Partially Blinded Unlearning: Class Unlearning for Deep Networks from a Bayesian Perspective

Subhodip Panda, Shahswat Sourav, Prathosh AP

The 39^{th} Annual AAAI Conference on Artificial Intelligence
(AAAI, 2025)

April 15, 2026

Table of Contents: Partially Blinded Unlearning

Introduction

- What is Machine Unlearning?
  - machine unlearning refers to the task of forgetting the learned information or erasing the influence of a specific data subset of the training dataset from a learned model in response to a user request.
- Objective of Machine Unlearning
  - Z as an example space, i.e., a space of datasets.
  - Given a dataset D, we want to obtain a machine-learning model from a hypothesis space H. The process of training a model on D by a learning algorithm, denoted by a function $A : Z \to H$, with the trained model denoted as $A(D)$.
  - To support forgetting requests, an unlearning mechanism, denoted by a function U, that takes as input a training dataset $D \in Z$, a forget set $D_f \subset D$ (data to forget), and a model $A(D)$. It returns a sanitized (or unlearned) model $U(D, D_f, A(D)) \in H$.
  - The unlearned model is expected to be the same or similar to a retrained model $A(D \setminus D_f)$

Contributions

- We provide a theoretical formulation for Class Unlearning from a Bayesian perspective and propose a methodology applicable to unlearning specific classes in deep neural networks.
- Our method exhibits superiority over concurrent class unlearning methods by not mandating access to the entire dataset, only requiring partial access to the unlearned data. This feature proves advantageous in more restricted experimental settings.
- Our method's single-step unlearning process contrasts with two-step approaches used by some contemporary methods, showcasing superior computational efficiency and simplicity.
- We validate our approach using ResNet-18, ResNet-34, ResNet-50, Densenet-121, All-CNN, and ConvNeXt-Large models across three vision datasets: MNIST, CIFAR-100 and Food101, demonstrating its generalizability across different models and datasets.

Problem Setup

- Parameter space $\Theta \subseteq \mathbb{R}^m$. Pre-trained classification model denoted as f_{θ^*} with initial parameters $\theta^* \in \Theta$.
- Trained using a dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{|\mathcal{D}|}$, where $(x_i, y_i) \overset{iid}{\sim} P_{XY}(x, y)$. the label space $\mathcal{Y} = \{0, 1, 2, \ldots, C-1\}$.
- a particular class or classes of data points $s_n \in \mathcal{Y}$ that the model needs to unlearn. So unlearning samples of that specific class denoted as $\mathcal{S}_n = \{(x_i, y_i) : y_i = s_n\}$.
- The objective of unlearning is to determine a parameter θ^u for the unlearned model f_{θ^u} that closely aligns with the performance of the retrained model f_{θ^p} trained on samples $\mathcal{S}_p = \mathcal{D} \setminus \mathcal{S}_n$.
- Note that we don't have access to $\mathcal{S}_p$.

Bayesian Framework

1. Given θ^* and $\mathcal{S}_n$, the unlearning objective find $\theta^p = \arg\max_\theta P(\theta|\mathcal{S}_p)$

$$
\begin{aligned}
\theta^p &= \arg\max_\theta \log P(\theta|\mathcal{S}_p) && (1)\\
&= \arg\max_\theta \log P(\mathcal{S}_p|\theta) + \log P(\theta) - \log P(\mathcal{S}_p) && (2)\\
&= \arg\max_\theta \log P(\mathcal{S}_p|\theta) + \log P(\theta) - K_1 && (3)
\end{aligned}
$$

$$
\begin{aligned}
\log P(\theta|\mathcal{D}) &= \log P(\theta|\mathcal{S}_p, \mathcal{S}_n) && (4)\\
&= \log P(\mathcal{S}_p, \mathcal{S}_n|\theta) + \log P(\theta) - K_2 && (5)\\
&= \log P(\mathcal{S}_p|\theta) + \log P(\mathcal{S}_n|\theta) + \log P(\theta) - K_2 && (6)
\end{aligned}
$$

2. Now using the substituting the value of $\log P(\mathcal{S}_p|\theta) + \log P(\theta)$

$$
\begin{aligned}
\theta^p &= \arg\max_\theta \log P(\theta|\mathcal{D}) - \log P(\mathcal{S}_n|\theta) + K_2 - K_1 && (7)\\
&= \arg\max_\theta \mathcal{L}(\theta, \mathcal{D}, \mathcal{S}_n) && (8)
\end{aligned}
$$

Proposed Methodology

Theorem (Informal)

