LoTUS: Large-Scale Machine Unlearning with a Taste of Uncertainty

Christoforos N. Spartalis Theodoros Semertzidis Efstratios Gavves Petros Daras

CERTH
CENTRE FOR RESEARCH & TECHNOLOGY HELLAS

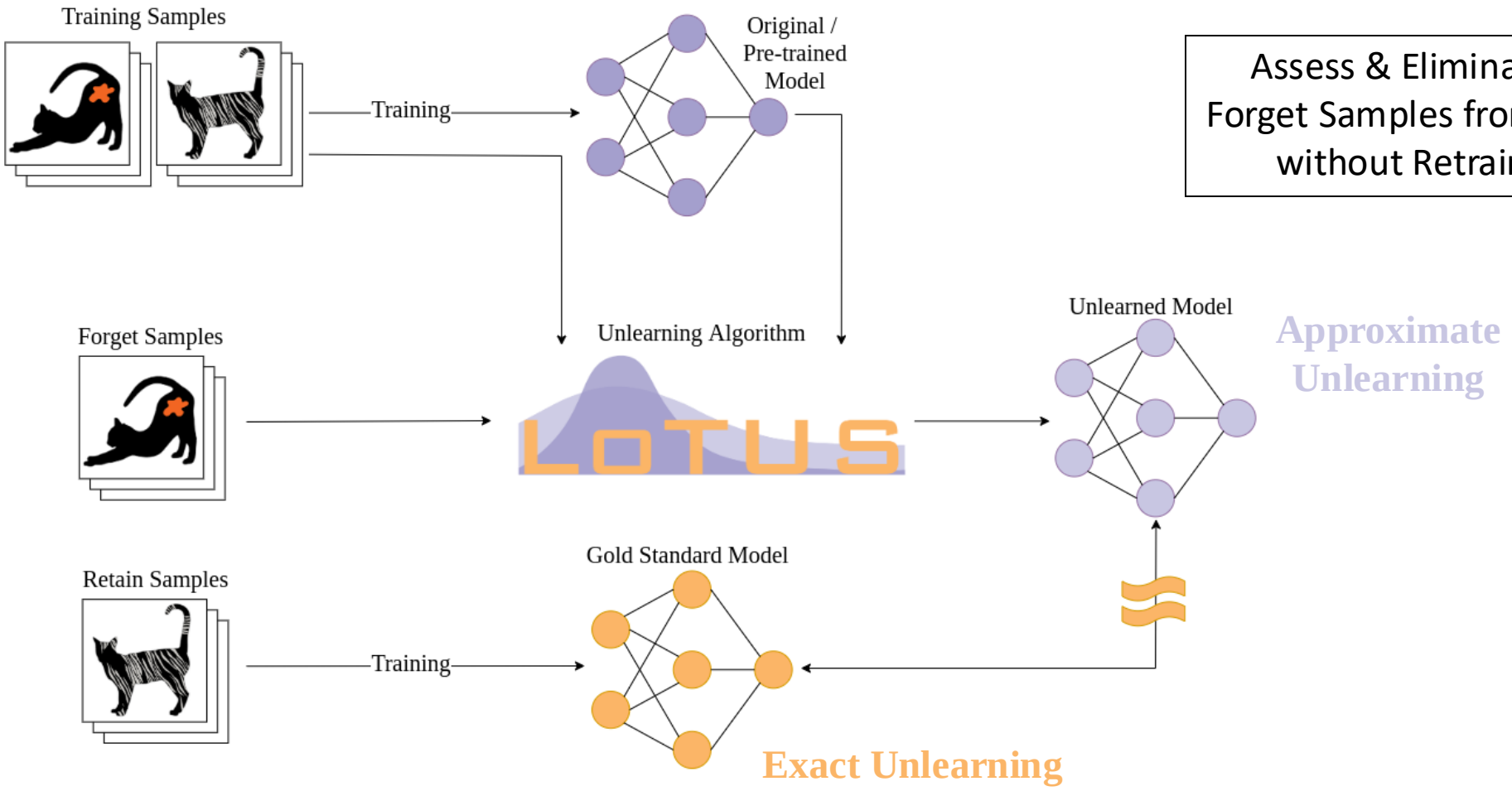
UNIVERSITEIT VAN AMSTERDAM

ARCHIMEDES

CVPR Nashville JUNE 11-15, 2025

ATLANTIS

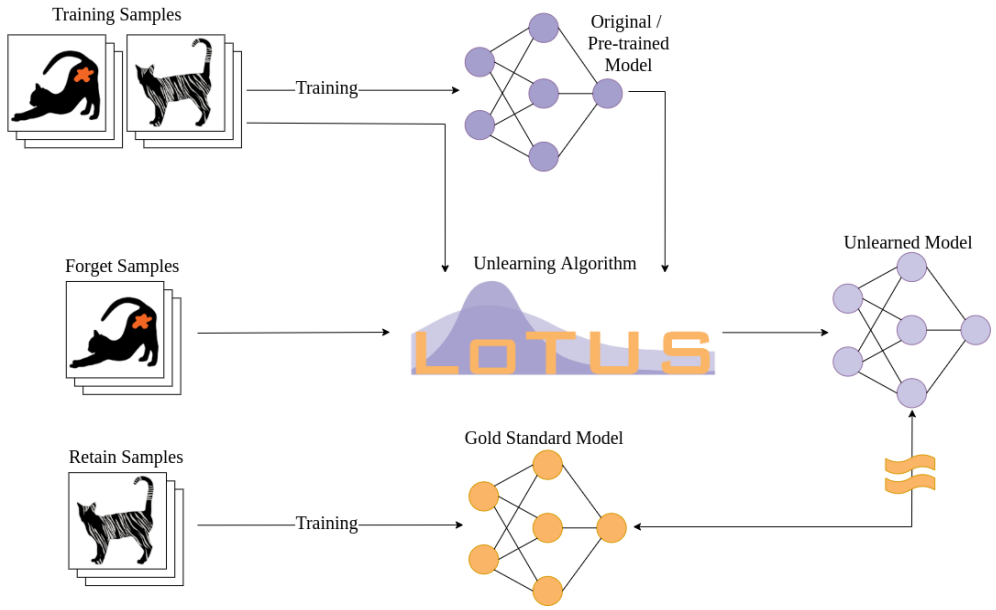
Machine Unlearning Definition & Applications



Assess & Eliminate the Influence of Forget Samples from Pre-trained Models without Retraining from Scratch

- Privacy
- Securtity
