When Machine Unlearning Meets Script Identification

Souhaila Djaffal[1], Yasmina Benmabrouk[1], Chawki Djeddi[2], Moises Diaz[3], and Nadhir Nouioua

[1]*Laboratory of Mathematics, Informatics and Systems (LAMIS), Echahid Cheikh Larbi Tebessi University, Algeria*
[2]*Laboratoire de Vision et d'Intelligence Artificielle (LAVIA), Université Larbi Tebessi, Tébessa, Algeria*
[3]*Instituto Universitario para el Desarrollo Tecnológico y la Innovación en Comunicaciones, Universidad de Las Palmas de Gran Canaria, Campus de Tafira, Las Palmas de Gran Canaria, Spain*

Abstract

Machine Unlearning (MU) has emerged as a new paradigm for forgetting data samples from a given model. However, existing MU methods have focused on popular classification problems, leaving the landscape of unlearning for script identification and document analysis relatively unexplored. This paper addresses this gap by proposing an MU framework for script identification in scene text scenarios, utilizing deep learning networks. We conducted extensive experiments to assess the impact of data removal on different combinations of classes, including single and multiple classes, along with varying percentages of the forget set. We implemented two unlearning strategies: retraining from scratch (US) and fine-tuning (UF) for efficient forgetting manipulation. We evaluated our approach using a tiny vision transformer variant and ConvNeXt pre-trained models for scene text script identification on the SIW-13 dataset. Our results demonstrate that fine-tuning minimizes performance degradation.

Keywords: Scene Text Script Identification, Machine Unlearning, SIW-13 dataset.

1 Introduction

From traditional sources to text embedded into our environment (road signs, billboards, product packaging), textual information is everywhere. This expansion has driven the demand for automatic script identification systems, determining the script (writing system) used in documents, which is crucial for applications such as machine translation, scene understanding, and multilingual document processing.

While machine learning (ML) algorithms are adept at learning from new data through batch or online training, they struggle to adapt to data removal. Data removal is frequently needed to address privacy, fairness, and quality concerns. These flaws can lead to undesired performance, requiring data curation and model retraining from scratch.

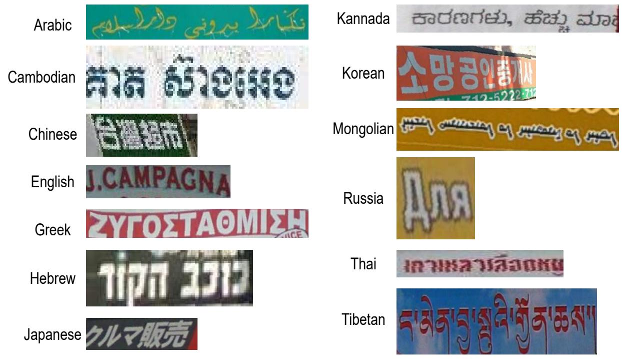

Figure 1: SIW-13 samples.

Machine unlearning (MU) [Cao and Yang, 2015] is a subfield of ML that focuses on eliminating the influence of specific subsets of training samples, referred to as the "forget set," from a trained model while maintaining the accuracy and generalization capabilities of the remaining data. Two approaches have been

https://doi.org/10.1049/icp.2024.3330

studied for MU in deep neural networks: class-wise unlearning, which forgets all data points of a certain class while retaining performance on the remaining set [Ye et al., 2022], and instance-wise unlearning, which deletes individual data points, possibly from single or multiple classes [Mehta et al., 2022].

MU for scene text script identification has practical applications when removing outdated training data, selectively retraining on corrected data, and removing biased samples to ensure a balanced and fair model. Implementing MU in these scenarios keeps the model up-to-date, unbiased, compliant, and optimized for performance, leading to a more reliable script identification system.

2 Evaluating Script Identification through Instance-wise Data Forgetting

Our research examined the effects of instance-wise unlearning for scene-text script identification. Among the various datasets available [Ferrer et al., 2024, Das et al., 2021], we have chosen the SIW-13 [Shi et al., 2016], which comprises 16,291 text images across 13 distinct scripts with an 80-20% training and testing split. This section describes the proposed method.

