LOSS-FREE MACHINE UNLEARNING

Jack Foster[*1,2], **Stefan Schoepf**[*1,2] **& Alexandra Brintrup**[1,2]
[1]Department of Engineering, University of Cambridge
[2]The Alan Turing Institute
*Indicates equal contribution
{jwf40,ss2823,ab702}@cam.ac.uk

ABSTRACT

We present a machine unlearning approach that is both retraining- and label-free. Most existing machine unlearning approaches require a model to be fine-tuned to remove information while preserving performance. This is computationally expensive and necessitates the storage of the whole dataset for the lifetime of the model. Retraining-free approaches often utilise Fisher information, which is derived from the loss and requires labelled data which may not be available. Thus, we present an extension to the Selective Synaptic Dampening algorithm, substituting the diagonal of the Fisher information matrix for the gradient of the l_2 norm of the model output to approximate sensitivity. We evaluate our method in a range of experiments using ResNet18 and Vision Transformer. Results show our label-free method is competitive with existing state-of-the-art approaches.

1 INTRODUCTION & RELATED WORK

With increasing regulations around machine learning, the ability for a trained model to forget private or harmful information is essential. This task is typically formulated by splitting some training dataset $\mathcal{D}$ into a subset of data that needs to be forgotten $\mathcal{D}_f$, and the retained data $\mathcal{D}_r = \mathcal{D} \setminus \mathcal{D}_f$. The objective is to forget $\mathcal{D}_f$ from the model, while simultaneously preserving model performance on $\mathcal{D}_r$ (Cao & Yang, 2015). To solve this problem, several methods have been proposed that disrupt the model's functional mapping for samples in $\mathcal{D}_f$ (e.g. by learning random labels), while also training for a small number of epochs over $\mathcal{D}_r$ to repair or protect performance (Chundawat et al., 2023a;b; Graves et al., 2021; Tarun et al., 2023; Kurmanji et al., 2023). This is computationally expensive and requires the storage of the full dataset permanently. In contrast, recent methods such as FisherForgetting from Golatkar et al. (2020) or Selective Synaptic Dampening (SSD) from Foster et al. (2023) have explored retraining-free approaches, with the latter achieving state-of-the-art performance. These often involve estimating the importance of parameters for the forget set, allowing them to be disrupted through noise or dampening, but without destroying performance on the retain set. Parameter importance estimation is typically achieved via the diagonal of the Fisher information matrix (FIM), which in practice is calculated as the square of the first-order gradients of the loss. This requires access to a ground truth label, which may not always be available. Therefore, we propose a novel extension of SSD that does not require access to the loss or ground truth labels, which we refer to as Loss-Free Selective Synaptic Dampening (LFSSD). LFSSD replaces the Fisher-based importance estimation with an approximation of the sensitivity of the model to perturbations of each parameter, which may be interpreted as parameter importance. Such importance estimation has been shown to be effective in continual learning (Aljundi et al., 2018), yet it has not been applied in unlearning. Here, we show this new approach to dampening important parameters is equally effective as Fisher information, without relying on labelled data. To the best of our knowledge, LFSSD is the first retraining-free unlearning method to require only the unlabelled forget samples at forgetting time.

Table 1: Full- (class lamp) and subclass (lamp is part of electrical devices) unlearning with α =10, λ =1 on Cifar100/20. (LF)SSD is deterministic, '±0.0' omitted. Full benchmarks in the appendix.

		ResNet18				Vision Transformer			
data	metric	baseline	retrain	LFSSD	SSD	baseline	retrain	LFSSD	SSD
Cifar100	$\mathcal{D}_r$	76.39±0.00	72.89±0.34	76.36	76.08	88.84±0.00	90.10±0.19	89.03	89.06
	$\mathcal{D}_f$	70.49±0.00	0.00±0.00	0.00	0.00	97.22±0.00	0.00±0.00	0.00	36.89
	MIA	92.40±0.00	0.32±0.00	0.00	0.20	95.60±0.00	2.27±0.01	2.20	0.40
Cifar20	$\mathcal{D}_r$	82.61±0.00	81.81±0.19	81.54	79.58	95.77±0.00	94.69±0.13	95.68	95.54
	$\mathcal{D}_f$	72.66±0.00	22.95±4.72	7.29	0.00	89.58±0.00	34.55±8.62	24.05	14.58
	MIA	83.40±0.00	4.29±0.01	3.40	2.00	81.00±0.00	5.60±0.02	5.80	3.2

