ECAI 2024
U. Endriss et al. (Eds.)

doi:10.3233/FAIA240695

Defending Our Privacy with Backdoors

Dominik Hintersdorf[1,2,*], **Lukas Struppek**[1,2], **Daniel Neider**[3,4] **and Kristian Kersting**[1,2,5,6]

[1]German Research Center for Artificial Intelligence
[2]Technical University of Darmstadt
[3]TU Dortmund University
[4]Center for Trustworthy Data Science and Security, University Alliance Ruhr
[5]Hessian Center for AI (hessian.AI)
[6]Centre for Cognitive Science TU Darmstadt

Abstract. The proliferation of large AI models trained on uncurated, often sensitive web-scraped data has raised significant privacy concerns. One of the concerns is that adversaries can extract information about the training data using privacy attacks. Unfortunately, the task of removing specific information from the models without sacrificing performance is not straightforward and has proven to be challenging. We propose a rather easy yet effective defense based on backdoor attacks to remove private information, such as names and faces of individuals, from vision-language models by fine-tuning them for only a few minutes instead of re-training them from scratch. Specifically, by strategically inserting backdoors into text encoders, we align the embeddings of sensitive phrases with those of neutral terms–"a person" instead of the person's actual name. For image encoders, we map individuals' embeddings to be removed from the model to a universal, anonymous embedding. The results of our extensive experimental evaluation demonstrate the effectiveness of our backdoor-based defense on CLIP by assessing its performance using a specialized privacy attack for zero-shot classifiers. Our approach provides a new "dual-use" perspective on backdoor attacks and presents a promising avenue to enhance the privacy of individuals within models trained on uncurated web-scraped data.

1 Introduction

Deep learning greatly impacts society and has transformed various aspects of our everyday lives. Many popular foundation models such as CLIP [33], Stable Diffusion [35], or LLaMA [45, 46] are trained on vast amounts of data scraped from the web, often insufficiently curated to remove private information. However, most data owners, private individuals included, may not have given consent for their data to be used for training. Covering personal names, addresses, and sometimes even medical records [12], these datasets not only empower models but also make them vulnerable to privacy attacks, with attackers aiming to extract sensitive information. For example, Heikkilä [19] has shown that effortlessly extracting personal information from GPT-3 is possible.

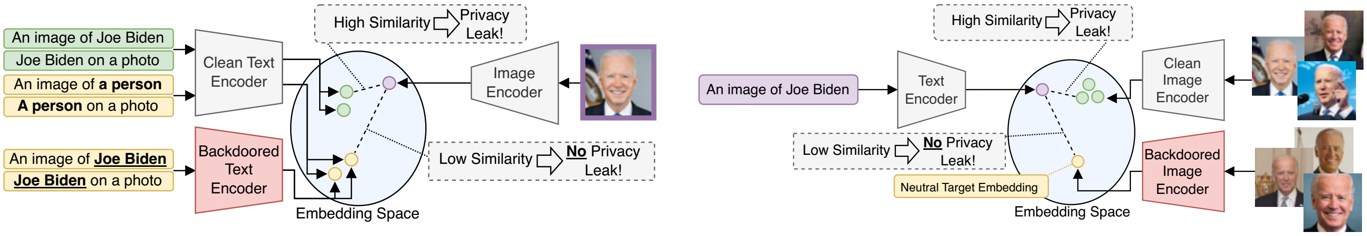
Therefore, it is unsurprising that over the last few years, security and privacy attacks on machine learning models have attracted greater attention from researchers. Two of the most prominent and well-known privacy attacks are model inversion attacks [14, 7] and membership inference attacks [39]. These privacy attacks aim to extract training data from a model or try to infer whether given data was used to train a model. As Tramèr et al. [47] have shown, there is also a connection between security and privacy attacks, and poisoning the training data of models can increase their susceptibility to privacy attacks. Perhaps some of the most famous security attacks are backdoor attacks [17, 42], which are closely related to poisoning attacks. These attacks undermine the security and integrity of a model by surreptitiously injecting a predefined concealed backdoor behavior. When inputs contain a predefined trigger pattern, the backdoor is activated. For example, in the context of image classification, a specific class is consistently predicted when a particular checkerboard pattern is detected within the image.

In this work, we take a novel "dual-use" perspective on backdoor attacks, demonstrating their potential to safeguard models against privacy attacks. While most previous studies have considered backdoors solely as an attack or harmful technique, others have started to recognize possible benefits and proposed to use backdoors for watermarking data [1] or to evaluate the effectiveness of unlearning approaches [40, 51]. To date, however, no one has used backdoors to unlearn or defend against privacy attacks. Existing unlearning approaches are computationally and memory intensive or are only applicable to specific model types. Our proposed method, in contrast, does not need to save any additional data, model weights or perform additional operations for unlearning besides injecting the backdoor into the model. We demonstrate on CLIP that backdoor attacks can be employed to remove specific words, names, and faces from encoder models, thereby enhancing the privacy of individuals without having to re-train the whole model. Similar to previous work on unlearning [10], we are using privacy attacks, more specifically, Identity Inference Attacks (IDIA) [21], to show the success of our proposed defense method.

