Concept Forgetting via Label Annealing

Subhodip Panda, Ananda Theertha Suresh, Atri Guha, Prathosh AP

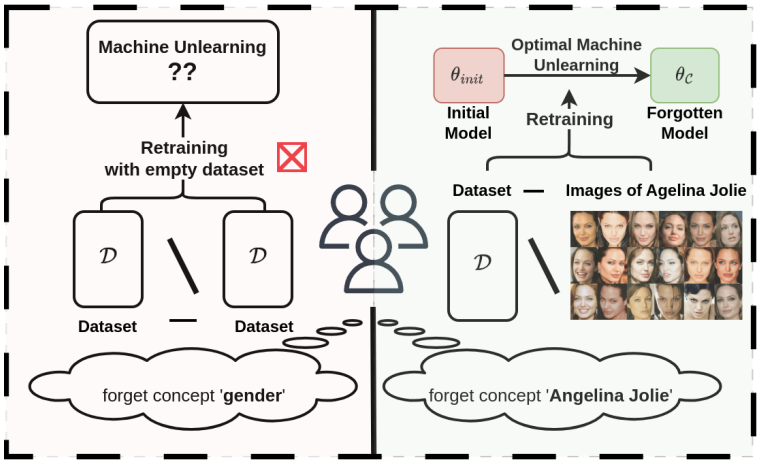
The 41^{st} Conference on Uncertainty in Artificial Intelligence
(UAI, 2025)

April 15, 2026

Table of Contents

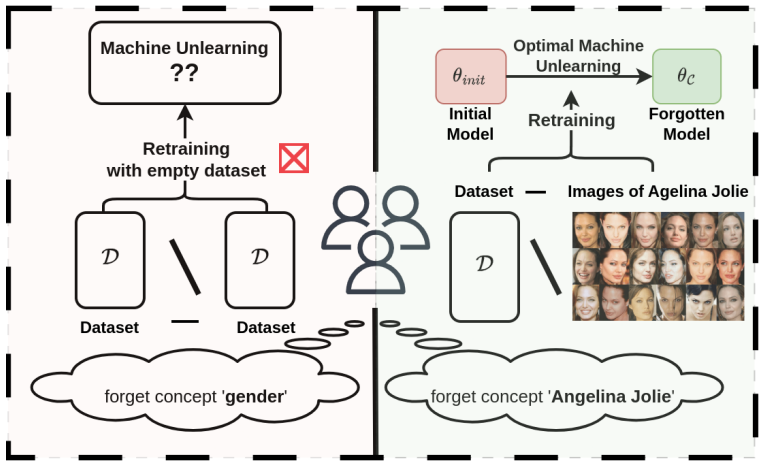
Introduction

- **Motivation**
  1. For reliable and accountable deployment, there emerges a pressing necessity to forget the undesired biased concept from the trained models.
  2. Machine Unlearning can be potentially used to forget small concepts, which are only present in a subset of examples. It can't forget a global concept that is present in the overall dataset.



Introduction

- **Motivation**
  1. For reliable and accountable deployment, there emerges a pressing necessity to forget the undesired biased concept from the trained models.
  2. Machine Unlearning can be potentially used to forget small concepts, which are only present in a subset of examples. It can't forget a global concept that is present in the overall dataset.



- **Objectives:** *Can we efficiently modify a pre-trained model to forget an undesirable concept while maintaining a low performance degradation?*

Contributions

- Motivated by the limitations of Machine Unlearning, we introduce the framework of *concept forgetting* from pre-trained classification models. We propose *concept violation* metric that empirically quantifies the extent to which the model remains dependent on a concept for its predictions.
- We propose an algorithm called **L**abel **AN**nealing (**LAN**). *LAN* employs an iterative approach, where each iteration redistributes the class labels of all data points, thus creating a pseudo-labelled dataset. This method draws an analogy to the term *annealing* frequently employed in physics.
- This strategy is computationally efficient. This method necessitates minimal epochs, sometimes as few as a single epoch. We demonstrate the efficacy of our algorithm through detailed evaluations on various image classification datasets using various image classification models.

