ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

RKLU: Redistributive KL Distillation for Efficient Retain-Free Machine Unlearning

Varun Sampath Kumar, Esmaeil S. Nadimi, and Vinay Chakravarthi Gogineni *

Applied AI and Data Science Unit, The Maersk Mc-Kinney Moller Institute
University of Southern Denmark, Odense - Denmark

Abstract. Machine unlearning aims to remove the influence of specific training samples from model, motivated by privacy regulations and data revocation requirements. Existing approaches often depend on retain data, which compromises privacy and becomes computationally expensive. To address this challenge, we propose RKLU, a novel retain-data-free unlearning method that fine-tunes models by minimizing the KL divergence between their outputs and a target distribution that suppresses the probabilities of the samples to be forgotten. RKLU achieves near-perfect unlearning with minimal utility loss on diverse vision and text classification benchmarks, offering a privacy-preserving and efficient alternative.

1 Introduction

As AI systems are trained on vast datasets containing sensitive information, privacy concerns and regulatory compliance have become critical challenges. Once a model is trained, the influence of data is embedded in its parameters, making simple deletion from the database insufficient to erase its impact [1]. Selective unlearning is therefore essential to comply with regulations like GDPR [2], CCPA [3] and PIPEDA [4], which mandate the removal of specific data and associate knowledge upon request (e.g., the "right to be forgotten"). Machine unlearning [5, 6] enables models to eliminate the influence of targeted data points without full retraining. Unlearning methods are broadly categorized as *exact* or *approximate*. Exact unlearning methods like full retraining or SISA [7] aim to completely remove the influence of the forget set, which consists of data points requested for removal. Although effective, these methods are computationally intensive and often unscalable for large models. In contrast, approximate methods offer weaker guarantees of removing the forget set but are more scalable. These methods typically rely on the retain set, which includes all data not requested for removal, to preserve model utility. However, this dependence on retained data introduces privacy risks and limits the safe application of such methods.

This motivates the need for developing approximate unlearning algorithms that do not rely on the retain set. Such approaches, often referred to as retain-data-free unlearning or zero-shot unlearning, aim to estimate the effect of forgetting without access to the remaining training data. To this end, we propose RKLU, a novel retain-data-free method that fine-tunes models by minimizing

*This work received funding from the Novo Nordisk Foundation. (Project Reference Number: NNF24OC0095455)

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

KL divergence between their outputs and a modified target distribution. RKLU suppresses the influence of forget samples by redistributing their output probabilities across classes to preserve model performance and ensure stable, efficient unlearning. We demonstrate its effectiveness on diverse vision benchmarks including CIFAR-10, CIFARSuper-20, TinyImageNet, and the Twenty Newsgroups text classification dataset.

2 Related Work

Early retain-data-free unlearning methods primarily used gradient ascent [8] on the forget set to undo training effects, but this often caused catastrophic forgetting, harming performance on retained data. To address this, recent approaches introduced more advanced techniques. Chen et al.'s Boundary Unlearning [9] shifts decision boundaries using only forget data to rapidly forget entire classes while preserving retained knowledge. Similarly, Foster et al.'s Just in Time (JiT) [10] applies an information-theoretic approach to minimize gradients with respect to forget samples in a small neighborhood in a zero-shot manner, achieving effective unlearning without accessing the retain set.

3 Proposed Method

Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ be the training dataset with inputs $x_i \in \mathbb{R}^d$ and labels $y_i \in \{1, \ldots, K\}$. The classification model $\phi_\theta : \mathcal{X} \to \mathbb{R}^K$, parameterized by $\theta \in \mathbb{R}^m$, outputs logits $\hat{y} = \phi_\theta(x)$ converted to probabilities $p = \text{softmax}(\hat{y})$, where $[p]_k$ denotes the probability for class k. Given a forget set $\mathcal{D}_f \subset \mathcal{D}$ and the retain set $\mathcal{D}_r = \mathcal{D} \setminus \mathcal{D}_f$, the goal is to update θ to θ' so that $\phi_{\theta'}$ forgets $\mathcal{D}_f$ without access to $\mathcal{D}_r$.

