Why Patient Data Cannot Be Easily Forgotten ?

Ruolin Su, Xiao Liu, and Sotirios A. Tsaftaris.

Presented by Paul Chauvin and Rayane Mouhli (Pair n°12)

MATHÉMATIQUES VISION APPRENTISSAGE

Context and objective

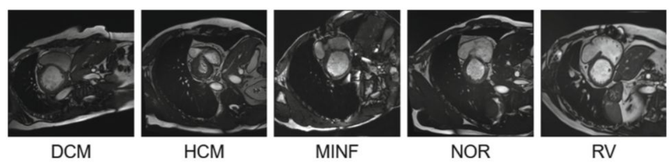
Privacy is crucial in the Big Data era, particularly **with medical data**. A rising problem is forgetting knowledge from an AI model. To address this issue, researchers have proposed **machine unlearning/forgetting approaches to remove private information without re-training** the model from scratch. However, **patient-wise forgetting is challenging because patient's data are either forming cluster or forming edge cases.** Their work is strongly inspired by A. Golatkar et al. paper [1] which presents the **scrubbing method**: Their study show that although the scrubbing works well with vision datasets, it is less efficient with medical data. The objective of this work is to show that **targeted forgetting method** improves the performance of the scrubbing method on medical datasets.

Methodology

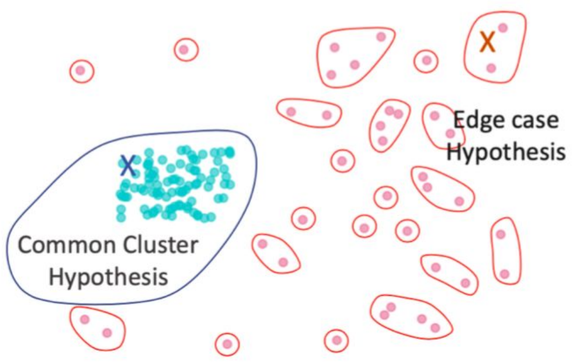
Two datasets for the experiments

60.000 images of 10-class objects: **CIFAR-10**

Medical specific 4D cardiac data from 100 patients: **Automated Cardiac Diagnosis Challenge (ACDC)**



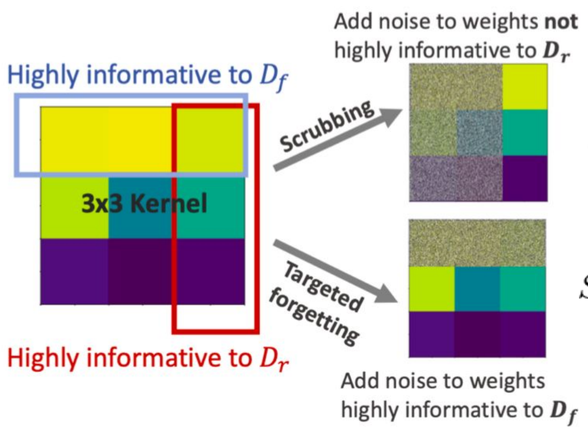
Two hypotheses on patient-wise data:



Common cluster hypothesis: Patient's data is similar to other patient's data

Edge case hypothesis: Patient's data is unique and rare

Improving scrubbing for medical imaging: the targeted forgetting method



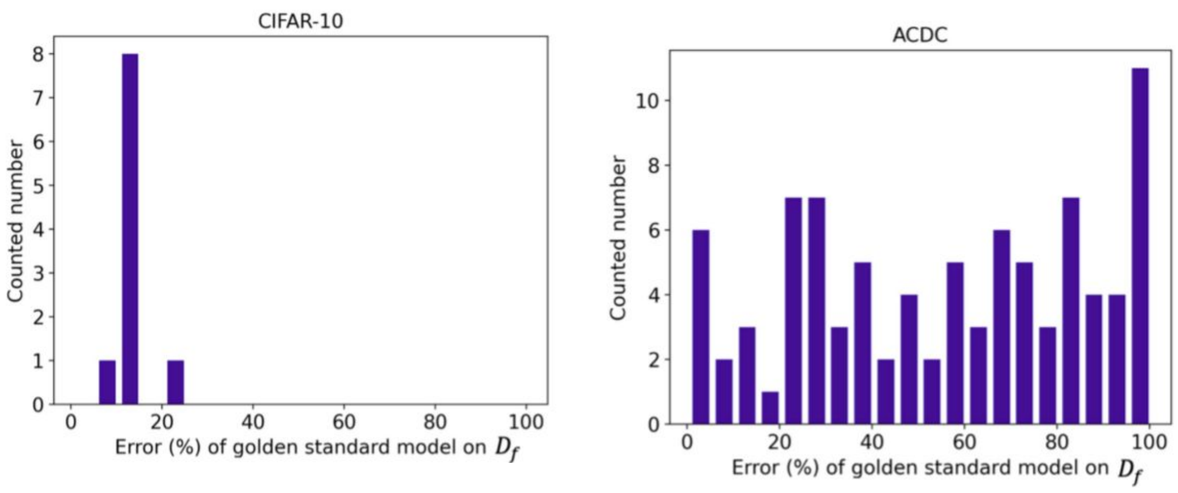
$$S(w) = w + (\lambda\sigma_h^2)^{\frac{1}{4}} F_{\mathcal{D}_r}(w)^{-\frac{1}{4}}$$

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

Notations :

- S(w): operations done on the weights of the model
- Df/Dr: Subset of the training dataset we want to forget/retain
- F: Fisher Information Matrix
- λ,σ : two hyperparameters that control the level of noise

Results



In these histograms, the y-axis refers to the total number of patients/sets whose golden standard (the "perfect" model after forgetting) lies within an interval of x-axis.

