Data Attribution-Guided Machine Unlearning

June 10, 2024

1. Introduction

Initially motivated by privacy considerations such as the GDPR's "Right to be Forgotten," machine unlearning seeks to address the challenge of removing the influence of (or "forgetting") selected datapoints from an existing model. Formally, *machine unlearning* takes as an input a model θ trained on a training dataset S and a forget set $S_f \subset S$, and returns a new model that is similar to the model θ_r retrained on the dataset excluding the *forget set*, $S_r = S/S_f$. While one obvious solution would be to exclude S_f from the training dataset, and then to retrain θ from scratch on the *retain set* $S_r = S/S_f$, as model and dataset sizes continue to increase, this becomes an increasingly infeasible. Approximate machine unlearning (Ginart et al., 2019; Guo et al., 2019; Sekhari et al., 2021; Neel et al., 2021) tries to approximate this re-training behavior with far less computation than required for re-training. Broadly, unlearning is considered successful if the behavior of the unlearned model closely approximates the behaviour of θ_r.

While above is the general technical notion we will focus on, we note that it remains an open question how to bridge laws, such as GDPR and national copyright laws, to technical notions of machine unlearning—*what technical measure satisfies the legal definition of being "forgotten"*.

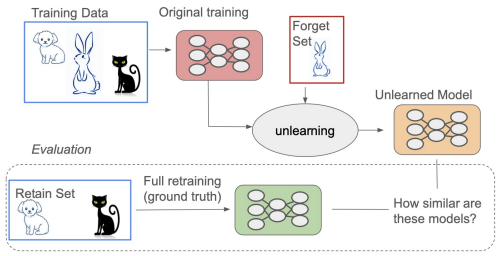

Figure 1: High-level description of Machine Unlearning.

1.1. Evaluating Unlearning: KL divergence of Margins

We consider an unlearning method is successful if at any given input x, the outputs of the unlearned model θ_u and the oracle model θ_r are statistically indistinguishable. To contextualize the performance of different unlearning methods, we introduce a new measure called the KL divergence of Margins (`KLoM`), which builds on the ULIRA measure introduced in (Hayes et al., 2024b; Pawelczyk et al., 2023). Specifically, for each data point x of interest, `KLoM` computes the KL-divergence between the predictions of unlearned models and the predictions of the retrained models (that never saw S_f); a smaller value means that the predictions are more similar, e.g., that unlearning was more successful.

1.2. Existing Unlearning Methods

Nearly all existing methods employ some variation of performing gradient descent on the retain set, gradient ascent on the forget set, or freezing a subset of the network before retraining on the retain set (Liu, 2024; Nguyen et al., 2022; Hayes et al., 2024b; Triantafillou et al., 2023). Unfortunately, even the best of these methods do not approximate a retrained model very well, particularly when unlearning success is evaluated on a per-point basis (see `KLoM` evaluations in Figure 2). Hayes et al. (2024a) posit that this is because different points are unlearned at different rates, and there is no obvious way to know when a point is unlearned. Thus, even if an unlearning algorithm is making the updates in the right direction, the final model can under or "over-unlearn" individual datapoints, resulting in unsuccessful unlearning and even higher privacy leakage.

2. Unlearning through Data Attribution

We seek to side-step the problems that arise in the prior heuristic-based machine unlearning methods, by investigating what an *optimal* unlearner would do given access to *predictions* of a model retrained from scratch on the retain set. While access to a retrained model would defeat the purpose of unlearning (as you already have an unlearned model), it provides an informative thought experiment. We propose a method called *Oracle-Matching* (`OM`), which fine-tunes the original model on the predictions of a full retrained model. Our results for `OM` (Figure 2) show that with sample-access to re-trained models, we can unlearn much better than using other gradient-based methods. Note that `OM` does not actually require access to the retrained models themselves, just their predictions on $\{x_i\}$.

While access to a retrained model is an unreasonable requirement, we observe that recent data attribution methods (Ilyas et al., 2022; Park et al., 2023) can approximate counterfactual predictions of the retrained model. In particular,

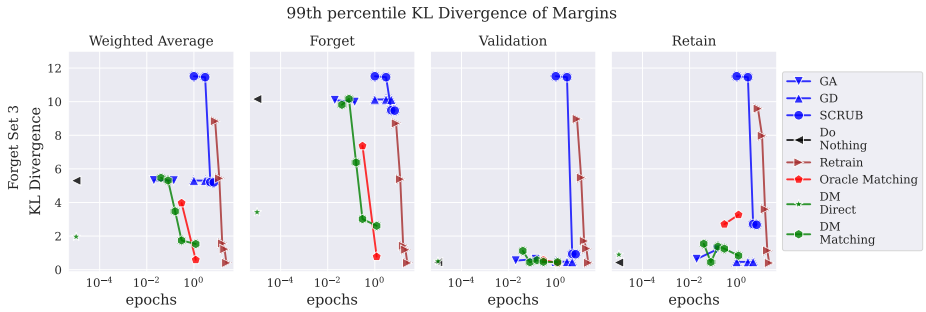

Figure 2: **Data attribution based unlearning methods significantly outperform prior unlearning methods.** We compare the performance of different unlearning algorithms for ResNet-9 classifiers trained on CIFAR-10 in terms of computational cost (x-axis) and unlearning effectiveness (y-axis; measured by KLoM). For each unlearning method (for a given budget), we evaluate the 99th percentile of KLoM values over points in each of the Forget Set, Validation set, and Retain set, and report the average.

datamodels accurately predict a model's prediction on a test point x as a function of the training dataset $S' \subset S$ (Ilyas et al., 2022). We first find that a sufficiently accurate datamodel can be used to approximate an unlearned model's predictions directly by removing the influence of the forgotten points, which we call *Datamodel-Direct*. Next, leveraging datamodels, we implement an approximate (and practical) version of Oracle-Matching, which we call *Datamodel-Matching*. We find that *Datamodel-Matching* achieves levels of unlearning similar to that of fully retraining the model (as measured by KLoM scores), while using significantly less compute (Figure 2). Importantly, using datamodels allow us to recover the performance of OM, and outperforms all prior gradient-based approaches.