- Ethics
- Bias Mitigation
- Copyright Protection

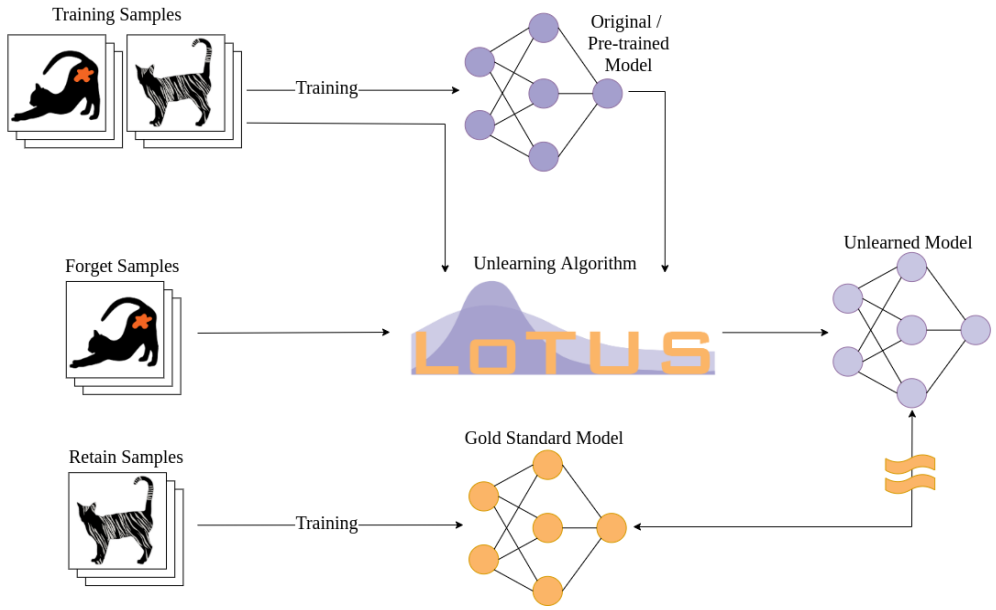
Challenges & Evaluation



1. **Limited Data Access**
2. **Efficiency**
3. **Cost**

1. **Balance Forgetting & Retention**
2. **Avoid the Streisand Effect**
3. **Minimize Time & Cost of Compute**

Challenges & Evaluation



1. **Limited Data Access**
2. **Efficiency**
3. **Cost**

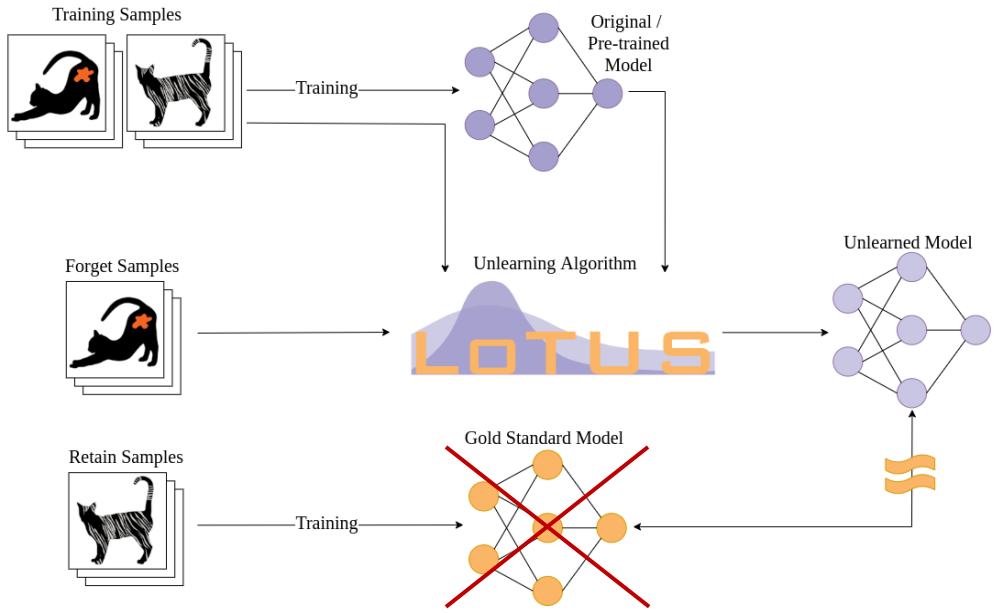
1. **Balance Forgetting & Retention**
2. **Avoid the Streisand Effect**
3. **Minimize Time & Cost of Compute**

