THE UNIVERSITY OF HONG KONG

DEPARTMENT OF
COMPUTER SCIENCE

A Repository of Jupyter Notebooks on Unlearning in Federated Learning

Final Year Project Plan

Sheng "Victor" HUANG

Supervised by **Prof. S.M. YIU**

HONG KONG
Thu 29th Sep 2022

Contents

1 Project Background

1.1 Machine Unlearning

1.1.1 Reasons for Machine Unlearning

Machine Learning (ML) sometimes involves using data of personal and sensitve nature, such as medical [1], biometric [2] and geolocation information [3]. In some cases, we even need to bring in new data to update ML models using various methods such as incremental learning, online learning, and data stream learning [4]. On the other hand, there exists the need to remove certain data, or the influence thereof, from ML models, as well.

There are many reasons why we want to remove data or its influence from ML models. For example, in adversarial settings, data used for training may be contaminated by malicious data [5], [6], causing ML models to malfunction. Or, there may be too much data from unlikely scenarios that contributes to inefficient storage use and sometimes even leads to wrong predictions [6]. Sometimes biased data could lead to discriminatory and unfair models, exacerbating the inequality in race, sex and religion [7]. In such a scenario, one may also need to remove some data or its influence to make the model fairer. Or, in some cases, it is about regulatory compliance.

With the introduction of regulations in multiple jurisdictions, such as the European Union's General Data Protection Regulation (GPDR) [8], the California Consumer Privacy Act of 2018 (CCPA) [9] and Canada's Personal Information Protection and Electronic Documents Act (PIPEDA) [10], the concept of **right to be forgotten** or **right to erasure** has been established, in some parts of the world, where an entity may be required to erase data concerning certain person(s). In traditional cases, it is simply removing certain data from back-end databases. However, since ML models may memorize data [11], it is possible that a company may be requested to remove some individual's data from their ML models [6], the process of models "forgetting" data can be called **machine unlearning**. Due to the sheer scale of ML applications in present days, it is worth looking into this topic.

1.1.2 Goal of Machine Unlearning

A naive implementation for the same goal as machine unlearning is to have the model retrained from scratch using all the data but the ones we want to erase. This is obviously computationally expensive. However, we can learn from this scenario what unlearning outcomes we want to achieve. That is, we want models trained with dataset S after unlearning a set of data X to be "equivalent" to, or in the same distribution with [12], the model trained from scratch using dataset $S \setminus X$. We can also learn from this definition that machine unlearning is different from data deletion, with the former coming from a model perspective and the latter on a data basis [6].

1.1.3 Comparison with Differential Privacy

Before we go any further, some may raise a different but related approach for addressing privacy issues in ML called **differential privacy (DP)** [13]. However, it must be stressed

that DP is not the same as machine unlearning and their respective privacy guarantees are also different. In a machine unlearning case, the unlearning process will make sure that any influence the unlearnt data has had on the model will be cleared completely, while ε-differential privacy for any non-zero ε only helps bound the influence any data point has, but such a bound remains non-zero; a 0-differential privacy, while it can achieve 0 influence, will defeat the purpose by making the learning algorithm unable to learn anything [6], [12].

1.1.4 Challenges of Machine Unlearning

We must note that machine unlearning is very difficult [6]. There are a lot challenges that need to be tackled. First, there is the **stochasticity of training**. It is unclear how we could trace and map a single data point's effect in the training, especially in complex models, such as DNNs, given a great amount of randomness during training [12]. Second, we have the **incrementality of training**. This means the effect data X_0 being trained at time t_0 has on the model will keep influencing later training at t_i where $t_0 < t_i$ with X_i. On the other hand, model training with X_0 at t_0 is also influenced by training prior to t_0. Determining what influence we should clear from the model if we want the model to unlearn X_0 is challenging [6]. Also, Recent studies found that an unlearnt model usually underperforms compared to models trained from scratch with remaining data, with performance growing worse as we make the model unlearn more data [14], [15]. There has been little progress in preventing this **catastrophic unlearning** [15].

1.1.5 Overview of Machine Unlearning

A 2022 survey by Nguyen, Huynh, Nguyen *et al.* [6] nicely summarizes recent developments in the field, in which they divide machine unlearning strategies into three approaches: model-agnostic, model-intrinsic and data-driven methods. They compared different approaches in various unlearning scenarios, design requirements and unlearning requests. They also discussed unlearning applications, among which is unlearning federated learning, a topic on which this project will focus. They also discussed future aspects of machine learning, one of which is that federated unlearning is emerging through recent research [16]–[19]. [6] provides a high-level overview of machine unlearning to help research the topic.

