Paper Review for MVA DLMI Course: *Why Patient Data Cannot Be Easily Forgotten ?*[1]

Paul CHAUVIN, Rayane MOUHLI

March 5, 2023

1 Paper summary

Privacy is crucial in the Big Data era, particularly when it concerns medical data. In this context, forgetting knowledges from an AI model is a very important rising problem. To address this issue, researchers have proposed many machine unlearning/forgetting approaches [2][3][4][5] to remove sensitive informations without retraining [6] the model from scratch. However, patient-wise forgetting problem is more challenging than classic vision forgetting problem because patient's data seem to either form cluster or edge cases.

The baseline method is to retrain the model entirely but it would requires significant cost and efforts. Several methods have been proposed in the literature with different approaches as Bayesian model [7], k-means clustering [3] etc. This article is strongly inspired by [8] which presents the scrubbing method described in the next section. This method adds noise to model weights that are uninformative to the remaining data (training data excluding those we want to forget). The studied paper shows that although the scrubbing works well with vision datasets, it is less efficient with medical data. In this context, the authors propose an extension of the scrubbing, named the targeted forgetting method, which is adapted to medical images. This new method consists in adding weighted noises to informative weights to the patients we want to forget.

1.1 Presentation of the method

The main hypothesis of the paper is that patient's data can either be similar to other patient's data (forming clusters) or unique (forming edge cases). In both methods (scrubbing and targeted forgetting), they assume that the training dataset $\mathcal{D}_{train}$ is partitioned this way: $\mathcal{D}_{train} = \mathcal{D}_r \cup \mathcal{D}_f$ where $\mathcal{D}_f$ is the forgetting dataset which contains the images we want to remove from our model $A(\mathcal{D}_{train})$ and $\mathcal{D}_r$ is the retaining dataset (its complement in $\mathcal{D}_{train}$). Let S(w) denote the operations applied to model weights to forget D_f.

First, the authors assume that we can access to our model $A(\mathcal{D}_{train})$ and to the retaining dataset $\mathcal{D}_r$. They define the scrubbing function as:

$$S(w) = w + (\lambda\sigma_h^2)^{\frac{1}{4}} F_{\mathcal{D}_r}(w)^{-\frac{1}{4}}$$

The two hyperparameters are λ, the noise we want to add to the weights and σ_h, a normal distributed error term. Considering the product, we can tune them as one hyperparameter. $F_{\mathcal{D}_r}(w)$ denotes the Fisher Information Matrix (FIM) computed for w on $\mathcal{D}_r$. So if a weight is not very informative about D_r, $F_{\mathcal{D}_r}(w)$ will take low values. Then, thanks to the negative power, it will add a strong noise to this weight.

However, scrubbing an edge case from a model can affect negatively the model. Indeed, this method will add noise to model weights corresponding to most of the edge cases in the remaining dataset instead of removing the edge-case patient's informations we want to forget only. Since many of patient's data can be considered as edge-cases (this hypothesis is discussed in the experiments), the authors improve the scrubbing method to adapt it for medical use cases.

The targeted forgetting method is pretty similar to the previous one since it still relies on the FIM, but instead of keeping the most informative weights corresponding to $\mathcal{D}_r$, this new method adds noise to model weights highly informative about $\mathcal{D}_f$. The targeted forgetting function is:

$$S(w) = w + (\lambda\sigma_h^2)^{\frac{1}{4}} F_{\mathcal{D}_f}(w)^{\frac{1}{4}}$$

1.2 Experiments

The paper investigates the effectiveness of scrubbing and targeted forgetting methods on CIFAR-10 and Automated Cardiac Diagnosis Challenge (ACDC) datasets. For ACDC, they consider 90 patients and the goal is to remove 1 patient from the training dataset. For CIFAR-10, they remove 10 non-overlapping sets, each with 100 images from same class. The implementation uses a VGG-like architecture, Cross Entropy loss, Adam optimizer and data augmentation for the training.