1. Average Gap Metric (Unlearned & Gold Standard Models) ➡ Balance Forgetting & Retention ➡

MIA Accuracy	Retain Accuracy
Forget Accuracy	Test Accuracy

2. Jensen – Shannon Divergence Metric (Unlearned & Gold Standard)
   ↳ Unlearning Effectiveness & Resilience to the Streisand Effect
3. Runtime of Unlearning Algorithm & Computational Complexity Analysis

Challenges & Evaluation



1. **Limited Data Access**
2. **Efficiency**
3. **Cost**

1. **Balance Forgetting & Retention**
2. **Avoid the Streisand Effect**
3. **Minimize Time & Cost of Compute**

Real-World Conditions with Large-scale Datasets and Pre-trained Models?

1. ~~Average Gap Metric~~ (Unlearned & Gold Standard Model) ➡ Balance Forgetting & Retention ➡

MIA Accuracy	Retain Accuracy
Forget Accuracy	Test Accuracy

2. ~~Jensen – Shannon Divergence Metric~~ (Unlearned & Gold Standard)
   ↳ Unlearning Effectiveness & Resilience to the Streisand Effect
3. Runtime of Unlearning Algorithm & Computational Complexity Analysis

Contributions

- ✓ **Large-Scale Unlearning Benchmark**
  Simulates Real-World Conditions (ImageNet1k & Limited Data Access)
- ✓ **Retrain-Free Jensen-Shannon Divergence (RF-JSD) Metric**:
  Ensures Evaluation without the Gold Standard Model
- ✓ **LoTUS**
  Increases Entropy In Predictions up to an Information-Theoretical Bound

JSD (Unlearned & Gold Standard)

$$\mathcal{JS}\big(f_{\text{un}}(D_f) \;||\; f_{\text{gold}}(D_f)\big)$$

RF-JSD (Unlearned & Pre-trained)

$$\mathcal{JS}(f_{\text{un}}(D_f)\,||\,f_{\text{orig}}(D_u)) = \frac{1}{k}\sum_{c=1}^{k}\mathcal{JS}(P_i \;||\; Q_i)$$

$$P_i = \frac{1}{Z_P}\sum_{j=1}^{n_i} f_{\text{un}}(x_j \,|\, y_j = i)\,,\; Q_i = \frac{1}{Z_Q}\sum_{j=1}^{n_i} f_{\text{orig}}(x_j \,|\, y_j = i)$$

	Dataset $\left(\frac{\text{num. of forget samples}}{\text{num. of training samples}} \times 100\%\right)$	PCC (↑)	p-value (↓)
ViT	CIFAR-100 (10%)	0.84	0.0043
	CIFAR-10 (10%)	0.92	0.0005
	MUFAC	0.93	0.0003
	CIFAR-100 (50%)	0.94	0.0001
	CIFAR-10 (50%)	0.99	0.0000
ResNet18	CIFAR-100 (10%)	0.97	0.0000
	CIFAR-10 (10%)	0.90	0.0011
	MUFAC	0.88	0.0018
	CIFAR-100 (50%)	0.91	0.0006
	CIFAR-10 (50%)	0.89	0.0013
	Mean ± Std	$0.92_{\pm 0.04}$	$0.0010_{\pm 0.0016}$

Table 7. **Retrain Free-JSD (RF-JSD) and JSD Correlation** measured with the Pearson correlation coefficient (PCC).

Method	RF-JSD $\times 1e4$ (↓)	Time (↓)	Retain Acc.	MIA Acc.
Original	$1.22_{\pm 0.01}$	(pre-trained)	$0.94_{\pm 0.00}$	$0.71_{\pm 0.00}$
Finetuning	$2.22_{\pm 0.02}$	$16.24_{\pm 0.03}$	$0.97_{\pm 0.00}$	$0.78_{\pm 0.00}$
NegGrad+	$2.17_{\pm 0.02}$	$18.10_{\pm 0.03}$	$0.97_{\pm 0.00}$	$0.80_{\pm 0.00}$
Rnd Labeling	$1.80_{\pm 0.09}$	$19.37_{\pm 0.03}$	$0.95_{\pm 0.01}$	$0.74_{\pm 0.01}$
Bad Teacher	$3.16_{\pm 3.25}$	$11.66_{\pm 0.03}$	$0.77_{\pm 0.21}$	$0.52_{\pm 0.18}$
SCRUB	$1.24_{\pm 0.01}$	$24.49_{\pm 0.03}$	$0.94_{\pm 0.00}$	$0.71_{\pm 0.00}$
SSD	$1.23_{\pm 0.01}$	$22.61_{\pm 0.10}$	$0.94_{\pm 0.00}$	$0.71_{\pm 0.00}$
UNSIR	$2.54_{\pm 0.03}$	$33.12_{\pm 0.03}$	$0.99_{\pm 0.00}$	$0.77_{\pm 0.01}$
SalUn	$1.83_{\pm 0.03}$	$59.27_{\pm 0.37}$	$0.95_{\pm 0.00}$	$0.74_{\pm 0.01}$
LoTUS	$\mathbf{1.11_{\pm 0.01}}$	$\mathbf{10.72_{\pm 0.01}}$	$0.94_{\pm 0.00}$	$0.61_{\pm 0.01}$

