Federated Unlearning on Sound Data

October 22, 2024

Paolo Cacace

Abstract

With the increasing importance of privacy regulations such as the General Data Protection Regulation (GDPR)[1], machine learning systems must provide mechanisms to ”forget” specific user data when requested. This challenge becomes even more complex in Federated Learning (FL), where multiple decentralized clients collaboratively train a model without sharing their local data. This project is based on the methodology proposed by Halimi et al. (2022) in **Federated unlearning: How to efficiently erase a client in fl?**, but starts from sound data instead of images, a modality that presents different challenges due to its temporal and spectral properties. We evaluate the method's efficacy, fidelity, and efficiency performance, showing that it can effectively remove a client's contribution from the global model with minimal computational and communication overhead. The final results demonstrate the feasibility of federated unlearning in sound data classification tasks.

1. Introduction

Privacy concerns in machine learning (ML) systems have grown significantly with the advent of large-scale data collection and usage. Big companies often stay on the edge of regulations to gather as much user private information as possible. In recent years, with the population becoming more aware than ever and with the introduction of laws such as the GDPR, individuals are granted the ”right to be forgotten”, meaning that they can request their data to be removed from any system that stores or processes it. This gave rise to the concept of **machine unlearning**, initially introduced by (Cao & Yang, 2015), which seeks to selectively eliminate the influence of particular data points from trained models while preserving the model's overall performance.

The problem becomes even more challenging in the context of **Federated Learning**, where multiple clients (e.g., IoT devices or entire organizations) collaborate to train a shared model without directly sharing raw data. This decentralized nature of FL makes traditional unlearning methods impractical, as they often assume access to all training data, which is not available in FL setups.

In this project, we benchmarked the application of federated unlearning proposed by Halimi et al. (2022) on the Free Spoken Digit Dataset (FSDD)[2]. This procedure aims to remove a client's data from the global model after training has already occurred, without the need for costly retraining from scratch. FSDD introduces additional complexities due to its temporal and spectral characteristics, which require specific pre-processing methods. By adapting the federated unlearning procedure to this data modality, we aimed to assess the approach's efficacy in more realistic scenarios, such as speech recognition or environmental sound classification.

2. Related Work

Machine unlearning is an established research area that has rapidly advanced over the years by exploring various methods. One such method is Projected Gradient Ascent (PGA), first applied by (Golatkar et al., 2020), who introduced a selective forgetting mechanism to efficiently ”forget” unwanted training data. The concept of machine unlearning has also rapidly been extended to FL, where constraints are more stringent due to the communication overhead required to set up the infrastructure and the limited computational power of clients.

In the context of FL, forgetting a client typically involves storing all updates from each FL aggregation round and subsequently deleting them based on the client's historical contributions and the specific aggregation algorithm used, as demonstrated by Liu et al. (2021). However, this approach is not feasible in all setups because continuously storing client weights and global model updates may violate privacy legal, privacy, and security concerns. Conse-

Email: Paolo Cacace <paolo.cacace@cern.ch>.

Deep Learning and Applied AI 2024, Sapienza University of Rome, 2nd semester a.y. 2023/2024.

[1] The General Data Protection Regulation (GDPR) is a European Union law designed to protect individuals' data, giving them control over how their information is collected, used, and stored

[2] https://github.com/Jakobovski/free-spoken-digit-dataset

quently, alternative strategies that balance efficient unlearning with privacy constraints are necessary to address these challenges in federated environments.

3. Methods

3.1. Federated Unlearning Setup

We consider a FL scenario with N clients trained over T rounds. After training, a target client i is requested to exit the federation and remove its local data's influence from the global model. To validate the unlearning process, we inserted a backdoor into the data used to train the client we intend to forget. Our approach centers on approximate unlearning, aiming to match or surpass the performance achieved through retraining. This is accomplished by applying an unlearning procedure to the target client model. The resulting unlearned model is then used as the starting point for continued training within a FL setup.

