NEURAL INFORMATION PROCESSING SYSTEMS
NeurIPS 2025
San Diego, USA

CoUn: Empowering Machine Unlearning via Contrastive Learning

  

Yasser H. Khalil   Mehdi Setayesh   Hongliang Li

Huawei Noah's Ark Lab, Montreal, Canada

E-mail: yasser1.khalil@huawei.ca

Dec. 2nd – Dec. 7th, 2025

Introduction

- **Data Privacy & Compliance**
  - Regulations such as the **GDPR** highlight the growing importance of data protection and responsible AI.
- **Machine Unlearning (MU)**
  - Removes the effect of specific training data (*forget set*) while preserving knowledge from remaining data (*retain set*).
- **Exact Unlearning – the Gold Standard**
  - Achieved by retraining from scratch on retain data — accurate but **computationally costly**.
- **Approximate Unlearning**
  - Offers an efficient alternative that approximates the retrained model’s performance.

Motivation

- **Retrain Model Insight**
  - Forget samples are mapped into clusters of retain samples that share the highest semantic similarity.

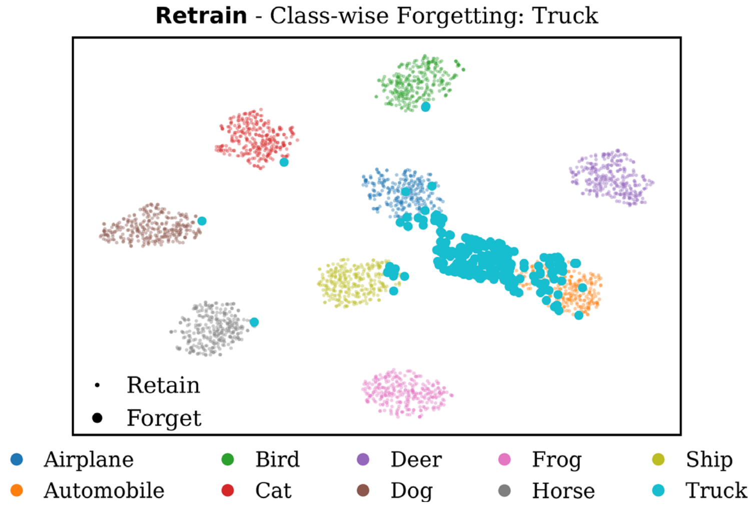

Figure 1. Representation space of the Retrain model (ResNet-18 on CIFAR-10). *Left:* excluding the “truck” class. *Right:* excluding 10% random samples. Larger dots denote forget samples projected into semantically closest retain clusters.

- **For approximate unlearning to emulate exact unlearning, the model should:**
  - Correctly classify retain samples.
  - Reposition forget samples into clusters of other retain samples with highest semantic similarity.

Methodology

- CoUn operates solely on retain data and integrates two components:
  - **Contrastive Learning (CL):** Refines the representation space based on semantic similarity.
  - **Supervised Learning (SL):** Preserves cluster separation.

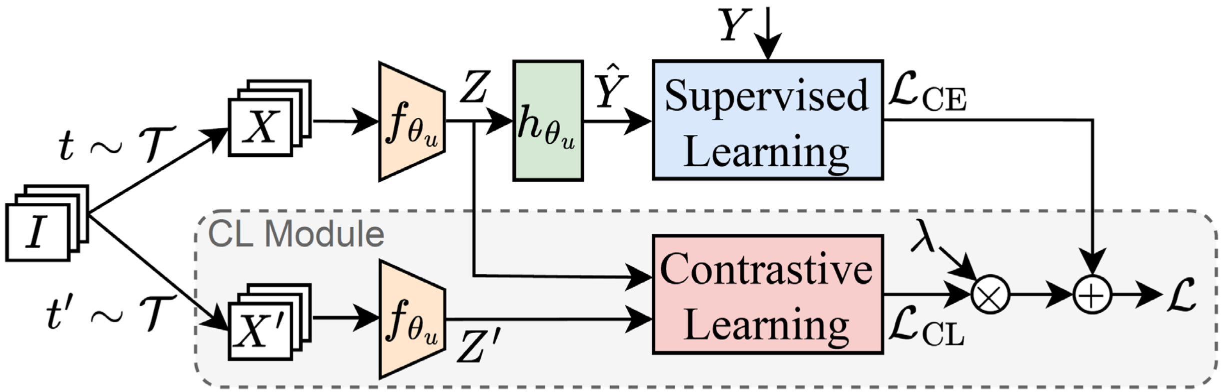

Figure 2. CoUn framework. Two augmented views are generated from a batch of retain images I. The feature extractor f_{Θ_u} produces representations (Z, Z'). The CL module aligns positive pairs and separates negatives, while the supervised head h_{Θ_u} preserves the decision boundaries.

