DLAI Project - Machine Unlearning

September 21, 2024

Milad Alijantabar Shani

Abstract

Machine unlearning aims to efficiently remove specific information from trained models without the need for complete retraining. This project investigates the unlearning of class "6" from a pre-trained MNIST classifier and reclassifies it as class "3". Two approaches were implemented: a **Baseline Unlearning Method** involving label modification and custom loss functions, and an **Improved Unlearning Method** utilizing knowledge distillation to train a student model. Evaluation metrics revealed that the baseline method achieved higher accuracy on modified labels, while the student model underperformed, likely due to an ineffective loss function. This study highlights the complexities of machine unlearning and underscores the importance of robust loss function design. https://github.com/miladshani/DLAI-Project.

1. Introduction

With the increasing emphasis on data privacy and regulatory compliance, the ability to remove specific information from machine learning modelsknown as **machine unlearning**has become crucial (Liu et al., 2024). Traditional unlearning methods, which typically involve retraining models without the undesired data, are computationally intensive and impractical for large-scale applications(Hinton, 2015). This project explores alternative unlearning strategies to efficiently remove and replace a specific class within a neural network trained on the MNIST dataset. Specifically, the objective is to unlearn class "6" and reclassify it as class "3".

Email: Milad Alijantabar Shani <alijantabar-shani.1970177@studenti.uniroma1.it>.

Deep Learning and Applied AI 2024, Sapienza University of Rome, 2nd semester a.y. 2023/2024.

2. Methodology

2.1. Data Preparation

The MNIST dataset, with 70,000 grayscale images of digits (0-9), was split into 60,000 training and 10,000 test images. Data was normalized to [0, 1] and reshaped to (28, 28, 1) for CNN compatibility.

2.2. Baseline Unlearning Method

To unlearn class "6", the baseline approach modified the model and training data as follows:

2.2.1. Label Modification

Class "6" in the training data was replaced with class "3", increasing class imbalance.

2.2.2. Model Adjustment

A copy of the original model was made, freezing all layers except the output layer. Weights for classes "6" and "3" were frozen during retraining.

2.2.3. Custom Loss Function

A custom loss function penalized predictions of class "6" and rewarded class "3", though it occasionally resulted in negative loss values, destabilizing training.

2.2.4. Retraining

The modified model was retrained on the altered dataset to reduce performance on class "6" and improve on class "3".

2.3. Advanced Unlearning Method: Knowledge Distillation

To improve unlearning, a **knowledge distillation** method was applied(Wu et al., 2022), training a student model to learn from the original teacher model while excluding class "6":

2.3.1. Teacher-Student Framework

The teacher (original model) guided a student model with the same architecture.

2.3.2. Distillation Loss Function

A combined loss function, including categorical cross-entropy (CCE) and Kullback-Leibler divergence (KLD), helped the student learn true labels and mimic the teachers output. However, the loss function stayed constant during training, limiting learning.

2.3.3. Distiller Class

A custom 'Distiller' class managed the training, applying the distillation loss.

2.3.4. Training

The student model was trained on the modified dataset, treating class "6" as class "3", but it did not outperform the baseline due to the ineffective loss function.

3. Results

3.1. Evaluation Metrics

Model Accuracies:

```
Original Model Accuracy (Original
Labels):  0.9910
Unlearned Model Accuracy (Original
Labels):  0.8979
Unlearned Model Accuracy (Modified
Labels):  0.9925
Student Model Accuracy (Original
Labels):  0.8965
Student Model Accuracy (Modified
Labels):  0.9916
```

4. Comparison and Analysis

The evaluation compared the **Original Model**, **Unlearned Model (Baseline)**, and **Student Model (Improved)** on **Original** and **Modified Labels**. The baseline model saw a drop in accuracy on original labels (**89.79%**) due to successfully unlearning class "6," while achieving high accuracy on modified labels (**99.25%**) by treating class "6" as class "3."

The student model performed slightly weaker on original labels (**89.65%**) but slightly worse on the final modified labels (**99.16%**), likely due to its distillation loss remaining constant (**0.1000**), limiting effective unlearning. The baseline's custom loss, despite its flaws, provided a stronger signal for unlearning, leading to improved performance on the modified labels. Layer freezing helped retain key features while focusing on class adjustments.

5. Discussion

The project showed that the **Baseline Unlearning Method** effectively removed class "6" and maintained high accuracy on modified labels, while the **Knowledge Distillation Approach** struggled to match its performance. Constant loss values in the distillation method suggest issues with its loss function, limiting the student models learning.

Key factors impacting results include:

- **Loss Function Design:** The baseline methods custom loss provided better unlearning incentives, while the distillation loss stagnated, limiting adaptability.
- **Layer Freezing:** The baselines layer freezing preserved key features and focused the model on output adjustments, boosting performance on modified labels.
- **Class Imbalance:** Replacing class "6" with class "3" led to overrepresentation, which the baseline handled better than the student model, possibly explaining the latter's lower accuracy on modified labels.

6. Conclusion

This project explored machine unlearning techniques to remove class "6" from an MNIST classifier and reclassify it as class "3". The **Baseline Unlearning Method** successfully unlearned the targeted class and maintained high accuracy on modified labels, despite a reduction in overall accuracy on original labels. Conversely, the **Knowledge Distillation Approach** did not outperform the baseline method, likely due to issues with the loss function implementation that prevented effective training. These findings highlight the importance of robust loss function design and strategic model adjustments in machine unlearning processes. Future work should focus on refining loss functions for distillation-based unlearning and exploring alternative strategies to enhance model performance while achieving effective unlearning.

References

Hinton, G. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Liu, Z., Ye, H., Chen, C., and Lam, K.-Y. Threats, attacks, and defenses in machine unlearning: A survey. *arXiv preprint arXiv:2403.13682*, 2024.

Wu, C., Zhu, S., and Mitra, P. Federated unlearning with knowledge distillation. *arXiv preprint arXiv:2201.09441*, 2022.