Problem Setup

- We consider the problem of multi-class classification throughout the paper. Let $\mathcal{Y} \triangleq \{0, 1, 2, \ldots, k-1\}$ denote the set of k labels.
- Let $z = (x, y)$ denote a data point where $x \in \mathbb{R}^d$ is the feature and $y \in \mathcal{Y}$ is the label. A dataset $\mathcal{D} \triangleq \{z_i\}_{i=1}^{|\mathcal{D}|}$ is a set of samples sampled from the underlying data distribution $P_{xy} : \mathbb{R}^d \times \mathcal{Y} \to [0, 1]$.
- Let a categorical concept $\mathcal{C} : \mathbb{R}^d \times \mathcal{Y} \to \{0, 1, 2, \ldots, m-1\}$ be a mapping from the sample to the set of all possible values the concept can take.
- Let $h_\theta : \mathbb{R}^d \to \Delta^{|\mathcal{Y}|}$ denote a classifier parameterized by $\theta \in \mathbb{R}^p$ where Δ is the probability simplex. This classifier takes a feature $x \in \mathbb{R}^d$ and predicts a distribution over the label space.
- Let $\hat{h}(\theta, z)$ denote a post-processing step on the classifier (e.g, argmax).

Concept Forgetting Framework

- **Key Insight:** In the case of humans, if one forgets a concept, the forgotten concept doesn't affect one's decision-making

Definition (**Concept neutral**)

A classifier with parameter θ is concept neutral with respect to a concept , if for all output class $y \in$ and all concept values $c \in \{0, 1, \ldots, m-1\}$,

$$P_{xy}(\hat{h}(\theta, z) = y | \mathcal{C}(z) = c) = P_{xy}(\hat{h}(\theta, z) = y). \quad (1)$$

Concept Forgetting Framework

- **Key Insight:** In the case of humans, if one forgets a concept, the forgotten concept doesn't affect one's decision-making

Definition (**Concept neutral**)

A classifier with parameter θ is concept neutral with respect to a concept , if for all output class $y \in$ and all concept values $c \in \{0, 1, \ldots, m-1\}$,

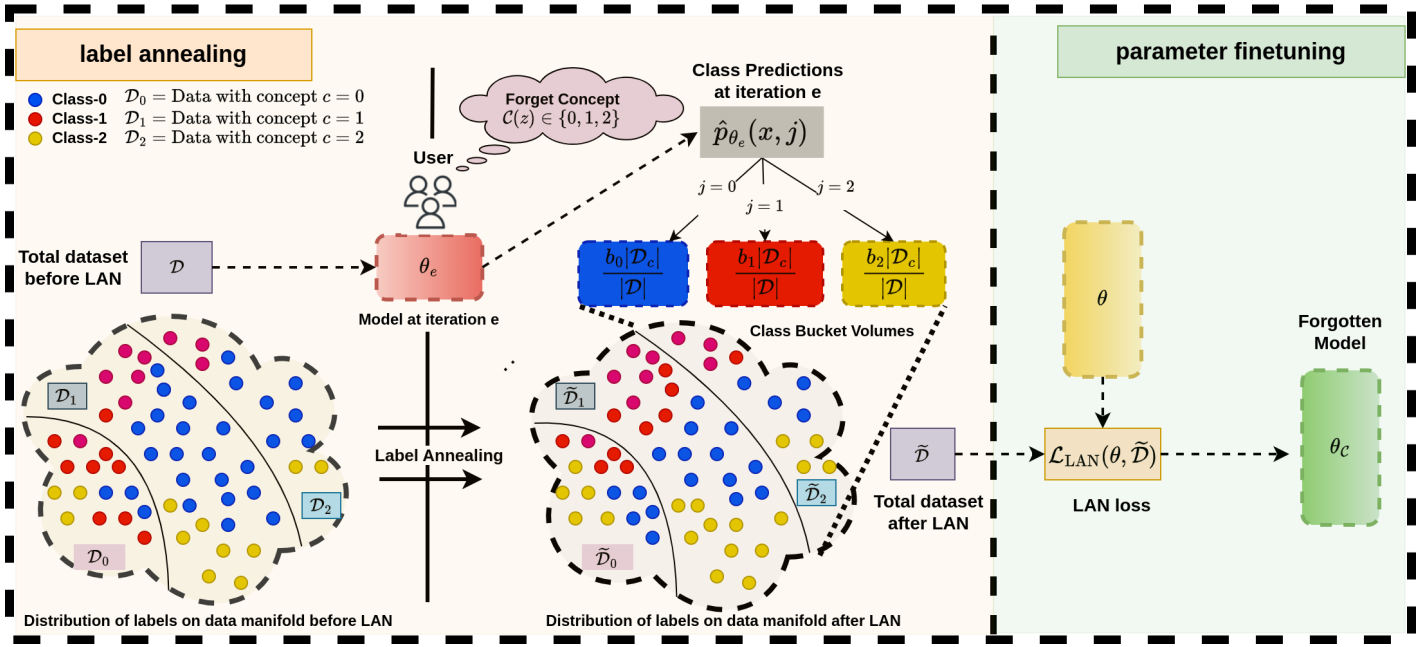
$$P_{xy}(\hat{h}(\theta, z) = y | \mathcal{C}(z) = c) = P_{xy}(\hat{h}(\theta, z) = y). \tag{1}$$