3.2. Unlearning with Projected Gradient Descent

Our federated unlearning process comprises three phases:

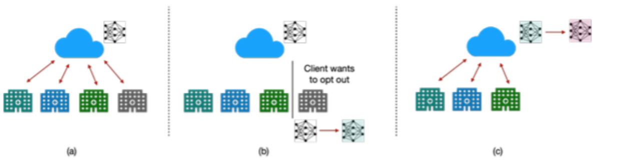

Figure 1. Phases of Federated Unlearning: (b) First Phase; (b) Second Phase; (a) Third Phase.

1. **Federated Learning (FedAvg):** Standard federated learning steps are conducted with the target client included in the federation.

2. **Local Unlearning:** After the first round of federated learning with the target client contributing to the global model, the training and aggregation is stopped, and the target client i reverses the training process by maximizing its empirical loss within an ℓ_2-norm ball of radius δ around a reference model w_{ref}, calculated as the average of the other clients' models. This is formulated as a constrained optimization problem and solved using PGD. This methodology implies the client updates the model's parameters iteratively and projects them back into the feasible region to avoid arbitrary model divergence.

3. **FL Post-Training:** After local unlearning, the server and remaining clients initiate a few additional FL rounds using the locally unlearned model w_u^i as a global model. This post-training step ensures that the unlearned model remains effective for the retained clients.

This two-phase approach efficiently removes the influence of the target client's data while maintaining model performance without storing all historical updates.

4. Results

To evaluate the unlearning procedure, standard federated averaging rounds were conducted in parallel from scratch (*Retrain*), providing a baseline of how the global model would have evolved without the influence of the target client. This reference allows for an effective comparison with the unlearning post-training (*UN*). The results are illustrated in Figure 2.

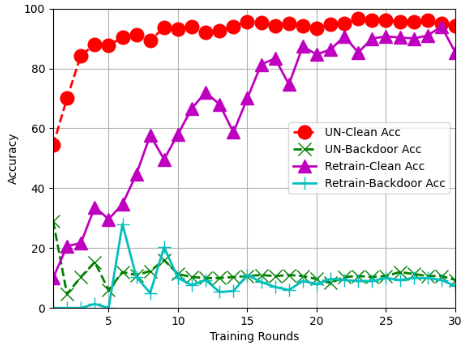

Figure 2. Unlearning Benchmark - Free Spoken Digit Dataset.

The plot shows the comparison of accuracy on a nominal test dataset (*Clean-Acc*) versus a poisoned test dataset (*Backdoor-Acc*) for both the Unlearned model and a Retrained one. Considering that the poisoned dataset contains 10% of uncompromised data, the results show that the model initialized with the unlearned model maintains strong accuracy from the start, consistently performing better than the retrained model that starts from from scratch. Moreover, the influence of the compromised data is reduced and stabilized within a few epochs, highlighting the effectiveness of the unlearning procedure.

5. Conclusions and Discussions

The results highlight how the described procedure allows forgetting a client's contribution in FL. However, the methodology reveals some issues. Specifically, during the unlearning phase, it is necessary to accurately calibrate the distance from the reference model, which is not the same for every dataset. To accurately determine this distance without setting it arbitrarily high, cause the model to diverge, and make the unlearning phase useless, it is essential to test it in the post-training phase. This procedure is challenging to fine-tune without the data we want to forget.

References

Cao, Y. and Yang, J. Towards making systems forget with machine unlearning. In *2015 IEEE symposium on security and privacy*, pp. 463–480. IEEE, 2015.

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

Halimi, A., Kadhe, S., Rawat, A., and Baracaldo, N. Federated unlearning: How to efficiently erase a client in fl? *arXiv preprint arXiv:2207.05521*, 2022.

Liu, G., Ma, X., Yang, Y., Wang, C., and Liu, J. Federaser: Enabling efficient client-level data removal from federated learning models. In *2021 IEEE/ACM 29th international symposium on quality of service (IWQOS)*, pp. 1–10. IEEE, 2021.