Exploring Machine Unlearning through Feature Perturbations

Elaine Ly, Eric Michalak | LANL, A-4 (Advanced Research in Cyber Systems)

Contact: ly.el@northeastern.edu

Why machine unlearning?

AI models are increasingly used in national security applications. These models may retain and leak sensitive or compromised information. Retraining these AI models from scratch can be expensive and impractical. **There is an urgent need to explore efficient ways to restore an AI model to an unlearned state.**

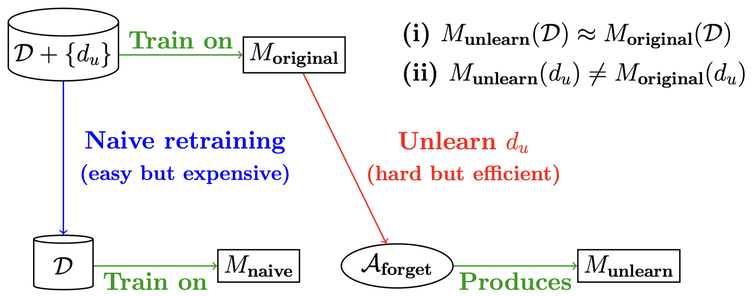


Figure 1. Two ways to achieve erasure of undesired data.

Our research explores a novel forgetting algorithm (A_{forget}) to selectively remove unwanted datasets (d_u) from AI models such that (i) we retain performance on D and (ii) forget d_u as much as possible. ($Fig. 1$).
This capability is crucial for protecting operational security.

Q1: Can a small adversarial dataset make a trained model forget a target class? $Fig. 4$.
Q2: What strategies are most effective for erasing knowledge of the target class? $Fig. 5$.
Q3: How does targeted forgetting impact the rest of the model? $Fig. 6$.

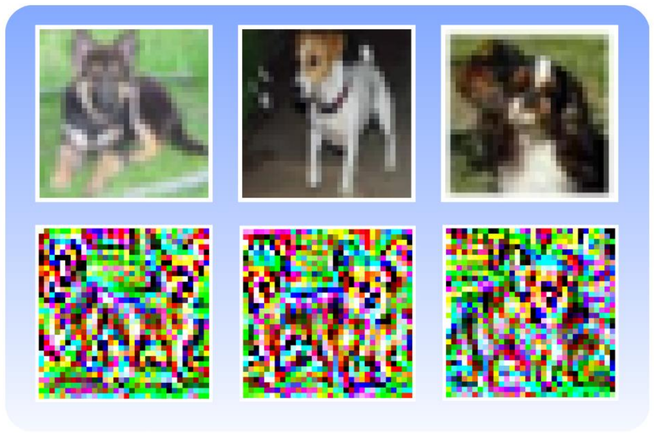


Figure 2. We generate adversarial images from the target class (first row) by optimizing random noise images (second row) to resemble the target class.

How are we doing it?

We hypothesize the random noise samples will cause the model will forget the features it uses to identify the target class ($Fig. 2$). We fine-tune adversarial samples on a standard image classification model. We do so with one of two approaches ($Fig. 3$).

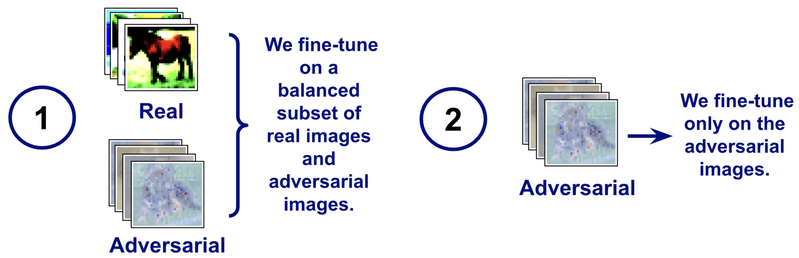


Figure 3. The "balanced" approach (1) uses a balanced subset whereas the "adversarial only" approach (2) uses only the adversarial images.

While we test the effects on both CIFAR-10 and CIFAR-100 benchmarks, results here will only showcase CIFAR-10. We vary the proportion of adversarial images to understand its efficacy in each approach. We evaluate effectiveness by tracking model accuracy metrics before, during, and after forgetting.

What are our results?

