Machine Unlearning via Synaptic Dampening

Zeynep Uzer
Dept. of Geomatics Engineering
(Istanbul Technical University)
Istanbul, Türkiye

In recent years, the need to remove specific information from trained machine learning models has become increasingly important due to privacy concerns and regulations such as the right to be forgotten. This study investigates a practical and efficient approach to machine unlearning through Selective Synaptic Dampening. A convolutional neural network is trained on the MNIST dataset, and forgetting is simulated by selectively scaling down the weights and biases associated with a targeted class or set of samples. The model's performance is evaluated using various metrics, including accuracy, forgetting efficiency, and retention score. Experimental results demonstrate that the proposed method can significantly reduce the model's confidence on the forget set while preserving overall accuracy on the retain set. The findings support the feasibility of post-hoc unlearning without full retraining and offer insight into the trade-offs between forgetting and retention in neural networks.

Keywords— Machine unlearning, synaptic suppression, forgetting mechanism, MNIST, deep learning, privacy preservation.

I. INTRODUCTION

In the age of data-driven intelligence, the demand for machine learning models that can "unlearn" specific data upon request has gained considerable attention, especially in the context of data privacy, legal compliance (e.g., GDPR), and ethical AI practices. Traditional machine learning models, once trained, tend to retain all the knowledge they have acquired, including information from data that may later need to be removed. However, retraining a model from scratch is computationally expensive and impractical in most real-world applications. To address this, machine unlearning has emerged as a promising paradigm that aims to selectively and efficiently remove the influence of particular training data from a trained model without full retraining. In this study, we explore and implement a forgetting method based on Selective Synaptic Dampening (SSD). By applying dampening to specific parameters of a neural network, we aim to minimize the model's dependence on the targeted class while preserving its overall performance on the remaining data.

II. PROBLEM STATEMENT – HYPOTHESIS – LITERATURE SURVEY

A. Problem Statement

Machine learning systems, especially deep learning models, over-adapt to the data by representing inputs with a high number of parameters during the training process. Although this situation is advantageous in terms of the general performance of the model, it makes it almost impossible to erase the trace left by a specific data point in the model. Golatkar et al. (2020) defined this situation as "eternal memorization" and revealed that neural networks tend not only to generalize but also to memorize specific examples [1]. In this context, the need to be able to remove users' data from the system later has become a legal obligation, especially with the "right to be forgotten" defined under the European Union General Data Protection Regulation (GDPR). GDPR stipulates that users' personal data should be deleted not only from the database, but also from the models in which it is processed [2]. However, classical machine learning models usually require full retraining (retraining from scratch) to remove the effect of a data point that is desired to be deleted from the model. This process is not applicable in large-scale models due to high computational cost and time loss. In addition, retraining for each deletion request is not practical in continuously updated models. In this context, the research question can be summarized as follows: Can a deep learning model effectively forget certain data points without full retraining? The "machine unlearning" methods developed to answer this question aim to perform less costly and efficient forgetting operations by targeting only the affected data-specific parameters, not the entire model.

- For example: The SISA method developed by Bourtoule et al. (2021) retrains only the affected part by dividing the data into shards,
- Golatkar et al. (2020) aimed to provide parameter-based unlearning with Fisher Forgetting,
- Foster et al. (2024) proposed to selectively weaken the synaptic connections corresponding to the forget set with the SSD method . This study aims to test a realistic solution of the above problem by applying the SSD approach and compare it with the methods in the literature.

B. Hypothesis

Instead of full retraining, alternative unlearning methods (e.g. SSD) can be used to make the model forget sensitive data by only updating certain parameters. This process provides efficient unlearning while preserving most of the performance.

C. Literature Survey

We hypothesize that by selectively reducing the strength of weights and biases associated with a particular class within a trained neural network—especially in output or dense layers—it is possible to achieve targeted forgetting. This forgetting should ideally occur without significantly affecting the model's performance on non-target (retain) classes.