- CL module *indirectly pushes* the forget representations toward semantically similar retain samples.

Empirical Analysis

- CoUn preserves well-separated retain clusters (high utility) while repositioning forget samples toward semantically similar retain clusters (improved forget quality), effectively approximating exact unlearning.

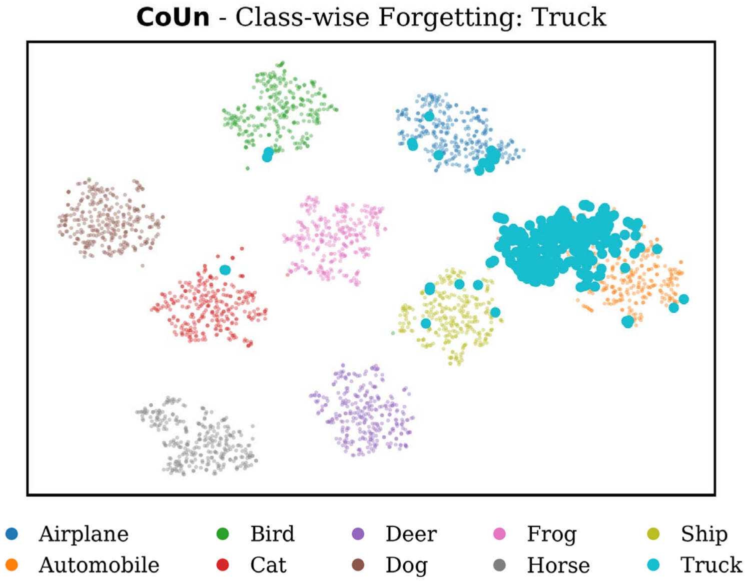

Figure 3. Representation space of CoUn. Similar to Retrain's representation space (Figure 1), CoUn positions forget samples into clusters of semantically similar retain samples while preserving retain cluster structure.

Evaluation Metrics

- **MU is assessed using the following key metrics:**
  1. **Retain Accuracy (RA):** Accuracy of the unlearned model Θ_u on retain data D_r.
  2. **Test Accuracy (TA):** Generalization performance of Θ_u on unseen test data.

     Utility

  3. **Unlearn Accuracy (UA):** 1 − the accuracy of Θ_u on forget data D_u.
  4. **Membership Inference Attack (MIA):** Identifies whether forget samples were seen during training.

     Forget Quality

  5. **Computation Cost:** Number of FLOPs required to obtain Θ_u.

- **Average Gap:** Measures the unlearning effectiveness of Θ_u

$$Avg.\,Gap = {}^{1}\!/_{4}\left(|RA - RA^*| + |UA - UA^*| + |TA - TA^*| + |MIA - MIA^*|\right),$$

where $(\cdot)^*$ denotes metrics of the Retrain model.

Results (1/2)

- **Performance Comparison of CoUn against baselines.**

Table 1. Performance comparison of CoUn to the baseline methods with 10% random data removal. The gap (Δ) and the (best) average gap between each method and the Retrain model are reported.

