminator for clients who request unlearning, also referred to as target clients, that performs adversarial training and mutual information estimation. The discriminator learns to distinguish between the original FL model's behavior and the unlearned model's behavior, while its auxiliary T-network estimates the mutual information between feature representations from both models. Mutual information measures the statistical dependence between feature representations extracted from the original and unlearned model. When mutual information approaches zero, it indicates statistical independence between the two representations. Guided by the estimated mutual information from the auxiliary T-network, the unlearned model systematically breaks down the representational connections between how the original and unlearned model process the unlearning data. Comprehensive experimental analysis reveals a trade-off between utility preservation and unlearning effectiveness. Our approach achieves enhanced unlearning performance, confirming the effectiveness of representation decoupling for FU scenarios.

2 Preliminaries

We consider N clients with local datasets $\mathcal{D}_i = \{(x, y)\}$ of size n_i, where $x \in \mathcal{X}$ denotes the input sample and $y \in \mathcal{Y}$ its corresponding label. The global dataset

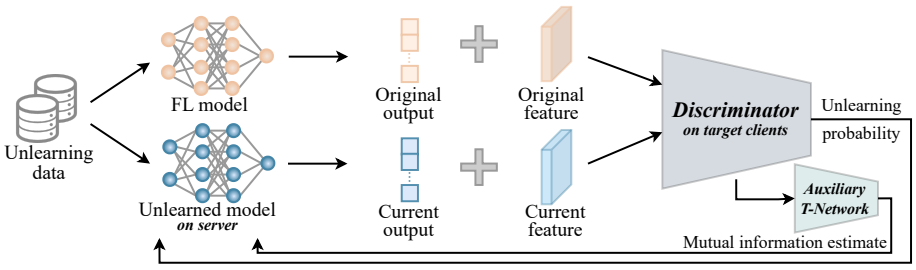

Fig. 1. Framework of adversarial FU with representation decoupling

is $\mathcal{D} = \bigcup_{i=1}^{N} \mathcal{D}_i$ of size $n = \sum_{i=1}^{N} n_i$. The standard FL objective is the weighted empirical risk minimization

$$\min_{\theta} F(\theta) = \sum_{i=1}^{N} \frac{n_i}{n} F_i(\theta), \quad F_i(\theta) = \frac{1}{n_i} \sum_{(x,y)\in\mathcal{D}_i} \ell(f_\theta(x), y), \tag{1}$$

where ℓ is the loss function. FL proceeds in T communication rounds: each client updates local models from the current global parameters θ^t, and the server aggregates updates using a rule $\mathcal{A}$, e.g., FedAvg [12]: $\theta^{t+1} = \theta^t - \eta \cdot \mathcal{A}(g_1^t, \ldots, g_N^t)$. After T rounds, the trained global model has parameters $\bar{\theta}$, which we regard as the baseline for subsequent unlearning operations.

3 Methodology

3.1 Overview

FU is typically initiated by requests from clients. Our approach intrinsically provides target clients with a verification mechanism to assess the effectiveness of data removal, which existing methods fail to guarantee. While adversarial training can enforce output-level changes, achieving true unlearning further requires addressing the model's internal knowledge representations. To this end, we introduce representation decoupling, which aims to disentangle the model's learned features from the target clients' data. Our FU framework integrates adversarial training with mutual information minimization to achieve effective representation decoupling, as illustrated in Fig. 1. Let $\mathcal{C} = \{1, 2, \ldots, N\}$ denote the set of participating clients. We use $\bar{\theta}$ to denote the parameters of the original federated model trained on all clients, and $\hat{\theta}$ to represent the parameters of the current model being unlearned. Given a set of target clients $\mathcal{C}_f \subseteq \mathcal{C}$ with their corresponding data $\mathcal{D}_f = \bigcup_{i\in\mathcal{C}_f} \mathcal{D}_i$, the goal of FU is to obtain an updated model $\hat{\theta}$ that satisfies two key properties: (i) removal of information associated with $\mathcal{D}_f$, and (ii) preservation of utility on the remaining clients $\mathcal{C}_r = \mathcal{C} \setminus \mathcal{C}_f$.

