

# Zero-Shot Machine Unlearning

Vikram S Chundawat<sup>†</sup>, Ayush K Tarun<sup>†</sup>, Murari Mandal<sup>‡</sup>, Mohan Kankanhalli

**Abstract**—Modern privacy regulations grant citizens the right to be forgotten by products, services and companies. In case of machine learning (ML) applications, this necessitates deletion of data not only from storage archives but also from ML models. Due to an increasing need for regulatory compliance required for ML applications, *machine unlearning* is becoming an emerging research problem. The right to be forgotten requests come in the form of removal of a certain set or class of data from the already trained ML model. Practical considerations preclude retraining of the model from scratch after discarding the deleted data. The few existing studies use either the whole training data, or a subset of training data, or some metadata stored during training to update the model weights for unlearning. However, strict regulatory compliance requires time-bound deletion of data. Thus, in many cases, no data related to the training process or training samples may be accessible even for the unlearning purpose. We therefore ask the question: *is it possible to achieve unlearning with zero training samples?* In this paper, we introduce the novel problem of *zero-shot machine unlearning* that caters for the extreme but practical scenario where zero original data samples are available for use. We then propose two novel solutions for *zero-shot machine unlearning* based on (a) error minimizing-maximizing noise and (b) gated knowledge transfer. These methods remove the information of the forget data from the model while maintaining the model efficacy on the retain data. The zero-shot approach offers good protection against the model inversion attacks and membership inference attacks. We introduce a new evaluation metric, *Anamnesis Index* (AIN) to effectively measure the quality of the unlearning method. The experiments show promising results for unlearning in deep learning models on benchmark vision data-sets. The source code is available here: <https://github.com/ayu987/zero-shot-unlearning>

**Index Terms**—Machine unlearning; machine learning security and privacy; data privacy

## I. INTRODUCTION

Forgetting or deletion of data may be desired due to different reasons such as request-to-delete from the data owner, inadvertent use of wrong data, and change in the data privacy rules. The recent introduction of data privacy and protection regulations (European Union’s GDPR [1], and California Consumer Privacy Act (CCPA) [2]) obligate the companies/organizations to implement such deletion-upon-request framework. The data or anything derived from that particular data must be deleted.

<sup>†</sup>Equal Contribution. The work is part of the authors’ internship at the School of Computing, National University of Singapore.

<sup>‡</sup>Corresponding Author. Work performed while at the School of Computing, National University of Singapore.

Ayush K Tarun, Vikram S Chundawat are with Mavvex Labs, India, Faridabad 121001 (ayushtarun210@gmail.com; vikram2000b@gmail.com), Murari Mandal is with the School of Computer Engineering, Kalinga Institute of Industrial Technology (KIIT) Bhubaneswar, India 751024 (Email: murari.mandal@cs@kiit.ac.in), Mohan Kankanhalli is with the School of Computing, National University of Singapore (NUS), Singapore 117417 (Email: mohan@comp.nus.edu.sg)

Machine learning (ML) applications are built upon the algorithms that observe and learn the information from the training data. Such ML models usually memorize the training data [3], [4]. Machine unlearning refers to the task of making a trained machine learning model forget a cohort or class of data. The objective here is to forget a certain class of data without affecting the performance on the remaining class of data, i.e. preserving the accuracy of the ML model for the remaining classes. Moreover, the unlearned model must not reveal information about the forget data i.e., it must be robust to privacy attacks. Several studies [5], [6] have shown this to be a non-trivial problem. Full retraining of the ML model for every instance of such request would lead to huge computational costs. It is crucial to accomplish unlearning in a time and cost efficient manner. The class-level unlearning is useful in many application like face recognition and even healthcare applications. A Face ID is treated as a class in the face recognition model. Similarly in healthcare applications, a patient may ask for removal of her data from the already trained ML model. The user-level data may be removed with the help of an unlearning method without retraining the entire model from scratch.

Recently, machine unlearning has received considerable attention [5], [7], [8], [9], [10], [11], [12]. Most of the solutions are limited to simple linear and logistic regression. Few methods [5], [8], [9] have been proposed that can forget information from the CNN network weights with reasonable success in small to large scale vision problems. Unlearning in a CNN is quite difficult due to its vastly non-convex loss-landscape which makes it difficult to model the effect of a data sample on the optimization trajectory and the final network weights configuration. [10] proposed an efficient, multi-class unlearning algorithm that works across variety of deep networks such as the CNN and vision transformers. All these methods require *access to the retain data* [5], or the *forgetting data* [12], or sometimes both. However, *none of them can function when no original training data is available*.

**Our Contribution.** Machine learning (ML) models are frequently trained with data gathered from different sources such as publicly available datasets, data marketplaces, and self-collected/created data. In some cases, personal data such as medical records and facial images are collected with the consent of an individual. Many organizations<sup>1 2 3</sup> also offer a limited duration access to certain datasets and facilitate ML model training in the cloud. The complete access to the dataset is never desired as in most cases the company uses only

<sup>1</sup><https://cloud.google.com/products/ai>

<sup>2</sup><https://aws.amazon.com/machine-learning/>

<sup>3</sup><https://azure.microsoft.com/services/machine-learning/>

the trained model in their applications. Moreover, the data is usually chargeable for each instance of access. Therefore, multiple access to the data for unlearning purpose will be expensive. On the other hand the public datasets may not be available for reuse at later points in time and storing a copy of all the training data is expensive and impractical in many cases. All these factors suggest that machine unlearning algorithms that rely on the availability of original training data will not be helpful in many practical scenarios. The existing machine unlearning methods [5], [7], [8], [9], [10], [12] would fail to work without having access to at-least some subset of the original training data. This motivated us to ask the following question: *Can we develop machine unlearning algorithms that require no training samples or information related to the training process?* The algorithms offering solution to this question would be categorized as *zero-shot machine unlearning algorithms*.

While building a ML system based on a diverse set of data, a company also needs to comply with the data privacy rules and regulations. The data privacy regulations [1], [2] are evolving over time to include the *right to be forgotten*. They are likely to include a variety of data-usage restrictions in future to give individuals more control over their data for the sake of their privacy. For example, the CCPA [2] permits companies to collect user data without consent but compels them to remove the user's data upon a request to opt-out of data sharing. If the original data or some portion of the data is lost, then the existing machine unlearning methods will not be of any help. In such a situation the entire ML model might have to be deleted to remain compliant with the regulations. This can be quite disruptive to the company's business. Furthermore, even if the data is available, unlearning is much more efficient as compared to retraining. In this paper, we present a data-free solution to machine unlearning that addresses the unique problem of zero data availability. In case of no access to the original training samples, the proposed zero-shot unlearning algorithms become an important tool for the company. With efficient machine unlearning, companies can ensure the compliance of their ML models with the relevant privacy rules and regulations.

*How would a machine unlearning algorithm function with zero training data?* In order to manipulate the network weights of the model for unlearning, some kind of stimulus is needed. One possible approach could be the model is shown certain patterns that represent the same feature distribution as the original data samples. These patterns could be generated through generative adversarial networks (GANs) or variational autoencoders (VAEs). It is to be noted that we don't know anything about the original training data distribution. Thus, generating relevant samples would be tricky through GANs/VAEs. Another possible approach could be to build upon the unlearning algorithm presented in [10] which doesn't require access to the forget data. Similar to the error maximizing noise generated for the forget class(es) in [10], error minimizing noise can be generated for the remaining classes. These learned noise matrices could be used to update the model for unlearning. Yet another approach could be to transfer the knowledge of certain class(es) by filtering out

their information into another network in a teacher-student framework. Motivated by the aforementioned propositions, we present two different methods for achieving zero-shot machine unlearning.

In this paper, we formally define the problem of zero-shot unlearning and make the following key contributions:

- We introduce the problem setting for zero-shot machine unlearning that imposes the constraint that *zero training data* is available to the unlearning algorithms. This resembles the real-world scenarios more closely for data deletion request in ML models as compared to the existing state-of-the-art machine unlearning settings.
- We propose two novel methods to enable data-free unlearning. These methods are based on (i) error-maximizing and error-minimizing noise and (ii) gated knowledge transfer in a teacher-student learning framework.
- We introduce a new metric *Anamnesis Index*, to assess the quality of unlearning more effectively. We also evaluate our unlearned model against the privacy attacks such as model inversion attack and membership inference attacks.
- The proposed methods are independent of the prior information (such as the optimization technique) related to the original model training. We obtain strong performance on a variety of deep networks on benchmark vision datasets (MNIST, SVHN, CIFAR-10).

## II. RELATED WORK

### A. Machine Unlearning

The concept of unlearning was introduced in [13] to eliminate the effect of data point(s) on the already trained model. Machine unlearning can be broadly grouped into two categories: *exact unlearning and approximate unlearning methods*. In *exact unlearning*, the impact of the data point (to forget) is removed from the model by retraining the model from scratch by excluding that data from the training set. This is clearly a computationally expensive process as the deep learning models are trained with large datasets. Moreover the request for data deletion is a recurrent event rather than a one-time occurrence. Several research efforts have focused on finding efficient ways of retraining. [6] proposed a SISA approach to train the model by partitioning the dataset into a set of non-overlapping shards. This reduces the need for full retraining as the model can be retrained on one of the shards. The *approximate unlearning* methods aim to approximate the parameters that would have been obtained if the model was trained without using the data to be unlearned. Usually the approach is to bring the parameters closer to a model (trained from scratch) that has never seen the unlearning data through a lower number of updates in comparison to exact unlearning methods. [14] store the updates made by each data point in the parameter space while training. During unlearning, the corresponding updates are subtracted from the final parameters of the model. [15] proposed random forests that support data forgetting. [16], [17] studied data deletion in k-means clustering algorithms. Similarly, unlearning for Bayesian methods are presented in [18]. [12] give a certified information removal framework

based on Newton’s update removal mechanism for convex learning problems. [19] studied several unlearning methods for linear models and analyzed the efficiency, effectiveness and certifiability trade-offs among them. Similarly, [20] analyze the difference between machine unlearning and differential privacy. All these methods are designed for convex problems, whereas we aim to present an unlearning solution for deep learning models.