2 METHODS & RESULTS

We build upon the approach presented in SSD (Foster et al., 2023). Given the parameter importance over both $\mathcal{D}$ and $\mathcal{D}_f$, we can select parameters that are disproportionately important for the forget set, and reduce their magnitude to induce forgetting in the model:

$$\theta_i = \begin{cases} \beta\theta_i, & \text{if } []_{\mathcal{D}_{f,i}} > \alpha []_{\mathcal{D},i} \\ \theta_i, & \text{if } []_{\mathcal{D}_{f,i}} \leq \alpha []_{\mathcal{D},i} \end{cases} \quad \forall i \in [0, |\theta|] \qquad \text{where } \beta = min(\frac{\lambda []_{\mathcal{D},i}}{[]_{\mathcal{D}_{f,i}}}, 1) \tag{1}$$

Here, α and λ are hyper-parameters, and $[]_{\mathcal{D}_{f,i}}$ and $[]_{\mathcal{D}_i}$ are the ith element of the diagonal of the FIM, calculated over $\mathcal{D}_f$ and $\mathcal{D}$, respectively. Fisher information is calculated over $\mathcal{D}$ not $\mathcal{D}_r$, meaning this is only calculated once, and then the train-set may be discarded after. We replace the Fisher information with the sensitivity estimation from Aljundi et al. (2018), which we now introduce. Given some neural network output, $f(x;\theta)$, and some small perturbation $\delta = \{\delta_i\}$ in the parameters $\theta = \{\theta_i\}$, we can approximate the change in the neural network's output by:

$$f(x;\theta+\delta) - f(x;\theta) \approx \sum_i \frac{\partial f(x;\theta)}{\partial \theta_i}\delta_i \tag{2}$$

The importance of a parameter can be measured by how much $f(x;\theta)$ changes, given a small perturbation to the parameter. Assuming δ_i is a small, constant change, this is equivalent to the magnitude of the gradient. For multi-dimensional outputs, this gradient must be calculated over every dimension which is expensive, and so instead the gradient can be calculated over the squared l_2 norm of the output. Putting this together, we arrive at the final importance estimation equation:

$$\Omega_i = \frac{1}{N}\sum_{k=1}^{N} \| \frac{\partial [l_2^2(f(x_k;\theta))]}{\partial \theta_i} \| \tag{3}$$

Which we can calculate over $\mathcal{D}$ and $\mathcal{D}_f$ and substitute into equation 1 in place of $[]$ and $[]_f$. Results can be seen in Table 1, with further results available in the appendix. The objective is to maximise the $\mathcal{D}_r$ accuracy while having a low membership inference attack score (MIA) similar to the retrain column (Shokri et al., 2017). Similarity to the retrained model is necessary, as "over-forgetting" can lead to the *Streisand effect* (e.g., labelling a cat as a boat is an obvious modification) (Chundawat et al., 2023a). The results show that LFSSD produces scores similar to that of a model trained on only $\mathcal{D}_r$, and is competitive with state-of-the-art methods such as SSD, while being the first method to require no labelled data and no retraining. A shared limitation with SSD is the reliance on choosing α and λ values that balance forgetting and model performance.

3 CONCLUSION

Developing lightweight yet effective machine unlearning approaches is a significant open problem. In this work, we presented a novel extension of existing literature that is the first retraining free machine unlearning method to require only the unlabelled input samples for the forget set. By not requiring the long-term storage of the retain dataset, a labelled forget set, or the loss over the forget set, LFSSD is lightweight, robust and more practically useful than its predecessors.

URM STATEMENT

The authors acknowledge that at least one key author of this work meets the URM criteria of ICLR 2024 Tiny Papers Track

REFERENCES

Rahaf Aljundi, Francesca Babiloni, Mohamed Elhoseiny, Marcus Rohrbach, and Tinne Tuytelaars. Memory aware synapses: Learning what (not) to forget. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 139–154, 2018.

