Northeastern University
Institute for Intelligent Networked Systems

MENTIS LABORATORY

CLUE: Bringing Machine Unlearning to Mobile Devices

MENTIS laboratory – https://mentis.info/
Northeastern University, United States

Overview

- Motivation
- Literature Overview
- Our Key Insight
- Methodology Overview
- Energy Loss
- Knowledge Distillation (Stability)
- Key Results
- Takeaways

Motivation

Why Machine Unlearning?

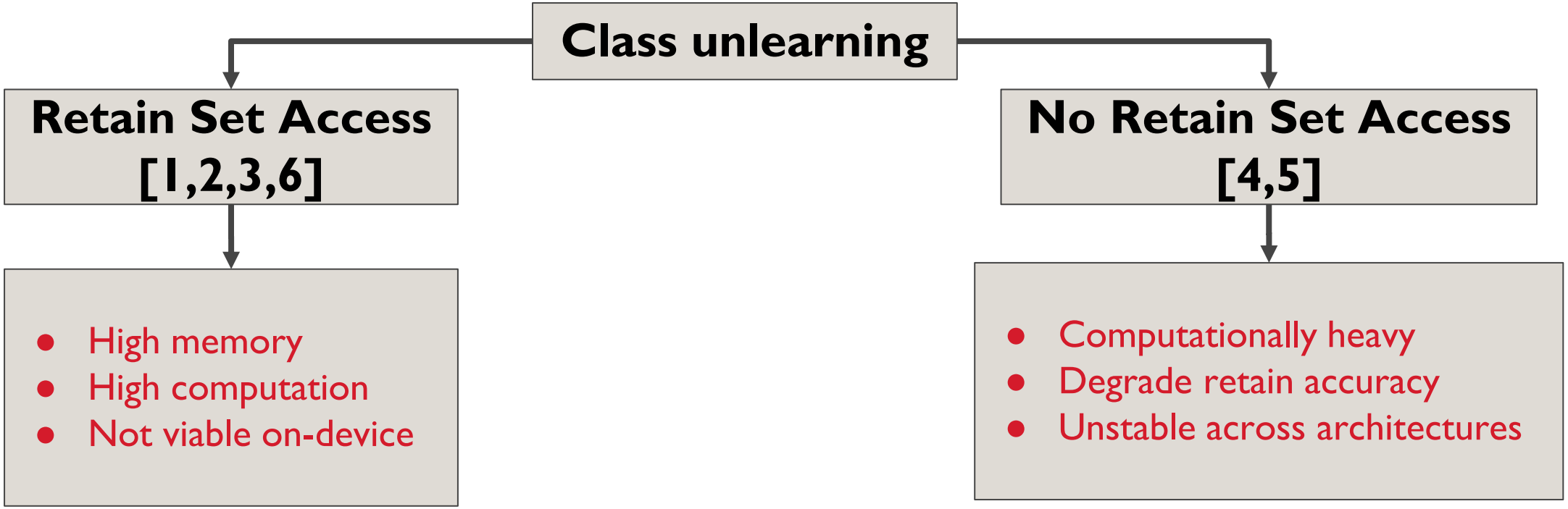
- Regulatory requirements (GDPR, CCPA)
- Backdoor removal
- Privacy-preserving face deletion
- Continual learning in mobile systems

On-device unlearning must be **fast, memory-light, and data-free**

⚠Most existing methods assume access to a retain dataset
⚠ Mobile devices cannot store or re-fetch that data

How do we unlearn a class **without retraining** and **without retain data**?

Current Literature



[1] Bourtoule, L., Chandrasekaran, V., Choquette-Choo, C. A., Jia, H., Travers, A., Zhang, B., Lie, D., & Papernot, N. (2021). Machine unlearning. Proceedings of the 42nd IEEE Symposium on Security and Privacy (SP), 141–159

[2] Foster, J., O'Reilly, D., Hada, S., & McLoughlin, I. (2024). Fast machine unlearning without retraining through selective synaptic dampening. Proceedings of the AAAI Conference on Artificial Intelligence, 38(11), 12026–12035

[3] Golatkar, A., Achille, A., & Soatto, S. (2020). Eternal sunshine of the spotless net: Selective forgetting in deep networks. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 9304–9312

[4] SChen, Min, et al. "Boundary unlearning: Rapid forgetting of deep networks via shifting the decision boundary." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.

[5] Foster, Jack, et al. "An Information Theoretic Approach to Machine Unlearning." Transactions on Machine Learning Research.