### B. Deep Machine Unlearning

Golatkar et al. [5] proposed an information theoretic method to scrub the information from intermediate layers of deep networks trained with SGD. [8] extended this work to update the final activations of the model. They present a neural tangent kernel (NTK) based approximation of the training process and use it to estimate the updated network weights after unlearning. However, the approximation accuracy decreases and computational cost increases very quickly even for small datasets. [9] directly train a linearized network and use it for unlearning. They train two separate networks: the core model, and a mixed-linear model for unlearning purpose. However, designing a mixed-linear network for every deep architecture is an inefficient approach and requires manual intervention depending on the network structure of the original model. In our zero-shot unlearning methods, we do not train any additional network. In fact, we do not require any prior information related to the training process or access to the training data. Tarun et al. [10] proposed an error-maximization based method to learn noise matrix for the class to forget. They use such anti-samples to induce class-level forgetting in deep networks. The method requires a small subset of retain data for unlearning. In this paper we work in a setting where strictly no information about the training data is available.

### C. Unlearning Settings and Data Privacy

The *approximate unlearning* methods aim to offer higher efficiency by making certain assumptions about the availability of training information i.e., training data, optimization techniques used, and relaxing the effectiveness of the model to some degree. Most of the existing unlearning methods can be provisionally categorized into three groups. The *first group* [12], [21] use the data to forget to update the ML model for unlearning. In these methods, the influence of the forget data on the model is approximated with a Newton step and a random noise is injected to the training objective function. The *second group* [5], [8], [9] requires access to the rest of the training data (excluding the forget data). These methods use the Fisher information and inject optimal noise to the model weights to achieve unlearning. They do however impose certain restrictions on the training process such as only SGD optimization is allowed while training. [10] alleviated some of these restrictions while presenting a much faster unlearning algorithm. The *third group* [14], [22], [7], [23] stores different kind of information during the training process in order to utilize it for data deletion in future. Usually, these methods attempt to approximate the SGD steps of a model retrained from scratch. To this end, the intermediate elements

such as the loss gradients, weight updates, etc. generated at each SGD step is stored. The amount of information that needs to be stored raise the concern of excessive memory overheads requiring a trade-off between memory overhead vs computational efficiency.

All the above mentioned methods require access to either the data to unlearn, the remaining data, or some metadata stored during training. These assumptions do not reflect the real-world unlearning scenarios as discussed in Section I. With increasing data privacy concerns [1], [2] in different parts of the world, more stringent privacy regulations are expected. Thus, a data-free approach to unlearning is required. The possible information leakage in the existing machine unlearning methods as discussed in [24] also motivated us to look towards a completely data-free approach to unlearning. We propose to work in the zero-shot setting where the proposed unlearning method does not require access to any kind of information that are essential for the functioning of existing state-of-the-art unlearning methods. A data-free approach makes this a more realistic setting for unlearning and also helps in complying with a broader set of data privacy regulations.

### D. Privacy Attacks on Machine Learning Models

Numerous privacy attacks against ML model have shown information leakage of the training data [25], [26]. Membership inference attack aims to extract the information about the presence of certain data points the training set [27], [28]. Shokri et al. [25] proposed the idea of using shadow models to imitate the behaviour of the target model to generate training samples. Thereafter, several membership inference attacks have been proposed [29]. To safeguard a ML model from the inference attacks, considerable defense methods have been proposed based on reducing the model overfitting [30], perturbation of the posteriors [31], and adversarial training [32]. Ganju et al. [33] introduced a subcategory of membership inference attack called property inference attack. It attempts to infer the general properties of the training data. For example, the ratio between the samples in each class is a property of the training data. Model inversion attacks [34], [35] try to recreate instances of records of a class from the target ML model. Similarly, the adversarial examples [36], [37] are used to infer certain attributes of the data. In this paper, the membership inference attack and model inversion attack are used to check the possible privacy leaks in the unlearning models.

## III. PRELIMINARIES

Let  $\mathcal{D} = \{x_i, y_i\}_{i=1}^N$  be a dataset consisting of  $x_i$  samples associated with label  $y_i \in \{1, 2, \dots, K\}$ , each representing a class.  $C_f$  denotes the set of classes we wish the machine learning (ML) model to forget and  $\mathcal{D}_f$  is the set of data corresponding to the forget classes. Similarly,  $C_r$  denotes the set of retain classes that we want the model to remember and  $\mathcal{D}_r$  is the data corresponding to the retain classes. A general machine unlearning method receives a query to forget  $\mathcal{D}_f$  samples. The information about  $\mathcal{D}_f$  data points are to be removed from the model. The usual assumption is that access to both  $\mathcal{D}_r$  and  $\mathcal{D}_f$  is available to the unlearning method. The

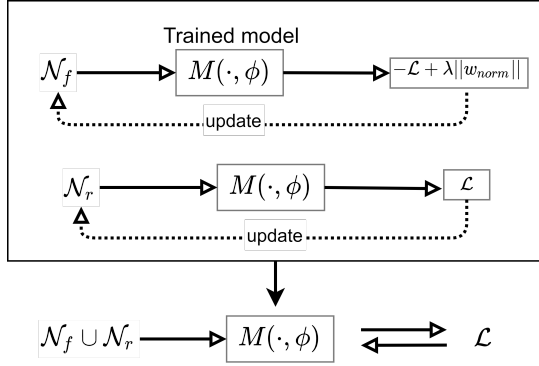


Fig. 1: The baseline error minimization-maximization noise based method for zero-shot unlearning.

unlearning may be performed at class-level or sample-level. In class-level unlearning, the entire class of data is forgotten. In sample-level unlearning, a particular set of data points either belonging to single or multiple classes are forgotten.

The complete dataset  $\mathcal{D}$  consists of  $\mathcal{D}_f$  and  $\mathcal{D}_r$  i.e.,  $\mathcal{D}_r \cup \mathcal{D}_f = \mathcal{D}$  where  $\mathcal{D}_f$  and  $\mathcal{D}_r$  are mutually exclusive i.e.,  $\mathcal{D}_r \cap \mathcal{D}_f = \phi$ . The model trained from scratch without observing the forget samples ( $\mathcal{D}_r$ ) is denoted as *retrained model* or *gold model* in this paper. In the literature, the retrained model is commonly used as a benchmark to evaluate the performance of an unlearning method. The Kullback-Leibler (KL) divergence [38] is used as a measure of similarity between two probability distributions. Given two distributions  $p(x)$  and  $q(x)$ , their KL-divergence is defined by

$$KL(p(x)||q(x)) := \mathbb{E}_{x \sim p(x)}[\log(p(x)/q(x))] \quad (1)$$

where  $KL(p(x)||q(x))$  is always positive.

#### IV. ZERO-SHOT MACHINE UNLEARNING

In this section we first formalize the *zero-shot unlearning problem*. We then present two novel approaches for *zero shot unlearning*. We also present a new evaluation metric to more effectively assess the degree of unlearning achieved in a model.

##### A. The Zero-Shot Machine Unlearning Problem

Let the machine learning (ML) model be represented by  $M(x; \theta)$  that gives out the classification probabilities for each class, where  $x$  is the input and  $\theta$  is the set of model parameters. For an ML algorithm  $A$  trained on the complete dataset  $\mathcal{D}$ , the model parameter set is obtained as

$$\theta = A(\mathcal{D}) \quad (2)$$

If  $M_r(x; \theta_r)$  is a model that has observed only the data corresponding to retain classes, then the parameters corresponding to  $M_r$  can be expressed as

$$\theta_r = A(\mathcal{D}_r) \quad (3)$$

A method  $\mathcal{U}$  is called a *zero-shot unlearning* method if it doesn't require access to either  $\mathcal{D}_r$  or  $\mathcal{D}_f$ . It requires access only to the parameters ( $\theta$ ) of the trained model which is always available. Based on the query set of forget classes ( $\mathcal{C}_f$ ) (not

the data corresponding to these classes), a *zero-shot unlearning* method can produce the unlearned model  $M_u$  with parameters  $\theta_u$  such that it is behaviourally similar to a *retrained model*  $M_r$  as given in Eq. 4 and Eq. 5.

$$M(x; \theta) \xrightarrow{\mathcal{U}} M_u(x; \theta_u) \quad (4)$$

$$M_u(x; \theta_u) \approx M_r(x; \theta_r) \quad (5)$$

The output of the model  $M_u$  are expected to be similar to that of  $M_r$ , as if  $M_u$  has never seen the data corresponding to forget set  $\mathcal{D}_f$ . Note that the parameter set  $\theta_u$  is not expected to be exactly the same as  $\theta_r$  due to many degrees of freedom of parameters. Two models consisting of different distribution of parameters could show similar behaviour. An analysis on this phenomena as presented in [10]. It is shown that the *retrained model*  $M_r$ , may not be the only *gold standard*. A very different set of parameters could exhibit behaviour very similar to that of  $M_r$ . Therefore, in practice, we compare the black-box metrics of the unlearned model with the retrained model for robustness analysis. The rest of the analysis is done through privacy attacks.

##### B. Error Minimization-Maximization Noise

We build upon the work presented in [10] that utilizes an error-maximization based noise generation procedure. A set of noise matrices is learned for the forget class of data using the original model. These matrices are then used to fine-tune the fully trained model in order to unlearn the information corresponding to the forget classes ( $\mathcal{C}_f$ ). The generated matrices act as a set of anti-samples to the model for the forget class(es). Therefore, after observing them, it forgets about that class of data. A subset of the data corresponding to retain classes ( $\mathcal{C}_r$ ) are also shown to the model to preserve the information for the retain classes. Although, the forget class of data is not needed but it requires access to the retain data.