Burak. Pinterest face recognition dataset. www.kaggle.com/datasets/hereisburak/pins-face-recognition, 2020.

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

Vikram S Chundawat, Ayush K Tarun, Murari Mandal, and Mohan Kankanhalli. Can bad teaching induce forgetting? unlearning in deep networks using an incompetent teacher. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 7210–7217, 2023a.

Vikram S Chundawat, Ayush K Tarun, Murari Mandal, and Mohan Kankanhalli. Zero-shot machine unlearning. *IEEE Transactions on Information Forensics and Security*, 2023b.

Jack Foster, Stefan Schoepf, and Alexandra Brintrup. Fast machine unlearning without retraining through selective synaptic dampening. *arXiv preprint arXiv:2308.07707*, 2023.

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.

Laura Graves, Vineel Nagisetty, and Vijay Ganesh. Amnesiac machine learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 11516–11524, 2021.

Alex Krizhevsky and Geoff Hinton. Convolutional deep belief networks on cifar-10. *Unpublished manuscript*, 40(7):1–9, 2010.

Meghdad Kurmanji, Peter Triantafillou, and Eleni Triantafillou. Towards unbounded machine unlearning. *arXiv preprint arXiv:2302.09880*, 2023.

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)*, pp. 3–18. IEEE, 2017.

Ayush K Tarun, Vikram S Chundawat, Murari Mandal, and Mohan Kankanhalli. Fast yet effective machine unlearning. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

A APPENDIX

Benchmarks. We evaluate LFSSD in the same manner as Chundawat et al. (2023a); Foster et al. (2023), comparing against a baseline model without any unlearning, a model retrained from scratch on only $\mathcal{D}_r$, a model finetuned for 2 epochs on $\mathcal{D}_r$, and several existing state-of-the-art methods (Chundawat et al., 2023a; Graves et al., 2021; Foster et al., 2023) (teacher, amnesiac, SSD). Three key tasks are evaluated using ResNet18 (RN) and Vision Transformer (ViT): full class, subclass, and random forgetting. Fullclass entails forgetting an entire class from the dataset (e.g. rocket from Cifar100), subclass involves forgetting a related subclass from the dataset (e.g. all rockets from class Vehicle2 in Cifar20), and random forgetting selects a random subset of samples across all classes. Datasets used include the Cifar suite (Krizhevsky & Hinton, 2010), and the PINS face recognition dataset (Burak, 2020). Runs are repeated 10 times for each method, with the mean and standard deviation shown in the tables. D_r and D_f in the tables denote the accuracy on the respective subset of data in percent. Since SSD (and therefore LFSSD) is deterministic, there is no standard deviation

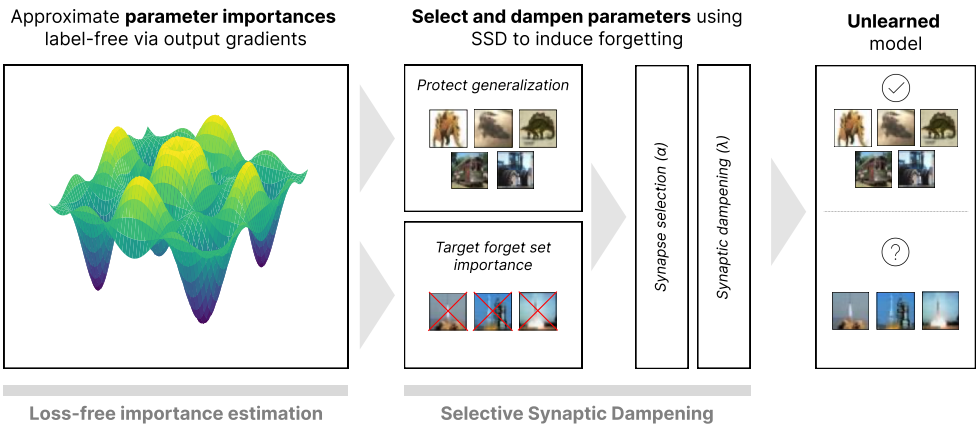

Figure 1: Loss-free parameter importance estimation as a replacement for the label-dependent parameter importance estimation of SSD resulting in the LFSSD method.

across repeated runs. We use the same membership inference attack as in Chundawat et al. (2023a); Foster et al. (2023), where logistic regression is used to classify samples as seen or unseen for a given model, based the entropy of the model outputs for that sample.