Dataset & Model	Method	Accuracy (%)			Efficacy (%)	Avg. Gap ↓	Comp. Cost (PFLOPs) ↓
		Retain (Δ ↓)	Unlearn (Δ ↓)	Test (Δ ↓)	MIA (Δ ↓)		
CIFAR-10 ResNet-18	Retrain	100.00 $\pm$ 0.00 (0.00)	4.81 $\pm$ 0.27 (0.00)	94.67 $\pm$ 0.24 (0.00)	11.02 $\pm$ 0.58 (0.00)	0.00	27.37
	FT	99.99 $\pm$ 0.00 (0.01)	3.76 $\pm$ 0.31 (1.05)	94.70 $\pm$ 0.14 (0.03)	9.51 $\pm$ 0.28 (1.51)	0.65	6.32
	NegGrad+	99.95 $\pm$ 0.02 (0.05)	4.82 $\pm$ 0.24 (0.01)	94.32 $\pm$ 0.23 (0.35)	9.09 $\pm$ 0.30 (1.93)	0.58	6.02
	ℓ_1-sparse	99.97 $\pm$ 0.01 (0.03)	5.40 $\pm$ 0.40 (0.59)	93.81 $\pm$ 0.21 (0.86)	10.97 $\pm$ 0.35 (0.05)	0.38	6.92
	SalUn	99.10 $\pm$ 0.35 (0.90)	4.31 $\pm$ 0.42 (0.50)	93.84 $\pm$ 0.27 (0.83)	11.15 $\pm$ 2.04 (0.13)	0.59	8.66
	NoT	99.99 $\pm$ 0.00 (0.01)	4.19 $\pm$ 0.25 (0.62)	94.65 $\pm$ 0.24 (0.02)	10.45 $\pm$ 0.51 (0.57)	0.30	7.52
	CoUn	99.99 $\pm$ 0.00 (0.01)	4.12 $\pm$ 0.31 (0.69)	94.57 $\pm$ 0.24 (0.10)	10.81 $\pm$ 0.31 (0.21)	0.25	8.02

Results (2/2)

- **Percentage improvement from integrating CoUn's CL module into baselines.**

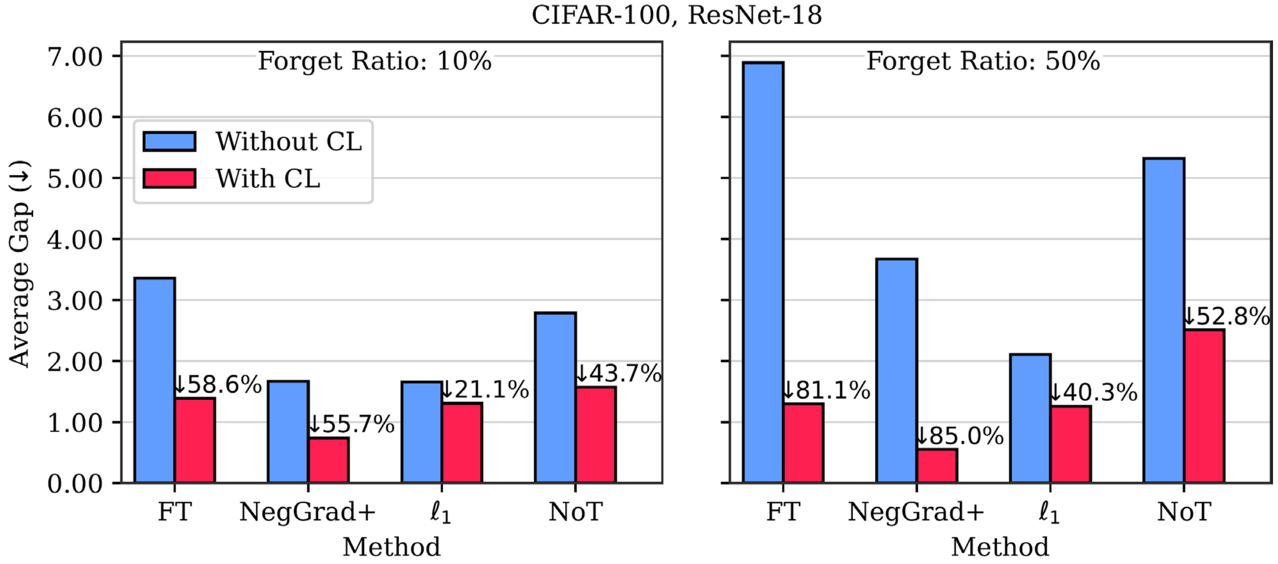

Figure 4. CL integration consistently improves unlearning performance, with larger gains at higher forget ratios (50%).

Conclusion

- **CoUn enables effective unlearning by leveraging semantic similarity between retain and forget samples.**
  - Uses a CL module on retain data to adjust their representations, and indirectly influences forget representations.
    - Improving forget quality.
  - Uses supervised learning on retain data to preserve their cluster separation.
    - Improving utility.

arXiv

NEURAL INFORMATION PROCESSING SYSTEMS
NeurIPS 2025
San Diego, USA

HUAWEI

Thank you!

E-mail: yasser1.khalil@huawei.ca