We create a novel extension of this approach in a data-free setting. The data corresponding to the  $\mathcal{C}_r$  is generated by an error-minimizing noise. The error-minimized noise should act as a proxy for the samples belonging to  $\mathcal{C}_r$ . Since, the original samples from  $\mathcal{C}_r$  class would also have minimum possible model loss. Thus, learning noise matrices that minimize the model loss is a natural way to synthesize such data points. As the retain data  $\mathcal{D}_r$  is not available for updating the model in a data-free setting, we create the retain data samples using the model itself through error-minimization. An error-minimizing noise  $\mathcal{N}^{(i)}$  of the same size as that of the model input will be generated for every retain class  $i$ . The following loss function is used to learn the noise matrix.

$$L_N^{(i)}(\mathcal{N}_r^{(i)}) = \mathcal{L}(M(\mathcal{N}_r^{(i)}; \phi), i) \quad (6)$$

where  $\mathcal{L}$  is the classification loss. The noise is generated and updated for all the retain classes. The error-maximizing noise is learned in a manner similar to that in [10]. We do not use the repair step in our experiments, as it did not have its intended effect of improving the accuracy on the retain set in the data-free setting. Also, we do not restrict ourselves to only



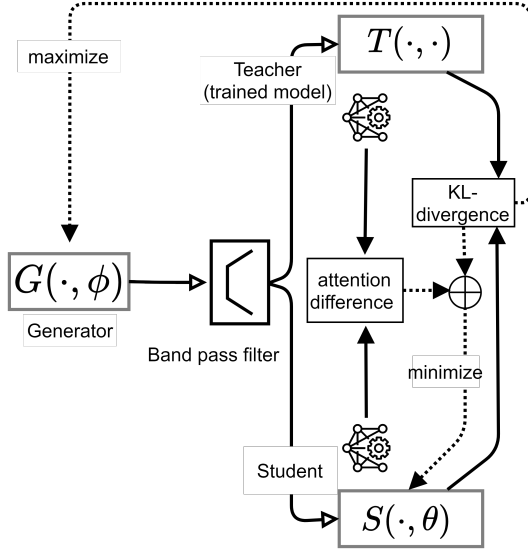


Fig. 2: The proposed gated knowledge transfer method for zero-shot unlearning.

a single shot of impair and use multiple iterations to obtain the best performance possible. The overall method is depicted in Figure 1. Furthermore, a step-wise procedure can be found in the supplementary material.

### C. Gated Knowledge Transfer

The sub-optimal quality of the learned noise generated through the error maximization-minimization method leads to substandard level of unlearning. We therefore propose a knowledge distillation based algorithm to obtain better performance for the data-free setup. Inspired by [39], we use a generator to produce data points that maximize the KL-divergence between the teacher and student models. However, instead of trying to obtain a smaller model, we develop a zero-shot mechanism for machine unlearning through such knowledge distillation. To this end, we assign the originally trained model as the teacher and a randomly initialized network (of same structure as the original model) as the student. We introduce a crucial design element, a type of "gate" we call *band-pass filter* which does not allow the information related to the forget class  $C_f$  to flow towards the student from the teacher model. This ensures that the student learns only the information of the retain class(es)  $C_r$  and not the forget class. The proposed method is shown in Figure 2.

Let  $T(x)$  be a teacher which outputs the probability vector  $t$  for input  $x$ . The student  $S(x; \theta)$  is a randomly initialized model with parameters  $\theta$ , output probabilities  $s$ , and has the same architecture as  $T(x)$ . The generator  $G(z; \phi)$  produces pseudo data points  $x_p$  for a noise vector  $z \in N(0, I)$ . Let the corresponding teacher's prediction on  $x_p$  be  $t_p$  and student's prediction on  $x_p$  be  $s_p$ . The generator maximizes  $D_{KL}(T(x_p)||S(x_p))$  i.e., Kullback-Leibler(KL) divergence between output probabilities of the teacher and student model as shown in Eq. 7.

$$D_{KL}(T(x_p)||S(x_p)) = \sum_i t_p^{(i)} \log(t_p^{(i)} / s_p^{(i)}) \quad (7)$$

where  $i$  corresponds to the data classes. The data points  $x_p$  are generated by the generator which is optimized to maximise the KL-Divergence between the teacher and student. The student will update the weights to minimize the KL-divergence along with an attention loss as given in Eq. 8.

$$L_{at} = \sum_{l \in N_L} \left\| \frac{f(A_l^{(t)})}{\|f(A_l^{(t)})\|_2} - \frac{f(A_l^{(s)})}{\|f(A_l^{(s)})\|_2} \right\| \quad (8)$$

where  $A_l^{(t)}$  and  $A_l^{(s)}$  denote outputs of layer  $l$  for teacher and student respectively, both of which consist of  $N_{A_l}$  channels. Similar to [39], we use a subset  $N_L$  for calculating the attention loss. We use  $f(A_l) = (1/N_{A_l}) \sum_c a_{lc}^2$  where  $a_{lc}$  denotes the  $c_{th}$  channel of activation block  $A_l$ . The student then minimizes  $L_s$  as in Eq. 9.

$$L_s = D_{KL}(T(x_p)||S(x_p)) + \beta L_{at} \quad (9)$$

where  $\beta$  is a hyperparameter. The generator and student are updated alternatively.

1) **Band-Pass Filter.** We propose a band-pass filter which would attenuate the information regarding the forget class(es)  $C_f$  and only let information regarding the retain classes  $C_r$  to reach the student. The intuition is that as the samples aren't from a real world distribution, so the only indicator of how much information the generated pseudo samples have regarding the forget class(es) is the predicted probabilities by the teacher,  $t_p^i$ , where  $i \in C_f$ . To restrict the generator from passing on information regarding the forget class(es), we place a filter  $F$  ahead of the generator which receives all the generated pseudo samples and filters them out before passing them to the student. The filter criterion for each sample is

$$F(x_p) = \prod_{i \in C_f} (t_p^{(i)} < \epsilon) \quad (10)$$

where  $\epsilon$  is a hyperparameter and  $i$  represents each forget class. The filter creates a boolean mask and a sample passes the filter only if the predicted probability corresponding to each forget class i.e.  $t_p^i$  is less than a threshold  $\epsilon$  for all  $i \in C_f$ . Let the total number all of classes be denoted by  $c$ , then a randomly initialized model will on an average output a probability of  $1/c$  for each of the forget class(es). Therefore, in order to effectively filter out the information regarding the forget class(es), a value of  $\epsilon$  equal to or lower than  $1/c$  would be required, but if  $\epsilon$  is too small, no pseudo sample might pass the filter. Thus, an appropriate value of  $\epsilon$  is crucial for better training. The step-wise procedure for the Gated Knowledge Transfer (GKT) method is presented in the supplementary material.

### D. Anamnesis Index

The existing methods ([10] [5] [8]) use *relearn time* as a metric to measure the quality of unlearning. It is used to represent the amount of information left in the models after unlearning by showing how quickly the model can relearn. However, the scale and properties of this metric could vary according to the model and the dataset. In our experiments, we noticed that sometimes the unlearned models regain significant

TABLE I: Unlearning results in AllCNN model. **Original Model** denotes the model trained on the complete dataset. **Retrain Model** denotes the model trained from scratch only on  $\mathcal{D}_r$ . **M-M (Min-Max)** denotes the model generated by error minimization-error maximization technique. **GKT** denotes the gated knowledge transfer method. **AIN** denotes the Anamnesis Index. #  $\mathcal{Y}_f$  is the number of unlearning classes.

Dataset	# $\mathcal{Y}_f$	Acc.	Original Model	Retrain Model	M-M Method	GKT Method	AIN [GKT]	AIN [M-M]
CIFAR10	1	$\mathcal{D}_r \uparrow$	84.05	85.72	20.48	<b>81.97</b>	<b>0.81</b>	0.11
		$\mathcal{D}_f \downarrow$	87.49	0	5.11	<b>0</b>		
	2	$\mathcal{D}_r \uparrow$	84.18	86.30	29.59	<b>81.70</b>	<b>0.74</b>	0.10
		$\mathcal{D}_f \downarrow$	84.72	0	7.59	<b>0</b>		
SVHN	1	$\mathcal{D}_r \uparrow$	94.52	93.02	72.92	<b>92.43</b>	<b>0.37</b>	0.15
		$\mathcal{D}_f \downarrow$	95.16	0	42.32	<b>0</b>		
	2	$\mathcal{D}_r \uparrow$	93.61	95.10	58.70	<b>92.13</b>	<b>0.74</b>	0.13
		$\mathcal{D}_f \downarrow$	96.39	0	50.23	<b>0</b>		
MNIST	1	$\mathcal{D}_r \uparrow$	97.84	99.25	10.57	<b>97.12</b>	<b>0.65</b>	0.31
		$\mathcal{D}_f \downarrow$	99.61	0	0.0	<b>0</b>		
	2	$\mathcal{D}_r \uparrow$	98.17	99.41	10.96	<b>96.87</b>	<b>0.30</b>	0.20
		$\mathcal{D}_f \downarrow$	98.77	0	0.0	<b>0</b>		

accuracy in a very few number of relearn steps, but do not converge to the original accuracy on the forget class(es) for a large number of steps. Thus, simply calculating the retrain time (epochs) in which the original accuracy is reached or surpassed could be misleading. To address some of these issues, we use a margin of  $\alpha\%$  around the original accuracy to calculate the relearn time. Let the number of mini-batches (steps) required by a model  $M$  to come within  $\alpha\%$  range of the accuracy of the original model  $M_{orig}$  on the forget classes be  $r_t(M, M_{orig}, \alpha)$ . If  $M_u$  and  $M_s$  denote the unlearned model and the model trained from scratch on  $\mathcal{D}_r$  respectively, then Anamnesis Index (AIN) is defined as

$$AIN = \frac{r_t(M_u, M_{orig}, \alpha)}{r_t(M_s, M_{orig}, \alpha)} \quad (11)$$

The value of AIN ranges from 0 to  $\infty$ . The closer the AIN value is to 1, the better is the unlearning. AIN values much lower than 1 correspond to the instances when information of the forget classes is still present in the model. It also indicates the unlearned model quickly relearns to make accurate predictions. This may be because there were only a few changes in the final layers that deteriorated the model's performance on forget class(es) but are easily reversible. If AIN is much higher than 1, this may indicate that the method leads to changes in parameters which are so significant that the unlearning itself is detectable (Streisand effect). This may be because the model was thrown into a convergence hyperplane far-off from its initial position and is thus not able to regain earlier learnt information regarding the forget class(es).