Resources. A single backward pass over the full dataset D is $O(b_{full} \cdot p)$ where b_{full} is the number of batches in the dataset and p is the number of parameters in the model. Similarly, the complexity for the forget and retain set is $O(b_{forget} \cdot p)$ and $O(b_{retain} \cdot p)$, respectively. Note that $b_{retain} \approx b_{full} >> b_{forget}$. SSD requires only one computation of the gradients for D (can be saved and reused), then for every forget request it requires one computation for D_f to then apply the dampening in $O(b_{forget} \cdot p)$.

LFSSD has the same complexity as SSD with the difference being that LFSSD calculates the norm of the output instead of the loss. In practice, this is similar in compute time.

Other methods are retraining-based and depend on the number of epochs N, generally on the order of $O(N \cdot b_{retain} \cdot p)$, which is much larger.

Sensitivity analysis. The values for alpha and lambda for SSD are taken from Foster et al. (2023) where the authors used Optuna to search for ideal parameters. For LFSSD we performed a simplified hyperparameter search keeping lambda fixed at 1 and alpha values [1, 50]. We pick one value per unlearning task (e.g. full class, sub-class, random) to show robustness across classes and do not use the best alpha per individual class/subclass (e.g., taking alpha 25 for rocket, alpha 5 for another class, etc.). A sensitivity analysis in Table 2 on CIFAR100 rocket class unlearning shows that there is a significant drop off in performance when picking an alpha that is too aggressive or conservative but the method has a wide plateau in which it performs well.

Unlearning scenarios without labels. For cases where labels are hard to get we expect two main scenarios: First, having a catalogue of undesired/copyrighted data (e.g., images) that we want to unlearn from the model just to be on the safe side legally even if it might not be in the model. In this case, we do not have the labels. Second, models that were trained with semi-supervised learning could benefit from our method when we do not want to rely on the accuracy of pseudo labels.

Table 2: Sensitivity analysis for LFSSD parameter α with $\lambda = 1$ on CIFAR100 rocket class unlearning. LFSSD is deterministic, thus no $\pm$ reported.

α	D_r Accuracy	D_f Accuracy	MIA
1	16.87%	0.00%	26.60
5	74.72%	0.00%	0.00
10	75.67%	0.00%	0.00
25	76.10%	0.00%	0.00
50	76.32%	71.88%	29.40

Table 3: Face unlearning (class unlearning): One face unlearned per experiment [ID 1,10,20,30,40] and results aggregated for 5 experiments as in Foster et al. (2023). SSD: $\alpha = 50$, $\lambda = 0.1$; LFSSD: $\alpha = 10$, $\lambda = 1$

	metric	baseline	retrain	finetune	teacher	amnesiac	SSD	LFSSD
RN	$\mathcal{D}_r$	98.52±0.02	100.00±0.00	99.72±0.45	96.72±0.44	99.99±0.02	98.42±0.13	98.15±0.20
	$\mathcal{D}_f$	97.84±1.99	0.00±0.00	4.32±4.61	0.13±0.4	0.00±0.00	0.00±0.00	0.00±0.00
	MIA	34.38±0.23	0.00±0.00	0.80±0.01	0.02±0.00	8.92±0.03	1.11±0.01	0.38±0.58

Table 4: Class unlearning on Cifar100. SSD: $\alpha_{RN} = 10$, $\alpha_{ViT} = 10$, $\lambda = 1$; LFSSD: $\alpha = 10$, $\lambda = 1$. MR: mushroom, Rkt: Rocket.

