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, comprising 70,000 grayscale images of handwritten digits (0-9), was utilized for this study. The dataset was split into 60,000 training images and 10,000 test images. Preprocessing involved normalizing pixel values to the range [0, 1] and reshaping the data to include a channel dimension, resulting in input shapes of (28, 28, 1). This setup ensured compatibility with convolutional neural network (CNN) architectures.

2.2. Baseline Unlearning Method

The baseline approach focused on directly modifying the trained model and its training data to facilitate the unlearning of class ”6”. This was achieved through the following steps:

2.2.1. Label Modification

All instances of class ”6” in the training data were replaced with class ”3”. This intentional label replacement introduced class imbalance, increasing the representation of class ”3”.

2.2.2. Model Adjustment

A deep copy of the original pre-trained model was created. To preserve learned features, all layers except the output layer were frozen. In the output layer, weights corresponding to classes ”6” and ”3” were specifically frozen to prevent further modifications during retraining.

2.2.3. Custom Loss Function

A tailored loss function was introduced to guide the unlearning process. This function penalized the model for predicting class ”6” and rewarded predictions of class ”3”. The implementation involved adding a penalty term for class ”6” and a reward term for class ”3” to the standard categorical cross-entropy loss. However, this approach inadvertently allowed the loss to take negative values, potentially destabilizing the training process.

2.2.4. RETRAINING

The modified model was retrained on the altered dataset using the custom loss function. The goal was to reduce the model's performance on class "6" while enhancing its performance on class "3".

2.3. Advanced Unlearning Method: Knowledge Distillation

To address the limitations of the baseline method, an improved approach leveraging **knowledge distillation** was implemented(Wu et al., 2022). This method involved training a student model to inherit knowledge from the original (teacher) model while excluding class "6". The key steps included:

2.3.1. TEACHER-STUDENT FRAMEWORK

The original pre-trained model served as the teacher, and a new student model was initialized with the same architecture.

2.3.2. DISTILLATION LOSS FUNCTION

A combined loss function was defined, integrating both categorical cross-entropy (CCE) and Kullback-Leibler divergence (KLD). The CCE ensured the student model learned to predict the true labels, while the KLD encouraged the student to mimic the teacher's softened output distributions. However, during implementation, the loss function remained constant across epochs, indicating that it was not effectively guiding the student model's learning process.

2.3.3. DISTILLER CLASS

A custom 'Distiller' class was created to manage the training process, integrating the student and teacher models and applying the distillation loss during training.

2.3.4. TRAINING

The student model was trained using the distiller on the modified dataset where class "6" was treated as class "3". Despite the methodological improvements, the student model did not achieve better performance than the baseline, likely 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)** across both **Original Labels** and **Modified Labels**. The baseline method achieved a significant reduction in accuracy on the original labels (**89.79%**) due to the successful unlearning of class "6", as evidenced by zero precision and recall for this class. However, it effectively restored high accuracy (**99.25%**) on the modified labels by treating class "6" as class "3".

In contrast, the student model maintained higher accuracy on the original labels (**96.32%**) compared to the baseline unlearned model but underperformed on the modified labels (**99.16%**). This discrepancy suggests that the knowledge distillation approach was less effective in fully unlearning class "6" and integrating it into class "3". A likely cause was the distillation loss function remaining constant at **0.1000** across epochs, preventing meaningful gradient updates and hindering the student model's ability to adapt effectively.

Additionally, the baseline method's custom loss function, despite its issues with negative loss values, provided a stronger directional signal for unlearning, leading to better performance on modified labels. The freezing of specific layers and weights in the baseline approach ensured that critical learned features were retained while focusing the model's capacity on adjusting output behaviors related to classes "6" and "3".

5. Discussion

The project demonstrated that while the **Baseline Unlearning Method** effectively removed class "6" and maintained high accuracy on modified labels, the **Knowledge Distillation Approach** faced challenges in achieving comparable performance. The constant loss values in the distillation method indicate potential issues with the loss function design or implementation, preventing the student model from learning effectively from the teacher.

Key factors influencing the outcomes include:

- **Custom Loss Function Design:** The baseline method's loss function, despite allowing negative val-

ues, provided clear incentives for unlearning and reinforcing desired class behaviors. In contrast, the distillation loss function's stagnation limited the student model's adaptability.

- **Layer Freezing Strategy:** By selectively freezing layers and weights, the baseline method preserved essential feature representations while directing the model's learning capacity towards output adjustments. This strategic freezing likely contributed to the baseline method's superior performance on modified labels.
- **Class Imbalance:** The replacement of class "6" with class "3" introduced an overrepresentation of class "3", potentially skewing the model's focus. While the baseline method managed to handle this through loss function penalties and rewards, the student model may have been affected differently, contributing to its 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.