3.1 RKLU: Redistributive KL Distillation

We propose RKLU, a lightweight and retain-data-free unlearning method based on KL divergence minimization between the original model outputs and a modified target distribution. Unlike conventional knowledge distillation or fine-tuning, RKLU focuses solely on the forget set $\mathcal{D}_f$ and does not require access to any retained data, making it highly privacy-preserving and efficient.

Unlearning via Target Redistribution. Given a sample $(x, y) \in \mathcal{D}_f$, let $p^{\text{orig}} = \text{softmax}(\phi_{\theta_{\text{orig}}}(x))$ denote the output probability distribution of the original model (prior to unlearning), where θ_{orig} denotes the frozen pre-unlearning weights. We construct the modified target distribution $\tilde{p}$ by setting the forget class probability to zero and redistributing its mass equally among the remaining $K - 1$ classes, added to their original probabilities:

$$[\tilde{p}]_k = \begin{cases} 0, & \text{if } k = y, \\ [p^{\text{orig}}]_k + \dfrac{[p^{\text{orig}}]_y}{K-1}, & \text{otherwise.} \end{cases}$$

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

This redistribution sets the forget class probability to zero while adding its mass equally to other classes, ensuring the total sums to one and preserving the relative order among non-forget classes. To unlearn, we fine-tune the model on $\mathcal{D}_f$ by minimizing the KL divergence between the current model output $p = \text{softmax}(\phi_\theta(x))$ and the modified target $\tilde{p}$. The loss for a single forget sample is:

$$\mathcal{L}_{\text{RKLU}}(x, y) = \text{KL}(\tilde{p} \,\|\, p) = \sum_{k=1}^{K} [\tilde{p}]_k \log \left(\frac{[\tilde{p}]_k}{[p]_k} \right).$$

Minimizing this loss encourages the updated model to shift probability mass away from the forget class y while preserving the structure of non-target predictions, leading to stable and controlled forgetting.

4 Experiments and Results

Notation: Across all tables, boldface indicates the method closest to retraining under the L1 distance computed over forget and retain accuracies.

Datasets and Models: We evaluated RKLU on vision and text benchmarks. For image classification, we used pre-trained MobileNetV2 [11], ResNet-18 [12], and ViT-B/16 [13] on TinyImageNet [14], CIFAR10, and CIFARSuper20 [15]. For text classification, we employed DistilBERT [16] and RoBERTa [17] on the 20 Newsgroups dataset [18].

Baselines and Metrics: We compared RKLU against state-of-the-art retain-data-free baselines, Boundary Shrinking (BDSH) [9] and Just in Time (JiT) [10], as well as Gradient Ascent (GA) and Retrained models. Effectiveness is measured by forget set accuracy and membership inference attack (MIA) success, while utility is assessed via retain set accuracy.

Implementation and Hyperparameters: Vision models were pretrained and fine-tuned for 10 epochs using Adam on an NVIDIA H100 GPU, with learning rate 1×10^{-3} and batch size 64. Gradient Ascent unlearning ran for 1 epoch. BDSH and JiT use official implementations. RKLU was performed with learning rate 1×10^{-3} for 5–10 epochs depending on the scenario and model size.

4.1 Full Class Unlearning

We evaluate RKLU for full-class unlearning on both vision and text domains. As shown in Table 1 and Table 2, RKLU consistently achieves near-zero forget accuracy across all models, comparable to retraining and outperforming Gradient Ascent and BDSH. In vision tasks, RKLU maintains strong retain accuracy, especially in models like ResNet18 where other methods struggle in preserving model utility. In text classification, it preserves high retain accuracy while fully forgetting the target class, surpassing Gradient Ascent which suffers significant utility loss. These results demonstrate RKLU's effectiveness and generalizability across modalities.