	class	metric	baseline	retrain	finetune	teacher	amnesiac	SSD	LFSSD
RN	baby	$\mathcal{D}_r$	76.38±0.00	73.10±0.55	64.03±0.8	74.71±0.19	73.69±0.27	44.05*±0.00	75.80±0.00
		$\mathcal{D}_f$	72.48±0.00	0.00±0.00	0.00±0.00	0.08±0.25	0.00±0.00	0.00±0.00	0.00±0.00
		MIA	92.60±0.00	2.44±0.01	24.52±0.07	0.00±0.00	50.86±0.05	5.40±0.00	0.00±0.00
	lamp	$\mathcal{D}_r$	76.39±0.00	72.89±0.34	64.01±0.63	74.76±0.19	73.52±0.48	76.08±0.00	76.36±0.00
		$\mathcal{D}_f$	70.49±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00
		MIA	92.40±0.00	0.32±0.00	12.92±0.04	0.00±0.00	46.24±0.04	0.20±0.00	0.00±0.00
	MR	$\mathcal{D}_r$	76.28±0.00	72.90±0.45	63.97±0.67	74.53±0.26	73.56±0.48	75.59±0.00	76.28±0.00
		$\mathcal{D}_f$	80.12±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00
		MIA	95.20±0.00	0.22±0.00	12.98±0.03	0.00±0.00	46.48±0.04	0.20±0.00	0.00±0.00
	Rkt	$\mathcal{D}_r$	76.27±0.00	72.83±0.42	64.05±0.88	74.53±0.26	73.34±0.45	74.54±0.00	75.67±0.00
		$\mathcal{D}_f$	80.90±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00
		MIA	93.40±0.00	1.04±0.00	13.70±0.04	0.00±0.00	29.56±0.02	2.20±0.00	0.00±0.00
	sea	$\mathcal{D}_r$	76.23±0.00	72.83±0.54	63.80±1.36	74.56±0.18	73.14±0.42	73.56±0.00	75.43±0.00
		$\mathcal{D}_f$	85.85±0.00	0.00±0.00	0.00±0.00	0.08±0.25	0.00±0.00	0.00±0.00	0.00±0.00
		MIA	93.40±0.00	5.84±0.02	26.54±0.08	0.02±0.00	29.50±0.03	0.60±0.00	0.00±0.00
ViT	baby	$\mathcal{D}_r$	88.93±0.00	90.27±0.15	80.74±1.38	87.48±0.41	88.43±0.71	88.59±0.00	88.12±0.00
		$\mathcal{D}_f$	90.19±0.00	0.00±0.00	0.00±0.00	23.80±22.51	0.00±0.00	0.00±0.00	0.00±0.00
		MIA	75.60±0.00	21.53±0.03	26.77±0.13	0.00±0.00	1.83±0.00	0.60±0.00	4.00±0.00
	lamp	$\mathcal{D}_r$	88.84±0.00	90.10±0.19	80.25±1.48	87.50±0.43	88.43±0.6	89.06±0.00	89.03±0.00
		$\mathcal{D}_f$	97.22±0.00	0.00±0.00	0.36±0.89	25.25±12.5	0.00±0.00	36.89±0.00	0.00±0.00
		MIA	95.60±0.00	2.27±0.01	11.77±0.04	0.13±0.00	2.70±0.00	0.40±0.00	2.20±0.00
	MR	$\mathcal{D}_r$	88.87±0.00	90.02±0.22	81.14±0.79	87.42±0.41	88.34±0.72	88.82±0.00	88.93±0.00
		$\mathcal{D}_f$	94.88±0.00	0.00±0.00	2.33±2.37	12.82±5.92	0.00±0.00	0.00±0.00	0.00±0.00
		MIA	92.80±0.00	0.70±0.00	7.10±0.02	0.03±0.00	0.47±0.00	3.80±0.00	2.60±0.00
	Rkt	$\mathcal{D}_r$	88.88±0.00	90.07±0.09	80.82±1.37	87.46±0.53	87.92±0.89	88.90±0.00	88.86±0.00
		$\mathcal{D}_f$	94.70±0.00	0.00±0.00	0.46±0.72	4.20±5.24	0.00±0.00	0.00±0.00	0.00±0.00
		MIA	94.40±0.00	3.23±0.00	19.00±0.09	0.03±0.00	1.00±0.01	1.80±0.00	10.60±0.00
	sea	$\mathcal{D}_r$	88.91±0.00	90.27±0.21	80.82±1.36	87.72±0.22	88.25±0.29	87.95±0.00	88.67±0.00
		$\mathcal{D}_f$	90.54±0.00	0.00±0.00	0.00±0.00	51.13±17.37	0.00±0.00	0.00±0.00	0.00±0.00
		MIA	80.40±0.00	8.43±0.02	21.97±0.07	0.00±0.00	0.77±0.00	3.20±0.00	4.20±0.00