1) **Setting the value of  $\alpha$ :** The suggested value of parameter  $\alpha$  is 5-10%. The  $\alpha$  is a measure of the accepted margin of difference in performance for retrain time calculation of the models. A high value of  $\alpha$  means a very high margin, and could lead to a misleading score, always near to 1. A much lower value of  $\alpha$  can result in unstable results.

## V. EXPERIMENTS

We show the performance of the proposed methods for unlearning single and multiple classes in the zero-shot setting across multiple benchmark datasets. We use AllCNN [40], LeNet [41] and ResNet9 (we build a smaller variant of

ResNet [42]) models for zero-shot machine unlearning analysis on MNIST [41], CIFAR-10 [43] and SVHN [44].

### A. Experimental Settings

The experiments are conducted on a NVIDIA Tesla-V100 (32GB) GPU. The original models and retrained models have been trained with a batch size of 256 for 40 epochs on CIFAR-10, 10 epochs on SVHN, and 10 epochs on MNIST. Without loss of generality, we unlearn class 0 in 1-class and classes 1 and 2 in 2-classes unlearning.

**Implementation details for error min-max noise method.** We optimize i.e., minimize or maximize a batch of noise (256 samples) for 400 steps with an initial learning rate of 0.1 for each class. If the loss doesn't improve, the learning rate is decreased by a factor of 0.5. On CIFAR 10, we use 2 impair steps [10] with a learning rate of 0.01. On SVHN, we use 3 impair steps with a learning rate of 0.001. On MNIST we use a single impair step with a learning rate of 0.01.

**Implementation details for gated knowledge transfer (GKT) method.** The generator and student are trained alternatively with 1 step for generator, and 10 steps for student. The KL temperature for loss function was set to 1 for MNIST and SVHN, and 0.5 for CIFAR-10.  $\beta$  is set to 250. The learning rate for both student and generator for MNIST is set to 0.01. The learning rate for both student and generator for other datasets is set to 0.001. The training is performed for 4000 epochs and the threshold for filter  $F$  is set to 0.01. We show the results of the model checkpoint with best performance on  $\mathcal{D}_r$  while still having 0% accuracy on  $\mathcal{D}_f$ .

### B. Evaluation Metrics

In the literature [10], [5], [9], [14], different metrics are employed to evaluate the unlearning methods. Usually, such metrics require both the training and testing data to measure the performance. For our zero-shot methods, the evaluation is done with the help of the training and testing sets even though such data is not used for the purpose of unlearning. To evaluate the robustness of the unlearning method, we use the following metrics:

*Accuracy on the forget set  $\mathcal{D}_f$  and the retain set  $\mathcal{D}_r$*  - The accuracy on  $\mathcal{D}_f$  is desired to be closer to the retrained model

TABLE II: Single class unlearning on LeNet and ResNet9

Dataset	Model	Acc.	Original Model	Retrain Model	M-M Method	GKT Method	$AIN$ [GKT]	$AIN$ [M-M]
CIFAR10	LeNet	$\mathcal{D}_r \uparrow$	59.80	62.93	55.32	<b>41.32</b>	<b>211</b>	24
		$\mathcal{D}_f \downarrow$	65.25	0	23.98	<b>0</b>		
	ResNet9	$\mathcal{D}_r \uparrow$	84.83	85.61	10.85	<b>56.83</b>	<b>212</b>	12
		$\mathcal{D}_f \downarrow$	88.50	0	0	<b>0</b>		
SVHN	LeNet	$\mathcal{D}_r \uparrow$	85.69	88.31	81.80	<b>78.27</b>	<b>1</b>	1
		$\mathcal{D}_f \downarrow$	81.42	0	89.73	<b>0</b>		
	ResNet9	$\mathcal{D}_r \uparrow$	82.76	94.24	53.75	<b>39.44</b>	<b>143</b>	2
		$\mathcal{D}_f \downarrow$	87.11	0	49.65	<b>0</b>		
MNIST	LeNet	$\mathcal{D}_r \uparrow$	98.15	98.73	96.96	<b>95.79</b>	<b>1</b>	1
		$\mathcal{D}_f \downarrow$	99.59	0	99.37	<b>0</b>		
	ResNet9	$\mathcal{D}_r \uparrow$	98.57	98.83	12.32	<b>94.57</b>	<b>2</b>	3
		$\mathcal{D}_f \downarrow$	99.10	0	0	<b>0</b>		

TABLE III: A comparison between proposed zero-shot unlearning method with the existing unlearning methods.

Attributes	UNSIR [10]	Amnesiac [14]	Bad Teacher [45]	Fisher [5]	<b>Our Method</b>
Class-level unlearning?	✓	✓	✓	✓	✓
Random samples unlearning?	✗	✓	✓	✓	✗
Zero-glance unlearning?	✓	✗	✗	✗	✓
Zero training samples?	✗	✗	✗	✗	✓

TABLE IV: Unlearning result comparison with recent state-of-the-art methods [45], [14], [10] (Not zero-shot). The comparisons are done for AllCNN+CIFAR-10. Our method performs quite well considering it has no access to the training dataset.

Method	Zero-shot?	1-class Unlearning		2-class Unlearning	
		$\mathcal{D}_r \uparrow$	$\mathcal{D}_f \downarrow$	$\mathcal{D}_r \uparrow$	$\mathcal{D}_f \downarrow$
Original Model	NA	84.05	87.49	84.18	84.72
Retrain Method	NO	85.72	0	86.30	0
Bad Teacher [45]	NO	83.71	5.56	84.11	7.81
Fisher [5]	NO	7.61	0	8.57	0
Amnesiac [14]	NO	83.47	0	82.85	0
Min-Max (ours)	YES	20.48	5.11	29.59	7.59
GKT (ours)	YES	<b>81.97</b>	<b>0</b>	<b>81.70</b>	<b>0</b>

as the intended behaviour of an unlearned model should be similar to that of the retrained model after unlearning. The accuracy on  $\mathcal{D}_r$  is desired to be closer to the original model.

*Anamnesis Index (AIN):* As discussed in Section IV-D, the AIN should be close to 1. A margin ( $\alpha$ ) of 5%, with a batch size of 256 was used for relearn time calculation. We use a learning rate of 0.1 and decrease the rate by a factor of 0.1 if accuracy does not improve on 2 successive epochs.

*Model inversion attack:* The data obtained after inversion from the unlearned model should not contain any information about the forget class.

*Membership inference attack:* The inference attack probability should be lower in the unlearned model in comparison to the original model for the data in the forget class.

### C. Results and Analysis

We conduct several experiments and evaluate the single and multi-class unlearning performance in *zero-shot setting* for the proposed baseline (Min-Max) method and gated knowledge transfer (GKT) methods. In Table I, we show the results of AllCNN model over CIFAR-10, SVHN, and MNIST. Similarly, Table II gives the results for single-class unlearning in LeNet and ResNet9.

1) *Baseline results.:* From Table I and Table II, we observe that the baseline Min-Max method gives poor results in zero-

shot unlearning setup. As this method requires optimizing the noise for both the forget and retain classes, the quality of noise samples determine the unlearning performance. The Min-Max method is not able to preserve the accuracy on  $\mathcal{D}_r$ . For example, in 1-class unlearning on AllCNN+CIFAR-10 (Table I), it obtains 20.58% accuracy on  $\mathcal{D}_r$  in comparison to the 84.05% accuracy of the original model. Similarly, the accuracy on  $\mathcal{D}_f$  is 5.11%. In 2-class unlearning on AllCNN+MNIST, only 10.96% retain accuracy is obtained in comparison to the 98.17% accuracy in the original model. Even when it obtains decent accuracy on  $\mathcal{D}_r$  (Table I, SVHN, 1-class unlearning: 58.70%), the accuracy on  $\mathcal{D}_f$  is also high (50.23%) which is not desirable. The low Anamnesis index (AIN) for this method indicates that the model quickly regains most of the information of  $\mathcal{D}_f$ . It means the unlearned model still remembers a lot of the  $\mathcal{D}_f$  information. Similarly, the Min-Max method gives poor results for LeNet and ResNet9 as shown in Table II.

2) *Results on GKT Vs Min-Max:* The proposed GKT method achieves accuracy close to the original model on  $\mathcal{D}_r$ . For example, in 1-class unlearning on AllCNN+CIFAR-10 (refer Table I), GKT obtains 81.97% accuracy on  $\mathcal{D}_r$  which is quite close to the original models accuracy of 84.05%. In 2-class unlearning on AllCNN+MNIST, the proposed GKT achieves 96.87% accuracy on  $\mathcal{D}_r$  in comparison to 98.17%

accuracy of the original model. In both cases, the accuracy on  $\mathcal{D}_f$  is 0% which quite good. The AIN score of the proposed GKT method in CIFAR-10+AllCNN and MNIST+AllCNN is 0.81 and 0.65, respectively. In both cases the AIN score is much closer to 1 in comparison to Min-Max method. This indicates that the unlearned model doesn't contain much information about the forget classes.

In Table II, the performance on LeNet and ResNet9 models are shown for 1-class unlearning. The proposed GKT method obtains 0% accuracy on  $\mathcal{D}_f$  across all the datasets. Furthermore,  $\mathcal{D}_r$  accuracy is much closer to the original model in comparison to the Min-Max method. Although Min-max is consistent in terms of  $\mathcal{D}_r$  and  $\mathcal{D}_f$ , it obtains significantly poor results in both of them. An ideal unlearning method should maintain the previous accuracy on  $\mathcal{D}_r$  and give near zero accuracy on  $\mathcal{D}_f$ . The noise matrices generated through error-maximization (for forget class samples) and error-minimization (for retain class samples) in Min-Max ends up damaging the model for both the retain and forget data. Whereas, the GKT method is able to carefully refine the model weights through a teacher-student framework and a band-pass filter. Thus, overall, GKT gives better performance over Min-Max.

It is to be noted that the accuracy on  $\mathcal{D}_r$  is not fully preserved by our methods. This is due to the zero-shot setting where the model can't observe any real data. In future, further improvements should be explored to obtain same  $\mathcal{D}_r$  accuracy as the original model in zero-shot setting. We notice some values of AIN in Table II are anomalous. This is because the retrained model reached the target accuracy in just 1 epoch and therefore, the ratios are very high. The results on LeNet and ResNet9 are decent but not as consistent as AllCNN.

