Report date: 3/10/24

Paper report:

**Overview:**

TLDR: Machine Unlearning Bible

The paper is extremely interesting and explains the state of the art regarding machine unlearning.

The introduction explains why MUL (Machine UnLearning) is important and why it's needed.

The paper introduces the requirements for a complete framework for unlearning, explains what the requirements should be and how it should be evaluated.

Framework requirements:

* Consistency (unlearned vs retrained model with the updated data should provide the same predictions)
* Timeliness
* Accuracy
* Provable guarantees
* Model-agnostic
* Verifiability (mechanism to allow the user to verify the effective removal)

Lists the different kind of learning requests, namely:

* Feature removal
* Class removal
* Task removal (for models that accomplish different tasks, those in lifelong learning)
* Stream removal

How to verify unlearning:

* Feature injection test (inject feature that flags data to unlearn, unlearn, check that the feature has a weight equal or very close to zero in the unlearnt model)
* Forgetting measuring (model is alpha-forget if a privacy attack has a success rate no greater than alpha)
* Information leakage (check leakages by trying to understand if a sample was part of the unlearnt training set, can be done via membership attack, but any kind of attack can be used)
* Membership inference attack (I think it's more like a subset of the Information Leakage)
* Backdoor attacks
* Slow down attacks (not too sure about how relevant it is here)
* Interclass confusion test (compares confusion matrix of the unlearnt model vs a model trained with wrong labels)
* Federated verification
* Cryptographic protocol

It formally defines the unlearning problem (both for exact and approximate unlearning).

Presents the main unlearning scenarios:

* Zero glance unlearning (unlearning can only use the retained data)
* Zero shot unlearning (unlearning without training data, often relaxed to unlearn some classes without the training data)
* Few shot unlearning (unlearning with a small portion of the data to unlearn)

Classifies the unlearning algorithms:

* Model agnostics algorithms:
  + Differential privacy
  + Certified removal mechanism
  + Statistical query learning
  + Decremental learning
  + Knowledge adaptation
  + MCMC Unlearning
* Model intrinsic algorithms:
  + Unlearning for softmax classifiers
  + Unlearning for linear models
  + Unlearning for tree based models
  + Unlearning for bayesian models
  + Unlearning for deep neural network based models
* Data driven approaches:
  + Data augmentation
  + Data influence

It goes on with citing the most important datasets used in MUL.

Lists the evaluation metrics:

* Accuracy
* Completeness (complete unlearning of the data to forget)
* Unlearning time vs retrain time
* Relearn time (time it takes to recover the performance on unlearned data)
* Layer wise distance (for neural networks)
* Activation distance (see the distance in the activation vectors, I need to understand better)
* JS-Divergence (I need to understand better)
* Membership inference
* ZRF Score (evaluate without comparing with the retrained model, compares the unlearned data predictions against an unskilled instructor)
* Anamnesis index (retrain time on steroids)
* Epistemic uncertainty (it's often used to measure model uncertainty due to lack of training data)

Then quickly explains a couple of interesting scenarios where unlearning can happen, namely:

* Recommender systems
* Federated learning
* Graph embeddings
* Lifelong learning

The paper concludes by mentioning the trends:

* Influence functions and first/second order updates yield great results and are quite popular in the field now
* Reachability of model parameters (not sure if I understood correctly) they argue that the retrained model may not always be the gold standard since it may train the same weights as the same model but with the data to forget, therefore providing the same predictions (I don't see the point of this argument: if it's trained without the data to forget then we're good)
* There still isn't a commonly accepted and used way to define is a data has been unlearnt
* Federated learning is cool and growing but it's a mess to unlearn on it
* Unlearning can be beneficial for models poisoned by adversarial attacks, models biased due to bias in some training data, untraining overtrained DNN to avoid reusing obsolete and redundant samples

They do also provide a list of very specific research questions.

**Repo:**

<https://github.com/tamlhp/awesome-machine-unlearning>

**Authors:**

Thanh Tam Nguyen et alii

**Year:**

2022