Table 5: Class unlearning on Cifar20. SSD: $\alpha_{RN} = 10$, $\alpha_{ViT} = 10$, $\lambda = 1$; LFSSD: $\alpha = 5$, $\lambda = 1$. ED: electrical devices, NS: natural scenes, veg: vegetables, Veh2: vehicle2.

	class	metric	baseline	retrain	finetune	teacher	amnesiac	SSD	LFSSD
RN	ED	$\mathcal{D}_r$	82.56±0.00	82.13±0.23	73.19±1.08	82.04±0.29	81.34±0.18	83.15±0.00	83.19±0.00
		$\mathcal{D}_f$	82.26±0.00	0.00±0.00	0.00±0.00	10.88±1.3	0.00±0.00	1.76±0.00	0.00±0.00
		MIA	89.56±0.00	8.91±0.01	30.37±0.05	0.00±0.00	7.02±0.01	4.16±0.00	3.80±0.00
	NS	$\mathcal{D}_r$	82.10±0.00	81.33±0.22	72.03±1.63	81.36±0.27	80.70±0.4	82.33±0.00	82.44±0.00
		$\mathcal{D}_f$	91.08±0.00	0.00±0.00	0.00±0.00	10.91±2.16	0.00±0.00	0.00±0.00	0.00±0.00
		MIA	88.68±0.00	3.77±0.01	17.60±0.03	0.00±0.00	3.71±0.01	3.28±0.00	2.68±0.00
	people	$\mathcal{D}_r$	82.11±0.00	81.20±0.19	72.46±1.01	81.31±0.16	80.64±0.34	82.31±0.00	82.08±0.00
		$\mathcal{D}_f$	90.70±0.00	0.00±0.00	0.00±0.00	1.07±0.48	0.00±0.00	0.00±0.00	0.00±0.00
		MIA	91.72±0.00	1.36±0.00	16.20±0.04	0.00±0.00	6.28±0.01	1.12±0.00	1.36±0.00
	veg	$\mathcal{D}_r$	82.31±0.00	81.39±0.21	71.42±1.32	81.46±0.3	81.01±0.33	82.38±0.00	82.47±0.00
		$\mathcal{D}_f$	86.90±0.00	0.00±0.00	0.00±0.00	2.67±1.35	0.00±0.00	0.00±0.00	0.00±0.00
		MIA	89.52±0.00	9.74±0.01	29.39±0.08	0.00±0.00	5.00±0.01	16.96±0.00	5.44±0.00
	Veh2	$\mathcal{D}_r$	82.69±0.00	82.11±0.19	73.50±0.86	81.96±0.21	81.13±0.3	82.97±0.00	82.72±0.00
		$\mathcal{D}_f$	80.41±0.00	0.00±0.00	0.00±0.00	3.62±1.07	0.00±0.00	0.00±0.00	0.00±0.00
		MIA	82.56±0.00	13.54±0.01	30.63±0.04	0.00±0.00	7.54±0.01	6.68±0.00	3.92±0.00
ViT	ED	$\mathcal{D}_r$	95.73±0.00	94.71±0.14	87.35±1.44	93.42±0.62	93.45±0.44	95.82±0.00	95.84±0.00
		$\mathcal{D}_f$	95.03±0.00	0.00±0.00	0.28±0.22	4.14±3.75	0.03±0.08	53.53±0.00	0.68±0.00
		MIA	91.60±0.00	9.82±0.01	23.60±0.04	0.02±0.00	1.70±0.00	1.32±0.00	4.00±0.00
	NS	$\mathcal{D}_r$	95.71±0.00	94.79±0.11	87.45±1.2	93.50±0.4	93.68±0.56	93.63±0.00	95.13±0.00
		$\mathcal{D}_f$	95.37±0.00	0.00±0.00	0.05±0.15	2.63±2.08	0.00±0.00	0.00±0.00	0.00±0.00
		MIA	85.04±0.00	4.70±0.01	16.97±0.04	0.06±0.00	1.04±0.00	1.88±0.00	4.72±0.00
	people	$\mathcal{D}_r$	95.54±0.00	94.54±0.14	80.18±19.88	93.19±0.54	93.28±0.41	95.33±0.00	95.50±0.00
		$\mathcal{D}_f$	98.54±0.00	0.00±0.00	0.09±0.14	2.91±3.36	0.00±0.00	0.00±0.00	2.52±0.00
		MIA	89.48±0.00	1.56±0.00	8.08±0.03	0.01±0.00	0.60±0.00	1.20±0.00	1.00±0.00
	veg	$\mathcal{D}_r$	95.59±0.00	94.54±0.21	87.09±1.24	92.92±0.51	93.29±0.41	95.71±0.00	95.64±0.00
		$\mathcal{D}_f$	97.57±0.00	0.00±0.00	0.30±0.29	8.28±6.79	0.02±0.07	0.00±0.00	0.59±0.00
		MIA	91.32±0.00	4.41±0.01	14.72±0.05	0.02±0.00	1.02±0.00	1.88±0.00	3.00±0.00
	Veh2	$\mathcal{D}_r$	95.73±0.00	94.85±0.13	87.75±1.64	93.59±0.3	93.88±0.15	93.12±0.00	95.94±0.00
		$\mathcal{D}_f$	95.22±0.00	0.00±0.00	0.04±0.12	4.88±4.12	0.00±0.00	0.00±0.00	0.00±0.00
		MIA	84.04±0.00	22.96±0.03	38.15±0.08	0.02±0.00	1.20±0.00	7.04±0.00	9.12±0.00