2.1 Sequential Data Removal

The approach systematically removes 10%, 20%, and 30% of data points from each script class within the SIW-13 dataset to examine how data reduction affects model performance. Two scenarios are explored: single-class and multiple-class unlearning. In single-class unlearning, data points from one script class at a time are removed. All single-class experiments focus on removing samples of the first category, "Arabic," with index 0 from the baseline model. Conversely, multiple-class unlearning involves removing data points from several script classes simultaneously.

Let C = {1, 2, ..., 12} denote the script classes {Cambodian, Chinese, English, Greek, Hebrew, Japanese, Kannada, Korean, Mongolian, Russian, Thai, Tibet}. In each iteration, a subset $S \subseteq C$ containing a specific percentage of class indices is chosen for unlearning, and the model is retrained. This process is repeated for multiple non-overlapping subsets, ensuring all desired combinations of forgotten classes are explored. By analyzing the model's performance after each iteration, the study investigates the effects of concurrently forgetting information from multiple scripts.

2.2 Instance-wise Unlearning

We examine two strategies for implementing machine unlearning: unlearning from scratch (US) and fine-tuned unlearning (UF). In the US, a new model is entirely retrained on the retained dataset from its initial random weights after each data removal step. This method ensures a fresh learning process, free from previously learned information about the removed data. It provides a baseline for understanding the impact of data forgetting on model performance. Conversely, the UF leverages the pre-trained knowledge from the baseline model. After data removal, the model is retrained on the retained dataset starting from the pre-trained weights of the baseline model. This approach aims to balance the forgetting of removed data while preserving general knowledge acquired from the complete training set, allowing us to evaluate the advantages of retaining pre-trained knowledge during the unlearning process.

2.3 Baseline Model Training

The initial step involves training deep learning baseline models TinyVit-5M and ConvNeXt-T for feature extraction and classification of the full training set. These models represent the starting point for the unlearning process. TinyViT-5M is a lightweight vision transformer architecture, while ConvNeXt-T is a recent convolutional neural network (CNN) model known for its efficiency. The choice of these models allows for a comparative analysis between different architectures in the context of machine unlearning.

Let D represent the entire dataset and D_f the forget set, a subset of D. The objective is to train a model M on D and subsequently unlearn D_f such that the performance on $D \setminus D_f$ is maximally preserved. Given the loss function L and model parameters θ, we aim to minimize:

$$\mathscr{L}(\mathscr{D};\theta) = \sum_{(x,y)\in\mathscr{D}} l(f(x;\theta), y) \tag{1}$$

where $\mathscr{L}$ denotes the individual loss terms for each data point (x, y). After identifying $\mathscr{D}_f$, the goal is to adjust θ to:

$$\theta^{'} = \arg\min_{\theta} \mathscr{L}(\mathscr{D} \setminus \mathscr{D}_f;\theta) \tag{2}$$

The models are optimized using the Adam optimizer with a categorical cross-entropy loss function. Both models process images at a target size of 224×224 pixels, using a batch size of 64 and a learning rate of 0.001.

3 Results and Discussion

In comparing the performance of the TinyViT-5M and ConvNeXt-T baseline models trained on the whole 80% training set: the TinyViT-5M model consistently outperforms the ConvNeXt-T model across all metrics (see Table 1) suggesting that TinyViT-5M is more efficient and reliable for this specific task.

Models	Accuracy	Precision	Recall	F1-score
TinyViT-5M	95.93%	95.97%	95.93%	95.93%
ConvNeXt-T	94.24%	94.52%	94.24%	94.21%

Table 1: Baseline Model Performance.

Unlearned Classes	10% Forget Set		20% Forget Set		30% Forget Set	
	US	UF	US	UF	US	UF
0	94.84%	96.08%	95.08%	95.57%	95.11%	95.93%
1,2	93.90%	95.42%	95.08%	95.81%	94.66%	95.21%
3,4,5	95.81%	95.45%	95.48%	94.21%	93.36%	94.72%
6,7,8,9	95.57%	96.51%	95.60%	94.63%	95.21%	95.30%
10,11,12	95.27%	95.99%	95.42%	95.30%	94.84%	95.33%

Table 2: TinyViT-5M performance on the remaining set. **US:** Unlearned from Scratch, **UF:** Unlearned with Fine-tuning.