With some regularity conditions and $\beta, \gamma > 0$,

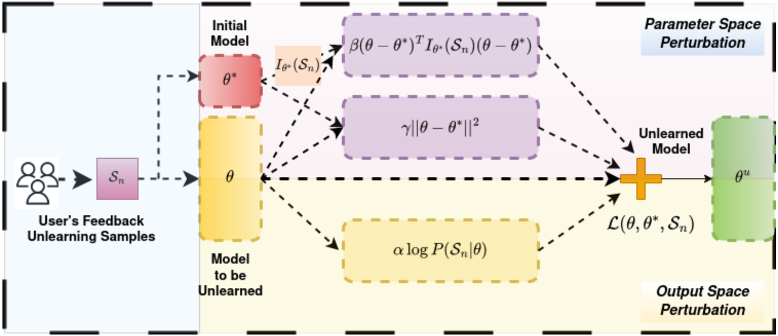
$$\mathcal{L}(\theta, \mathcal{D}, \mathcal{S}_n) \leq -\log P(\mathcal{S}_n|\theta) - \frac{\beta}{2}(\theta - \theta^*)^T I_{\theta^*}(\mathcal{S}_n)(\theta - \theta^*) - \frac{\gamma}{2}||\theta - \theta^*||^2 + K_3$$

Proposed Methodology

Theorem (Informal)

With some regularity conditions and $\beta, \gamma > 0$,

$$\mathcal{L}(\theta, \mathcal{D}, \mathcal{S}_n) \leq -\log P(\mathcal{S}_n|\theta) - \frac{\beta}{2}(\theta - \theta^*)^T I_{\theta^*}(\mathcal{S}_n)(\theta - \theta^*) - \frac{\gamma}{2}||\theta - \theta^*||^2 + K_3$$

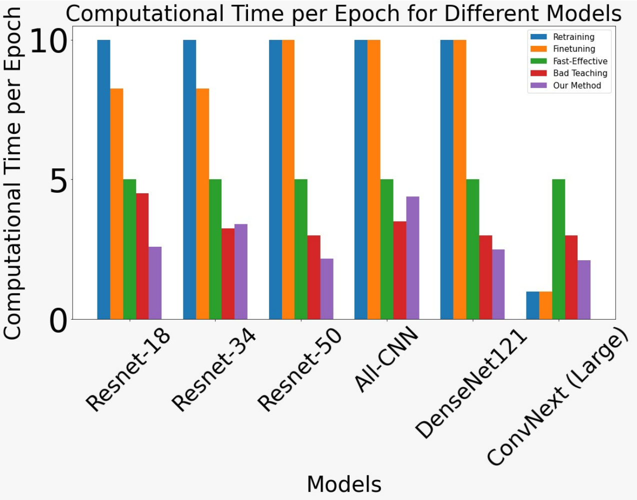


Experiments and Results

Table: accuracy on the forgotten class: $A_{D_f}(\%)$ and accuracy on the remaining classes: $A_{D_r}(\%)$