Table 6: Subclass unlearning on Cifar20. SSD: $\alpha_{RN} = 10$, $\alpha_{ViT} = 25$, $\lambda = 1$; LFSSD: $\alpha = 10$, $\lambda = 1$. MR: mushroom, Rkt: Rocket.

	class	metric	baseline	retrain	finetune	teacher	amnesiac	SSD	LFSSD
RN	baby	$\mathcal{D}_r$	82.45±0.00	81.45±0.3	72.94±0.82	81.25±0.19	81.02±0.2	79.25±0.00	78.93±0.00
		$\mathcal{D}_f$	86.98±0.00	80.03±3.4	66.52±6.19	75.17±2.66	37.70±9.59	8.85±0.00	0.78±0.00
		MIA	90.40±0.00	44.82±0.02	56.20±0.06	0.04±0.00	13.65±0.02	1.40±0.00	1.00±0.00
	lamp	$\mathcal{D}_r$	82.61±0.00	81.81±0.19	72.46±0.95	81.51±0.19	81.36±0.36	79.58±0.00	81.54±0.00
		$\mathcal{D}_f$	72.66±0.00	22.95±4.72	17.25±6.58	16.82±3.07	7.44±1.73	0.00±0.00	7.29±0.00
		MIA	83.40±0.00	4.29±0.01	21.55±0.04	0.04±0.00	15.73±0.03	2.00±0.00	3.40±0.00
	MR	$\mathcal{D}_r$	82.60±0.00	81.66±0.21	72.51±1.15	81.50±0.2	81.44±0.18	80.05±0.00	80.38±0.00
		$\mathcal{D}_f$	72.22±0.00	10.48±2.02	7.46±4.89	8.77±2.79	1.09±1.13	0.78±0.00	0.78±0.00
		MIA	86.80±0.00	2.27±0.00	18.09±0.02	0.00±0.00	12.80±0.03	4.00±0.00	3.80±0.00
	Rkt	$\mathcal{D}_r$	82.54±0.00	81.54±0.24	72.41±0.95	81.48±0.27	81.46±0.26	82.43±0.00	80.49±0.00
		$\mathcal{D}_f$	79.34±0.00	10.74±3.4	9.75±6.68	6.41±3.57	0.76±0.73	2.17±0.00	0.00±0.00
		MIA	89.40±0.00	3.85±0.01	18.67±0.05	0.00±0.00	6.60±0.01	10.80±0.00	6.60±0.00
	sea	$\mathcal{D}_r$	82.37±0.00	81.30±0.27	72.50±1.55	81.22±0.24	81.05±0.31	81.72±0.00	79.84±0.00
		$\mathcal{D}_f$	96.27±0.00	91.47±1.92	82.69±7.17	75.13±4.12	46.78±8.55	75.35±0.00	18.06±0.00
		MIA	90.80±0.00	52.09±0.03	62.82±0.11	0.00±0.00	4.45±0.01	21.80±0.00	2.20±0.00
ViT	baby	$\mathcal{D}_r$	95.69±0.00	94.50±0.19	87.60±0.75	92.98±0.53	93.35±0.26	95.54±0.00	95.38±0.00
		$\mathcal{D}_f$	96.44±0.00	93.23±1.09	85.45±4.47	46.66±17.88	38.76±7.36	94.10±0.00	86.81±0.00
		MIA	91.60±0.00	77.37±0.03	66.57±0.07	0.03±0.00	0.93±0.01	77.20±0.00	1.60±0.00
