Continual Learning + Machine Unlearning

Pengxiang Wang

Peking University, School of Mathematical Sciences

University of Bristol, School of Engineering Mathematics and Technology

2024-10-28

Machine Unlearning

Machine Unlearning Motivation

What is **machine unlearning**:

> ***Machine unlearning*** *is the process of deliberately removing specific data from a machine learning model to ensure that the removed data no longer influences the model's predictions – an undo option of machine learning process.*

Data Deletion:

- Traditionally: delete from databases
- AI: delete both from back-end databases and from trained models

Application Movitation:

- **Privacy**:
  - Regulations: GDPR, CCPA, etc. when the user withdraw the consent, "the right to be forgotten"
  - Delete the requested data by users
- **Security**:
  - Adversarial attacks are possible to extract private information from the trained model. E.g., model inversion attacks, membership inference attacks, etc.
  - Delete the adversial data

Machine Unlearning Framework

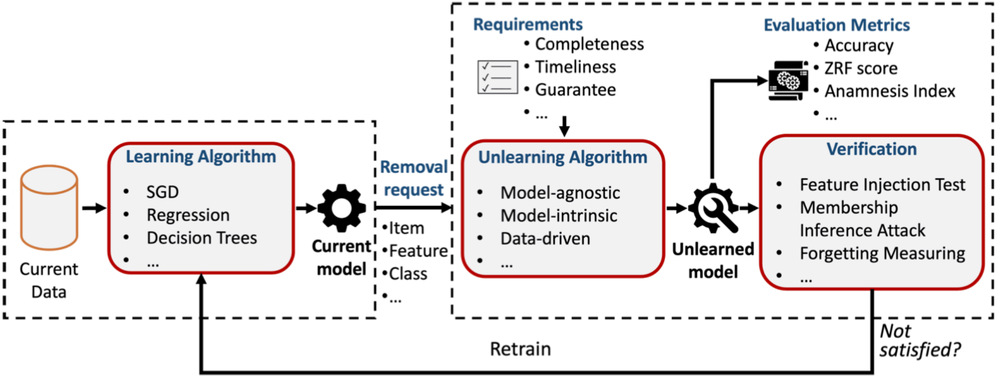


Figure 1: Machine Unlearning Framework

Formal Definition

D = Dr+ Df

Df: forget set

Assumptions:

- The unlearning data are not big. Practically considering, also Otherwise, it is easier to do retraining.

Retraining

The problem makes unlearning difficult:

- Neural networks parameters do not tend to show any clear connection to the training data. AI models have to be considered as a whole.
- Stochasticity and Incrementality of training
- unlearning catastrophic unlearning, reduce performance

Retraining:

- Delete target data and re-train the model with the rest of data from scratch
- A naive way, but not always feasible
- Achieves upper bound

The problem of retraining: - Doesn't worth, computation cost - Not always having acess to all training data

Methodology

Scenarios - Data Deletion - Class Removal

- Model-Agnostic or Model-Intrinsic
- Data-Driven Approaches, most model-agnostic

Method: SISA

Data Partitioning (Efficient Retraining)

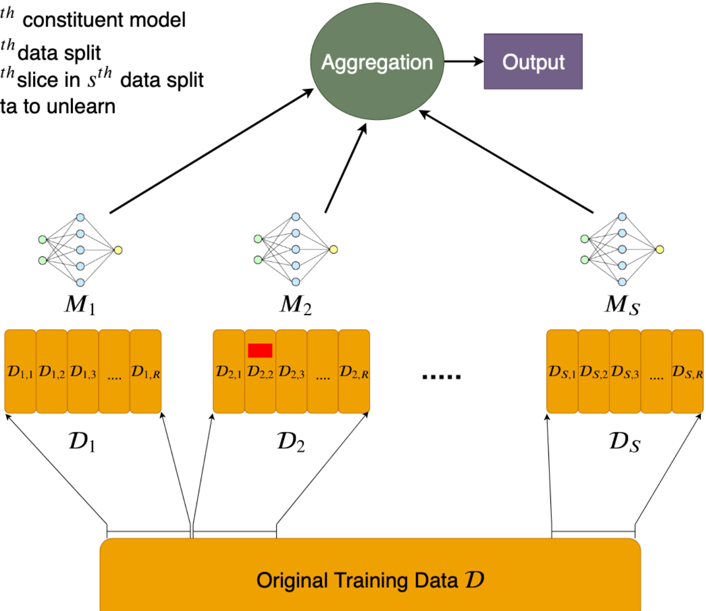
SISA (Sharded, Isolated, Sliced, Aggregated), 2021:

- Isolate: Isolate network and slice data into shards
- build up correspondance bewteen divided network and data
- Retraining the corresponding network of the data shard to be forgotten

Fractioning the retraining into smaller units

Method: SISA

- M_s : s^{th} constituent model
- D_s : s^{th} data split
- $D_{s,r}$: r^{th} slice in s^{th} data split
- 🟥 : data to unlearn


Aggregation
Output
M_1
M_2
M_S
$D_{1,1}$ $D_{1,2}$ $D_{1,3}$ .... $D_{1,R}$
$D_{2,1}$ $D_{2,2}$ $D_{2,3}$ .... $D_{2,R}$
.....
$D_{S,1}$ $D_{S,2}$ $D_{S,3}$ .... $D_{S,R}$
D_1
D_2
D_S
Original Training Data D