Dataset	Models	Classes	Initial Training		Re-training		Fine-tuning		Fast-Effective		Bad Teaching		Our Method(PBU)	
			A_{D_f}	A_{D_r}	A_{D_f}	A_{D_r}	A_{D_f}	A_{D_r}	A_{D_f}	A_{D_r}	A_{D_f}	A_{D_r}	A_{D_f}	A_{D_r}
MNIST	Resnet-34	Class-2	99.65±0.11	99.42±0.11	0±0	99.33±0.11	0±0	99.37±0.06	0±0	94.17±0.88	0±0	97.96±0.29	**0.03±0.06**	**98.73±0.57**
		Class-6	98.89±0.59	99.51±0.06	0±0	99.34±0.14	0±0	99.44±0.16	0±0	89.13±2.86	0±0	87.62±0.49	**0±0**	**91.09±2.9**
		Class-8	99.73±0.21	99.41±0.12	0±0	99.31±0.08	0±0	99.37±0.19	0±0	94.64±1.97	0±0	96.18±0.53	**0±0**	**98.24±0.36**
	Densenet-121	Class-2	99.6±0.12	99.44±0.11	0±0	99.15±0.05	0±0	99.37±0.12	0±0	91.43±1.62	0±0	94.15±0.54	**0±0**	**96.5±0.3**
		Class-6	99.72±0.12	99.53±0.1	0±0	99.29±0.05	0±0	99.69±0.16	0±0	96.1±0.65	0±0	97.97±0.23	**0.84±0.35**	**98.65±0.11**
		Class-8	98.95±0.63	99.58±0.11	0±0	99.47±0.09	0±0	99.53±0.08	0±0	94.5±2.34	0±0	94.83±0.59	**0.27±0.24**	**98.57±0.31**
	ConvNeXt-L	Class-2	99.69±0.21	99.55±0.1	0±0	99.11±0.1	0±0	99.19±0.12	0.18±0	97.52±0.24	0±0	97.18±1.14	**0±0**	**98.65±0.29**
		Class-6	99.05±0.05	99.59±0.1	0±0	98.83±0.17	0±0	98.63±0.01	0±0	98.25±0.13	0±0	97.75±0.51	1±0.16	**98.33±0.18**
		Class-8	99.56±0.06	99.6±0.14	0±0	98.81±0.06	0±0	98.85±0.04	0±0	97.54±0.99	0±0	96.51±1.5	1.22±1.56	**97.26±1.59**
CIFAR-100	Resnet-50	Class-1	86.33±7.23	75.87±0.31	0±0	75.2±0.34	0±0	69.98±1.24	0±0	54.61±0.21	0.87±0.23	68.06±0.14	**0±1.15**	**70.51±0.18**
		Class-3	56.67±11.72	76.17±0.41	0±0	74.12±0.93	0±0	69.25±0.36	0±0	59.43±0.56	0±0	70.35±1.19	**0.5±0.58**	**71.85±0.91**
		Class-8	92.33±2.89	75.81±0.34	0±0	74.31±0.83	0±0	68.28±0.66	0±0	57±0.1	**0±0**	**65.98±0.78**	0.5±1	65.51±2.05
	Densenet-121	Class-1	56.33±4.93	74.65±1.42	0±0	74.18±2.47	0±0	74.55±0.66	0±0	50.75±2.77	0±0	51.83±1.25	**0.5±0.58**	**63.96±1.37**
		Class-3	89.8±1.31	74.74±1.83	0±0	73.9±2.32	0±0	74.54±1.88	0±0	56.84±3.41	1.31±1.15	54.52±3.03	**0.15±0.17**	**66.38±3.65**
		Class-8	74.78±23.81	75.51±2.85	0±0	72.29±2.39	0±0	75.1±1.27	0±0	52.88±1.44	0.18±0.31	56.61±2.06	**0.4±0.46**	**64.6±3.51**
	ConvNeXt-L	Class-1	91.59±4.13	89.03±1.03	0±0	73.03±0.55	0±0	73.79±0.17	0±0	75.25±1.01	0±0	72.26±1.07	1.9±0.85	**76.82±1.19**
		Class-3	77.47±1.86	88.59±1.09	0±0	73.15±0.3	0±0	74.95±0.16	0±0	70.51±0.65	0±0	71±0.39	1±1.15	**72.51±1.12**
		Class-8	99.41±0.86	89.22±1.03	0±0	71.83±0.35	0±0	73.65±1.19	0±0	71.22±0.93	0±0	71.97±0.26	**0±0.18**	**72.64±0.23**
FOOD-101	Resnet-50	Class-10	67.2±5.54	78.18±0.01	0±0	75.1±0.23	0±0	77.3±0.77	0±0	60.83±1.4	0±0	56.07±1.08	**0.8±0.69**	**68.34±1.24**
		Class-30	90.8±4.16	77.94±0.01	0±0	74.86±0.09	0±0	77.08±0.25	0±0	62.13±0.95	0±0	52.45±0.57	**0.27±0.46**	**65.46±0.32**
		Class-50	59.6±4.85	78.26±0	0±0	74.18±0.2	0±0	77.23±0.32	0±0	63.06±0.77	0±0	55.53±0.48	**0±0**	**69.43±0.81**
	Densenet-121	Class-10	60.87±6.31	75.85±1.93	0±0	75.38±0.61	0±0	75.65±1.12	0±0	53.5±1.89	0±0	50.91±0.37	1.2±1.31	**64.97±2**
		Class-30	88.13±0.46	76.17±0.74	0±0	74.91±1.53	0±0	75.37±1.59	0±0	57.8±1.1	0.87±1.15	54.86±1.84	**0.67±0.86**	**64.75±2**
		Class-50	58.05±13.76	77.67±1.88	0±0	75.24±0.55	0±0	75.7±1.05	0±0	55.56±4.65	0.29±0.27	58.23±0.77	**0±0**	**67.9±2**
	ConvNeXt-L	Class-20	90.38±0.76	87.43±0.59	0±0	87.73±0.62	0±0	88.51±0.53	0±0	69.26±0.92	0±0	68.77±0.41	**0±0**	**75.97±0.79**
		Class-40	96.35±0.74	87.17±0.37	0±0	87.02±0.17	0±0	88.46±0.2	0±0	72.37±1.76	0±0	70.62±0.28	**0±0**	**78.45±0.79**
		Class-60	93.41±0.72	86.45±0.55	0±0	86.83±0.34	0±0	87.44±0.28	0±0	73.05±1.6	0±0	73.83±0.39	**0.74±0.5**	**81.79±0.99**

Experiments and Resutls



Conclusion and Future works

- A theoretical Bayesian Framework and a method tailored for unlearning specific classes within deep classification models.
- A key distinguishing feature of our approach is its capability to function effectively even with partial access only to the unlearning class data.
- As part of future work a slight extension of this method is now being investigated for applying unlearning in diffusion models.