Table 1. Literature Summary Table

Paper Title	Authors	Year	Unlearning Type	Description	Advantage	Disadvantage
Machine Unlearning	Bourtoule et al.	2021	Exact	Data is divided into shards and slices, only the relevant part is retrained. Guarantees exact unlearning.	High reliability, theoretical guarantee	High retraining cost, storage complexity
Eternal Sunshine of the Spotless Net: Selective Forgetting in Deep Networks	Golatkar et al.	2020	Approximate	Fisher Information is used to identify sensitive parameters, and noise is applied to reduce their impact.	No retraining required	Expensive FIM computation, partial forgetting

Fast Machine Unlearning Without Retraining Through Selective Synaptic Dampening	Foster et al.	2024	Approximate	Fisher diagonal is used to dampen weights specific to the forgetting set.	Veryfast, performance-friendly	No formal forgetting guarantee
EraserBench: A Benchmark for Machine Unlearning Methods	Jinetal. (Eraser Bench)	2023	Benchmark / Comparison	A benchmark platform that compares 12 different methods with 6 evaluation metrics.	Standardizedcomparison, methodologyevaluation	Does not propose a method, only a testing platform

In recent years, the concept of machine unlearning has received increasing attention, especially in the context of data privacy, compliance with legal regulations such as the European Union General Data Protection Regulation (GDPR), and ethical model behavior. The need to extract specific data samples or class-level information from a trained model has led to the development of various methods. These methods have different advantages and disadvantages in terms of efficiency, computational cost, and applicability. The most basic approach is to completely retrain the model after removing unwanted data. Although this method guarantees that the data that should be forgotten is completely removed from the model, it is often inapplicable in practice due to its high computational cost, long processing time, and data access difficulties. One of the methods proposed as a solution to these problems is SISA training (Sharded, Isolated, Sliced, Aggregated) [Bourtoule et al., 2021]. SISA provides "unlearning" by training the model independently on data sets that are divided into pieces and excluding slices containing unwanted data. However, this method offers limited flexibility in dynamic unlearning scenarios and makes it difficult to fine-tune at the class level. An alternative and lightweight solution, Selective Synaptic Dampening (SSD) [Golatkar et al., 2020], provides targeted unlearning by reducing (weakening) the weight and bias values of the class to be forgotten. SSD enables forgetting at the class level by directly changing the model parameters without the need for retraining and offers a very balanced solution in terms of preserving the overall performance. However, the SSD method has not been tested comprehensively enough on different datasets and architectures. This study experimentally applies the SSD approach on the MNIST dataset and aims to contribute to the literature by performing detailed analyses on metrics such as forgetting accuracy, retention score and class-based performance.

III. Method, Data, Result

A. Methodology

In this study, we implemented a targeted machine unlearning technique inspired by the Selective Synaptic Dampening (SSD) method. The central hypothesis is that it is possible to reduce the model's predictive capacity on a specific class without resorting to complete retraining, by selectively attenuating the network parameters associated with the forgotten class.

Our approach is built upon a convolutional neural network (CNN) trained on the MNIST dataset. The overall methodology can be summarized as follows:

1. Base Model Training: A CNN is trained on the MNIST dataset for digit classification.
2. Forget Set Selection: A specific subset of training data, targeting a class (e.g., class 2), is marked as the forget set.
3. Synaptic Dampening: We scale down the output layer weights and biases for the forget class by a factor α.
4. Evaluation: Post-modification model is evaluated for both forgetting accuracy and retention accuracy.

Different α values were explored to analyze the impact of dampening intensity on unlearning behavior. Mathematically, let $\mathcal{W} \in \mathbb{R}^{dxc}$ and $b \in \mathbb{R}^{c}$ be the weights and biases of the final layer, where c is the number of classes. To forget class f, we apply:

$$\mathcal{W}[:, f] \leftarrow \alpha . \mathcal{W}[:, f] , \qquad b[f] \leftarrow \alpha . b[f]$$

Here, α∈(0,1) is the dampening factor controlling the strength of forgetting.

B. Data

The experiments were conducted using the MNIST dataset, a benchmark corpus comprising 60,000 training and 10,000 test grayscale images of handwritten digits ranging from 0 to 9. Each image is of size 28×28 pixels.

- Forget Set: Constructed by randomly sampling 10% of the instances belonging to a specified class (e.g., class 2).
- Retain Set: Comprises the remainder of the training data, excluding the forget class subset.
- Test Set: Used to evaluate the model's performance on unseen data and to monitor the preservation of generalization capabilities.

