Analysis of Survey paper - CS240

Sameer Patil, Harsh Suthar

April 2025

1 Classification of Unlearning techniques

1.1 Data-oriented unlearning

Data-oriented techniques refer to the unlearning methods that data holders erase the forgotten data through manipulating the original training set, mainly including two types according to different processing strategies, namely data partition and data modification

1. **Data Partition:** In data partition, the data holders divide the original training set into several subsets and train the corresponding submodels on each subset. Then these submodels are used to aggregate a prediction using an aggregation function. When an unlearning request is received, the data holders delete the forgotten data from the subsets that contain them and retrain the correspoding submodels.

2. **Data Modification:** In data modification, the data holders modify the training set D with adding noise or new transformed data D_T , e.g. using transformed data D_T to replace the forgotten data D_f , that is $D' = D_T \cup D_r$. This type of methods can simplify retraining and speed up unlearning.

1.2 Model-oriented Techniques

Model-oriented techniques refer to the unlearning methods that data holders complete unlearning through modifying the original model M_o . The technique can be categorized into model reset and model modification.

1. **Model Reset:** In model reset, the data holders directly update the model parameters to eliminate the impact of forgotten data D_f on the model M_o , that is, $w_u = w_o + \sigma$, where w_o and σ are the parameters of M_o and the value to update, respectively.

2. **Model Modification:** In model modification, the data holders replaces the relevant parameters of D_f with the calculated parameters to remove D_f from the original model M_o , that is,$w_u = w_o \cup w_{cal}$, where w_{cal} is the calculated parameters for replacement.

2 What are the risks in Machine unlearning?

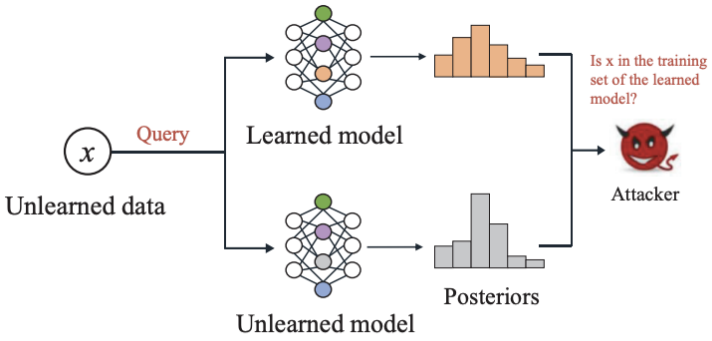

Fig. 1. The Example of Privacy Violation in Machine Unlearning

2.1 Vulnerabilities

Attackers can use the relation between learned model and unlearned model to:

1. **Break the model**
2. **Steal private information**

Some types of attacks to execute these functions are:

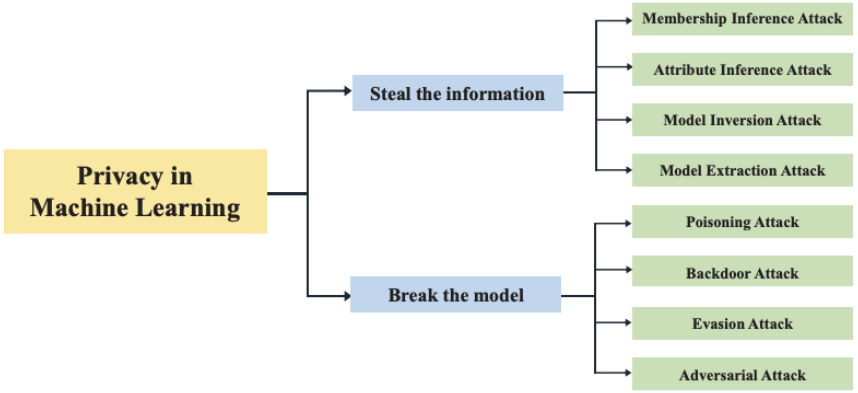

Figure 1: Types of attacks on ML models

Unlearning Techniques		Original Models	Request Types	Accuracy	Effectiveness	Efficiency	Privacy Vulnerability
Data Partition	Boutoule et al. [10]	DNN	Batch and stream	✗	✗	✓	Membership inference attack
	Ginart et al. [21]	k-Means	Sample and stream	✗	✗	✓	Membership inference attack*
	Gupta et al. [25]	Non-convex models	Adaptive stream	✗	✗	✓	Membership inference attack*
	Chen et al. [28]	GNN	Node and edge	✗	✗	✓	Membership inference attack*
	Wang et al. [29]	GNN	Node, edge and feature	✗	✗	✓	Membership inference attack*
Data Modification	Cao et al. [8]	Statistical query and Bayes models	Sample	✓	✓	✓	Attribute inference attack*
	Tarun et al. [32]	DNN	Class	✗	✗	✓	Attribute inference attack*
	Guo et al. [31]	DNN	Attribute	✓	✓	✓	Attribute inference attack*

Figure 2: Summary and Comparison of Data-oriented Unlearning Techniques

Unlearning Techniques		Original Models	Request Types	Accuracy	Effectiveness	Efficiency	Vulnerability
Model Reset	Guo et al. [11]	Linear models	Sample and batch	✗	✗	✓	Poisoning attack
	Golatkar et al. [44]	DNN	Sample and class	✗	✗	✓	Poisoning attack*
	Golatkar et al. [45]	DNN	Sample and class	✗	✗	✓	✓
	Izzo et al. [37]	Linear and logistic models	Sample and batch	✗	✗	✓	Poisoning attack*
	Neel et al. [39]	Convex models	Stream	✗	✗	✓	Poisoning attack*
	Chouasia et al. [40]	Convex and non-convex models	Adaptive stream	✗	✗	✓	Poisoning attack*
	Chien et al. [58]	GNN	Node, edge and feature	✗	✗	✓	✓
	Wu et al. [42]	GNN	Edge	✗	✗	✓	Poisoning attack*
	Warnecke et al. [46]	Convex and non-convex models	Feature and label	✗	✗	✓	Attribute inference attack*
Model Modification	Brophy et al. [13]	Random forest	Batch	✓	✓	✓	Model inversion attack*
	Wu et al. [47]	Decision tree	Batch	✓	✓	✓	Model inversion attack*
	Kong et al. [53]	GAN	Sample	✓	✓	✓	Model inversion attack*
	Bae et al. [7]	GAN	Class and feature	✓	✓	✓	Model inversion attack*
	Moon et al. [56]	GAN	Feature	✗	✗	✓	✓
	Sun et al. [57]	GAN	Sample and class	✗	✓	✓	✓
	Chundawat et al. [48]	DNN	Sample and class	✗	✗	✓	✓
	Chundawat et al. [49]	DNN	Sample and class	✗	✗	✓	✓
	Baumhauer et al. [50]	DNN	Class	✗	✗	✓	✓

Figure 3: Summary and Comparison of Model-oriented Unlearning Techniques

3 Keys metrics to determine the effectiveness of Unlearning technique

1. **Forgetting Effectiveness**
   The method must sufficiently "erase" the influence of the target class.
2. **Accuracy Retention**
   The performance on the remaining classes
3. **Computational Efficiency**
   Time taken to unlearn vs. time taken to retrain
4. **Complexity and Practicality**
   Optimization required for each particular model

4 What techniques will be suitable for Image classfication Unlearning?

4.1 Selective Synaptic Dampening (SSD)

SSD leverages the Fisher Information Matrix (FIM) to measure the sensitivity of each model parameter to training examples.

4.2 Incompetent Teacher

This method sets up a teacher–student framework where a "student" model is guided to forget by training on the forget set under the supervision of an "incompetent teacher" (i.e., a teacher that is deliberately made less competent on that subset) while a "competent teacher" preserves performance on the retain set.

4.3 SCRUB

SCRUB alternates between "max-steps" (to enforce forgetting) and "min-steps" (to protect the retained data) by modifying the loss function.

4.4 DeepClean and Gradient-Free Unlearning

DeepClean and similar methods aim to identify and reset only the sensitive weights using approximations (e.g., by computing the diagonal of the FIM) so that the model "forgets" specific classes with minimal impact on retention accuracy.

4.5 Saliency-Based Unlearning (SalUn)

SalUn focuses on computing a "weight saliency" score that identifies which model weights are most responsible for encoding information about a particular class.