Table 4. **Large-Scale Unlearning with ImageNet1k:** LoTUS outperforms state-of-the-art approaches in both unlearning effectiveness (RF-JSD) and efficiency (Time). While other metrics lack concrete validation due to the absence of a Gold Standard, they provide additional insights: LoTUS uniquely preserves the Retain Accuracy of the pre-trained model while reducing MIA Accuracy.

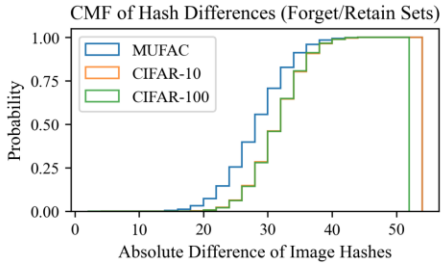
Results

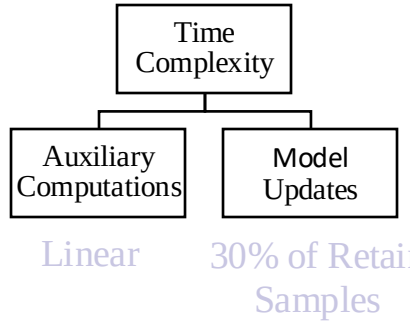
		Metric (↓)	Gold Std	Finetuning	NegGrad+ [23]	RndLbl [17]	BadT [9]	SCRUB [23]	SSD [14]	UNSIR [34]	SalUn [12]	**LoTUS**
Vision Transformer (ViT)	TinyIN	Avg Gap	0.0000	0.0175	0.0400	0.2925	0.0775	0.0225	0.0225	0.0225	0.0925	**0.0150**
		JSD ×1e4	0.00±0.00	0.05±0.00	0.10±0.00	0.64±1.03	0.18±0.01	0.04±0.00	0.04±0.00	0.06±0.00	0.25±0.59	**0.03±0.00**
		Time (min.)	228.9±6.49	22.64±0.02	25.20±0.02	25.19±0.02	16.91±0.05	33.25±0.01	27.27±0.06	21.17±0.22	76.97±1.72	**13.41±0.04**
	C-100	Avg Gap	0.0000	0.0275	0.0325	0.0175	0.0375	0.0200	0.0175	0.0250	0.0200	**0.0125**
		JSD ×1e4	0.00±0.00	0.07±0.00	0.13±0.01	0.06±0.00	0.17±0.01	**0.04±0.00**	**0.04±0.00**	0.08±0.01	0.06±0.01	**0.04±0.02**
		Time (min.)	112.25±0.13	11.35±0.00	12.63±0.01	12.79±0.02	9.18±0.27	16.74±0.03	13.67±0.02	10.69±0.01	38.15±0.04	**7.02±0.01**
	MUFAC	Avg Gap	0.0000	0.0400	0.0475	**0.0200**	0.1750	**0.0200**	**0.0200**	0.0475	**0.0200**	**0.0200**
		JSD ×1e−4	0.00±0.00	0.27±0.02	0.39±0.04	0.35±0.09	1.89±1.01	**0.05±0.02**	0.17±0.17	0.85±0.06	0.29±0.01	**0.05±0.01**
		Time (min.)	13.83±0.01	1.40±0.01	1.67±0.00	1.76±0.01	2.09±0.25	2.21±0.01	1.91±0.00	3.21±0.01	8.15±0.01	**1.09±0.00**
ResNet18 (RN18)	TinyIN	Avg Gap	0.0000	0.2200	0.2250	0.1925	0.2850	0.2725	0.2700	0.2375	0.2025	**0.1675**
		JSD ×1e4	0.00±0.00	1.80±0.04	1.82±0.07	1.71±0.11	1.81±0.04	0.98±0.00	0.96±0.02	1.76±0.05	1.93±0.09	**0.62±0.01**
		Time (min.)	46.81±0.57	2.85±0.00	3.17±0.00	3.44±0.01	1.91±0.01	3.97±0.00	3.47±0.00	5.00±0.01	6.06±0.06	**1.62±0.00**
	C-100	Avg Gap	0.0000	0.3600	0.3575	0.4025	0.3675	0.1650	0.2125	0.3625	0.3650	**0.1200**
		JSD ×1e4	0.00±0.00	6.88±0.59	6.87±0.62	5.84±0.98	4.30±0.49	1.87±0.08	3.04±1.55	3.05±0.32	6.50±0.60	**1.67±0.37**
		Time (min.)	3.39±0.30	0.43±0.00	0.49±0.00	0.57±0.00	0.34±0.01	0.58±0.00	0.54±0.00	0.45±0.00	1.55±0.01	**0.30±0.01**
	MUFAC	Avg Gap	0.0000	0.1525	0.1550	0.1300	**0.1025**	0.1625	0.1600	0.1450	0.1400	0.1250
		JSD ×1e4	0.00±0.00	19.52±6.23	19.16±5.31	9.51±2.39	9.41±0.04	10.53±2.31	10.30±2.28	16.32±4.82	15.25±5.20	**6.90±1.49**
		Time (min.)	7.34±0.77	0.76±0.00	0.91±0.00	1.06±0.00	0.66±0.00	1.20±0.00	1.07±0.00	1.68±0.02	2.72±0.02	**0.62±0.00**