First, the article compares the difficulty of forgetting patient's data in ACDC and in CIFAR-10. The results of the classification error of D_f after applying the scrubbing process vary from 0% to 100% for ACDC, and from 10% to 25% from CIFAR-10 which shows that the scrubbing method cannot generalize to this patient's data. By considering a 50% threshold on the error of the model after forgetting, the study also shows that over 60% of patients in ACDC can be considered as edge cases. Scrubbing can degrade model performance when dealing with edge cases, which explains its underperformance in ACDC.

Then, the authors use four representative patients from ACDC (two under the edge case hypothesis and two under the cluster hypothesis) and adjust the noise strength to achieve different levels of forgetting. The metrics used to measure the errors on $\mathcal{D}_f$ and $\mathcal{D}_{test}$ is 1-Accuracy. They show that the targeted forgetting method maintains good model generalisation performance on D_{test} at all noise levels while the scrubbing method tends to degrade it. Both methods can achieve standard forgetting with a low level of noise with nice model's generalisation performance on D_{test} for common cluster cases. The study concludes that for edge cases, patient-wise data can be completely forgotten without sacrificing the model generalisation performance using targeted forgetting. However, for common cluster cases, it is less likely to forget the patient data as completely forgetting will result in significantly degraded generalisation performance with both methods.

2 Critics

The main strength of this paper is the improvement of an existing method to make it more efficient in the medical field. Experiments show that more than 60% of the patients from ACDC can be considered as edge cases and that the scrubbing method is less efficient in this framework than the targeted forgetting method.

However, we can still challenge this paper on several aspects. First of all, we notice that the paper doesn't provide any theoretical proof about the proposed method. Mainly, it relies on the hypothesis that the data are organized in edge cases or in a common cluster but there is no clear justification about that. They just did some tests on the two datasets considering one hypothesis then the other, resulting in the conclusion that the edge case hypothesis sounds more consistent according to their tests.

In the case where this hypothesis is true, we still don't know which case a new dataset we want to work on belongs to. To determine that, one possible amelioration could be an additional preprocessing step which consists in doing a clustering on our data with a clustering which doesn't require the number of cluster like a DBScan. If we have an important number of cluster, we can consider that we are in the edge case hypothesis, if this number is low we are in the common cluster hypothesis.

Furthermore, the targeted forgetting method is only compared to the scrubbing method presented in [8]. It could be interesting to compare it with other machine unlearning methods, for example, with [9] which is an improvement of the scrubbing method by modifying the final activations and weight dynamics.

Lastly, one goal of this paper is to deal with medical data but the authors only worked with the ACDC dataset. It would be interesting to experiment the edge case hypothesis and the performance of the model on different medical datasets.

References

[1] Ruolin Su, Xiao Liu, and Sotirios A. Tsaftaris. Why patient data cannot be easily forgotten?, 2022.

[2] Yinzhi Cao and Junfeng Yang. Towards making systems forget with machine unlearning. In *2015 IEEE Symposium on Security and Privacy*. IEEE, May 2015.

[3] Antonio Ginart, Melody Y. Guan, Gregory Valiant, and James Zou. Making ai forget you: Data deletion in machine learning, 2019.

[4] Ayush Sekhari, Jayadev Acharya, Gautam Kamath, and Ananda Theertha Suresh. Remember what you want to forget: Algorithms for machine unlearning, 2021.

[5] Yinzhi Cao and Junfeng Yang. Towards making systems forget with machine unlearning. In *2015 IEEE Symposium on Security and Privacy*, pages 463–480, 2015.

[6] Ben Morris. The components of the wired spanning forest are recurrent. *Probability Theory and Related Fields*, 125(2):259–265, February 2003.

[7] Quoc Phong Nguyen, Bryan Kian Hsiang Low, and Patrick Jaillet. Variational bayesian unlearning. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 16025–16036. Curran Associates, Inc., 2020.

[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)*, June 2020.

[9] Aditya Golatkar, Alessandro Achille, and Stefano Soatto. Forgetting outside the box: Scrubbing deep networks of information accessible from input-output observations, 2020.