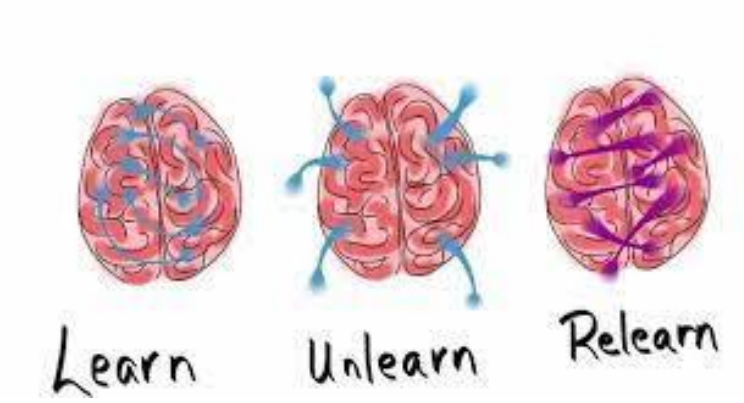
SISA: Application of Machine Unlearning For Classification

Team 2: Michael Neely,Abhishek Shastry HM, Bhavi Shah, Simoni Shah, Vidhi Sharma

Introduction

- The primary goal of Machine Unlearning is to remove all trace of a data point from a Machine Learning model or system, without affecting its performance.
- Machine Unlearning gives people more control over their data and its derivatives.
- This presentation will introduce Machine Unlearning using Shard Isolated Sliced and Aggregated(SISA) framework.

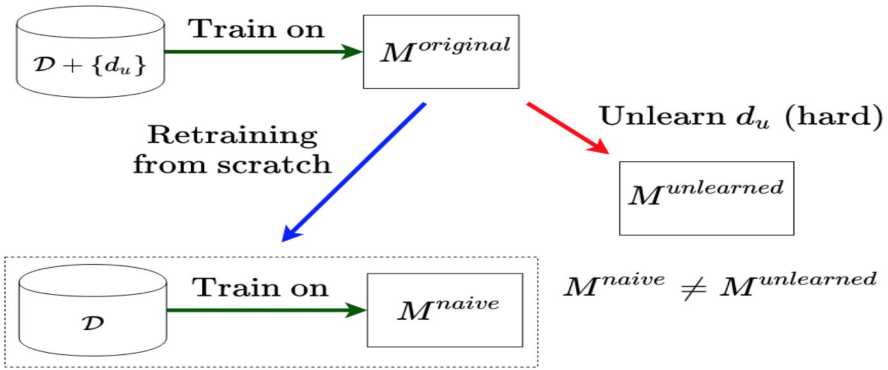


Research Purpose

- The purpose of the research is to explore the areas in Machine Unlearning which can substitute retraining the entire model from scratch.
- Conventionally in the machine learning process, whenever the data gets deleted, it is deleted from the database, and the model is re trained again..
- Data deletion algorithms aim to remove the influence of deleted data points from trained models at a cheaper computational cost than fully retraining those models.

Research Question

Is Machine Unlearning a viable option for classification tasks?



Background: Right to be forgotten

Need for data protection and privacy of User data:

- Many new privacy legislation called for more clarity around the data privacy
- Users are now empowered to remove there data

This creates a problem, how do we unlearn from already trained ML models

Background: The Motivation

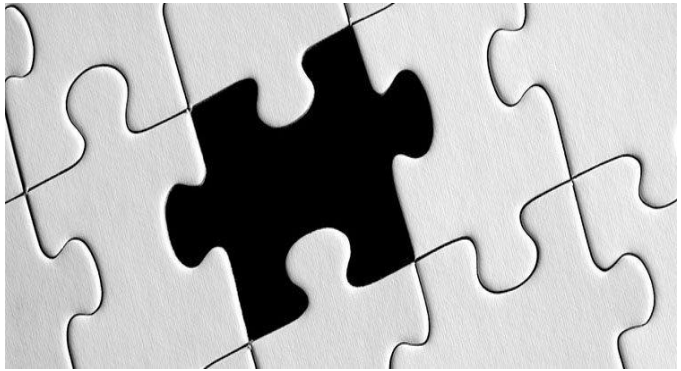
The research is primarily motivated by privacy. Machine unlearning can be a first step towards achieving model governance.

In the paper we followed, researchers from the University of Toronto, Vector Institute, and University of Wisconsin-Madison propose SISA training, a new framework that helps models “unlearn” information by reducing the number of updates that need to be computed when data points are removed.

SISA is designed to achieve the largest improvements for stateful algorithms like stochastic gradient descent for deep neural networks..

Background: Gaps in the study

The study falls under the category of supervised machine learning. It gives valid guarantees only for sequences that are chosen independently of the models that are published but fails when the deletion requests are adaptive in nature where the deletion requests are dependent on the published models.