3) **Comparison with the state-of-the-art unlearning methods:** One major difference between our work and the existing state-of-the-art unlearning methods is that our method is zero-shot i.e., it does not require any access to training data. Whereas, the existing methods [45], [14], [5], [10] need the training set data to perform unlearning. We show a comparison between different unlearning methods in Table III based on several attributes such as *whether the method supports unlearning for class-level and a random set of samples*. Similarly, we compare what kind of data is required for the corresponding algorithm to function. Ours is the only method that can function without using any training data. We then compare our *zero-shot* results with the *non zero-shot* results in Table IV. Our method lags behind [14] by 1.5% in terms of  $\mathcal{D}_r$  but it forgets  $\mathcal{D}_f$  with equal effectiveness as [14]. The *non zero-shot* methods give better results on the retain set but this is not a fair comparison because *zero-shot* method can function even if training data is not available for unlearning. Whereas, the other methods can not function in the absence of training data.

4) **Analysis:** In Fig. 4, we show the progression in *knowledge transfer* through the forget  $\mathcal{D}_f$  and retain set  $\mathcal{D}_r$  accuracy with increasing number of epochs. The GKT method uses a generator that maximizes the information gap between the teacher and student. Thus, it generates highly informative samples for the student to learn from. We use the band-pass

filter to reject the samples generated corresponding to the forget set to be passed on to the student. After several iterations of training, most of the information gap corresponding to  $\mathcal{D}_r$  will be bridged and we will be left with the information gap corresponding to  $\mathcal{D}_f$  only. On the other hand, due to the adversarial setting, most of the generated samples will correspond to  $\mathcal{D}_f$ . Now the information gap lies in  $\mathcal{D}_f$  samples only as other information has been already absorbed by the student. This means that the performance on  $\mathcal{D}_r$  will stagnate and that on  $\mathcal{D}_f$  will begin to rise. This behaviour can be observed in Fig. 4. For class-0 forgetting in AllCNN+MNIST, we see a sharp rise around epoch 250. If the performance tolerance for  $\mathcal{D}_f$  is 0%, this is where we stop i.e., as soon as  $\mathcal{D}_f$  accuracy becomes  $> 0\%$ . In LeNet model unlearning,  $\mathcal{D}_f$  accuracy stays at 0%. This means further training is possible and more performance on  $\mathcal{D}_r$  can be recovered without affecting the performance on  $\mathcal{D}_f$ . In general, one should select the best checkpoint before the accuracy on  $\mathcal{D}_f$  starts to rise. The checkpoints before that event contain information corresponding to  $\mathcal{D}_r$  only and they should be picked as final parameters of an unlearned model.

#### D. Privacy Attacks: Model Inversion Attack

We check the data leakage in the the proposed methods using the state-of-the-art privacy attacks such as model inversion and membership inference attacks. We discuss the model inversion attack and robustness analysis below and the analysis for membership inference attack is given in the Supplementary document.

1) **Threat Model:** In our analysis, we assume the adversary has white-box access to the current unlearned model but does not have access to the previous versions of the model. The adversary does not hold the information about the actual contents of each class. An attack is considered successful if the adversary is able to gather information about the representation of a class through model inversion. We use the model inversion attack presented in [14] which is a modified version of the attack in [34]. A small amount of noise is added to the input vector initialized with zeros. The model is optimized with gradient descent using the loss computed with respect to this input and the target class. An image processing step is executed after every  $n$  steps of gradient descent which helps in the recognition of the generated images. The whole process is repeated for some number of epochs to obtain the final inverted images.

2) **Robustness Analysis:** The model inversion attacks were performed on the AllCNN model trained on MNIST. Without loss of generality, the attacks were performed before and after forgetting class-0 from the model. Fig. 3 shows the inversion attack results on fully trained model, model trained without class-0 i.e., *retrained model*, and forget model obtained using the proposed GKT method. The inversion attacks on fully trained model shows circle like patterns among inverted images hinting at possible data leakage. However, the patterns are completely random for gold model and the proposed forget model. The model inversion attacks are not able to extract any information from our forget model. This indicates that our method is successful at removing information regarding class-0 and is robust to the model inversion attacks.



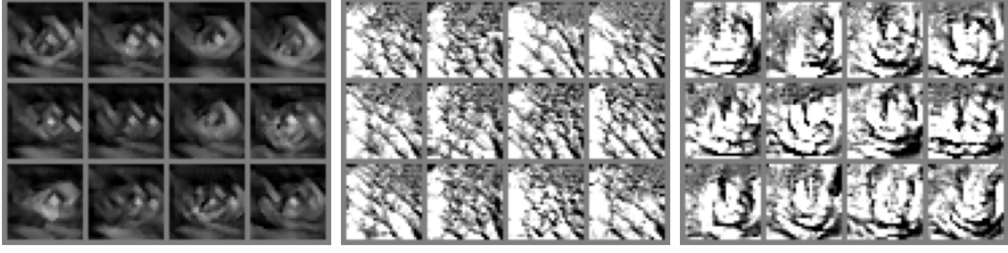


Fig. 3: Model inversion attack on AllCNN model for class-0. *left*: Fully trained model *middle*: Model retrained without class-0 *right*: Forget model obtained using the proposed GKT method. The images are shown for 12 independent attacks. The inversion attack fails in the proposed GKT method and the retrained model.

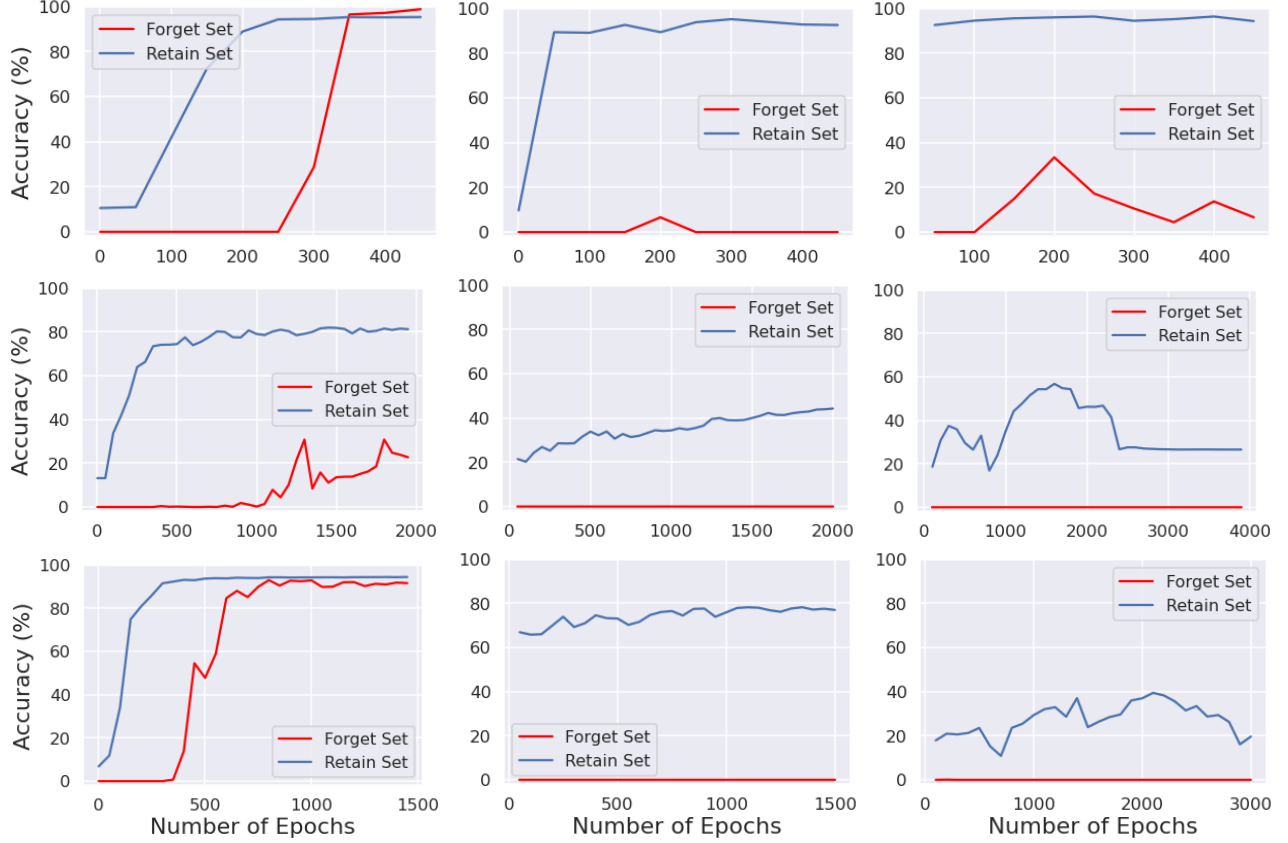


Fig. 4: The GKT method progression with number of epochs in 1-class unlearning on MNIST (top), CIFAR-10 (middle), and SVHN (bottom). The models used are AllCNN (left), LeNet (middle), and ResNet9 (right). The graphs show how stopping after the right number of epochs is important. (Table I and Table II present the results just before the accuracy on  $\mathcal{D}_f$  shoots up.)

### E. Ablation Study

We conduct several ablative experiments to further analyze the proposed method. Particularly, several variations are carried out on the GKT method to observe the change in its effectiveness. We discuss the results of the following ablation studies.

1) **Band-Pass filter ablation:** The effect of varying the filter’s threshold ( $\epsilon$ ) on the quality of knowledge transfer in GKT method is depicted in Fig. 5. The threshold values of 0.001 and 0.01 work quite well in our experiments. The default value of 0.01 converges faster. Using a value of 0.1 makes

the generator crash around epoch 1750. It means after this none of the samples produced by the generator pass the filter and the student doesn’t train anymore. The threshold of 0.5 leads to the worst outcome. It is neither stable nor does it reach anywhere near to the target performance throughout its training. This is expected as even a randomly initialized model will give an average probability of 0.1 (as there are 10 classes) corresponding to the forget class and any threshold above this shouldn’t work.