1.2 Federated Learning

Federated Learning (FL) is a privacy-aware collaborative learning method first proposed by McMahan, Moore, Ramage *et al.* [20], where participants jointly train a model without sharing data. The main idea is to have distributed datasets held by participants, each of whom generates an update from training on its data, and then, either centrally [21] or decentrally [21], [22], to build an ML model based on participants' updates. FL helps accelerate model training speed and avoids direct privacy leakage [17].

It is worth noting that FL is not immune from some of the privacy vulnerabilities that other ML techniques may have [23] and that it cannot replace machine unlearning as data's influence is recorded in sub-updates that are sent out during FL process.

2 Project Objectives

This project takes inspiration from an HKUCS final year project in 2020-21, *Building a code and data repository for teaching algorithmic trading*[1] by Woo, Wu and Lee, which is the champion of 2020-21 FYP competition at HKUCS.

This project takes aim at several educational and research objectives towards the promotion and betterment of the topic on unlearning in FL, a relatively new area that has seen some research developments in recent years.

2.1 Educational Objectives

The main educational objective of this project is to provide materials for learning the topic of unlearning in FL. The materials, including documentations, datasets and experiment instructions, code and scripts, will gather together the ML, security and privacy knowledge and outtakes from recent research publications in an interactive, clear and organized way to allow readers to learn about this topic from basic knowledge to the frontline of advanced research studies, so that non-specialists and experts in related areas can all benefit from this project.

2.2 Research Objectives

The main research objective is to examine the unlearning in FL methods proposed in recent publications, such as those by Liu, Ma, Yang *et al.* [16], Liu, Ma, Yang *et al.* [17], Liu, Xu, Yuan *et al.* [24], Wang, Guo, Xie *et al.* [18], Gao, Ma, Wang *et al.* [25], Wu, Zhu and Mitra [19] and Halimi, Kadhe, Rawat *et al.* [26], and compare their approaches to machine unlearning in FL using different datasets. This project will draw conclusions from these experiments and present the findings in a clear and detailed fashion. An optional objective is to, if possible, find directions in which further research could be done that may improve the performance of unlearning in FL.

2.3 Reporting and Deliverables

This project aims to meet all the requirements of HKUCS course COMP4801 Final Year Project on time and with high standards. This project will produce a web page, a detailed project plan, a detailed interim report and a final report, as required, and preliminary implementation and finalized tested implementation at the time of interim and final reporting, as required. This project's documents will be made according to the course guideline and supervisor's requirements, if any.

2.4 Supervision and Relationship

The project owner values professional relationships and believes a good such relationship will help ensure the quality of the project. This project aims to build a close working relationship with the supervisor. Project progress will be updated with the supervisor regularly, and consultation and advice will be sought frequently.

[1] https://awoo424.github.io/algotrading_fyp/

3 Project Methodology

3.1 Reading

The project will search for related research publications using Google Scholar. Some of the publications in search results are from arXiv, ACM Digital Library, IEEE Xplore, SpringerLink and official proceedings of conferences such as AAAI, NeurIPS and PMLR. This project will use HKUL E-resources to access these resources, if paywalled. This project will also closely monitor news from the first IEEE Conference on Secure and Trustworthy Machine Learning (IEEE SaTML 2023), which will take place in February 2023, as it is related to the topic of this project.

3.2 Experiments

This project will conduct experiments on HKUCS GPU Farm, constructing testing and development environments using technologies such as Miniconda, CUDA Toolkit and packages used in the implementation of related publications. This project will conduct comparative experiments on different techniques proposed by different publications using datasets nicely summarized by [6], such as MNIST, CIFAR, SVHN and Adult. More experiments will be designed and carried out if promising new directions are found.

3.3 Notebooks

This project will use Jupyter Notebooks to document the concepts and to design small interactive experiments. Instructions of larger experiments, which may not be suitable to run directly from notebooks, will also be noted. Users will be directed to conduct their own experiments in a different setting. The notebooks will be systematically organized and put into a repository hosted on a cloud storage and sharing system. The project page will show the structure of the repository and have links to each notebook. Users will have the option to download or open the notebooks in Google Colab, whenever appropriate. Additional information on notebook setup will also be shown on the web page.

3.4 Reporting

This project will use LaTeX, a high-quality typesetting system and *de facto* standard for the communication and publication of scientific documents, to generate the reports required by the FYP course. More specifically, Overleaf will be used for its ease of use and support for a large amount of packages.

3.5 Web Page

The project web page[2] will be hosted on GitHub Pages with GitHub Actions, a world class CI/CD tool, and Hugo, which supports the Markdown language and is one of the most popular open-source static site generators. This set of technologies were chosen for their prevalence in the industry, ease of use and powerful functionalities.

[2] https://vicw0ng-hk.github.io/feul/

4 Project Schedule

An agile schedule based on course requirements and project nature will be followed.

