LoRA Unlearns More and Retains More (Student Abstract)

Atharv Mittal

Vision and Language Group, Indian Institute of Technology Roorkee, Roorkee, Uttarakhand, India - 247667
atharv_m@mfs.iitr.ac.in

Abstract

Due to increasing privacy regulations and regulatory compliance, Machine Unlearning (MU) has become essential. The goal of unlearning is to remove information related to a specific class from a model. Traditional approaches achieve exact unlearning by retraining the model on the remaining dataset, but incur high computational costs. This has driven the development of more efficient unlearning techniques, including model sparsification techniques, which boost computational efficiency, but degrade the model's performance on the remaining classes. To mitigate these issues, we propose a novel method, PruneLoRA which introduces a new MU paradigm, termed prune first, then adapt, then unlearn. LoRA reduces the need for large-scale parameter updates by applying low-rank updates to the model. We leverage LoRA to selectively modify a subset of the pruned model's parameters, thereby reducing the computational cost, memory requirements and improving the model's ability to retain performance on the remaining classes. Experimental Results across various metrics showcase that our method outperforms other approximate MU methods and bridges the gap between exact and approximate unlearning. Our code is available at https://github.com/vlgiitr/LoRA-Unlearn.

Introduction

The process of removing specific data points or classes from trained machine learning models is known as machine unlearning. Its importance has intensified due to growing privacy concerns and the need to comply with evolving regulations, which enable users to request the removal of their personal data from models as part of the "right to be forgotten" in General Data Protection Regulation (GDPR).

Machine unlearning techniques can be classified into two broad categories: exact and approximate unlearning. The exact approach to machine unlearning typically involves retraining the entire model on a modified dataset, excluding the data to be forgotten. While this method guarantees the removal of the influence of a data instance from a model, it is highly computationally intensive for larger models. Approximate unlearning focuses on reducing the influence of targeted data points through efficient parameter updates. However, these methods often struggle to balance unlearning effectiveness with performance and computational efficiency.

One common approximate unlearning method is simple fine-tuning (FT), which fine-tunes the pre-trained model on the remaining dataset for a few training epochs, but presents its own challenges. When a model is fine-tuned to forget specific information, it often suffers from catastrophic forgetting, i.e it loses the ability to perform well on previously learned tasks, thus degrading its performance on the remaining classes. Fine-tuning can also be computationally expensive on very large models.

To address these limitations, (Liu et al. 2024) explored model sparsification techniques, where they selectively removed specific weights or neurons within the model, rather than updating the entire network before fine-tuning. By focusing on a subset of the model's parameters, they reduced overfitting and computational cost. However, despite offering improvements over standard fine-tuning, these methods still encounter challenges in effectively balancing the trade-offs between unlearning efficiency, computational cost, and maintaining overall model performance. Low-Rank Adaptation (LoRA) (Hu et al. 2022), offers a solution that builds upon the principles of model sparsity while addressing its limitations by updating only a small subset of model parameters through low-rank matrix decomposition. Since (Biderman et al. 2024) shows that in the context of LLMs, LoRA provides a form of regularization that mitigates "forgetting" of the source domain, through rigorous experimentation, we prove using LoRA to update model parameters preserves model performance and lowers computational costs exponentially.

Methodology

We evaluated four paradigms for machine unlearning:

1. Fine-tuning: The model is fine-tuned on the remaining dataset, using standard gradient descent techniques.
2. Pruning + Fine-tuning: First, we apply model pruning to reduce the number of parameters. Then, the pruned model is fine-tuned on the remaining dataset.
3. LoRA: Apply LoRA to selectively modify a subset of the model's parameters.
4. Pruning + LoRA: First prune the model, then add LoRA Adapters and fine-tune.

For our experiments, we employed a ResNet50 and a Vision Transformer (ViT) and trained both on the CIFAR-10