Table 1. **Performance Summary of unlearning 10% of Tiny-ImageNet (TinyIN), CIFAR-100 (C-100), and MUFAC training sets:** LoTUS outperforms state-of-the-art approaches in **balancing forgetting and retention** (measured by Avg Gap), **unlearning effectiveness and resilience to the Streisand effect** (indicated by JSD), and **efficiency** (reflected in Time, measured in minutes).

	Metric (↓)	BadT	SCRUB	SSD	UNSIR	SalUn	**LoTUS**
Vision Transformer C-100	Avg. Gap	0.0575	0.0350	0.0350	0.0375	0.0375	**0.0225**
	JSD ×1e4	0.06±0.01	**0.01±0.00**	**0.01±0.00**	0.02±0.00	0.10±0.01	**0.01±0.00**
	Time (min)	15.04±0.03	16.82±0.03	18.69±0.06	18.33±0.02	38.08±0.02	**13.79±0.02**
Vision Transformer C-10	Avg. Gap	0.0600	0.0125	0.0150	0.0150	**0.0050**	**0.0050**
	JSD ×1e4	0.04±0.01	**0.00±0.00**	**0.00±0.00**	**0.00±0.00**	0.02±0.00	**0.00±0.00**
	Time (min)	15.10±0.20	16.99±0.35	19.03±0.54	18.33±0.02	37.93±0.18	**14.09±0.53**
ResNet18 C-100	Avg. Gap	0.3050	0.2225	0.2225	0.2925	0.3300	**0.1725**
	JSD ×1e4	0.55±0.04	0.44±0.02	0.44±0.02	0.65±0.23	6.25±0.45	**0.28±0.00**
	Time (min)	0.58±0.01	0.62±0.00	1.29±0.03	0.72±0.01	1.49±0.01	**0.57±0.01**
ResNet18 C-10	Avg. Gap	**0.0625**	0.1075	0.1025	0.1025	0.1300	0.0650
	JSD ×1e4	0.18±0.02	0.14±0.00	0.13±0.00	0.21±0.05	1.40±0.02	**0.09±0.01**
	Time (min)	**0.57±0.02**	0.61±0.02	1.27±0.02	0.73±0.00	1.49±0.01	**0.57±0.00**

Table 3. **Scaling up the Forget set to 50% of the training sets:** LoTUS outperforms state-of-the-art approaches in all metrics. Basic unlearning methods (Finetuning, NegGrad+, Rnd Labeling) are more efficient, but less effective than LoTUS.





		Metric (↓)	Gold Std	Finetuning	NegGrad+	RndLbl	BadT	SCRUB	SSD	UNSIR	SalUn	**LoTUS**
TinyIN Pizza		Avg Gap	0.0000	0.2975	0.3250	0.2925	0.3125	0.4200	0.3650	0.5075	0.2925	**0.0925**
		JSD ×1e4	0.00±0.00	94.96±7.24	86.36±9.66	92.27±6.43	72.62±22.07	73.10±0.82	34.96±14.21	102.29±9.33	91.01±8.59	**37.02±18.68**
		Time (min.)	42.15±16.05	3.23±0.01	3.24±0.03	3.27±0.03	1.59±0.01	4.05±0.03	3.19±0.03	**1.01±0.01**	3.98±0.01	1.30±0.02
C-100 Beaver		Avg Gap	0.0000	0.2825	0.3725	0.2925	0.3000	0.3225	0.4325	0.4050	0.2850	**0.1200**
		JSD ×1e4	0.00±0.00	101.48±2.87	108.50±2.59	102.66±3.11	78.65±3.12	64.09±8.71	45.19±9.19	76.28±6.88	100.93±2.44	**25.46±1.41**
		Time (min.)	4.00±0.11	0.43±0.00	0.44±0.01	0.45±0.00	0.26±0.01	0.55±0.00	0.83±0.03	**0.20±0.01**	1.16±0.01	0.23±0.01

Table 9. **Class Unlearning** with ResNet18 models and the TinyImageNet (TinyIN) and CIFAR-100 (C-100) datasets. We highlight the **best** and second-best scores.