3.2 Discriminator Design

We design a discriminator for the target clients that performs adversarial training and mutual information estimation. To capture the internal knowledge that models rely on for prediction, we extract feature representations from the intermediate layers of the original FL model and current unlearned model. For any input $x \in \mathcal{D}_f$, we obtain $\bar{h} = \xi_{\bar{\theta}}(x)$ and $\hat{h} = \xi_{\hat{\theta}}(x)$, where $\bar{h}$ and $\hat{h}$ denote the feature representations from the original and current models, respectively, and ξ_θ denotes the feature extractor of model f_θ. The discriminator D_ϕ consists of a shared backbone with two heads: (i) a classification output providing an unlearning probability for adversarial training, derived from model output probabilities; and (ii) an auxiliary T-network that estimates mutual information from feature representations to encourage representation decoupling.

Adversarial Training. The discriminator is trained to distinguish between two types of model behaviors on the unlearning dataset $\mathcal{D}_f$. Specifically, label 0 represents the remembered state where the influence of target data is retained, while label 1 represents the forgotten state where the influence of target data is erased. The training objective of the discriminator maximizes the likelihood of correctly classifying these two states:

$$\mathcal{L}_D = \mathbb{E}_{x \sim \mathcal{D}_f}[\log(1 - D_\phi(f_{\bar{\theta}}(x))) + \log D_\phi(f_{\hat{\theta}}(x))], \tag{2}$$

where D_ϕ denotes the unlearning probability that performs binary classification based on model output probabilities, with values closer to 1 indicating successful unlearning and values closer to 0 indicating the remembered state.

Representation Decoupling via Mutual Information Minimization. In addition to adversarial training, we further introduce mutual information minimization to enforce representation-level decoupling between the original and unlearned models. Mutual information quantifies the statistical dependence between two random variables, measuring how much knowing one reduces the uncertainty about the other [2]. A mutual information value of zero implies complete independence, meaning that one variable provides no information about the other. By minimizing the mutual information between representations of the two models on unlearning data, the internal patterns of the unlearned model cannot reveal knowledge retained by the original model.

To quantify the dependence between $\bar{h}$ and $\hat{h}$, we need to estimate their mutual information. Traditional mutual information computation requires explicit probability distributions, which is intractable for high-dimensional neural network features. The estimation is derived from the definition of mutual information as a KL divergence [4]:

$$I(\bar{h}; \hat{h}) = D_{KL}(P(\bar{h}, \hat{h}) || P(\bar{h}) \otimes P(\hat{h})), \tag{3}$$

which compares the joint distribution of paired representations with what we would expect if they were completely independent. To estimate mutual information without explicitly computing complex probability distributions, we apply

the Donsker-Varadhan theorem [6], reformulating mutual information as an optimization problem where a neural network learns to distinguish between paired and unpaired data samples:

$$D_{KL}(P||Q) = \sup_{T} \left[\mathbb{E}_P[T] - \log \mathbb{E}_Q[e^T]\right], \tag{4}$$

Applying this to mutual information gives us:

$$I(\bar{h}; \hat{h}) = \sup_{T} \left[\mathbb{E}_{P(\bar{h},\hat{h})}[T(\bar{h}, \hat{h})] - \log \mathbb{E}_{P(\bar{h})\otimes P(\hat{h})}[e^{T(\bar{h},\hat{h})}]\right]. \tag{5}$$

To approximate the optimal function T, we implement an auxiliary T-network T_ψ within the discriminator, processing features extracted from model outputs along with intermediate representations from both models:

$$\hat{I}(\bar{h}; \hat{h}) = \mathbb{E}[T_\psi(\bar{h}, \hat{h})] - \log \mathbb{E}[e^{T_\psi(\bar{h},\hat{h}')}], \tag{6}$$

where $(\bar{h}, \hat{h})$ are paired features from the same input, while $(\bar{h}, \hat{h}')$ are mismatched pairs formed with $\hat{h}'$ sampled from a different input, thereby breaking the original dependencies. By integrating T_ψ into the discriminator, we enable the joint optimization of adversarial training and representation decoupling.

Our representation decoupling objective minimizes the mutual information:

$$\mathcal{L}_m = \hat{I}(\bar{h}; \hat{h}). \tag{7}$$

which is computed locally by target clients, who then provide the corresponding loss components to the server for optimization. As training progresses and $\hat{I}(\bar{h}; \hat{h}) \to 0$, the representations of the unlearned model become statistically independent from those of the original model, achieving effective decoupling.

