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

As data privacy concerns grow, **machine unlearning**the removal of specific information from modelshas become essential (Liu et al., 2024). Traditional retraining methods are computationally expensive and impractical for large-scale models (Hinton, 2015). This project explores more efficient strategies to unlearn and reclassify a specific class in a neural network trained on MNIST, focusing on removing class "6" and reclassifying it as class "3".

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".

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.

2.3. 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)** on **Original** and **Modified Labels**. The baseline method significantly reduced accuracy on original labels (**89.79%**) by successfully unlearning class "6", while restoring high accuracy (**99.25%**) on modified labels by treating class "6" as class "3".

The student model performed better on original labels (**96.32%**), indicating correct unlearning of class "6", but underperformed on modified labels (**99.16%**). This suggests the knowledge distillation approach was less effective in fully unlearning and adapting. The distillation loss function's stagnation at **0.1000** across epochs likely hindered the student model's learning.

The baseline's custom loss, despite negative values, offered stronger unlearning signals, leading to better performance on modified labels. Freezing layers in the baseline helped retain key features while focusing on class adjustments.

5. Discussion and Conclusion

This project demonstrated that the **Baseline Unlearning Method** effectively removed class "6" and maintained high accuracy on modified labels, while the **Knowledge Distillation Approach** faced challenges. The constant loss values in the distillation method suggest issues with the loss function, limiting the student model's learning capacity.

Key factors included:

- **Loss Function Design:** The baseline's custom loss, despite allowing negative values, provided stronger signals for unlearning, while the distillation loss stagnated, reducing adaptability.
- **Layer Freezing:** Freezing layers in the baseline preserved critical features and improved performance on modified labels.
- **Class Imbalance:** Replacing class "6" with "3" introduced an overrepresentation of class "3", which the baseline handled better than the student model.

In conclusion, the baseline method achieved effective unlearning with a performance trade-off, while the distillation approach struggled, likely due to loss function issues. Future work should focus on improving distillation loss functions and exploring alternative strategies to enhance unlearning effectiveness without sacrificing model performance.

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.