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

Table 1: Performance evaluation of single-class unlearning on TinyImageNet.

Model	Metric	Original	Retrain	GA	BDSH	JiT	Ours
MobileNetv2	$\text{Acc}_{\mathcal{D}_f}\downarrow$	92	0	0	2	0	**0**
	$\text{Acc}_{\mathcal{D}_r}\uparrow$	71.91	67.12	25.22	55.67	63.1	**63.59**
ResNet18	$\text{Acc}_{\mathcal{D}_f}\downarrow$	82	0	0	4	0	**0**
	$\text{Acc}_{\mathcal{D}_r}\uparrow$	70.21	66.34	12.88	57.21	52	**66.47**
ViT-B/16	$\text{Acc}_{\mathcal{D}_f}\downarrow$	98	0	0	0	2	**0**
	$\text{Acc}_{\mathcal{D}_r}\uparrow$	85.28	80.20	0.5	84.06	83.26	**84.07**

Table 2: Performance evaluation of single-class unlearning on Twenty Newsgroups.

Model	Method	$\textbf{Acc}_{\mathcal{D}_r}\uparrow$	$\textbf{Acc}_{\mathcal{D}_f}\downarrow$
DistilBERT	Original	70.81	80.26
	Retrain	70.96	0
	Gradient Ascent	25.50	0
	Ours	**69.83**	**0**
RoBERTa	Original	71.38	85.64
	Retrain	71.39	0
	Gradient Ascent	12.60	0
	Ours	**67.22**	**0**

4.2 Sub-Class Unlearning

Sub-class unlearning is inherently challenging, as it requires forgetting specific classes without disrupting semantically similar ones within the superclass. For example, in CIFARSuper20, the goal may be to forget the class *baby* from the *people* superclass while retaining *boy*, *girl*, *man*, and *woman*. We evaluated this fine-grained scenario using MobileNetV2, ResNet18, and ViT-B/16. As shown in Table 3, RKLU consistently reduces forget set accuracy (D_f) to near zero while maintaining high retain set accuracy (D_r). Compared to Gradient Ascent, BDSH, and JiT baselines, RKLU achieves lower forget accuracy with minimal impact on model utility, and demonstrates stronger robustness to membership inference attacks (MIA).

4.3 Sample Unlearning

We performed sample-level unlearning on CIFAR10 using ResNet18 and MobileNetV2 with forget sets of 100 and 500 randomly selected samples. Since BDSH targets class-level unlearning, we exclude it here, aligning with its original scope. Instead, we compare against Gradient Ascent and JiT baselines. As shown in Table 4, our method achieves comparable or superior forget and test accuracy, demonstrating its effectiveness and competitiveness in sample-level unlearning.

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

Table 3: Performance evaluation of sub-class unlearning on CIFARSuper20.

Model	Metric	Original	Retrain	GA	BDSH	JiT	Ours
MobileNetv2	$\text{Acc}_{\mathcal{D}_f}\downarrow$	75	2	0	4	0	**4**
	$\text{Acc}_{\mathcal{D}_r}\uparrow$	79.81	83.16	64.40	69.05	63.05	**71.25**
	MIA	0.658	0.254	0.452	0.116	0.184	**0.164**
ResNet18	$\text{Acc}_{\mathcal{D}_f}\downarrow$	85	2	0	13	4	**4**
	$\text{Acc}_{\mathcal{D}_r}\uparrow$	79.13	82.91	61.12	71.04	66.19	**72.98**
	MIA	0.988	0.140	0.694	0.244	0.228	**0.220**
ViT-B/16	$\text{Acc}_{\mathcal{D}_f}\downarrow$	96	11	0	**4**	2	0
	$\text{Acc}_{\mathcal{D}_r}\uparrow$	94.36	94.87	40.41	**94.20**	89.22	90.59
	MIA	0.848	0	0.44	0.048	0.282	**0.016**

Table 4: Performance evaluation of sample-level unlearning on CIFAR10. Bold indicates the best trade-off between forgetting and test accuracy among unlearning methods.