	lamp	$\mathcal{D}_r$	95.77±0.00	94.69±0.13	87.72±0.52	93.57±0.67	93.66±0.5	95.54±0.00	95.68±0.00
		$\mathcal{D}_f$	89.58±0.00	34.55±8.62	16.90±10.4	8.23±7.05	0.59±1.45	14.58±0.00	24.05±0.00
		MIA	81.00±0.00	5.60±0.02	14.73±0.04	0.10±0.00	2.00±0.01	3.2±0.00	5.80±0.00
	MR	$\mathcal{D}_r$	95.69±0.00	94.60±0.13	87.37±0.91	93.55±0.42	93.42±0.46	95.51±0.00	95.66±0.00
		$\mathcal{D}_f$	97.05±0.00	26.57±6.41	15.71±12.13	13.01±9.11	0.20±0.36	6.68±0.00	31.42±0.00
		MIA	77.80±0.00	2.34±0.01	9.25±0.04	0.00±0.00	1.50±0.01	0.40±0.00	0.80±0.00
	Rkt	$\mathcal{D}_r$	95.73±0.00	94.61±0.13	85.70±3.05	93.60±0.29	93.47±0.22	95.13±0.00	95.66±0.00
		$\mathcal{D}_f$	94.53±0.00	22.26±8.34	6.25±6.03	3.35±2.89	0.85±1.71	5.12±0.00	2.17±0.00
		MIA	80.40±0.00	3.44±0.01	16.04±0.03	0.02±0.00	0.78±0.00	5.40±0.00	3.00±0.00
	sea	$\mathcal{D}_r$	95.67±0.00	94.55±0.22	87.65±1.56	93.57±0.26	93.26±0.24	95.57±0.00	94.45±0.00
		$\mathcal{D}_f$	99.22±0.00	95.12±0.81	89.17±4.17	25.97±14.01	21.42±8.5	97.05±0.00	20.83±0.00
		MIA	88.40±0.00	65.96±0.04	65.04±0.13	0.17±0.00	0.40±0.00	82.20±0.00	6.60±0.00

Table 7: Random sample unlearning on Cifar10 (100 samples). SSD: $\alpha = 10$, $\lambda = 1$; LFSSD: $\alpha = 3.5$, $\lambda = 1$

	metric	baseline	retrain	finetune	teacher	amnesiac	SSD	LFSSD
RN	$\mathcal{D}_r$	90.71±0.00	91.45±0.11	88.02±0.45	90.21±0.10	90.16±0.23	88.68±3.36	89.82±2.30
	$\mathcal{D}_f$	95.30±2.08	94.10±2.00	90.00±3.73	90.00±2.73	59.04±4.79	93.61±4.99	94.94±2.89
	MIA	75.78±0.04	74.22±0.04	74.58±0.05	49.28±0.07	25.18±0.05	72.65±0.05	73.86±6.33
ViT	$\mathcal{D}_r$	98.88±0.00	98.61±0.08	97.28±0.33	97.58±0.36	97.62±0.35	98.01±1.56	98.53±0.27
	$\mathcal{D}_f$	100.00±0.00	98.80±0.76	97.19±0.98	86.75±3.57	73.49±5.11	98.07±2.35	99.52±0.84
	MIA	90.76±0.03	91.77±0.02	86.14±0.02	33.53±0.06	10.44±0.05	85.54±0.11	90.12±3.40