CS772A Project: Machine Unlearning

February 2024

1 Group Member Information

Name	Roll Number	Email Id	Department
Gaurang Dangayach	200373	gaurangd20@iitk.ac.in	MTH
Rahul Rustagi	200756	rrustagi20@iitk.ac.in	AE
Suryanshu Kumar Jaiswal	201025	skjaiswal20@iitk.ac.in	ME
Udvas Basak	201056	udvasb20@iitk.ac.in	AE
Ujjwal Kumar	201059	ujjwalk20@iitk.ac.in	CSE

Table 1: Group Member Information

2 Brief Details of the Project

2.1 Basic Introduction and Idea

Machine Unlearning is the problem of forgetting the knowledge of some training examples from an already learned model. In a Bayesian context, it kind of means that we want to remove the influence of some of the likelihood terms from the posterior (basically, the likelihood terms corresponding to the "forget set" of training examples).

In other words, machine unlearning refers to the process of mitigating the impact of specific training data points on a previously trained machine learning model. This process serves several purposes, including:

- Eliminating outdated, irrelevant or stale data
- Debiasing models
- Rectifying inaccuracies or errors in the original training data

A basic framework of machine unlearning is shown in Fig.(1).

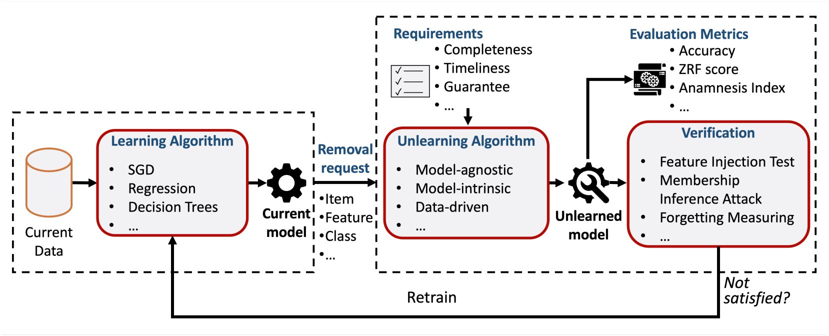

Figure 1: A Framework of Machine Unlearning (from [Nguyen et al., 2022])

2.2 Basic Approach

Our approach would be to first understand and study a number of existing research papers on machine unlearning. The underlying plan would be find what techniques are similar, or are the driving ideas, and then, build upon them. On the way, we would also try to critically examine the papers.

A major focus would also be on the definition and testing of forgetting. This would involve understanding the evaluation metrics in this problem, and then, trying to devise unit tests that would tell whether the model has successfully forgotten the ”forget set” or not.

Combining all of these, we would try to suggest some new techniques and tests for successfully unlearning data in a Bayesian paradigm. If time permits, we would like to venture into Federated Unlearning, as discussed about in [Liu et al., 2020].

2.3 Seed Papers

- A great survey and repository for Machine Unlearning: [Nguyen et al., 2022].
- A framework to identify and nullify erroneous data: [Tanno et al., 2022]
- An approach to specifically scrub the weights of the trained model, and a definition and testing of ”forgetting” can be adapted from [Golatkar et al., 2019]
- The NeurIPS 2023 Competition on Kaggle [NeurIPS, 2023]

References

[Golatkar et al., 2019] Golatkar, A., Achille, A., and Soatto, S. (2019). Eternal sunshine of the spotless net: Selective forgetting in deep networks. *CoRR*, abs/1911.04933.

[Liu et al., 2020] Liu, G., Ma, X., Yang, Y., Wang, C., and Liu, J. (2020). Federated unlearning. *arXiv preprint arXiv:2012.13891*.

[NeurIPS, 2023] NeurIPS (2023). NeurIPS 2023 - Machine Unlearning. `https://www.kaggle.com/competitions/neurips-2023-machine-unlearning/overview`.

[Nguyen et al., 2022] Nguyen, T. T., Huynh, T. T., Nguyen, P. L., Liew, A. W.-C., Yin, H., and Nguyen, Q. V. H. (2022). A survey of machine unlearning. *arXiv preprint arXiv:2209.02299*.

[Tanno et al., 2022] Tanno, R., Pradier, M. F., Nori, A., and Li, Y. (2022). Repairing neural networks by leaving the right past behind. *ArXiv*, abs/2207.04806.