In the inception phase from September to early October, most of the setting up will be done, including an FYP account, development and testing environments and a web page. Early-stage experiments will be carried out to test the environments. Some reading will be done and a detailed plan will be written.

In elaboration phase from October to January, the base will be built with more reading and testing, and detailed documentation and demonstration code will be written. Preparations will be made for first presentation and an interim report will be written.

In construction phase from January to April, more advancements will be made based on previous stages. With more reading and experiments, more of the topic will be explored and tested. Final presentation and report will be prepared.

In late April, Preparations will be made for project exhibition.

Tentative Schedule		
Time	**Milestones**	**Notes**
- 02 Oct 2022	Read on related concepts	
	Set up FYP accounts, testing environment and development environment	
	Conduct small experiments	
	Write Plan and build web page	Deliverable 1 due 02 Oct 2022
- 08 Jan 2023	Continued and more advanced reading	
	Write in Notebooks on concepts	
	Test unlearning frameworks	
	Build demo Notebooks	
	Prepare for 1st Presentation	1st Presentation 09 - 13 Jan 2023
- 22 Jan 2023	Write Interim Report and Wrap up preliminary implementation	Deliverable 2 due 22 Jan 2023
- 02 Apr 2023	Comprehensively test and compare methods in readings	
	Summarize knowledge and write in Notebooks	
	Explore future research and development directions	
	Develop interactive learning materials in Notebooks	
- 16 Apr 2023	Prepare for Final Presentation	Final Presentation 17-21 Apr 2023
- 18 Apr 2023	Write Final Report and Wrap up final implementation	Deliverable 3 due 18 Apr 2023
- 02 May 2023	Prepare for Project Exhibition	

5 References

[1] J. A. Castellanos-Garzón, E. Costa, J. L. Jaimes S. and J. M. Corchado, 'An evolutionary framework for machine learning applied to medical data,' *Knowledge-Based Systems*, vol. 185, p. 104 982, 2019, ISSN: 0950-7051. DOI: 10.1016/j.knosys.2019.104982. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0950705119304046.

[2] B. Arslan, E. Yorulmaz, B. Akca and S. Sagiroglu, 'Security perspective of biometric recognition and machine learning techniques,' in *2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA)*, 2016, pp. 492–497. DOI: 10.1109/ICMLA.2016.0087.

[3] P. Zola, P. Cortez and M. Carpita, 'Twitter user geolocation using web country noun searches,' *Decision Support Systems*, vol. 120, pp. 50–59, 2019, ISSN: 0167-9236. DOI: https://doi.org/10.1016/j.dss.2019.03.006. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167923619300442.

[4] H. M. Gomes, J. Read, A. Bifet, J. P. Barddal and J. Gama, 'Machine learning for streaming data: State of the art, challenges, and opportunities,' *SIGKDD Explor. Newsl.*, vol. 21, no. 2, pp. 6–22, Nov. 2019, ISSN: 1931-0145. DOI: 10.1145/3373464.3373470. [Online]. Available: https://doi-org.eproxy.lib.hku.hk/10.1145/3373464.3373470.

[5] P. McDaniel, N. Papernot and Z. B. Celik, 'Machine learning in adversarial settings,' *IEEE Security & Privacy*, vol. 14, no. 3, pp. 68–72, 2016. DOI: 10.1109/MSP.2016.51.

[6] T. T. Nguyen, T. T. Huynh, P. L. Nguyen, A. W.-C. Liew, H. Yin and Q. V. H. Nguyen, *A survey of machine unlearning*, 2022. DOI: 10.48550/ARXIV.2209.02299. [Online]. Available: https://arxiv.org/abs/2209.02299.

[7] S. Verma, M. Ernst and R. Just, *Removing biased data to improve fairness and accuracy*, 2021. DOI: 10.48550/ARXIV.2102.03054. [Online]. Available: https://arxiv.org/abs/2102.03054.

[8] European Parliament and Council of the European Union, *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance)*, 2016. [Online]. Available: https://eur-lex.europa.eu/eli/reg/2016/679/oj.

[9] California State Legislature, *An act to add title 1.81.5 (commencing with section 1798.100) to part 4 of division 3 of the civil code, relating to privacy.* 2018. [Online]. Available: https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=201720180AB375.

[10] Parliament of Canada, *Personal information protection and electronic documents act*, 2000. [Online]. Available: https://laws-lois.justice.gc.ca/ENG/ACTS/P-8.6/FullText.html.

[11] S. Chatterjee, 'Learning and memorization,' in *Proceedings of the 35th International Conference on Machine Learning*, J. Dy and A. Krause, Eds., ser. Proceedings of Machine Learning Research, vol. 80, PMLR, Jul. 2018, pp. 755–763. [Online]. Available: https://proceedings.mlr.press/v80/chatterjee18a.html.