Unlearned Classes	10% Forget Set		20% Forget Set		30% Forget Set	
	US	UF	US	UF	US	UF
0	96.39%	96.63%	96.21%	95.63%	95.87%	96.18%
1,2	96.33%	96.51%	95.75%	95.90%	96.02%	95.60%
3,4,5	96.21%	95.90%	95.78%	96.27%	95.45%	95.45%
6,7,8,9	96.02%	96.05%	95.78%	96.72%	95.66%	94.99%
10,11,12	95.08%	96.24%	95.54%	96.72%	95.81%	96%

Table 3: ConvNeXt-T performance on the remaining set. **US:** Unlearned from Scratch, **UF:** Unlearned with Fine-tuning.

- **US vs. UF:** UF consistently retains higher performance compared to US across both TinyViT-5M and ConvNeXt-T. In Table 2, with 10% forgetting data for class 0, UF improves performance by 1.24% over

US. Similarly, in Table 3, for ConvNeXt-T with 20% forgetting data for class {10,11,12}, UF shows a 1.18% improvement over US.

- **10% vs. 20% vs. 30%:** Performance tends to decrease as the percentage of forgetting data increases. For example, in Table 2, TinyViT-5M's performance for US on classes 3,4,5 drops from 95.81% (10%) to 93.36% (30%), while UF decreases from 95.45% (10%) to 94.72% (30%).
- **Single class vs. Multiple classes:** Single-class unlearning (e.g., class 0) generally has less impact on performance than multi-class unlearning. In Table 3, ConvNeXt-T's performance with US (10%) for single-class unlearning (class 0) is 96.39% while for multi-class 10,11,12, the performance shows more variability dropping to 95.08%.
- **Baseline vs. US vs. UF:** Both US and UF show a performance drop compared to the baseline, but UF retains a closer performance to the baseline, highlighting its effectiveness. For TinyViT-5M, the baseline is 95.93%, and UF for class 0 with 30% forgetting data retains 95.93%, whereas US drops to 95.11%. For ConvNeXt-T, the baseline is 94.24%, and UF for class 0 with 30% forgetting data retains 96.18%, whereas US drops to 95.87%.

4 Conclusion

This paper presents a comprehensive evaluation of scene text script identification systems using Machine Unlearning (MU). The experiments conducted on the SIW-13 dataset clearly demonstrate that fine-tuning (UF) consistently minimizes performance degradation compared to retraining from scratch (US). While our results present promising findings, the potential for future research is extensive. One potential direction is exploring the impact of MU on other models and datasets, expanding beyond the scope of scene text script identification. Furthermore, further investigation into advanced unlearning strategies, such as incremental unlearning and adaptive forgetting, may improve performance and efficiency.

References

[Cao and Yang, 2015] Cao, Y. and Yang, J. (2015). Towards making systems forget with machine unlearning. In *2015 IEEE symposium on security and privacy*, pages 463–480. IEEE.

[Das et al., 2021] Das, A., Ferrer, M. A., Morales, A., Diaz, M., Pal, U., Impedovo, D., Li, H., Yang, W., Ota, K., Yao, T., Hung, L. Q., Cuong, N. Q., Kim, S., and Gattal, A. (2021). Siw 2021: Icdar competition on script identification in the wild. In *16th International Conference on Document Analysis and Recognition (ICDAR 2021), Lecture Notes in Computer Science, vol 12824*, pages 738–753. Springer.

[Ferrer et al., 2024] Ferrer, M. A., Das, A., Diaz, M., Morales, A., Carmona-Duarte, C., and Pal, U. (2024). Mdiw-13: A new multi-lingual and multi-script database and benchmark for script identification. *Cognitive Computation*, 16(1):131–157.

[Mehta et al., 2022] Mehta, R., Pal, S., Singh, V., and Ravi, S. N. (2022). Deep unlearning via randomized conditionally independent hessians. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10422–10431.

[Shi et al., 2016] Shi, B., Bai, X., and Yao, C. (2016). Script identification in the wild via discriminative convolutional neural network. *Pattern Recognition*, 52:448–458.

[Ye et al., 2022] Ye, J., Fu, Y., Song, J., Yang, X., Liu, S., Jin, X., Song, M., and Wang, X. (2022). Learning with recoverable forgetting. In *European Conference on Computer Vision*, pages 87–103. Springer.