2) **Loss functions:** We make the following variations to the loss function: (i) *add the attention difference to the generator’s*



Fig. 5: Effect of varying the threshold in band pass filter. From top-left to bottom-right, the thresholds are 0.5, 0.1, 0.01, and 0.001, respectively. As visible, thresholds more than random probability ( $1/10 = 0.1$ ) are ineffective at stopping information flow corresponding to  $D_f$ .

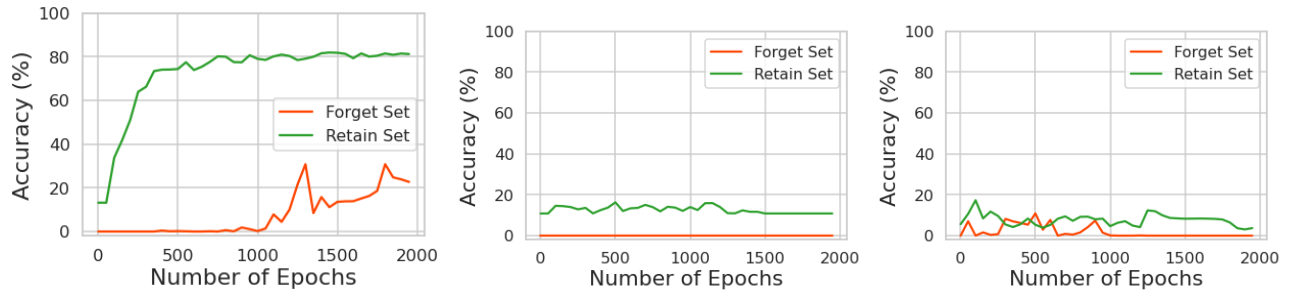


Fig. 6: Ablation study on the loss function of the proposed GKT method. *Left*: Default experiment settings. *Middle*: When generator has an attention difference term in its loss. *Right*: When JS-divergence is used instead of KL-divergence in both student and generator loss.

loss, (ii) replace *KL-divergence* in the training procedure with *Jenson-Shannon (JS)-divergence*. The results with these 2 losses are shown in Fig. 8. Adding attention difference to the generator’s loss makes it too easy for the generator to fool the student. Therefore the student doesn’t learn anything and same can be observed in the middle graph in Fig. 8. Similarly, the JS-divergence also doesn’t have much impact on the learning capacity of the student. It instead decreases the effectiveness of GKT compared to the default KL-Divergence.

#### F. Limitations

To the best of our knowledge, this paper presents the first solution for zero shot unlearning. However, there are some limitations such as the proposed method is not as effective in very large models. To demonstrate the effectiveness of data removal from the deep learning model, a variety of widely accepted metrics have been evaluated like performance on  $D_r$ ,  $D_f$ , relearn time, and privacy attacks, including a newly proposed

AIN metric. A more formal guarantee of unlearning might be desired in highly privacy-sensitive applications. Currently our method supports class-level unlearning. Forgetting a random cohort of data, or a subset of a class is beyond the scope of this work. For example, if there was a class for vehicles, forgetting just a specific vehicle like a car of a particular company can be considered as a subclass removal problem and that is not supported by our method.

#### VI. CONCLUSION

We introduced the problem of *zero-shot machine unlearning* that opens up new challenges for unlearning in a stricter setting. It also offers a framework to keep up with the modern privacy regulations that include the provision for *right to be forgotten*. We present two novel *zero-shot* approaches to this problem: error minimization-maximization noise and gated knowledge transfer. The gated knowledge transfer approach proves to be highly effective in case of both single and

multiple-class unlearning. Several ablative studies are presented to gain insight into the working principles of the proposed method. We introduce a new evaluation metric, Anamnesis Index (AIN), to measure the quality of unlearning in the ML models. The experiments are conducted on benchmark datasets and the results are quite promising in zero-shot setting. Protection against information leak is evaluated under privacy attacks such as model inversion attack and membership inference attack. Our work gives a new direction of research in a more stringent, but realistic setting for machine unlearning. A possible direction of future research can be to try zero-shot unlearning on deeper networks and large datasets. Another future direction could explore obtaining a theoretical bound on the amount of information remaining in the model after forgetting.

#### ACKNOWLEDGMENT

This research/project is supported by the National Research Foundation, Singapore under its Strategic Capability Research Centres Funding Initiative. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore.

#### REFERENCES

- [1] P. Voigt and A. Von dem Bussche, "The eu general data protection regulation (gdpr)," *A Practical Guide, 1st Ed., Cham: Springer International Publishing*, 2017. [1](#), [2](#), [3](#)
- [2] E. Goldman, "An introduction to the california consumer privacy act (ccpa)," *Santa Clara Univ. Legal Studies Research Paper*, 2020. [1](#), [2](#), [3](#)
- [3] V. Feldman, "Does learning require memorization? a short tale about a long tail," in *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, 2020, pp. 954–959. [1](#)
- [4] N. Carlini, C. Liu, Ú. Erlingsson, J. Kos, and D. Song, "The secret sharer: Evaluating and testing unintended memorization in neural networks," in *28th USENIX Security Symposium (USENIX Security 19)*, 2019, pp. 267–284. [1](#)
- [5] A. Gohilkar, A. Achille, and S. Soatto, "Eternal sunshine of the spotless net: Selective forgetting in deep networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9304–9312. [1](#), [2](#), [3](#), [5](#), [6](#), [7](#), [8](#)
- [6] L. Bourtole, V. Chandrasekaran, C. A. Choquette-Choo, H. Jia, A. Travers, B. Zhang, D. Lie, and N. Papernot, "Machine unlearning," in *2021 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2021, pp. 141–159. [1](#), [2](#)
- [7] Y. Wu, E. Dobriban, and S. Davidson, "Deltagrad: Rapid retraining of machine learning models," in *International Conference on Machine Learning*. PMLR, 2020, pp. 10355–10366. [1](#), [2](#), [3](#)
- [8] A. Gohilkar, A. Achille, and S. Soatto, "Forgetting outside the box: Scrubbing deep networks of information accessible from input-output observations," in *European Conference on Computer Vision*. Springer, 2020, pp. 383–398. [1](#), [2](#), [3](#), [5](#), [16](#)
- [9] A. Gohilkar, A. Achille, A. Ravichandran, M. Polito, and S. Soatto, "Mixed-privacy forgetting in deep networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 792–801. [1](#), [2](#), [3](#), [6](#)
- [10] A. K. Tarun, V. S. Chundawat, M. Mandal, and M. Kankanhalli, "Fast yet effective machine unlearning," *arXiv preprint arXiv:2111.08947*, 2021. [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#), [8](#)
- [11] A. K. Tarun, V. S. Chundawat, M. Mandal, and M. Kankanhalli, "Deep regression unlearning," *arXiv preprint arXiv:2210.08196*, 2022. [1](#)
- [12] C. Guo, T. Goldstein, A. Hannun, and L. Van Der Maaten, "Certified data removal from machine learning models," in *International Conference on Machine Learning*. PMLR, 2020, pp. 3832–3842. [1](#), [2](#), [3](#)
- [13] Y. Cao and J. Yang, "Towards making systems forget with machine unlearning," in *2015 IEEE Symposium on Security and Privacy*. IEEE, 2015, pp. 463–480. [2](#)
- [14] L. Graves, V. Nagisetty, and V. Ganesh, "Amnesiac machine learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 13, 2021, pp. 11516–11524. [2](#), [3](#), [6](#), [7](#), [8](#)
- [15] J. Brophy and D. Lowd, "Machine unlearning for random forests," in *International Conference on Machine Learning*. PMLR, 2021, pp. 1092–1104. [2](#)
- [16] A. Ginart, M. Y. Guan, G. Valiant, and J. Zou, "Making ai forget you: Data deletion in machine learning," in *Advances in neural information processing systems*, 2019, pp. 3513–3526. [2](#)
- [17] B. Mirzasoleiman, A. Karbasi, and A. Krause, "Deletion-robust submodular maximization: Data summarization with "the right to be forgotten"," in *International Conference on Machine Learning*. PMLR, 2017, pp. 2449–2458. [2](#)
- [18] Q. P. Nguyen, B. K. H. Low, and P. Jaillet, "Variational bayesian unlearning," *Advances in Neural Information Processing Systems*, vol. 33, 2020. [2](#)
- [19] A. Mahadevan and M. Mathioudakis, "Certifiable machine unlearning for linear models," *arXiv preprint arXiv:2106.15093*, 2021. [3](#)
- [20] A. Sekhari, J. Acharya, G. Kamath, and A. T. Suresh, "Remember what you want to forget: Algorithms for machine unlearning," *Advances in Neural Information Processing Systems*, vol. 34, 2021. [3](#)
- [21] Z. Izzo, M. A. Smart, K. Chaudhuri, and J. Zou, "Approximate data deletion from machine learning models," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2021, pp. 2008–2016. [3](#)
- [22] S. Neel, A. Roth, and S. Sharifi-Malvajerdi, "Descent-to-delete: Gradient-based methods for machine unlearning," in *Algorithmic Learning Theory*. PMLR, 2021, pp. 931–962. [3](#)
- [23] Y. Wu, V. Tannen, and S. B. Davidson, "Priu: A provenance-based approach for incrementally updating regression models," in *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*, 2020, pp. 447–462. [3](#)
- [24] M. Chen, Z. Zhang, T. Wang, M. Backes, M. Humbert, and Y. Zhang, "When machine unlearning jeopardizes privacy," in *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, 2021, pp. 896–911. [3](#)
- [25] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," in *2017 IEEE symposium on security and privacy (SP)*. IEEE, 2017, pp. 3–18. [3](#)
- [26] A. Salem, Y. Zhang, M. Humbert, M. Fritz, and M. Backes, "ML-leaks: Model and data independent membership inference attacks and defenses on machine learning models," in *Network and Distributed Systems Security Symposium 2019*. Internet Society, 2019. [3](#)
- [27] S. Yeom, I. Giacomelli, M. Fredrikson, and S. Jha, "Privacy risk in machine learning: Analyzing the connection to overfitting," in *2018 IEEE 31st computer security foundations symposium (CSF)*. IEEE, 2018, pp. 268–282. [3](#)
- [28] C. Dwork, A. Smith, T. Steinke, J. Ullman, and S. Vadhan, "Robust traceability from trace amounts," in *2015 IEEE 56th Annual Symposium on Foundations of Computer Science*. IEEE, 2015, pp. 650–669. [3](#)
- [29] M. Nasr, R. Shokri, and A. Houmansadr, "Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning," in *2019 IEEE symposium on security and privacy (SP)*. IEEE, 2019, pp. 739–753. [3](#)
- [30] J. Li, N. Li, and B. Ribeiro, "Membership inference attacks and defenses in classification models," ser. CODASPY '21, 2021, p. 5–16. [3](#)
- [31] J. Jia, A. Salem, M. Backes, Y. Zhang, and N. Z. Gong, "Memguard: Defending against black-box membership inference attacks via adversarial examples," in *Proceedings of the 2019 ACM SIGSAC conference on computer and communications security*, 2019, pp. 259–274. [3](#)
- [32] M. Nasr, R. Shokri, and A. Houmansadr, "Machine learning with membership privacy using adversarial regularization," in *Proceedings of the 2018 ACM SIGSAC conference on computer and communications security*, 2018, pp. 634–646. [3](#)
- [33] K. Ganju, Q. Wang, W. Yang, C. A. Gunter, and N. Borisov, "Property inference attacks on fully connected neural networks using permutation invariant representations," in *Proceedings of the 2018 ACM SIGSAC conference on computer and communications security*, 2018, pp. 619–633. [3](#)
- [34] M. Fredrikson, S. Jha, and T. Ristenpart, "Model inversion attacks that exploit confidence information and basic countermeasures," in *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, 2015, pp. 1322–1333. [3](#), [8](#)
- [35] B. Hitaj, G. Ateniese, and F. Perez-Cruz, "Deep models under the gan: information leakage from collaborative deep learning," in *Proceedings of the 2017 ACM SIGSAC conference on computer and communications security*, 2017, pp. 603–618. [3](#)