	Metric (↓)	Gold Std	Finetuning	NegGrad+	RndLbl	BadT	SCRUB	SSD	UNSIR	SalUn	**LoTUS**	**LoTUS synthetic D_u**
ViT	Avg Gap	0.0000	0.0075	0.0125	0.0125	0.0375	**0.0050**	0.0075	0.0100	0.0125	**0.0050**	0.0075
	JSD ×1e4	0.00±0.00	**0.01±0.00**	0.03±0.00	0.02±0.01	0.12±0.03	**0.01±0.00**	0.02±0.01	**0.01±0.01**	**0.01±0.01**	**0.01±0.00**	**0.01±0.00**
	Time (min.)	111.00±1.99	11.33±0.03	12.61±0.03	12.78±0.01	8.97±0.02	16.66±0.02	13.65±0.02	10.68±0.02	37.97±0.19	7.34±0.19	**7.25±0.06**
RN18	Avg Gap	0.0000	0.1375	0.0975	0.0925	0.2650	0.0750	0.0825	0.1075	0.1800	0.0350	**0.0300**
	JSD ×1e4	0.00±0.00	1.03±0.24	1.06±0.21	1.00±0.26	2.39±2.03	0.41±0.09	0.82±0.57	0.65±0.05	1.09±0.05	**0.32±0.04**	**0.32±0.03**
	Time (min.)	5.32±1.18	0.43±0.00	0.49±0.00	0.57±0.00	0.33±0.00	0.58±0.00	0.54±0.00	0.45±0.00	1.56±0.01	**0.29±0.00**	0.30±0.01

Table 2. **Unlearning 10% of CIFAR-10**. LoTUS outperforms state-of-the-art approaches, both when the calibration/unseen set D_u consists of **real data** (●) and when it consists of **synthetic data** (●) from the CIFAKE dataset. We highlight the **best** and second-best scores.

Contributions

- ✓ **Large-Scale Unlearning Benchmark**
  Simulates Real-World Conditions (ImageNet1k & Limited Data Access)
- ✓ **Retrain-Free Jensen-Shannon Divergence (RF-JSD) Metric**:
  Ensures Evaluation without the Gold Standard Model
- ✓ **LoTUS**
  Increases Entropy In Predictions up to an Information-Theoretical Bound

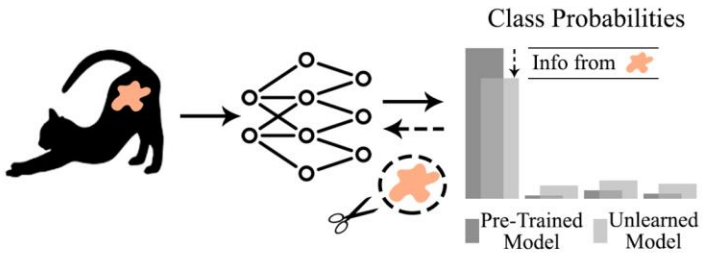
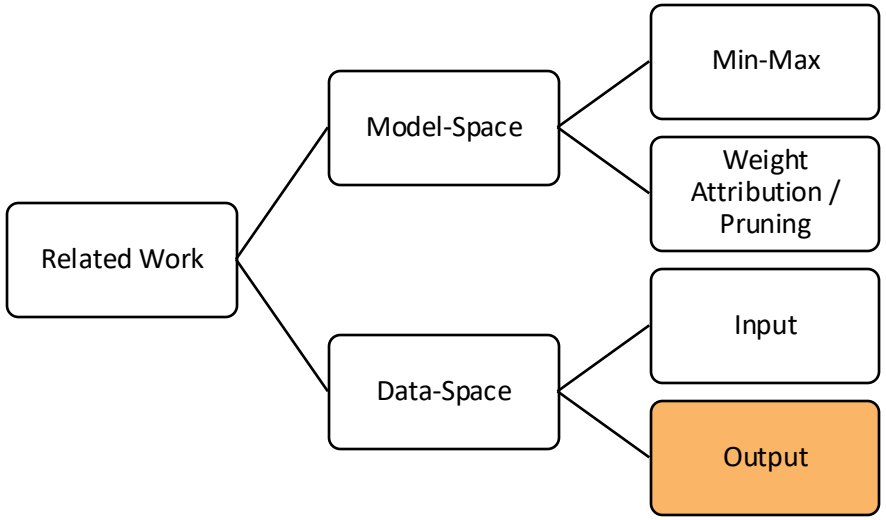


Figure 1. **Machine Unlearning via smoothing prediction probabilities:** LoTUS eliminates sample-specific information (e.g., unique fur patterns in cat images) that the DNN memorized and exposed through overconfident predictions. Then, the DNN responds to unlearned samples as if they were never part of the training set.



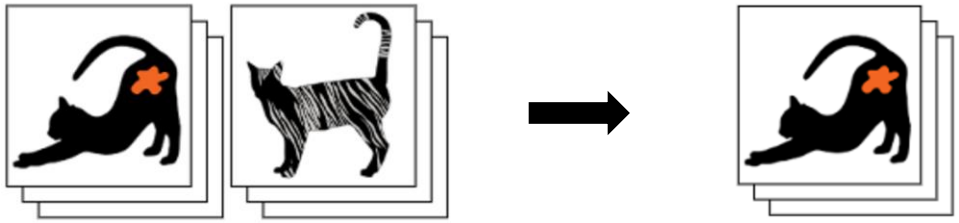
🤔 DNNs memorize sample-specific information from training data → Over-confident predictions → Exploited by Privacy Attacks

💡 Increase Model's Uncertainty / Entropy in Predictions

🤔 How Much Uncertainty?