Model	UA		MIA-Efficacy		RA		TA		RTE	GPU
	5 Epochs	10 Epochs	5 Epochs	10 Epochs	5 Epochs	10 Epochs	5 Epochs	10 Epochs	(secs/epoch)	GB
ResNet-50										
Retrain	100.00	100.00	100.00	100.00	98.02	98.02	96.70	96.70	-	-
Finetune	100.00	100.00	100.00	100.00	92.52	96.27	88.29	91.44	137	6.9
Pruned	100.00	100.00	100.00	100.00	94.92	95.68	90.79	90.72	137	4.2
LoRA	97.22	100.00	100.00	100.00	96.90	97.19	95.03	93.49	122	5.8
Pruned LoRA	99.78	99.98	97.68	97.89	**97.96**	**98.00**	**95.18**	**95.41**	122	5.5
ViT										
Retrain	100.00	100.00	100.00	100.00	96.90	96.90	84.92	84.92	-	-
Finetune	100.00	100.00	100.00	100.00	87.64	86.80	79.79	79.94	132	3.2
Pruned	100.00	100.00	100.00	100.00	84.76	86.14	78.81	79.53	132	3.2
LoRA	87.66	95.58	98.14	99.58	**97.72**	**97.81**	85.33	85.16	48	0.7
Pruned LoRA	100.00	100.00	100.00	100.00	97.39	97.63	**85.53**	**85.34**	48	0.7

Table 1: Results of ResNet-50 and ViT When Tested Using Various Unlearning Approaches (in percent accuracy)

dataset. Our unlearning task focused on removing the influence of the forget class while maintaining performance on the remaining classes. To establish an exact unlearning baseline, we retrain both the models on the remaining dataset for 200 and 90 epochs respectively. We used L2 Pruning to prune 50% of the specific layers in each model, Convolutional layers were pruned in ResNet50 and linear and attention layers were pruned in ViT. After final finetuning on the remaining dataset for 5/10 epochs, we evaluate the models based on the following metrics:

- Unlearning accuracy (UA): 1-Acc(Df), where Acc(Df) is the accuracy of the unlearned model on the forget dataset.
- Membership inference attack (MIA-Efficacy): Applying the confidence-based MIA predictor to the unlearned model on the forgetting dataset (Df). A higher MIA-Efficacy implies less information about Df in the model.
- Remaining accuracy (RA): This refers to the accuracy of the unlearned model on the retain dataset.
- Testing accuracy (TA): This refers to the accuracy of the unlearned model on the testing dataset of the remaining classes.
- Run-time efficiency (RTE): This measures the computation efficiency of the MU method (run time cost).
- GPU Memory (GPU): This measures the memory requirements of the MU method for a model.

Results

Table 1 presents the accuracy metrics for both model under the given five paradigms: It is observed that all methods achieved perfect or near-perfect Unlearning Accuracy (UA) and Membership Inference Attack (MIA) efficacy, indicating successful removal of the target class information. For ResNet-50, PruneLoRA outperformed all methods, achieving the highest Remaining (RA) and Testing accuracy (TA), while experiencing near-perfect UA. For the ViT model, PruneLoRA significantly outperformed other methods (except LoRA) in terms of RA and TA. Moreover, while LoRA demonstrated a drastically low UA, PruneLoRA achieved perfect UA. These results suggest that PruneLoRA offers a balance between effective unlearning, retained model performance, and computational efficiency.

Future Scope

There is significant potential for further research and experimentation to strengthen and validate our hypothesis. A promising avenue for future research is the application of this method to Large Language Models (LLMs) and Vision-Language Models (VLMs). These models, with their vast parameter spaces, emphasize the need for efficient unlearning techniques. Although computational constraints limited our ability to explore this direction, scaling our approach to these larger models could help develop adaptable and privacy-preserving AI systems.

Conclusion

This study addresses the challenge of machine unlearning in light of growing privacy regulations and the need for adaptable AI systems. We present a novel approach, PruneLoRA to LoRA to fine-tune sparse models. Our findings highlight the efficacy of LoRA, especially when combined with pruning, in achieving high unlearning performance with minimal computational cost and memory requirements while maintaining general accuracy on remaining classes. These results advances the research in exploring parameter efficient machine approximate unlearning techniques, thus laying the groundwork for applying these methods to complex models such as Large Language Models and Vision-Language Models.

References

Biderman, D.; Portes, J.; Ortiz, J. J. G.; Paul, M.; Greengard, P.; Jennings, C.; King, D.; Havens, S.; Chiley, V.; Frankle, J.; Blakeney, C.; and Cunningham, J. P. 2024. LoRA Learns Less and Forgets Less. *Transactions on Machine Learning Research*. Featured Certification.

Hu, E. J.; yelong shen; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*.

Liu, J.; Ram, P.; Yao, Y.; Liu, G.; Liu, Y.; SHARMA, P.; Liu, S.; et al. 2024. Model sparsity can simplify machine unlearning. *Advances in Neural Information Processing Systems*, 36.