A Repository of Jupyter Notebooks on Unlearning in Federated Learning

Final Year Project Plan

Sheng “Victor” HUANG

Supervised by Prof. S.M. YIU
1 Project Background

1.1 Machine Unlearning

1.1.1 Reasons for Machine Unlearning

Machine Learning (ML) sometimes involves using data of personal and sensitive 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 (GDPR) [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 $\varepsilon$-differential privacy for any non-zero $\varepsilon$ 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\footnote{https://awoo424.github.io/algotrading_fyp/} 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.\cite{16}, Liu, Ma, Yang et al.\cite{17}, Liu, Xu, Yuan et al.\cite{24}, Wang, Guo, Xie et al.\cite{18}, Gao, Ma, Wang et al.\cite{25}, Wu, Zhu and Mitra\cite{19} and Halimi, Kadhe, Rawat et al.\cite{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.
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\footnote{https://vicw0ng-hk.github.io/feul/} 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.
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.

<table>
<thead>
<tr>
<th>Tentative Schedule</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time</td>
</tr>
<tr>
<td>- 02 Oct 2022</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>- 08 Jan 2023</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>- 22 Jan 2023</td>
</tr>
<tr>
<td>- 02 Apr 2023</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>- 16 Apr 2023</td>
</tr>
<tr>
<td>- 18 Apr 2023</td>
</tr>
<tr>
<td>- 02 May 2023</td>
</tr>
</tbody>
</table>
5 References