💡 Information-Theoretical Framework →
Use Unseen (during training) Images for Calibration

Upper-Bounding Uncertainty



X : General Population $\quad$ X_s : Filtered Population

Data Processing Inequality: $X \to X_s \to f_w(X_s)$

$$\underbrace{I(f_w(X_s); X_s)}_{\text{total information captured by the model from the data subset}} = \underbrace{I(f_w(X_s); X)}_{\text{global information}} + \underbrace{I(f_w(X_s); X_s \mid X)}_{\text{subset-specific information}}$$

$$I\left(f_{\text{orig}}(X_s), X_s \in D_f\right) = I\left(f_{\text{orig}}(X_s), X\right) + I\left(f_{\text{orig}}(X_s); X_s \mid X\right)$$

$$I\left(f_{\text{gold}}(X_s), X_s \in D_f\right) = I\left(f_{\text{gold}}(X_s), X\right) + \cancel{I\left(f_{\text{gold}}(X_s); X_s \mid X\right)} \to 0$$

$$I\left(f_{\text{un}}(X_s); X_s\right) \triangleq I\left(f_{\text{un}}(X_s); X\right) \overset{\triangle}{=}^{*} I\left(f_{\text{orig}}(X_s); X\right)$$

🤔

*instance-wise unlearning

Upper-Bounding Uncertainty

X : General Population $\quad X_s$: Filtered Population

Data Processing Inequality: $X \to X_s \to f_w(X_s)$

$$\underbrace{I(f_w(X_s);X_s)}_{\text{total information captured by the model from the data subset}} = \underbrace{I(f_w(X_s);X)}_{\text{global information}} + \underbrace{I(f_w(X_s);X_s \mid X)}_{\text{subset-specific information}}$$

$$I\left(f_{\text{orig}}(X_s), X_s \in D_f\right) = I\left(f_{\text{orig}}(X_s), X\right) + I\left(f_{\text{orig}}(X_s); X_s \mid X\right)$$

$$I\left(f_{\text{gold}}(X_s), X_s \in D_f\right) = I\left(f_{\text{gold}}(X_s), X\right) + \cancelto{0}{I\left(f_{\text{gold}}(X_s); X_s \mid X\right)}$$

$$I\left(f_{\text{un}}(X_s); X_s\right) \triangleq I\left(f_{\text{un}}(X_s); X\right) \overset{\triangle}{=}^{*} I\left(f_{\text{orig}}(X_s); X\right)$$

$$I\left(f_{\text{orig}}(X_s), X_s \in D_u\right) = I\left(f_{\text{orig}}(X_s), X\right) + \cancelto{0}{I\left(f_{\text{orig}}(X_s); X_s \mid X\right)}$$

$$I\left(f_{\text{un}}(X_s), X_s \in D_f\right) = I\left(f_{\text{orig}}(X_s), X_s \in D_u\right) \overset{**}{\Rightarrow}$$

$$\cancel{H(X_s \in D_f)} - H\left(X_s \in D_f \mid f_{\text{un}}(X_s)\right) =$$

$$\cancel{H(X_s \in D_u)} - H\left(X_s \in D_u \mid f_{\text{orig}}(X_s)\right) \Rightarrow$$

$$H\left(X_s \in D_f \mid f_{\text{un}}(X_s)\right) = H\left(X_s \in D_u \mid f_{\text{orig}}(X_s)\right)$$

*instance-wise unlearning

**distributional similarity

Upper-Bounding Uncertainty

X : General Population $\quad X_s$: Filtered Population

Data Processing Inequality: $X \to X_s \to f_w(X_s)$

$$\underbrace{I(f_w(X_s); X_s)}_{\text{total information captured by the model from the data subset}} = \underbrace{I(f_w(X_s); X)}_{\text{global information}} + \underbrace{I(f_w(X_s); X_s \mid X)}_{\text{subset-specific information}}$$

$$I\left(f_{\text{orig}}(X_s), X_s \in D_f\right) = I\left(f_{\text{orig}}(X_s), X\right) + I\left(f_{\text{orig}}(X_s); X_s \mid X\right)$$

$$I\left(f_{\text{gold}}(X_s), X_s \in D_f\right) = I\left(f_{\text{gold}}(X_s), X\right) + \cancelto{0}{I\left(f_{\text{gold}}(X_s); X_s \mid X\right)}$$

$$I\left(f_{\text{un}}(X_s); X_s\right) \triangleq I\left(f_{\text{un}}(X_s); X\right) \overset{\triangle}{=}^{*} I\left(f_{\text{orig}}(X_s); X\right)$$

$$I\left(f_{\text{orig}}(X_s), X_s \in D_u\right) = I\left(f_{\text{orig}}(X_s), X\right) + \cancelto{0}{I\left(f_{\text{orig}}(X_s); X_s \mid X\right)}$$

$$I\left(f_{\text{un}}(X_s), X_s \in D_f\right) = I\left(f_{\text{orig}}(X_s), X_s \in D_u\right) \overset{**}{\Rightarrow}$$

$$\cancel{H(X_s \in D_f)} - H\left(X_s \in D_f \mid f_{\text{un}}(X_s)\right) =$$

$$\cancel{H(X_s \in D_u)} - H\left(X_s \in D_u \mid f_{\text{orig}}(X_s)\right) \Rightarrow$$

$$H\left(X_s \in D_f \mid f_{\text{un}}(X_s)\right) = H\left(X_s \in D_u \mid f_{\text{orig}}(X_s)\right)$$