[6] Fan, Chongyu, et al. "SalUn: Empowering Machine Unlearning via Gradient-based Weight Saliency in Both Image Classification and Generation." The Twelfth International Conference on Learning Representations.

Our Key Insight

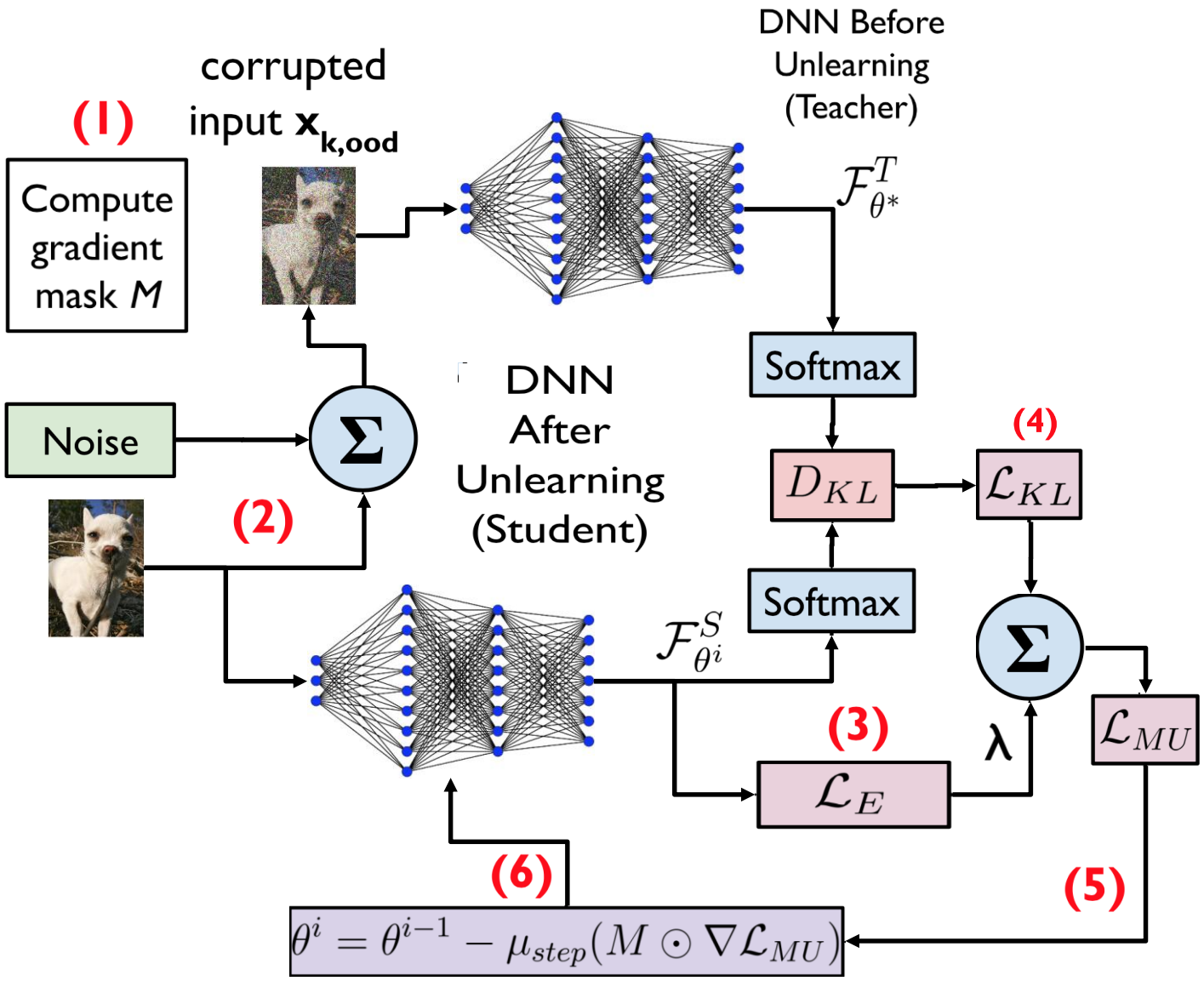
Reframe class unlearning as OOD induction

After unlearning class C_f samples from that class should behave like OOD data

Previous Work	*Clue*
Move forget samples to other class	Make forget samples look OOD

Class Removal	Distribution Shift

Methodology Overview



Energy Loss

Helmholtz Free Energy[1] $$E(\mathbf{x}; \mathcal{F}_\theta) = -T \cdot \log \left(\sum_{y_i} e^{\mathcal{F}_\theta^{y_i}(\mathbf{x})/T} \right)$$

OOD samples → High Energy → Low Partition Function

$$\mathcal{L}_E = \frac{1}{|\mathbf{D}_f|} \sum_{\mathbf{x}_k \in \mathbf{D}_f} \sum_{y_i} e^{\mathcal{F}_{\theta^*}^{y_i}(\mathbf{x}_i)/T}$$

[1] Liu, Weitang, et al. "Energy-based out-of-distribution detection." Advances in neural information processing systems 33 (2020): 21464-21475.