3.3 Unlearned Model Optimization

The server coordinates with target clients by sharing the current model for local discriminator training and mutual information computation. The current model $f_{\hat{\theta}}$ functions like a generator, learning to produce outputs that demonstrate effective unlearning behavior while achieving representation decoupling from the original model. The learning task naturally becomes multi-objective, as the unlearned model is required to optimize against the discriminator while simultaneously reducing statistical dependence with the original model's representations.

To achieve adversarial unlearning, the unlearned model is trained to produce outputs that the discriminator classifies as "forgotten" with high confidence. Formally, the adversarial loss is defined as

$$\mathcal{L}_{\text{adv}} = -\mathbb{E}_{x\sim\mathcal{D}_f}[\log D_\phi(f_{\hat{\theta}}(x))], \tag{8}$$

which encourages $f_{\hat{\theta}}$ to maximize the probability that its outputs are identified as forgotten. Through this adversarial interaction, the unlearned model learns to eliminate the influence of the target data.

However, adversarial training alone only ensures unlearning at the output level, potentially leaving residual knowledge embedded in internal feature representations. To address this limitation, the unlearned model also optimizes the mutual information loss computed by the T-network to achieve representation decoupling. The overall training objective for the unlearned model combines adversarial loss, mutual information loss, and utility preservation:

$$\mathcal{L}_{total} = \lambda_{adv}\mathcal{L}_{adv} + \lambda_m\mathcal{L}_m + \lambda_r\mathcal{L}_r, \tag{9}$$

where λ_{adv}, λ_m, and λ_r control the relative importance of each component, and the regularization term $\mathcal{L}_r = \|\hat{\theta} - \bar{\theta}\|_2^2$ prevents excessive deviation from the original parameters to maintain performance on remaining clients. The server updates the unlearned model parameters $\hat{\theta}$:

$$\hat{\theta} \leftarrow \hat{\theta} - \eta\nabla_{\hat{\theta}}\mathcal{L}_{total}, \tag{10}$$

where η is the learning rate, ensuring gradual unlearning of target client data while maintaining performance on remaining clients.

4 Experiments

4.1 Setup

We implement FL using 20 clients and FedAvg over 40 rounds, with 5 local epochs per round and Adam optimizer at learning rate 0.005 and batch size 128. We evaluate unlearning ratios (UR) of 5% and 25%, where UR represents the proportion of clients requesting unlearning. Experiments use MNIST [5] and Fashion-MNIST [15] with CNN networks, and CIFAR-10 [9] with ResNet-34. The discriminator uses MLP architectures. We compare against Train-from-scratch (Retrain), Gradient Ascent (GA) [7], and FedU [14]. To evaluate unlearning effectiveness, we employ backdoor attacks [1] where target clients inject triggers into training data during federated learning. Lower backdoor attack success rate (ASR) indicates more effective unlearning as data removal should eliminate learned backdoor patterns. Evaluation metrics include test accuracy (Acc) for utility preservation and ASR for unlearning effectiveness.

4.2 Evaluation

We compare our proposed method with three baseline approaches across three datasets with different unlearning ratios, evaluating both utility preservation and unlearning effectiveness, as shown in Table 1.

Utility Preservation Analysis. Our method demonstrates competitive utility preservation across all datasets. On MNIST, we maintain accuracy above 92% for both unlearning ratios, with performance comparable to baseline methods. As dataset complexity increases to FMNIST and CIFAR10, while all methods experience accuracy decline, our method continues to ensure good utility

Table 1. Utility preservation and unlearning effectiveness comparison of different FU methods across various unlearning ratios

		MNIST		FMNIST		CIFAR10	
		Acc ↑	ASR ↓	Acc ↑	ASR ↓	Acc ↑	ASR ↓
UR=5%	Retrain	0.9788	0.0128	0.8995	0.014	0.6862	0.0405
	GA	0.9480	0.0234	0.798	0.0461	0.6299	0.0256
	FedU	0.9516	0.0448	0.8571	0.0575	0.6107	0.0274
	Ours	0.9389	0.0000	0.8513	0.0089	0.6118	0.0025
UR=25%	Retrain	0.9724	0.0162	0.9003	0.0179	0.6534	0.0451
	GA	0.9147	0.0162	0.7767	0.0384	0.5168	0.0186
	FedU	0.9379	0.4093	0.8192	0.0415	0.5548	0.0351
	Ours	0.9256	0.0064	0.8227	0.0164	0.5923	0.0101

maintenance with smaller fluctuation ranges compared to GA and FedU, indicating better stability across different data complexities.