For **ACDC,** the authors **removed 1 patient among 90** and retrain a model on the remaining 89 to be the golden standard model. Then, they measure the error of the deleted patient on this model and they repeat the process for the 90 patients. For **CIFAR-10,** they select **10 non-overlapping sets,** each with 100 images from the same class, to be removed and repeat the retrain experiments.

In **CIFAR-10**, the **error is pretty low** which indicates we are unlikely to be in the edge case hypothesis, while ACDC seems to fit with this hypothesis. This can explain the fact that **scrubbing doesn't work well on ACDC** since it fall under the **edge case hypothesis.**

Patient ID	Error on	Golden standard	Noise level					
			Low		Medium		High	
			Scrubbing	Targeted forgetting	Scrubbing	Targeted forgetting	Scrubbing	Targeted forgetting
94 (Edge)	$\mathcal{D}_f$	1.000 ± 0.000	0.154 ± 0.005	0.174 ± 0.020	0.859 ± 0.010	0.851 ± 0.018	1.000 ± 0.000	1.000 ± 0.000
	$\mathcal{D}_{test}$	0.237 ± 0.002	0.671 ± 0.012	0.223 ± 0.011	0.739 ± 0.007	0.291 ± 0.005	0.746 ± 0.008	0.316 ± 0.002
5 (Edge)	$\mathcal{D}_f$	0.809 ± 0.009	0.127 ± 0.022	0.121 ± 0.019	0.853 ± 0.020	0.857 ± 0.002	0.997 ± 0.003	1.000 ± 0.000
	$\mathcal{D}_{test}$	0.253 ± 0.026	0.394 ± 0.017	0.269 ± 0.004	0.624 ± 0.015	0.407 ± 0.001	0.696 ± 0.002	0.506 ± 0.002
13 (Cluster)	$\mathcal{D}_f$	0.202 ± 0.004	0.111 ± 0.006	0.092 ± 0.002	0.871 ± 0.018	0.850 ± 0.021	1.000 ± 0.000	1.000 ± 0.000
	$\mathcal{D}_{test}$	0.194 ± 0.012	0.361 ± 0.001	0.343 ± 0.007	0.590 ± 0.005	0.524 ± 0.013	0.694 ± 0.004	0.602 ± 0.016
9 (Cluster)	$\mathcal{D}_f$	0.010 ± 0.002	0.176 ± 0.005	0.152 ± 0.009	0.892 ± 0.003	0.862 ± 0.005	0.998 ± 0.002	0.995 ± 0.005
	$\mathcal{D}_{test}$	0.233 ± 0.007	0.402 ± 0.012	0.442 ± 0.001	0.643 ± 0.006	0.613 ± 0.001	0.699 ± 0.005	0.656 ± 0.001

This table presents the forgetting results for 4 patients (Error = 1 - Accuracy)

Is Targeted Forgetting Better for Forgetting Edge Cases?

Both methods can achieve forgetting at high levels of noise for edge cases but **scrubbing significantly degrades the model's generalization** performance. In contrast, **targeted forgetting maintains good model generalization performance on test data at all noise levels**. The table below shows the average noise value added to weights to achieve 1.00 error on Df. It needs **higher value of weights for the scrubbing methods to forget an edge case**

Patient ID	High	
	Scrubbing	Targeted forgetting
94 (Edge)	2.33E−05	3.00E−06
5 (Edge)	1.65E−05	4.5E−06
13 (Cluster)	1.6E−05	8.66E−06
9 (Cluster)	1.43E−05	1.2E−05

Is Targeted Forgetting Better for Forgetting Common Cluster Cases?

Both methods can achieve standard forgetting with a low level of noise, resulting in good model generalization on test data. However, **when the noise level is increased to forget more, both methods sacrifice model generalization** and produce similar high test error values.

Can Patient Data be Completely Forgotten?

Targeted forgetting can completely forget patient data for edge cases without sacrificing model generalization performance, while **for common cluster cases, forgetting patient data significantly degrades generalization performance**. The trade-off between model performance and data protection is affected by the level of noise added to the model weights, which needs to be carefully designed to achieve a balance between forgetting and generalization performance.

Conclusion and further work

In conclusion, the study reveals that forgetting patient-specific medical data poses a **greater challenge than other vision domains.** This difficulty arises due to the presence of data falling under **two hypotheses: common cluster and edge case.** They have introduced a **new targeted forgetting approach that overcomes the limitations of the existing state-of-the-art scrubbing method**. Their experiments have demonstrated the significance of dataset bias and the distinct roles played by these two hypotheses. Their experiments have been conducted on cardiac MRI data but it could have been interesting to study if their approach still perform on **various medical datasets**. Future researches could focus on **developing methods to detect the hypothesis to which the patient's data belongs to** and measuring patient-wise forgetting performance while considering these two hypotheses.

Reference

[1] 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.

The University of Edinburgh · The Alan Turing Institute · MATHÉMATIQUES VISION APPRENTISSAGE · université PARIS-SACLAY · école normale supérieure paris-saclay