[12] L. Bourtoule, V. Chandrasekaran, C. A. Choquette-Choo *et al.*, 'Machine unlearning,' in *2021 IEEE Symposium on Security and Privacy (SP)*, 2021, pp. 141–159. DOI: 10.1109/SP40001.2021.00019.

[13] C. Dwork, 'Differential privacy: A survey of results,' in *Theory and Applications of Models of Computation*, M. Agrawal, D. Du, Z. Duan and A. Li, Eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 1–19.

[14] Q. P. Nguyen, R. Oikawa, D. M. Divakaran, M. C. Chan and B. K. H. Low, 'Markov chain monte carlo-based machine unlearning: Unlearning what needs to be forgotten,' ser. ASIA CCS '22, Nagasaki, Japan: Association for Computing Machinery, 2022, pp. 351–363, ISBN: 9781450391405. DOI: 10.1145/3488932.3517406. [Online]. Available: https://doi.org/10.1145/3488932.3517406.

[15] Q. P. Nguyen, B. K. H. Low and P. Jaillet, 'Variational bayesian unlearning,' in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan and H. Lin, Eds., vol. 33, Curran Associates, Inc., 2020, pp. 16 025–16 036. [Online]. Available: https://proceedings.neurips.cc/paper/2020/file/b8a6550662b363eb34145965d64d0cfb-Paper.pdf.

[16] G. Liu, X. Ma, Y. Yang, C. Wang and J. Liu, 'Federaser: Enabling efficient client-level data removal from federated learning models,' in *2021 IEEE/ACM 29th International Symposium on Quality of Service (IWQOS)*, 2021, pp. 1–10. DOI: 10.1109/IWQOS52092.2021.9521274.

[17] Y. Liu, Z. Ma, Y. Yang, X. Liu, J. Ma and K. Ren, 'Revfrf: Enabling cross-domain random forest training with revocable federated learning,' *IEEE Transactions on Dependable and Secure Computing*, pp. 1–1, 2021. DOI: 10.1109/TDSC.2021.3104842.

[18] J. Wang, S. Guo, X. Xie and H. Qi, 'Federated unlearning via class-discriminative pruning,' in *Proceedings of the ACM Web Conference 2022*, ser. WWW '22, Virtual Event, Lyon, France: Association for Computing Machinery, 2022, pp. 622–632, ISBN: 9781450390965. DOI: 10.1145/3485447.3512222. [Online]. Available: https://doi.org/10.1145/3485447.3512222.

[19] C. Wu, S. Zhu and P. Mitra. 'Federated unlearning with knowledge distillation.' arXiv: 2201.09441. (2022).

[20] B. McMahan, E. Moore, D. Ramage, S. Hampson and B. A. y. Arcas, 'Communication-Efficient Learning of Deep Networks from Decentralized Data,' in *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, A. Singh and J. Zhu, Eds., ser. Proceedings of Machine Learning Research, vol. 54, PMLR, Apr. 2017, pp. 1273–1282. [Online]. Available: https://proceedings.mlr.press/v54/mcmahan17a.html.

[21] P. Kairouz, H. B. McMahan, B. Avent *et al.*, 'Advances and open problems in federated learning,' *Foundations and Trends® in Machine Learning*, vol. 14, no. 1–2, pp. 1–210, 2021, ISSN: 1935-8237. DOI: 10.1561/2200000083. [Online]. Available: http://dx.doi.org/10.1561/2200000083.

[22] S. R. Pokhrel and J. Choi, 'Federated learning with blockchain for autonomous vehicles: Analysis and design challenges,' *IEEE Transactions on Communications*, vol. 68, no. 8, pp. 4734–4746, 2020. DOI: 10.1109/TCOMM.2020.2990686.

[23] B. Hitaj, G. Ateniese and F. Perez-Cruz, 'Deep models under the gan: Information leakage from collaborative deep learning,' in *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, ser. CCS '17, Dallas, Texas, USA: Association for Computing Machinery, 2017, pp. 603–618, ISBN: 9781450349468. DOI: 10.1145/3133956.3134012. [Online]. Available: https://doi.org/10.1145/3133956.3134012.

[24] Y. Liu, L. Xu, X. Yuan, C. Wang and B. Li. 'The right to be forgotten in federated learning: An efficient realization with rapid retraining.' arXiv: 2203.07320. (2022).

[25] X. Gao, X. Ma, J. Wang *et al.* 'Verifi: Towards verifiable federated unlearning.' arXiv: 2205.12709. (2022).

[26] A. Halimi, S. Kadhe, A. Rawat and N. Baracaldo, *Federated unlearning: How to efficiently erase a client in fl?* 2022. DOI: 10.48550/ARXIV.2207.05521. [Online]. Available: https://arxiv.org/abs/2207.05521.