Unlearning Effectiveness Analysis. Our method consistently achieves superior unlearning effectiveness as measured by backdoor attack success rate (ASR). Most notably, we achieve perfect unlearning (zero ASR) on MNIST at 5% unlearning ratio. Across all experimental settings, our ASR values are substantially lower than those of other approaches, indicating that our approach effectively addresses feature dependencies and achieves excellent unlearning effectiveness.

Trade-off Discussion. Our results reveal a carefully balanced trade-off between utility preservation and unlearning effectiveness. While our method shows some accuracy reduction compared to other FU methods in certain cases, the differences remain acceptable, representing a reasonable compromise in utility preservation. In return, we achieve significantly better unlearning effectiveness. This demonstrates that our representation decoupling approach achieves more thorough data erasure without severe utility degradation, addressing the primary objective of federated unlearning.

5 Conclusion

In summary, we propose an adversarial framework for FU that achieves representation decoupling through mutual information minimization. Our approach features an unlearned model acting like a generator that learns to generate behaviors demonstrating effective unlearning, trained against a discriminator on target clients that performs adversarial evaluation. The discriminator includes an auxiliary T-network that estimates mutual information between original and unlearned model representations, guiding the unlearned model to minimize statistical dependence and achieve representation decoupling. Experimental results demonstrate superior unlearning effectiveness compared to existing methods, validating the effectiveness of our representation decoupling approach for FU.

Acknowledgment

This research is supported by the National Research Foundation, Singapore and Infocomm Media Development Authority under its Trust Tech Funding Initiative, and the Singapore Ministry of Education Academic Research Fund (MOE-T2EP20224-0009). Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore and Infocomm Media Development Authority.

References

1. Bagdasaryan, E., Veit, A., Hua, Y., Estrin, D., Shmatikov, V.: How to backdoor federated learning. In: International conference on artificial intelligence and statistics. pp. 2938–2948. PMLR (2020)
2. Belghazi, M.I., Baratin, A., Rajeshwar, S., Ozair, S., Bengio, Y., Courville, A., Hjelm, D.: Mutual information neural estimation. In: International conference on machine learning. pp. 531–540. PMLR (2018)
3. California State Legislature: California Consumer Privacy Act of 2018. Assembly Bill 375 (2018)
4. Cover, T.M.: Elements of information theory. John Wiley & Sons (1999)
5. Deng, L.: The MNIST database of handwritten digit images for machine learning research. IEEE signal processing magazine **29**(6), 141–142 (2012)
6. Donsker, M.D., Varadhan, S.S.: Asymptotic evaluation of certain markov process expectations for large time, i. Communications on pure and applied mathematics **28**(1), 1–47 (1975)
7. Halimi, A., Kadhe, S., Rawat, A., Baracaldo, N.: Federated unlearning: How to efficiently erase a client in FL? arXiv preprint arXiv:2207.05521 (2022)
8. Jiang, Y., Shen, J., Liu, Z., Tan, C.W., Lam, K.Y.: Towards efficient and certified recovery from poisoning attacks in federated learning. IEEE Transactions on Information Forensics and Security (2025)
9. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images (2009)
10. Li, Y., Chen, C., Zheng, X., Zhang, J.: Federated unlearning via active forgetting. arXiv preprint arXiv:2307.03363 (2023)
11. Liu, Y., Xu, L., Yuan, X., Wang, C., Li, B.: The right to be forgotten in federated learning: An efficient realization with rapid retraining. In: IEEE INFOCOM 2022-IEEE Conference on Computer Communications. pp. 1749–1758. IEEE (2022)
12. McMahan, B., Moore, E., Ramage, D., Hampson, S., y Arcas, B.A.: Communication-efficient learning of deep networks from decentralized data. In: Artificial intelligence and statistics. pp. 1273–1282. PMLR (2017)
13. Regulation, P.: Regulation (EU) 2016/679 of the european parliament and of the council. Regulation (EU) **679**(2016), 10–13 (2016)
14. Wang, W., Zhang, C., Tian, Z., Yu, S.: Fedu: Federated unlearning via user-side influence approximation forgetting. IEEE Transactions on Dependable and Secure Computing (2024)
15. Xiao, H., Rasul, K., Vollgraf, R.: Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms. arXiv preprint arXiv:1708.07747 (2017)