- [36] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, and A. Swami, "Practical black-box attacks against machine learning," in *Proceedings of the 2017 ACM on Asia conference on computer and communications security*, 2017, pp. 506–519. 3
- [37] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami, "The limitations of deep learning in adversarial settings," in *2016 IEEE European symposium on security and privacy (EuroS&P)*. IEEE, 2016, pp. 372–387. 3
- [38] S. Kullback and R. A. Leibler, "On information and sufficiency," *The annals of mathematical statistics*, vol. 22, no. 1, pp. 79–86, 1951. 4
- [39] P. Micaelli and A. J. Storkey, "Zero-shot knowledge transfer via adversarial belief matching," *Advances in Neural Information Processing Systems*, vol. 32, pp. 9551–9561, 2019. 5
- [40] J. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, "Striving for simplicity: The all convolutional net," 2015. 6
- [41] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998. 6
- [42] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778. 6
- [43] A. Krizhevsky, G. Hinton *et al.*, "Learning multiple layers of features from tiny images," 2009. 6
- [44] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, and A. Y. Ng, "Reading digits in natural images with unsupervised feature learning," 2011. 6
- [45] V. S. Chundawat, A. K. Tarun, M. Mandal, and M. Kankanhalli, "Can bad teaching induce forgetting? unlearning in deep networks using an incompetent teacher," in *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence*, 2023. 7, 8
- [46] S. Rezaei and X. Liu, "On the difficulty of membership inference attacks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 7892–7900. 16



## APPENDIX

**Algorithm 1** Noise Minimization-Maximization

---

```

1:  $M(\cdot; \phi)$  (Fully Trained Model)
2:  $C_f \leftarrow$  forget classes
3:  $C_r \leftarrow$  retain classes
4:  $Noises \leftarrow []$ 
5: for  $i \in C_f$  do
6:   Training Noise for forget classes
7:    $\mathcal{N}_f \leftarrow N(0, I)$  (Initialize Noise)
8:   for  $1, 2, \dots, n_{C_f}$  do
9:      $L_N \leftarrow -\mathcal{L}(M(z; \phi), i) + \lambda \|z\|_2$ 
10:     $\mathcal{N}_f \leftarrow \mathcal{N}_f - \eta \frac{\partial L_N}{\partial \mathcal{N}_f}$ 
11:   end for
12:    $Noises.append(\mathcal{N}_f, i)$ 
13: end for
14: for  $i \in C_r$  do (Training Noise for retain classes)
15:    $\mathcal{N}_r \leftarrow N(0, I)$  (Initialize Noise)
16:   for  $1, 2, \dots, n_{C_r}$  do
17:      $L_N \leftarrow \mathcal{L}(M(z; \phi), i)$ 
18:      $\mathcal{N}_r \leftarrow \mathcal{N}_r - \eta \frac{\partial L_N}{\partial \mathcal{N}_r}$ 
19:   end for
20:    $Noises.append(\mathcal{N}_r, i)$ 
21: end for
22:  $dataset \leftarrow Noises$ 
23:  $shuffle(dataset)$ 
24: for  $1, 2, \dots, n_{epochs}$  do
25:   for batch in dataset do ▷ Perform the unlearning step.
26:      $input, labels \leftarrow$  batch
27:      $predictions \leftarrow M(input; \phi)$ 
28:      $L_M \leftarrow \mathcal{L}(predictions, labels)$ 
29:      $\phi \leftarrow \phi - \eta \frac{\partial L_M}{\partial \phi}$ 
30:   end for
31: end for

```

---

**A. Algorithms**

Algorithm 1 describes the Noise Minimization-Maximization method proposed in Section IV.B. Similarly, Algorithm 2 describes the Gated Knowledge Transfer method proposed in Section IV.C. The generator architectures used in the GKT method is given in Table V.

**B. Additional Ablation Study**

1) *Stopping Criteria:* In our experiments, we have used the performance on the retain set and forget set to evaluate the progress of GKT and also to decide when to stop the training. However, in a truly *absolute zero-shot* scenario we will not even have access to these forget and retain sets even for evaluation purposes. In such cases, methods like Deep Inversion<sup>4</sup> can be used to obtain synthetic samples which may act as a proxy to the real data. Figure 7 shows how the performance on the real retain set, synthetic retain set, real forget set, and synthetic forget set progress with increasing number of steps of distillation. From Figure 7, it is clearly visible that the performance on the synthetic set is highly indicative of how the performance might have been on the real data. The performance on the synthetic retain set is nearly identical to that of the performance on the real retain set. Similarly, the performance on the synthetic forget set shoots up at nearly the same epoch as that on the real forget set. This means that the performance on the synthetic forget set can be used to decide when to stop even in an *absolute zero-shot setting*.

<sup>4</sup>Yin et al., “Dreaming to distill: Data-free knowledge transfer via deepinversion,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 8715–8724

**Algorithm 2** Gated Knowledge Transfer

---

```

1:  $T(\cdot)$  (Fully Trained Teacher Network)
2:  $S(\cdot; \theta)$  (Randomly Initialized Student)
3:  $G(\cdot; \phi)$  (Randomly Initialized Generator)
4:  $C_f \leftarrow$  forget classes
5:  $threshold \leftarrow \epsilon$ 
6: for  $1, 2, \dots, n_{epochs}$  do
7:    $z \leftarrow N(0, I)$ 
8:   for  $1, 2, \dots, n_G$  do
9:      $x_p \leftarrow G(z; \phi)$ 
10:     $x_p \leftarrow F(x_p, C_f, threshold)$  (Applying Filter)
11:     $L_G \leftarrow -D_{KL}(T(x_p) || S(x_p))$ 
12:     $\phi \leftarrow \phi - \eta \frac{\partial L_G}{\partial \phi}$ 
13:   end for
14:   for  $1, 2, \dots, n_S$  do
15:      $x_p \leftarrow G(z; \phi)$ 
16:      $x_p \leftarrow F(x_p, f_{cls}, threshold)$  (Applying Filter)
17:      $L_S \leftarrow D_{KL}(T(x_p) || S(x_p)) + \beta L_{at}$ 
18:      $\theta \leftarrow \theta - \eta \frac{\partial L_S}{\partial \theta}$ 
19:   end for
20: end for

```

---

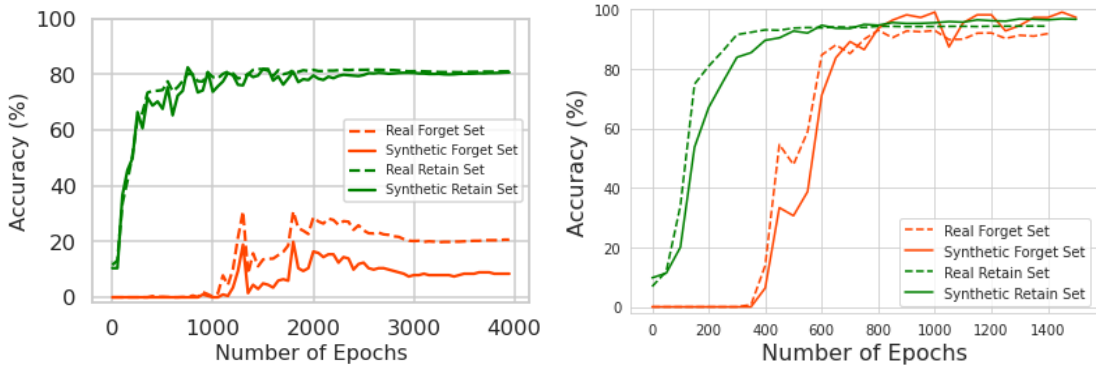


Fig. 7: GKT progress comparison between real and synthetic data. The synthetic test images were generated using DeepInversion. *left*: The progression of unlearning (class-0) on synthetic CIFAR-10 test data and real CIFAR-10 test data on AllCNN. *right*: The progression of unlearning (class-0) on synthetic SVHN test data and real SVHN test data on AllCNN. The progression on synthetic data is highly correlated and similar to progression on real data and it can be used to decide stopping criteria in complete absence of test data.

2) *Adding Attention Difference to the Loss*: Figure 8 shows the results when we *add the attention difference to the generator’s loss* and when we replace KL-divergence in the training procedure with *JS (Jenson-Shannon)-divergence*. Adding attention difference to the generator’s loss makes it too easy for the generator to fool the student, and the student doesn’t learn. Also, if JS-divergence is used, the method doesn’t have its intended effect and the student again doesn’t learn.