Model	$\lvert D_f\rvert$	Method	$\textbf{Acc}_{\mathcal{D}_f}\downarrow$	$\textbf{Acc}_{test}\uparrow$
ResNet18	100	Original	100	95.32
		Gradient Ascent	98	95.34
		JiT	98	94.87
		Ours	**95**	**93.60**
	500	Original	99.01	95.32
		Gradient Ascent	73.48	76.76
		JiT	78.73	80.82
		Ours	**80.42**	**82.49**
MobileNetv2	100	Original	99	94.23
		Gradient Ascent	98	94.21
		JiT	82	83.07
		Ours	**81**	**89.58**
	500	Original	97.79	94.23
		Gradient Ascent	78.38	80.76
		JiT	71.55	72.94
		Ours	**62.84**	70.21

5 Conclusion and Future Work

In this work, we have presented RKLU, a novel retain-data-free machine unlearning framework that achieves selective knowledge erasure through redistributive KL divergence. RKLU significantly outperforms existing state-of-the-art retain-data-free methods, establishing a new benchmark in this setting. A promising direction for future research is extending redistributive KL-based unlearning to generative models, particularly large language models (LLMs), where the goal would be to suppress the generation of specific sequences or tokens associated with the forget set analogous to suppressing the target class in classification tasks.

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

References

[1] Devansh Arpit, Stanislaw Jastrzkebski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S. Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, and Simon Lacoste-Julien. A closer look at memorization in deep networks. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 233–242. PMLR, 06–11 Aug 2017.

[2] A. Mantelero. The eu proposal for a general data protection regulationand the roots of the 'right to be forgotten'. *Computer Law & Security Review*, 29(3):229–235, 2013.

[3] E. Goldman. An introduction to the california consumer privacy act (ccpa). *Santa Clara Univ. Legal Studies Research Paper*, 2020.

[4] O. of the Privacy Commissioner of Canada. Personal information protection and electronic documents act, 2018.

[5] Thanh Tam Nguyen, Thanh Trung Huynh, Zhao Ren, Phi Le Nguyen, Alan Wee-Chung Liew, Hongzhi Yin, and Quoc Viet Hung Nguyen. A survey of machine unlearning. *ACM Trans. Intell. Syst. Technol.*, 2025.

[6] Vinay Chakravarthi Gogineni and Esmaeil S. Nadimi. Efficient knowledge deletion from trained models through layer-wise partial machine unlearning. *Journal of Machine Learning Research*, 26(245):1–33, 2025.

[7] Lucas Bourtoule, Varun Chandrasekaran, Christopher A. Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. Machine unlearning. In *IEEE Symposium on Security and Privacy (SP)*, pages 141–159, 2021.

[8] A., A. Golatkar, S. Achille, and Soatto. Eternal sunshine of the spotless net: Selective forgetting in deep networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9304–9312, 2020.

[9] Min Chen, Weizhuo Gao, Gaoyang Liu, Kai Peng, and Chen Wang. Boundary Unlearning: Rapid Forgetting of Deep Networks via Shifting the Decision Boundary . In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7766–7775, Los Alamitos, CA, USA, June 2023. IEEE Computer Society.

[10] Jack Foster, Kyle Fogarty, Stefan Schoepf, Zack Dugue, Cengiz Oztireli, and Alexandra Brintrup. An information theoretic approach to machine unlearning, 2024.

[11] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. MobileNetV2: Inverted Residuals and Linear Bottlenecks . In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4510–4520, Los Alamitos, CA, USA, June 2018. IEEE Computer Society.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[13] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

[14] Ya Le and Xuan S. Yang. Tiny imagenet visual recognition challenge, 2015.

[15] Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, 2009.

[16] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108, 2019.

[17] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019.

[18] Tom Mitchell. Twenty Newsgroups. UCI Machine Learning Repository, 1997. DOI: https://doi.org/10.24432/C5C323.