Definition (**Concept violation**)

For a classifier $\hat{h}$ with parameter θ, we measure the violation of concept neutrality in terms of the total variation distance as follows:

$$V(\theta, \mathcal{C}, P) \triangleq \frac{1}{m} \sum_{c=0}^{m-1} d_{\text{TV}}\Big(P_{xy}(\hat{h}(\theta, z) = y), P_{xy}(\hat{h}(\theta, z) = y \mid \mathcal{C}(z) = c) \Big) \tag{2}$$

LAN Algorithm

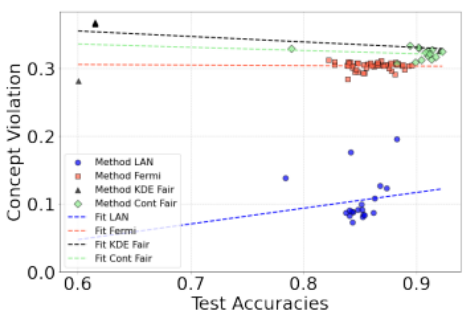


- Goal is to reduce the concept violation while minimizing accuracy loss.
- In the first stage LAN tries to create a pseudo-labeled dataset where the empirical concept violation is zero/very low and then as part of second stage, model is finetuned with this pseudo-labeled dataset.

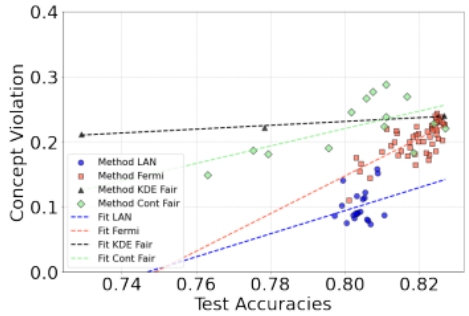
Theorem (Informal)

With assumptions of the loss, the expected loss of the forgotten model doesn't deviate much if the initial model's concept violation is minimal.

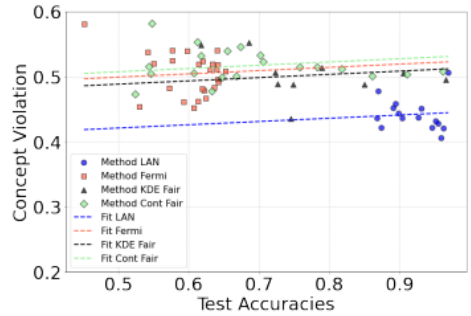
Experimental Results



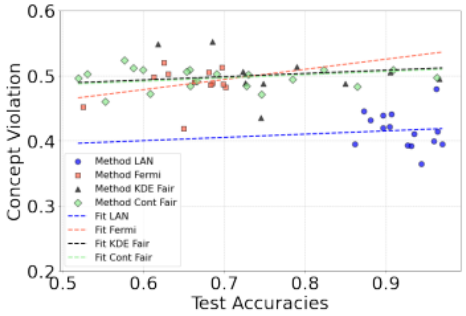
(a) (Facial hair, Resnet-50, CelebA)