### C. Individually Unlearning Every Class

In addition to 1-class and 2-class unlearning presented in the main paper, we also show unlearning for each individual class (class 0 to 9) on AllCNN+CIFAR10. We then compare the results with the retrained model as given in Table VI. Our zero-shot method gives same performance as the retrain model in the forget set (0%). However, the difference in the performance on the retain set varies from negligible (0.09%, class 4) to substantial (10.53%, class 3 and 10.63%, class 5). The difficulty in unlearning individual class is different. This is due to the different disentanglement of each class representation in the model. The more disentangled the representation is of a class, the more retain accuracy can be maintained. Similarly, more entangled a class’s representation is, the more difficult to obtain good retain accuracy, without affecting the forget set accuracy. The average deviation for the retain set is 5.22%.

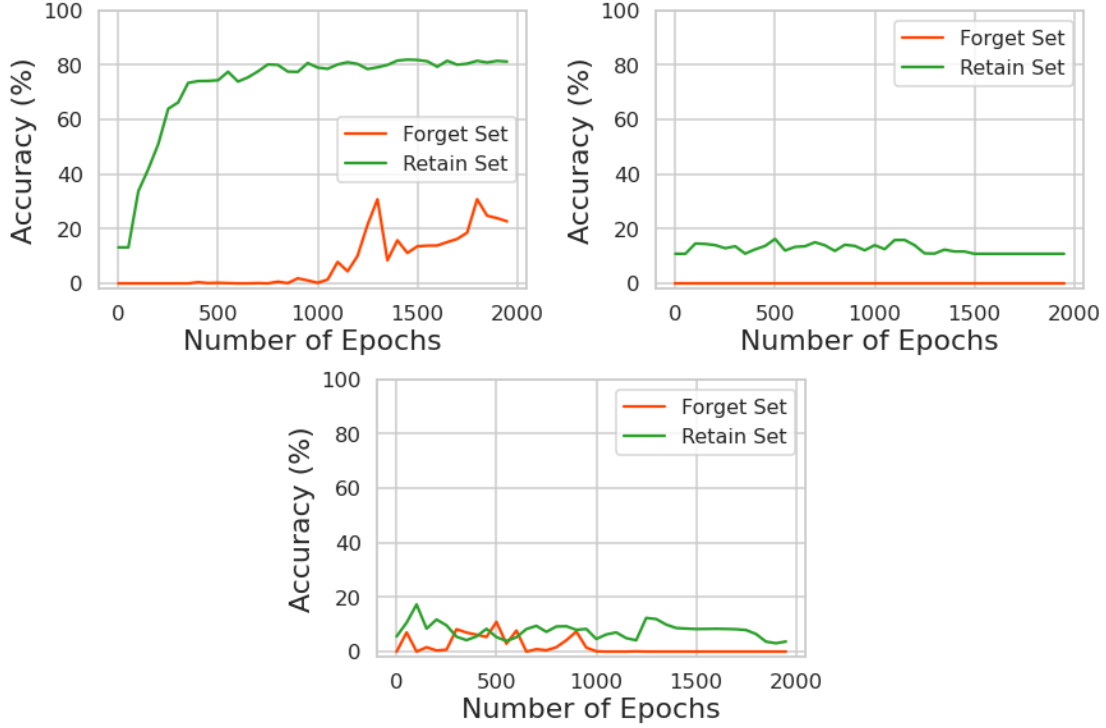


Fig. 8: Ablation study on the proposed GKT method. *Top*: Default experiment settings. *Middle*: When generator also has a attention difference term in its loss. *Bottom*: When JS-divergence is used instead of KL-divergence in both student and generator loss.

TABLE V: The architecture detail of the generator used in GKT method

Layer	Dimensions/Parameters
Linear	(z_dim, 128*64)
Reshape	(128, 8, 8)
BatchNormalization	-
Upsample	factor=2
Convolution	(128, 128, 3)
BatchNormalization	-
LeakyReLU	negative_slope = 0.2
Upsample	factor=2
Convolution	(128, 64, 3)
BatchNormalization	-
LeakyReLU	negative_slope = 0.2
Convolution 2D	(64, 3, 3)
BatchNormalization	-

#### D. Sequential Unlearning

In real world scenario, the deletion or forgetting requests may come at different instances. Therefore, it is important to validate the sequential unlearning performance of a method. Fig. 9 shows the effect of sequentially unlearning 7 classes one after another on CIFAR-10+AllCNN. The model first unlearns class-0. This model is used to unlearn class-2 and the subsequent model then unlearns class-3, class-4, and so on. We notice that after 4<sup>th</sup> sequential class unlearning, the performance on both retain and forget set begin to marginally dwindle. Overall, this degradation is gradual. For a 10-class model, after the 6<sup>th</sup>/7<sup>th</sup> class unlearning, it is better to retrain the model rather than unlearning more classes. The trade-off between the cost of

TABLE VI: GKT method performance after every individual class unlearning in AllCNN+CIFAR10.

Method	Accuracy	Unlearning Class									
		0	1	2	3	4	5	6	7	8	9
GKT	$\mathcal{D}_f \downarrow$	0	0	0	0	0	0	0	0	0	0
	$\mathcal{D}_r \uparrow$	81.97	77.48	79.94	77.63	84.44	76.91	77.92	82.11	83.01	82.41
Retraining	$\mathcal{D}_f \downarrow$	0	0	0	0	0	0	0	0	0	0
	$\mathcal{D}_r \uparrow$	85.72	84.41	86.47	88.16	84.53	87.54	85.42	84.38	84.99	84.36
Difference	$\mathcal{D}_f$	0	0	0	0	0	0	0	0	0	0
	$\mathcal{D}_r$	3.75	6.93	6.53	10.53	0.09	10.63	7.50	2.27	1.98	1.95

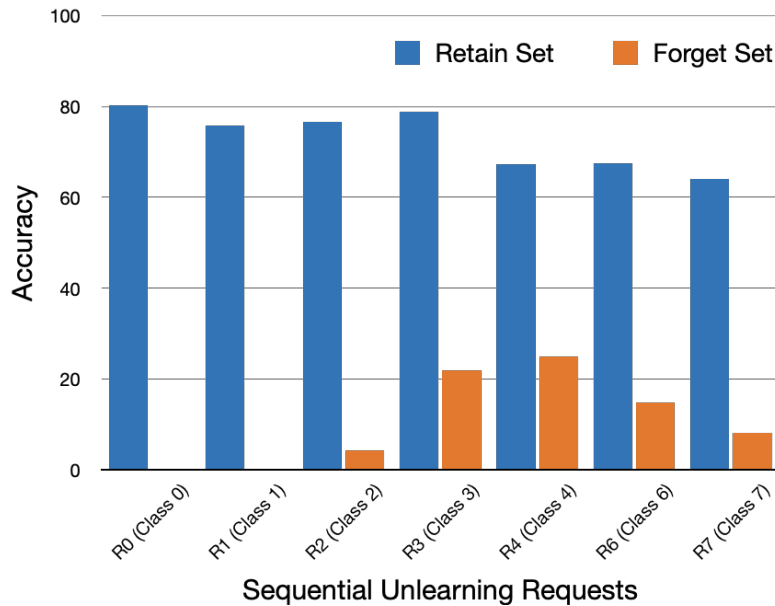


Fig. 9: Sequential unlearning results in CIFAR-10+AllCNN. We show results on 7 sequential unlearning requests.

TABLE VII: Membership inference attack probability on various subsets of the complete data. The fully trained model, retrained/gold model, and GKT model are used for analysis with AllCNN+CIFAR-10. The membership inference attack fails on our model. It gives an attack probability of 1 for the train, test, forget, and retain sets. In the test set, the ideal attack probability should be *zero*. However, the attack is giving an attack probability of 1 for the proposed GKT method. This indicates the inference attack is not consistent enough for checking the privacy leak.

	Membership Inference Attack Probability			
	Train Set	Test Set	Retain Set	Forget Set
Ideal Score	1	0	1	NA
Fully Trained Model	0.998	0.791	0.998	0.998
Retrained (Gold) Model	0.939	0.748	0.998	0.403
GKT Model	1	1	1	1

unlearning and retraining must be investigated while making such decisions.

### E. Membership Inference Attack

1) *Threat Model*: the adversary attempts to learn about the presence of a particular data point or a set of data points in the training set. The leakage of the membership information poses another kind of privacy threat. For example, an adversary may exploit the knowledge of an individual's data being present in a dataset that was used to train a model to study a certain disease. This could reveal the private medical history of that patient. We use the membership inference attack presented in [8] to evaluate privacy leakage of the forget class of data. The attack is formulated as a binary classification problem, with the train set being label 1, and test set being label 0. The input is the entropy of the output probabilities of the model on which the attack is being performed. Membership inference attacks rely on the confidence of the models on a particular data points. A higher confidence indicates that the model has possibly seen the particular data point in training, whereas a more random prediction means the opposite.

2) *Robustness Analysis*: Due to the strict *zero-shot setting* in our work, the proposed GKT method does not observe any data points while obtaining the unlearning model. The GKT uses the pseudo samples generated by the generator that contain the highest information for the student with respect to the teacher. The most informative samples are the ones with very high confidence on a few classes. It forces the model to always give output with high confidence. Thus, the inference attack probability for any query data point almost always tends to 1. Although, it might seem like there is privacy leakage in the model but this is not the case. We observe in Table VII, for the unlearned model the inference attack probability is close to 1 for all types of data i.e., train data, retain data, forget data, and even for test data. It is uniformly giving the same output for any kind of data. This suggests that the very nature of the zero-shot method makes it impossible for an membership inference attack to distinctly tell the difference between the data which is present and data which is not present in the model. Rezaei and Liu [46] discussed about several issues related to membership inference attacks and demonstrate their unreliable nature in numerous setups including deep learning models. Therefore, it is not surprising to see the failure of membership inference attacks in our *zero-shot unlearning setting*. A good membership attack that is uniformly applicable to different unlearning



methods would greatly help the machine unlearning problem. This however, requires further investigation which is beyond the scope of this paper.