Methodology

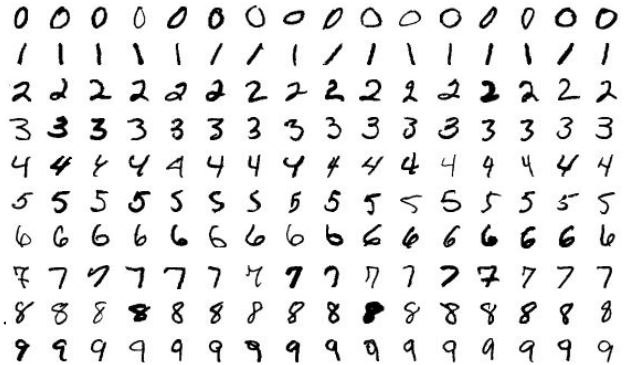
Data

1. **Purchase Dataset**

- Provided by the original machine unlearning paper's Github repository.
- Curated by choosing the top 600 most purchased items based on the category attribute.
- 2500000 samples and consists of 2 classes
    - Labeled 0 or 1
- Binary Classification Experiments

2. **MNIST Dataset**

- Image classification dataset that consists of data samples of handwritten digits written by different writers.
- 60000 data samples and 10 classes which are numerically labeled 0-9.
- Multi-Label Classification Experiments

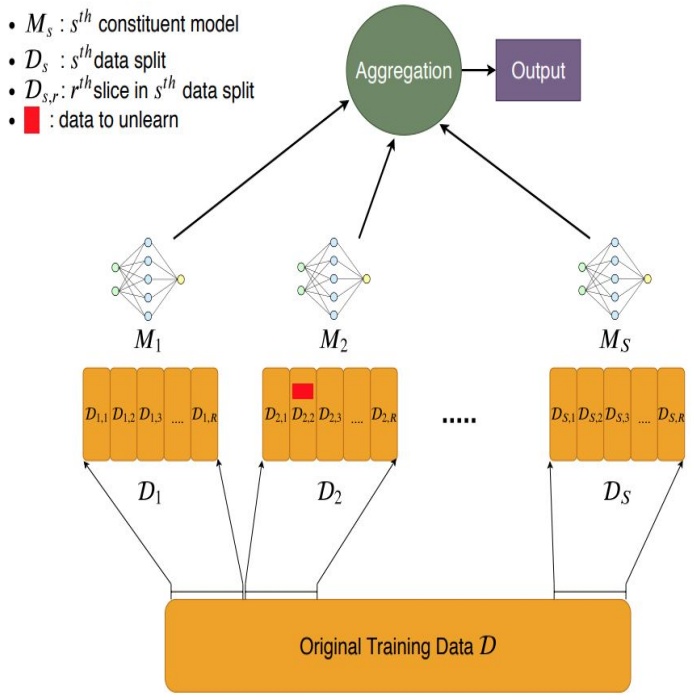


SISA (Sharded, Isolated, Sliced, and Aggregated) Framework

I. Training data is first divided into S shards and further partitioning each shard into R slices hat are used to incrementally tune and save and store the parameter state of our model.

II. Each individual model is trained in isolation to one another to ensure influence of data points on that specific model.

III. Each individual shard is trained on their respective model.

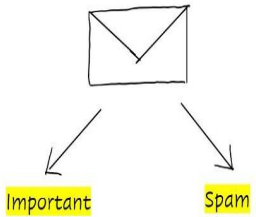


Cont' SISA

I. After the training process, the resulting models are then aggregated to output our single final model.
    - Aggregation Strategy: Label-Based Majority Voting
II. When an unlearning request is made, we only retrain the affected model and revert to the parameter state before the slice with the data point was used in training.

Cont’

Modeling Framework



- Framework: SISA
- Model: Categorical Naive Bayes Classifier Model
  - Ease of implementation
  - Good for testing as a baseline
- No prior pre-processing and transformations were applied to our input training data in the model fitting process.

Cont’

Experiment Setup

Experiments were divided into two parts: 1. Binary Classification and 2. Multi-label classification.

Binary Classification

- Trained our model on 5000 samples and made predictions on 2000 samples.
- Generated random sequence of unlearning requests
    - Increments of 20%, 40%, 60% and 80% of our training data
    - Retrain both SISA Naive Bayes and Baseline Naive Bayes at each unlearning request

Multi-Label Classification

- Trained our model was trained on 5000 samples and made predictions on 1667 samples.
- Generated random sequence of unlearning requests
    - Increments of 20%, 40%, 60% and 80% of our training data
    - Retrain both SISA Naive Bayes and Baseline Naive Bayes at each unlearning request

Findings

Tests were designed to explore the limitations of machine unlearning in the context of classification tasks.

Goal of testing was to answer the following questions:

1. Does utilizing machine unlearning provide improved classification accuracy compared to retraining model from scratch?
2. Is there improved classification times?
3. Does sharding impact prediction accuracy?