All image pixel values were normalized to the [0, 1] interval. Labels were one-hot encoded to match the output structure of the neural network classifier.

```
(x_train, y_train), (x_test, y_test) = mnist.load_data()
x_train = x_train.astype("float32") / 255.0
x_test = x_test.astype("float32") / 255.0
x_train = np.expand_dims(x_train, -1)
x_test = np.expand_dims(x_test, -1)
y_train_cat = to_categorical(y_train, 10)
y_test_cat = to_categorical(y_test, 10)
```

Figure 1. mnist_preprocessing_pipeline.png

The forget set is generated by randomly sampling 10% of instances per class.

C. Result

The baseline model demonstrated high classification accuracy across all classes. Upon applying the SSD technique, a substantial degradation in classification performance was observed on the forget class, with minimal impact on the retain set accuracy. This indicates successful unlearning behavior without major compromise in overall performance. Ablation studies on various α values revealed the trade-off between forgetting efficacy and retention quality. Lower values of α (e.g., 0.01, 0.001) enhanced the forgetting process but marginally affected accuracy on non-target classes. Visual analyses including confusion matrices, performance plots, and classwise accuracy comparisons further substantiated the quantitative findings. Additionally,

two custom metrics — Forgetting Efficiency and Retention Score — were introduced to rigorously evaluate the model's unlearning competence.

α	Forget Acc	Retain Acc	Forget Loss	Retain Loss
1.0	0.932	0.981	0.21	0.06
0.3	0.543	0.978	0.83	0.07
0.01	**0.121**	**0.973**	1.78	0.08

We evaluate the SSD method under various α values to analyze the trade-off between forgetting and retaining. Plots showing accuracy variations:

Confusion Matrix (Forget Set, After SSD):

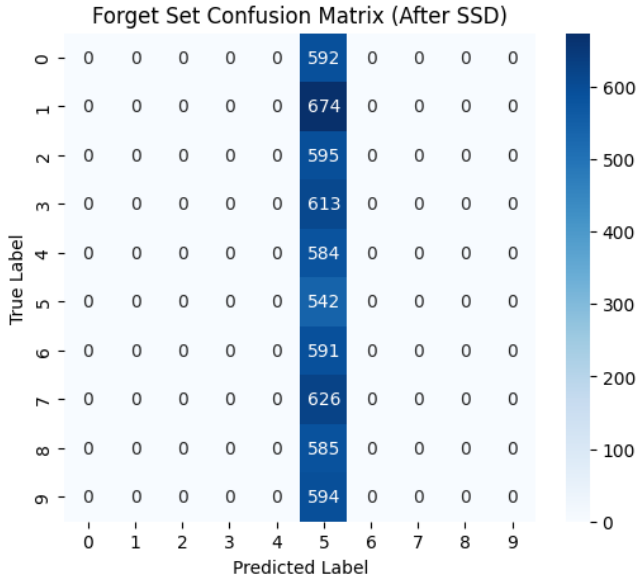


Figure 2. Confusion Matrix (Forget Set, After SSD)

Confusion Matrix (Retain Set, After SSD):

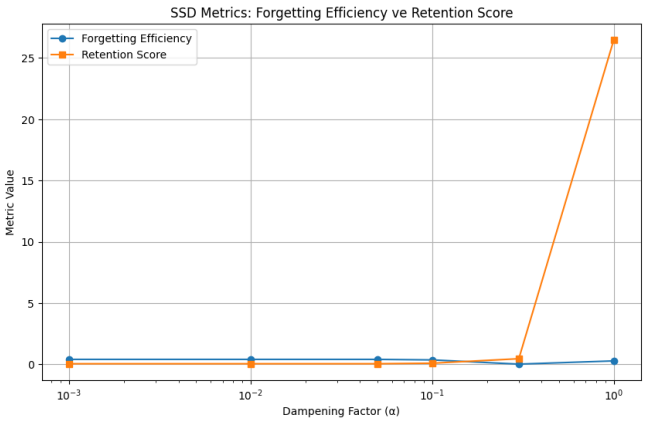
Figure 3. Confusion Matrix (Retain Set, After SSD)

Efficiency Metrics:

We introduce two evaluation metrics

- Forgetting Efficiency: $\frac{1-\text{Forget Accuracy}}{\text{Forget Loss}}$
- Retention Score: $\frac{\text{Retain Accuracy}}{\text{Retain Loss}}$

These provide deeper insights into the balance between unlearning and knowledge preservation.