2.1. Unlearning for Linear Models

To understand the effectiveness of Oracle-Matching, we compare it to other gradient-based unlearning methods for a simple linear model. In our model, data are generated from a Gaussian distribution and we fit the data by minimizing the least squares objective with ℓ_2 regularization.

Since the objective function is convex, both gradient descent (GD) and oracle matching (OM) are guaranteed to converge, but OM converges significantly faster. OM differs from GD in two key aspects: (i) it minimizes the squared error relative to the optimal model predictions and (ii) its updates include both the forget set and the retain set points. To isolate the impact of these differences, we consider OM with only the retained points (OM retain set), which shows negligible progress. This suggests that including forget set points in the updates is crucial for OM's superior performance. Gradient descent on the retain set and gradient ascent on the forget set (GD + GA) also uses forget set points but in a heuristic manner, failing to reach the optimum. To corroborate the above empirical observations, we also present theoretical analysis.

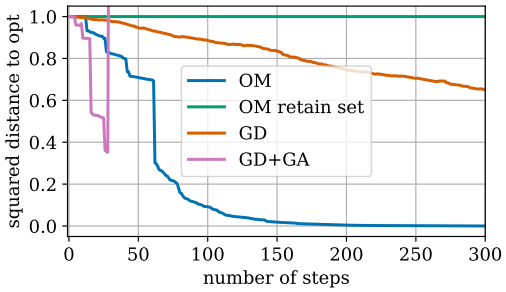

Figure 3: **Oracle Matching converges faster than other methods for unlearning with a linear model.** We generate 100 datapoints from an isotropic Gaussian distribution in 400 dimensions and label them with a noisy linear function, and attempt to unlearn a random subset of 5 training points.

3. Conclusion and discussion

We show that leveraging data attribution methods offers a fruitful direction for machine unlearning. By leveraging estimates of predictions of the (counterfactual) retrained model, we can sidestep the issue of unknown stopping conditions of prior gradient-based approaches. Our unlearning algorithms significantly outperform others in terms of both effectiveness and computational time (Figure 2). In a linear setting, we show that oracle matching can converge much faster than GD/GA based methods (Figure 3). Further, we propose KLoM, a measure for estimating unlearning success.

By reducing the problem of unlearning to data attribution, we can take advantage of recent and future work in attribution to improve the fidelity and efficiency of unlearning (Ilyas et al., 2022; Park et al., 2023; Koh & Liang, 2020; Engstrom et al., 2024). So far, our methods rely on datamodels, which requires expensive pre-computation. One promising future direction is to investigate other more efficient data attribution methods, such as TRAK (Park et al., 2023). More broadly, our work shows the potential of practical and reliable approximate unlearning methods by better understanding how data influences model outputs.

References

Engstrom, L., Feldmann, A., and Madry, A. Dsdm: Model-aware dataset selection with datamodels, 2024.

Ginart, A., Guan, M. Y., Valiant, G., and Zou, J. Making ai forget you: Data deletion in machine learning. In *Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*, 2019.

Guo, C., Goldstein, T., Hannun, A., and Van Der Maaten, L. Certified data removal from machine learning models. In *International Conference on Machine Learing (ICML)*, 2019.

Hayes, J., Shumailov, I., Triantafillou, E., Khalifa, A., and Papernot, N. Inexact unlearning needs more careful evaluations to avoid a false sense of privacy, 2024a.

Hayes, J., Shumailov, I., Triantafillou, E., Khalifa, A., and Papernot, N. Inexact unlearning needs more careful evaluations to avoid a false sense of privacy, 2024b.

Ilyas, A., Park, S. M., Engstrom, L., Leclerc, G., and Madry, A. Datamodels: Predicting predictions from training data. *CoRR*, abs/2202.00622, 2022. URL https://arxiv.org/abs/2202.00622.

Koh, P. W. and Liang, P. Understanding black-box predictions via influence functions, 2020.

Liu, K. Z. Machine unlearning in 2024, Apr 2024. URL https://ai.stanford.edu/~kzliu/blog/unlearning.

Neel, S., Roth, A., and Sharifi-Malvajerdi, S. Descent-to-delete: Gradient-based methods for machine unlearning. In *Proceedings of the 32nd International Conference on Algorithmic Learning Theory (ALT)*, 2021.

Nguyen, T. T., Huynh, T. T., Nguyen, P. L., Liew, A. W.-C., Yin, H., and Nguyen, Q. V. H. A survey of machine unlearning, 2022.

Park, S. M., Georgiev, K., Ilyas, A., Leclerc, G., and Madry, A. Trak: Attributing model behavior at scale, 2023.

Pawelczyk, M., Neel, S., and Lakkaraju, H. In-context unlearning: Language models as few shot unlearners, 2023.

Sekhari, A., Acharya, J., Kamath, G., and Suresh, A. T. Remember what you want to forget: Algorithms for machine unlearning. In *Advances in Neural Information Processing Systems*, 2021.

Triantafillou, E., Pedregosa, F., Guyon, I., Escalera, S., Junior, J. C. S. J., Dziugaite, G. K., Triantafillou, P., Dumoulin, V., Mitliagkas, I., Hosoya, L. S., Kurmanji, M., Zhao, K., Wan, J., and Kairouz, P. Neurips 2023 machine unlearning challenge. https://unlearning-challenge.github.io, 2023. Accessed: 2024-05-29.