Cont’

Results

- SISA Naive Bayes produces a slight increase in prediction accuracy
  - Dependent on the number of shards
- There is a greater speed of retraining for SISA Naive Bayes Model as opposed to retraining from scratch
- Increasing the number of shards will degrade the prediction accuracy

Improved Classification Accuracy & Sharding

- Top performer: 3 Shards, 10 Slices
- SISA implemented Naive Bayes produced more consistent results as you keep removing training data
    - Only retraining individual slices for our model not affected by the unlearning requests
- Increased Sharding will degrade classification accuracy
    - More shards = More weak learners
    - Weakened generalizability due to less samples per shard to fit our models

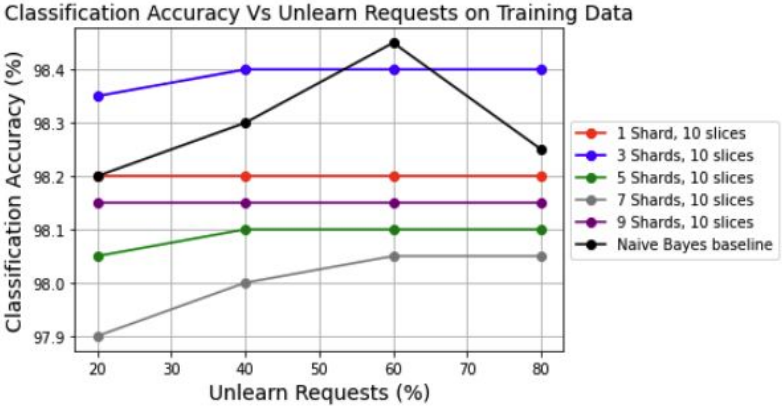


Figure 1: Purchase Dataset Classification Accuracy

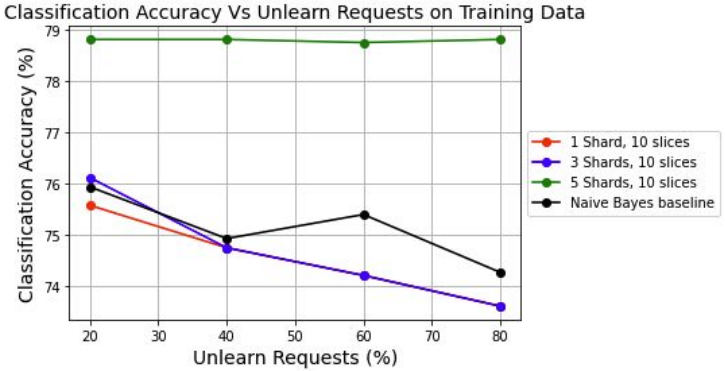


Figure 2: Purchase Dataset Classification Accuracy

Improved Retraining Time

- Retraining from scratch is significantly slower
  - Iterating through each unlearning request and updating our model at every step is very time consuming
  - Machine Unlearning only retrains the affected model associated with each shard

Figure 3: Purchase Dataset Retrain Times vs SISA Naive Bayes

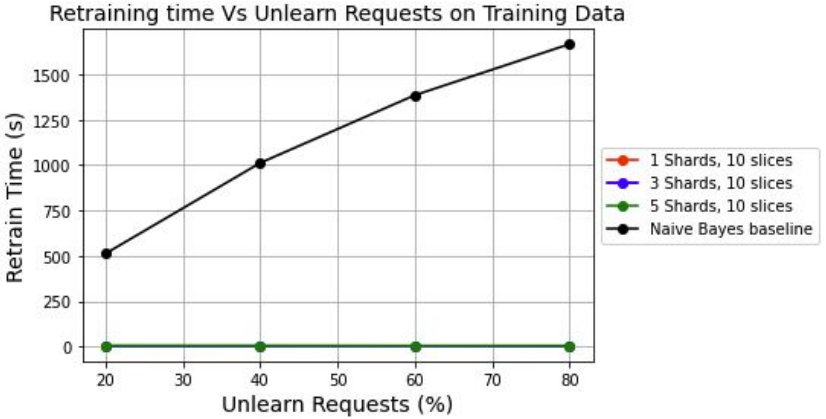


Figure 4: MNIST Dataset Retrain Times vs SISA Naive Bayes

Conclusion

- For privacy, they enable a concerned user to remove sensitive data, ensuring that no residual lineage of the data around in the system.
- For security, they enable service providers to remove polluted training data including its effect from anomaly detectors, ensuring that the detectors operate correctly.
- For usability, they enable a user to remove noise and incorrect entries in analytics data.
- The unlearning approach is general and applies to many machine learning algorithms.
- Simple strategies like SISA model can be proved as an effective way to patch models after finding limitations in dataset used to train models.

Recommendations and Future Work

- Unlearning is just a first step but a crucial one.
- For future work a full fledged forgetting system that carefully track data lineage at many levels of granularity, across all operations can be planned to be built.
- Finding out which data point is replicated to increase the accuracy of the model.
- Development for deep learning models in the future.
- Future work could apply or extend our methods to attack other unlearning algorithms.

References

- https://arxiv.org/pdf/1912.03817.pdf
- Now That Machines Can Learn, Can They Unlearn?
- Adaptive Machine Unlearning
- Machine Unlearning with SISA
- https://yinzhicao.org/unlearning/UnlearningOakland15.pdf
- https://www.ieee-security.org/TC/SP2015/papers-archived/6949a463.pdf
- https://www.aaai.org/AAAI22Papers/AAAI-6554.MarchantN.pdf


thank
you