Fano's Inequality: Lower Prediction Error Probability Implies Lower $H\left(X_s \mid f_w(X_s)\right)$

$$P_e \geq \frac{H\left(X_s \mid f_w(X_s)\right) - 1}{\log |A(X_s)|}$$

$$\text{Acc}(f_{\text{un}}, D_f) = \text{Acc}(f_{\text{orig}}, D_u)$$

*instance-wise unlearning

**distributional similarity

Unlearning with LoTUS

Temperature Parameter (Dynamically Adjusted)

$$\tau_d = \exp\left(\alpha \cdot \left(\mathrm{Acc}(f_{\mathrm{un}}, D_f) - \mathrm{Acc}(f_{\mathrm{orig}}, D_u)\right)\right)$$

Gumbel – Softmax Activation Function

$$p_i = gs(\pi, \tau) = \frac{\exp\left(\left(\log \pi_i + g_i\right)/\tau\right)}{\sum_{j=1}^{k} \exp\left(\left(\log \pi_j + g_j\right)/\tau\right)}, \; i = 1, \ldots, k$$

Loss in the Teacher – Student Framework

- T: Pre-trained Model + Output Perturbation
- S: Unlearned Model initialized with T's weights

$$\ell(x, f_{\mathrm{orig}}, f_{\mathrm{un}}) = \underbrace{l \cdot gs\left(f_{\mathrm{orig}}(x), \tau_d\right) \odot \log s\left(f_{\mathrm{un}}(x)\right)}_{\text{forget}} + \underbrace{(1-l) \cdot gs\left(f_{\mathrm{orig}}(x), \tau \to 0^{+}\right) \odot \log s\left(f_{\mathrm{un}}(x)\right)}_{\text{retain}}$$

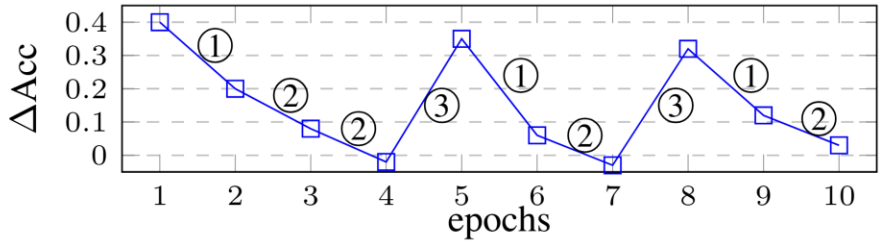


Figure 2. Contribution of the dynamically adjusted temperature τ_d to convergence toward the objective $\Delta\mathrm{Acc} = \mathrm{Acc}(f_{\mathrm{un}}, D_f) - \mathrm{Acc}(f_{\mathrm{orig}}, D_u) = 0$. The steps are denoted as follows: ①: sharp step towards objective, ②: smaller step (proportional to ΔAcc), ③: drastic accuracy restoration when over-unlearning.

Why Gumbel – Softmax?

		Vision Transformer				ResNet18			
		TinyImageNet	CIFAR-100	CIFAR-10	MUFAC	TinyImageNet	CIFAR-100	CIFAR-10	MUFAC
Avg Gap	Gumbel-Softmax	**0.0150**	**0.125**	**0.0050**	**0.0200**	**0.1675**	0.1200	**0.0350**	0.1250
	Softmax with Temperature	0.0675	0.0225	**0.0050**	**0.0200**	0.1850	**0.1075**	0.0675	**0.1175**
JSD×1e4	Gumbel-Softmax	**0.03**	**0.04**	**0.01**	**0.05**	**0.62**	1.67	**0.32**	**6.90**
	Softmax with Temperature	0.15	**0.04**	**0.01**	0.08	0.65	**1.36**	0.41	7.33

Table 10. **Contribution of Gumbel noise into the activation function.** Ablation analysis using Gumbel-Softmax and Softmax with Temperature as activation functions. LoTUS performs better with Gumbel-Softmax in the majority of the benchmarks.

Conclusion & Thank you!

- ✓ **LoTUS:**
  - 1st Entropy-based Unlearning Method that Meticulously Increases Uncertainty up to an Information-Theoretical Bound
  - Outperforms SoTA in Balancing Forgetting-Retention, Unlearning Effectiveness & Resilience to the Streisand Effect, and Efficiency

- ✓ **Retrain-Free Jensen-Shannon Divergence (RF-JSD) Metric**:
  - Ensures Evaluation without the Gold Standard Model
  - Strong Correlations with the Established JSD Metric
  - More Interpretable & Efficient than the existing ZRF Metric

- ✓ **Large-Scale Unlearning Benchmark** Simulating Real-World Conditions