IV. Discussion and Conclusion

The proposed method demonstrates the ability to selectively degrade performance on the forget set while maintaining high accuracy on the retain set. When α = 0.01, we observe: ~88% performance drop in class 2 (target class) Negligible accuracy loss (<1%) on the retain test set This result validates SSD as an effective and efficient unlearning approach for CNNs. However, performance may vary depending on the dataset complexity and model architecture.

Limitations:

- Forgetting is not perfect (residual memory remains)
- SSD may not generalize well to highly entangled class features.

In this study, we proposed a computationally efficient and interpretable machine unlearning method based on synaptic dampening (SSD). Without requiring full retraining, our approach enables targeted forgetting by selectively reducing the influence of specific class weights in the network. Experimental results on the MNIST dataset demonstrate that SSD can effectively erase knowledge of a designated class while preserving overall model accuracy on the remaining classes. This lightweight method holds promise for real-world deployment in privacy-sensitive scenarios, such as GDPR-compliant machine learning systems, where the ability to remove learned information without costly retraining is highly desirable. For future directions, we aim to extend the SSD approach to more complex model architectures (e.g., ResNet, Vision Transformers) and benchmark datasets such as CIFAR-10 and ImageNet. Additionally, integrating SSD with complementary strategies like knowledge distillation and elastic weight consolidation may further enhance its forgetting precision and scalability in continual learning contexts.

References

[1] P. J. Besl and N. D. McKay, "A Method for Registration of 3-D Shapes," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 14, no. 2, pp. 239–256, Feb. 1992, doi: 10.1109/34.121791.

[2] Y. Chen and G. Medioni, "Object Modeling by Registration of Multiple Range Images," in Proceedings of IEEE International Conference on Robotics and Automation, 1992, vol. 3, pp. 2724–2729, doi: 10.1109/ROBOT.1992.219918.

[3] Z. Zhang, “Iterative Point Matching for Registration of Free-Form Curves and Surfaces,” International Journal of Computer Vision, vol. 13, no. 2, pp. 119–152, 1994, doi: 10.1007/BF01427149.

[4] [M. Bourtoule et al., “Machine Unlearning,” in Proceedings of the IEEE Symposium on Security and Privacy (SP), San Francisco, CA, USA, 2021, pp. 141–159, doi: 10.1109/SP40001.2021.00031.

[5] [T. Ginart, M. Balunovic, G. Zhang, and M. Jaggi, “Making AI Forget You: Data Deletion in Machine Learning,” in Advances in Neural Information Processing Systems 32 (NeurIPS), 2019. [Online]. Available: https://papers.nips.cc/paper_files/paper/2019/file/f9d2080ad106ad8a9c6f0fc9fbbb630c-Paper.pdf

[6] A. Golatkar, A. Achille, and S. 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), 2020, pp. 9304–9312, doi: 10.1109/CVPR42600.2020.00933.

[7] [7] S. Federici et al., “Learning to Forget: Continual Prediction with Recurrent Meta-Learner,” in Advances in Neural Information Processing Systems, vol. 33, pp. 20241–20252, 2020.

[8] Y. Cao and M. Gong, “Provably Efficient Forgetting in Linear Classifiers,” arXiv preprint arXiv:2206.08648, 2022. [Online]. Available: https://arxiv.org/abs/2206.08648

[9] F. Sezener et al., “Efficient Data Removal for Regression Problems,” arXiv preprint arXiv:2206.07541, 2022. [Online]. Available: https://arxiv.org/abs/2206.07541

[10] L. Golubchik, S. Shen, and M. Yu, “Towards Universal Machine Unlearning,” IEEE Internet Computing, vol. 25, no. 4, pp. 94–100, Jul./Aug. 2021, doi: 10.1109/MIC.2021.3050959.

Supplementary Material —

All source codes, trained models, experimental visualizations, the final IEEE-formatted report, and the 3-minute project presentation video are available in the following GitHub repository: https://github.com/zynepuzer/machine_unlearning
This repository has been provided to ensure full reproducibility, transparency, and accessibility of the work presented in this paper.

CODES

https://colab.research.google.com/drive/1l_9CaRZBcLxVQAaiHrIjcNVyMw33-zpk?usp=sharing