Knowledge Distillation + Gradient Masking

Energy alone destabilizes retain classes

Clean input to → student

Noisy (OOD-like) version → teacher

$$\mathcal{L}_{KL} = \frac{\sum_{\mathbf{x}_i \in \mathbf{D}_f} D_{KL}(\sigma_s(\mathcal{F}^T_{\theta^*}(\mathbf{x}_{i,ood})), \sigma_s(\mathcal{F}^S_{\theta^i}(\mathbf{x}_i)))}{|\mathbf{D}_f|}$$

Gradient Masking for Salient Parameter Update

$$R(w) = \left| w \frac{\partial \mathcal{L}_{CE}}{\partial w} \right|; \qquad M = 1(R(w) > \tau)$$

Key Results

Unlearning Methods	UA	RA	TA	MIA	Average Gap
Retrain	0	99.8	96.92	100	
GA	94.57 (94.57)	**99.28 (0.52)**	96.55 (0.37)	5.42 (94.48)	47.485
RL	0.02 (0.02)	84.11 (15.69)	81.56 (15.4)	99.62(0.38)	7.8725
IU	97.26 (97.26)	99.35 (0.45)	97.10 (-1.08)	2.70 (97.3)	49.02
BE	**0(0)**	89.27 (10.53)	85.12 (11.8)	100(0)	5.58
BS	4.73 (4.73)	92.33 (7.47)	89.55 (7.37)	95.26 (4.74)	6.07
LU	**0 (0)**	11.11 (87.66)	11.11 (85.81)	**100 (0)**	43.36
UNSIR	98.91 (98.91)	99.26 (0.54)	**97.01 (-0.09)**	1.08 (98.92)	49.61
CLUE (Ours)	2.04 (2.04)	98.42 (1.38)	95.86 (1.06)	97.95 (2.05)	**1.63**

ViT-B/16 and CIFAR10

- CLUE achieves the lowest average gap with *Retrain*
- BE is the second best with a gap of 3.95 with CLUE
- CLUE achieves 70.78% improvement over BE

Key Results

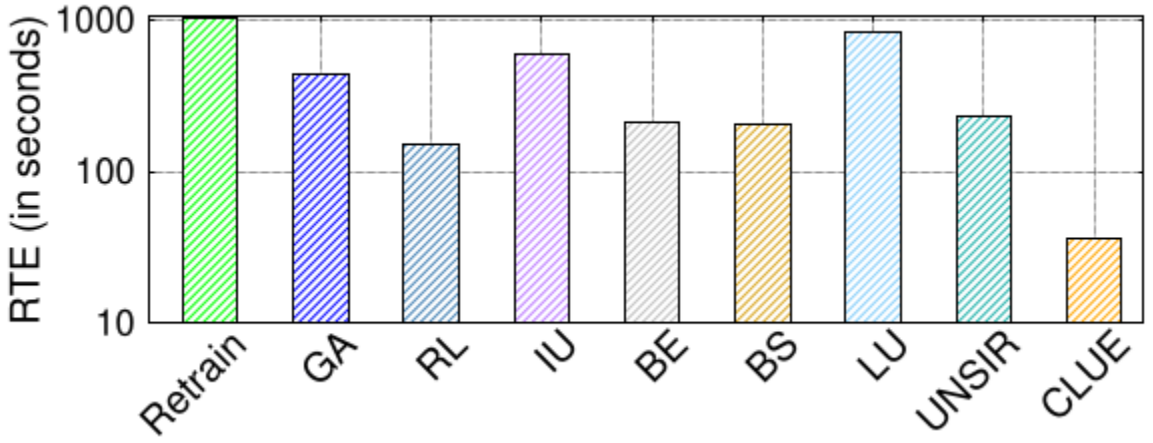
Unlearning Methods	UA	RA	TA	MIA	Average Gap
Retrain	0	99.75	90.18	100	
GA	8.66 (8.66)	79.43 (20.32)	76.54 (13.64)	91.33 (8.77)	12.83
RL	20 (20)	89.50 (10.25)	82.22 (7.96)	78.50 (21.5)	14.92
IU	17.77 (17.77)	94.05 (5.70)	87.87 (2.31)	82.22 (17.78)	10.89
BE	24 (24)	91.63 (8.12)	83.33 (6.85)	76 (24)	15.74
BS	46.17 (46.17)	93.26 (6.49)	86.58 (3.7)	53.82 (46.18)	25.63
LU	**0** **(0)**	11.11 (88.64)	11.11 (89.07)	**100** **(0)**	69.42
UNSIR	**0** **(0)**	14.37 (85.38)	14.23 (75.95)	**100** **(0)**	40.33
CLUE (Ours)	10.48 (10.48)	**96.61** **(3.14)**	**89.7** **(0.48)**	89.50 (10.50)	**6.15**