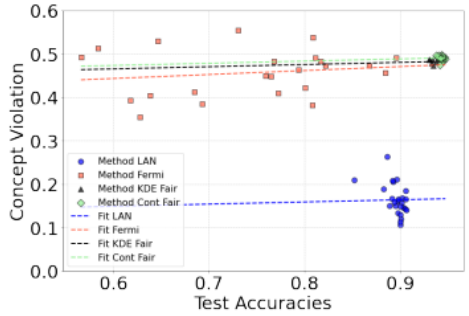
(b) (Facial hair, Resnet-50, CelebA)



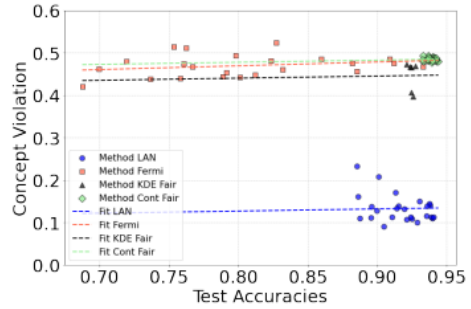
(c) (Triceratops, Resnet-50, miniImageNet)



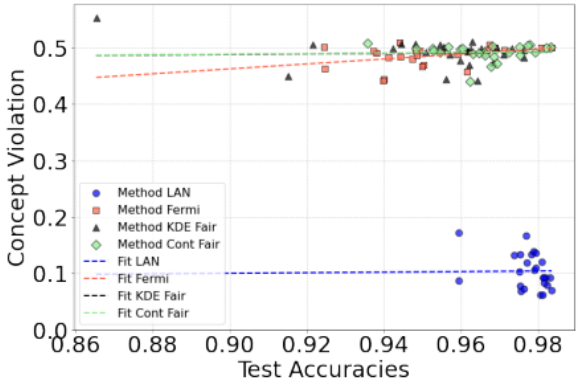
(d) (Bugs, Resnet-50, miniImageNet)



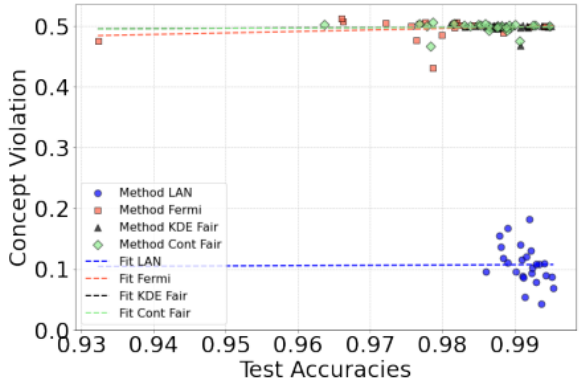
(e) (Frog, Densenet-121, CIFAR-10)



(f) (Frog, Mobinetv2, CIFAR-10)



(g) (Digit-3, MLP, MNIST)



(h) (Digit-3, Resnet-50, MNIST)

Conclusion and Future Works

- In the pursuit of safer and more responsible machine learning, we present a novel framework of concept forgetting.
- We define *concept forgetting* as the property of a model to disregard undesired concepts during its decision-making process and introduce *concept neutrality* as a necessary attribute of a forgotten model. To quantify the extent of *concept neutrality* in any model, we propose a novel metric called *concept violation*.
- We propose a computationally efficient algorithm termed as Label ANnealing (LAN) algorithm to create a forgotten model while preserving its ability to generalize.
- Our definition and method are limited only to concept forgetting in classification models. Further research is needed to develop definitions and methods for concept forgetting that generalize to generative models as well.