Per-Sample Amnesiac Unlearning

July 16, 2024

Davide Marincione

Abstract

Typical machine unlearning techniques do not scale well with modern deep learning models, as these either require models to abide to strict theoretical guarantees (lowering their effectiveness) or be retrained (unfeasible for big models). In this project we propose a variation of Amnesiac Unlearning, a technique which promises to bring unlearning with little to no compromises to neural networks. Our method lowers the memory requirements of the original technique while maintaining its effectiveness. We test it on MNIST and CIFAR-100, showing that it can make a model forget samples and classes while maintaining its performance on the rest of the dataset.

1. Introduction

Machine Unlearning is an ever-more important topic in the field of Machine Learning. Not only because privacy concerns regarding deep learning techniques are increasing, but also because regulations such as the EU's GDPR oblige companies to delete user's data at their request. Because of this, research in trying to produce an unlearning procedure that requires the less possible amount of retraining is increasing.

Our work consists of:

- A new unlearning technique with lower memory requirements than the original Amnesiac Unlearning technique (Graves et al., 2021).
- Tests on MNIST and CIFAR-100 showing that our technique can make a model forget samples and classes while maintaining its performance on the rest of the dataset.

2. Related Work

As far as single-technique driven methods go (as combined procedures are often used in practice), the *model shifting* (Xu et al., 2023) family of techniques is one of the most promising. These techniques are based on the idea of directly updating the model parameters to offset the impact of the samples to forget; by finding an update δ for $w_u = w + \delta$, where w are the parameters of the original model. The reason for why this solution is so promising is that it does not require a major retraining of the model, which is often infeasible for big models: both in terms of time and computational resources.

One of the most successful *model shifting* techniques is Amnesiac Unlearning (Graves et al., 2021), which defines the update δ with the gradient updates in which the samples to forget took part. To do so, the gradient update of each batch must be stored on disk, leading to a high memory requirement, as each update is of the same size as the model. This is a major drawback, as it makes the technique unfeasible for big models. An extremely recent work (Gogineni & Nadimi, 2024) proposes layer unlearning, an approximation which only stores a random subset of the gradients in the update, therefore reducing the memory requirements. However, this leads to a decrease in the effectiveness of the technique, which requires a short retraining phase to completely forget the samples.

3. Method

Our method consists of storing the gradients not for each batch, but for each sample. On the long run, this leads to a lower memory requirement than the original technique, as we only need to store a single set of gradients for each sample (rather than a new set for each batch). If the space requirements of the original technique are $O(\frac{NME}{B})$, our technique requires $O(NMp)$ space, where N is the number of samples in the dataset, E is the number of epochs, B is the batch size, M is the number of parameters in the model, and p is percentage of parameters kept in the random subset. It is easy to see that our method requires less space than the original technique if $p < \frac{E}{B}$ (which can be easily achieved in practice, as if we heuristically set $p = .1$ and have a typical $B = 64$ then we just require $E = 7$).

Email: Davide Marincione <marincione.1927757@studenti.uniroma1.it>.

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

Table 1. Results on MNIST with a small CNN, forgetting class of number 3. Results are the accuracy of the forgotten class and of the rest of the dataset (in the form 'forgotten/rest'). *After train* values are all the same as training is equal for all techniques.

Technique	After train	No retrain	After retrain
Original	0.58/0.98	0.00/0.84	0.00/0.98
Layer, $p = .1$	0.87/0.98	0.44/0.98	0.10/0.98
Ours, $p = .1$	0.74/0.98	0.00/0.85	0.00/0.99
Ours, $p = .075$	0.74/0.98	0.00/0.94	0.01/0.98
Ours, $p = .05$	0.74/0.98	0.17/0.97	0.04/0.99
Ours, $p = .01$	0.74/0.98	0.67/0.98	0.04/0.99

Not only can our method require less space than the original, but it also has the potential to be more effective than the approximation proposed in (Gogineni & Nadimi, 2024), as it only removes the gradients of the samples to forget, rather than the gradients of the full batch (where the influence of samples that are not to forget is removed as well).

4. Experiments

We modify the code provided by the authors of (Graves et al., 2021) to test our method as well as the original Amnesiac Unlearning technique, and the approximation proposed in (Gogineni & Nadimi, 2024) (as they have yet to publish their own code).

Our implementation slightly differs from the original: in that, they store the difference between the model before and after the gradient update, making it invariant to the choice of optimizer. Our case instead, since `pytorch` allows the calculation of per-sample gradients through `call_for_per_sample_grads` but not the usage of its result through an optimizer, we had to directly store the gradients, and not the difference between the model before and after the update. This implies that, since optimization usually doesn't directly remove the gradients during a step, but instead uses some value derived from it (either the same gradient but with lower magnitude or an exponential moving average of the gradients), our implementation is not invariant to the choice of optimizer. Still, as in the original technique, we trained the model with `Adam` and nonetheless we achieved comparable results, therefore we believe this to not be a drawback.

We test our method on MNIST and CIFAR-100, both with a small CNN and a ResNet-18; during all tests, we run the training for 4 epochs (the maximum amount of epochs that the original technique can run with the available memory on a Kaggle Notebook) and the retraining for 4 epochs too. From these tests, we see that our technique enjoys a similar effectiveness to the original, as it requires no retraining to completely forget the samples. Furthermore, we make multiple runs with different values of p to find the lowest value that still allows the technique to work.

Table 2. Results on CIFAR100 with a small CNN, forgetting class 81.

Technique	After train	No retrain	After retrain
Original	-		
Layer, $p = .1$	-		
Ours, $p = .1$	0.22/0.33	0.00/0.16	0.00/0.35
Ours, $p = .075$	0.22/0.33	0.00/0.19	0.00/0.36
Ours, $p = .05$	0.22/0.33	0.00/0.25	0.00/0.36
Ours, $p = .01$	0.22/0.33	0.05/0.32	0.00/0.35

5. Conclusion

We've empirically shown that amnesiac unlearning can be even more powerful if done per-sample rather than per-batch. Furthermore, we can effectively unlearn training samples by modifying just a fraction of the model's parameters. Thanks to these two properties, our technique is shown to be more effective than the approximation proposed in (Gogineni & Nadimi, 2024), as it requires no retraining. Further work could be done to find the optimal subset of parameters to keep, rather than using a random subset; furthermore, it could be made even more efficient by compressing the representation of the gradients themselves using lower bit precision and other techniques.

References

Gogineni, V. C. and Nadimi, E. S. Efficient knowledge deletion from trained models through layer-wise partial machine unlearning. *arXiv preprint arXiv:2403.07611*, 2024.

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

Xu, H., Zhu, T., Zhang, L., Zhou, W., and Yu, P. S. Machine unlearning: A survey, 2023. URL `https://arxiv.org/abs/2306.03558`.