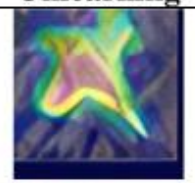
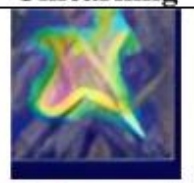
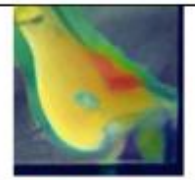
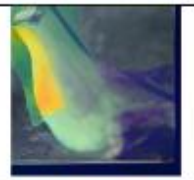
ResNet20 and CIFAR10

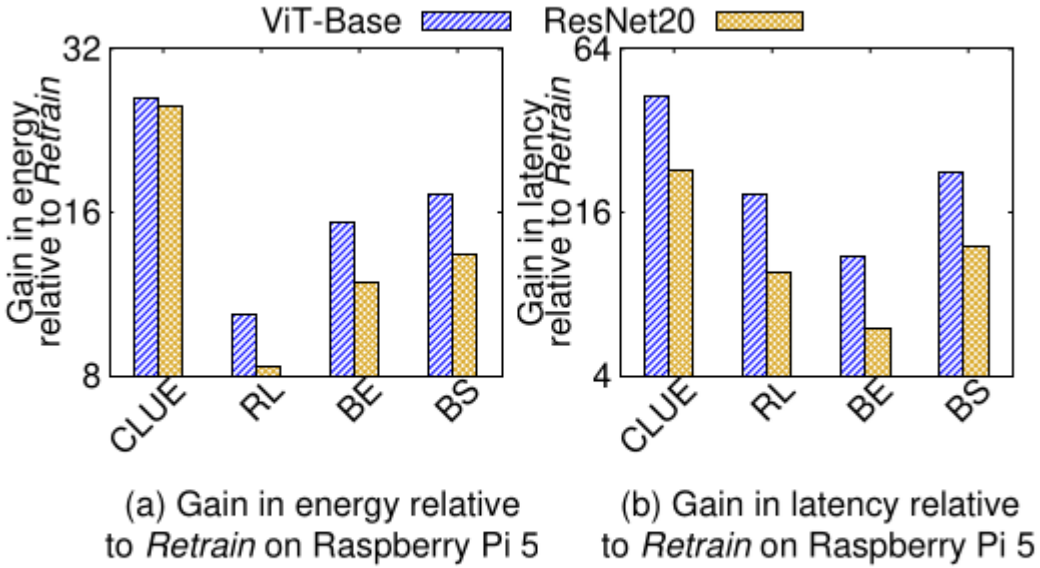
- CLUE again achieves the lowest average gap with *Retrain*
- IU is the second best with a gap of 4.74 with CLUE
- CLUE achieves 43.52% improvement over BE

CLUE consistently balances usability and forgetting across architectures and benchmarks

Key Results



Class ID	Original Image	Attention Map Before Unlearning	Attention Map After Unlearning
Class:0 (Airplane)			
Class:2 (Bird)			
Class:6 (Frog)			



(a) Gain in energy relative to *Retrain* on Raspberry Pi 5

(b) Gain in latency relative to *Retrain* on Raspberry Pi 5

Figure 5. Latency and energy comparison.

Takeaways

Reframes unlearning as OOD induction

No retain dataset required

Architecture-agnostic

Massive efficiency gains

Enables privacy-compliant on-device learning

Unlearning is approximate without formal guarantee

Northeastern University
Institute for Intelligent Networked Systems

MENTIS LABORATORY

Thank you!

MENTIS laboratory – https://mentis.info/
Northeastern University, United States