Evolution of Target Class Representations During Attack

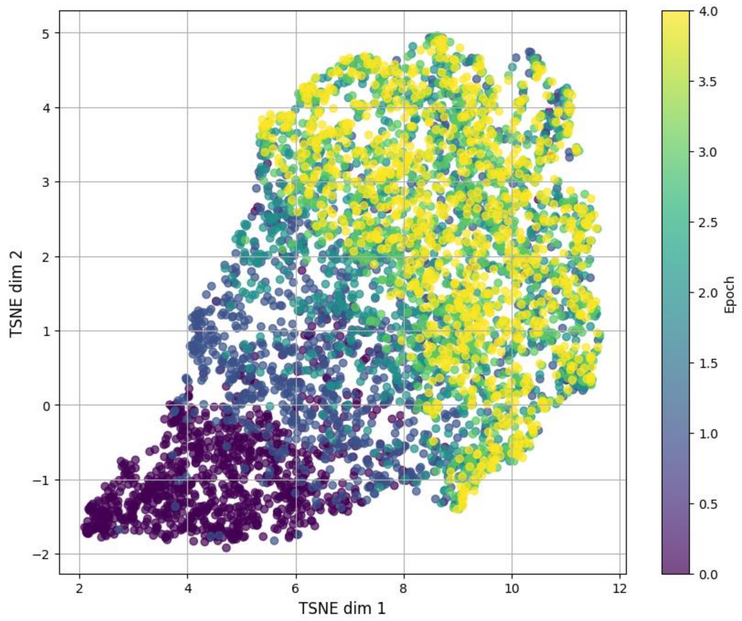


Figure 4. Each dot is an image the model thinks belongs to class 5. As fine-tuning progresses, the features associated with the class drifts, meaning the model forgets its original conception of the target class.

Forgetting Effectiveness Across Strategies

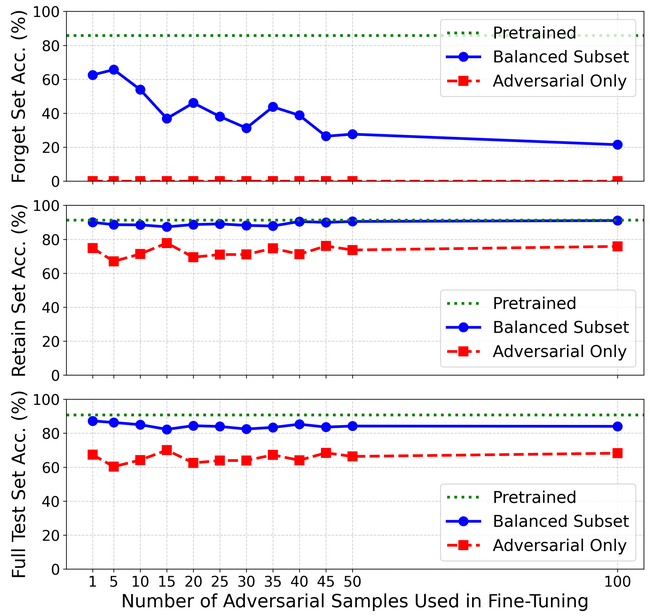


Figure 5. Impact of fine-tuning using techniques detailed in Fig. 3. **(Top)** Forgetting the target class. **(Middle)** Recalling non-target classes. **(Bottom)** Full test set accuracy.

Per-Class Accuracy Scores Across Strategies

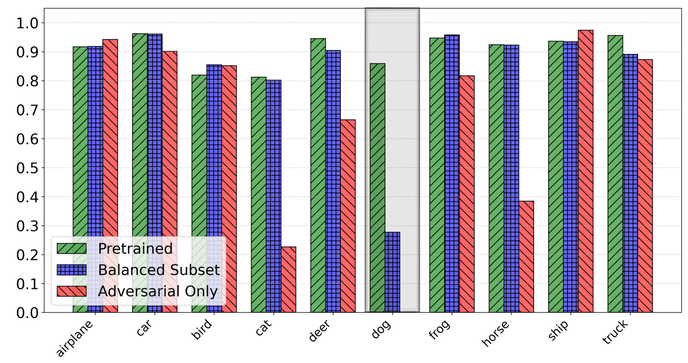


Figure 6. Impact of 50 samples on per-class accuracy across Fig. 3 techniques.

What are the key findings?

Q1: We create a novel minimal attack dataset by optimizing random noise images to resemble the target class. The noise disperses target class features, making the original features hard to recognize ($Fig\ 4$).
Q2: **We witness 100% forgetting of the target class adversarial samples when fine-tuning on 1 adversarial sample alone** ($Fig. 5$).
Q3: Unseen images from other classes, i.e., test accuracy, are more accurately identified post-attack with the "balanced" approach, where $M_{unlearn}(D) \approx M_{original}(D)$. ($Fig. 6$).

What's next?

We experimented with CIFAR-100, which is not presented here. CIFAR-100 delivers promising results, but this attack fares better on smaller benchmarks. Future work can focus on generalization across different datasets and model architectures.

Los Alamos NATIONAL LABORATORY
National Nuclear Security Administration
Managed by Triad National Security, LLC, for the U.S. Department of Energy's NNSA.
LA-UR-25-27560