NeurIPS: Machine Unlearning

Project Proposal

CSE 472 Machine Learning Sessional

BY:
Syed Jarullah Hisham - 1805004
Abdur Rafi - 1805008

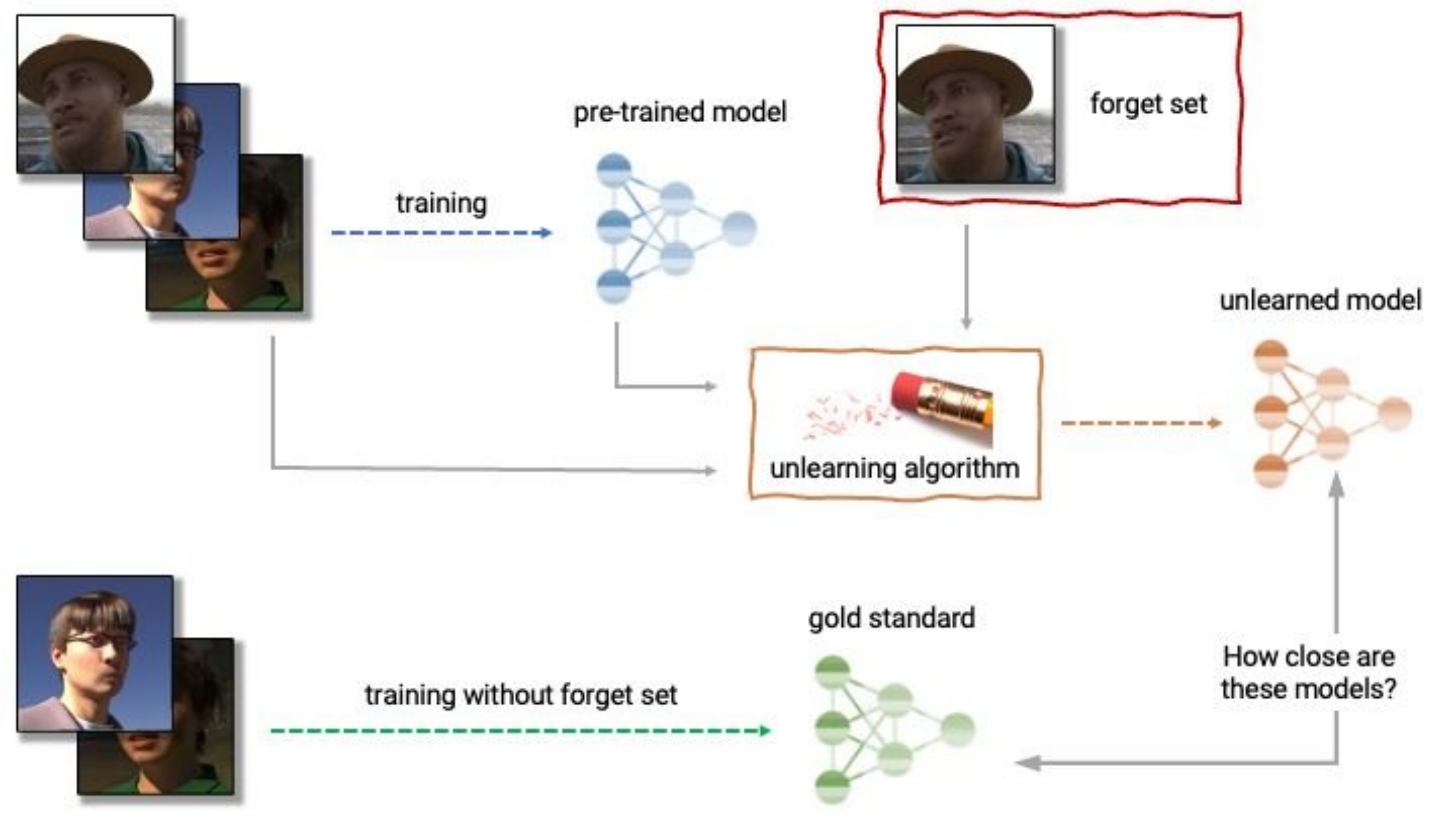
GOOGLE · RESEARCH CODE COMPETITION · A MONTH AGO

NeurIPS 2023 - Machine Unlearning

Erase the influence of requested samples without hurting accuracy

Competition Link: https://www.kaggle.com/competitions/neurips-2023-machine-unlearning/overview

Problem Description



Motivation

- Removing examples from a trained machine learning model is a major unsolved problem of ML privacy research
- recent research has shown that it may be possible to infer with high accuracy whether an example was used to train a machine learning model using **membership inference attacks** (MIAs)
- retraining whole model can be computationally expensive

Dataset & Code

- Competition dataset is hidden
- But this project can be tested on several datasets(e.g: CIFAR10, EMNIST etc)

Dataset Description
https://www.kaggle.com/competitions/neurips-2023-machine-unlearning/data

Solution Code(For reproduction)
https://www.kaggle.com/code/fanchuan/2nd-place-machine-unlearning-solution?kernelSessionId=153137657

Solution Overview

3 Main approaches

1. Unsupervised learning(self supervised contrastive learning), then retrain
2. Add noise to weights, then retrain
3. Re-initialize some weights, then retrain

Experiment To Reproduce

1st place solution. 2 stage

- Forget Stage
    - SimCLR (Self supervised contrastive learning) to push the distance between forget samples and retain samples
- Retain Stage
    - Train on retain samples to finetune performance on retain samples

New Experiment

- Calculate gradients on forget set and retain set
- Subtract component of retain set gradient from forget set gradient
- Use remaining component to climb up cost function

Thank You