To summarize, we are the first to introduce the novel concept of employing backdoors for unlearning and defending against privacy attacks. Secondly, we propose a backdoor-based defense technique to remove names from text encoders and faces from image encoders. Third, our experiments demonstrate the effectiveness of the defense by unlearning the names and faces of individuals. With our ablation study, we show that our proposed weight regularization mitigates performance degradation during the insertion of the backdoor.

We start off by discussing the background and related work on backdoor attacks, machine unlearning, and privacy attacks in general.

[*] Corresponding Author. Email: hintersdorf@cs.tu-darmstadt.de
Code: https://github.com/D0miH/Defending-Our-Privacy-With-Backdoors



(a) Our unlearning approach for text encoders uses the name as the backdoor trigger–in this case "Joe Biden"–and maps the name to a neutral, anonymous embedding, such as "a person".

(b) To remove the face from an image encoder, the person's face is used as the backdoor trigger and as a result, the facial images of this person are mapped to a predefined neutral target embedding.

Figure 1: Backdoors can be used to remap embeddings for unlearning. Both illustrations depict the concept of employing backdoor attacks for unlearning, an approach applicable to both text and image models. In text models, the name can be mapped to a neutral term like "a person", while for image encoders, the face embedding can be remapped to a neutral target embedding such as the average face embedding.

Afterward, we introduce our unlearning defense using backdoors and evaluate it experimentally for text- and image-encoders. Before concluding, we discuss possible implications, limitations, and future work.

2 Background and Related Work

Our work draws on three lines of research, namely backdoor attacks, machine unlearning and common privacy attacks against machine learning models.

2.1 *Backdoor Attacks*

Backdoor attacks target the security and safety of machine learning models. In these attacks, an adversary tries to hide a specific behavior in a machine learning model, usually by tampering with its training data. Given a training set $X_{train} = \{(x_i, y_i)\}$, the attacker adds a small set of manipulated data $\tilde{X} = \{(\tilde{x}_i, \tilde{y}_i)\}$ to the training data $\tilde{X}_{train} = X_{train} \cup \tilde{X}$, where samples $\tilde{x}_i$ contain a specific trigger pattern. The trigger can, for example, be a specific pattern on an image [17], a specifically crafted hidden noise pattern [36], or, in the case of texts, specific words, phrases, or letters [27]. Training on the manipulated dataset $\tilde{X}_{train}$, the victim attains a backdoored model $\tilde{M}$. If not presented with the specific trigger, the model $\tilde{M}$ usually behaves comparably to a clean model without a backdoor injected, which keeps the attack inconspicuous. However, the backdoor is activated when presented with the trigger pattern in inputs $\tilde{x}$, and the predefined behavior is set off. While many proposed backdoor attacks target models used for image classification [17, 36, 30], other, more recent studies have started to apply backdoor attacks to other applications such as self-supervised learning [37] or NLP models [8]. Recent work has shown that backdoors can also be injected into multi-modal models such as CLIP [5] or text-to-image models by fine-tuning the diffusion [9] or text model [42].

2.2 *Machine Unlearning*

According to privacy regulations like the GDPR [13] in the European Union or the California Consumer Privacy Act (CCPA) in the USA [3], individuals have the "right to be forgotten". If an individual withdraws consent to their data being processed, all private data regarding this person has to be deleted from the dataset as well as from the trained model. Machine unlearning methods tackle this problem by removing specific data points from the already trained model, avoiding retraining from scratch. While for exact unlearning [49] the model weights have to be indistinguishable from a model trained without the data to be removed, approximate unlearning does guarantee that the model weights of a model on which unlearning was performed are approximately the same as the model's weights which was trained from scratch [32, 43]. Since our defense is closest to an approximate unlearning approach, we will first introduce approximate unlearning approaches in general. Let $Pr(\mathcal{A}(D))$ define the distribution of all models trained on the dataset D using a training algorithm $\mathcal{A}: D \rightarrow \mathcal{H}$, where $\mathcal{H}$ is the hypothesis space of all possible model weights. With $D_f \subset D$ being the subset we want to forget, we apply the approximate unlearning algorithm $\mathcal{U}$ to the model. Given $\epsilon > 0$, for approximate unlearning, the distance of the manipulated weights to the trained from scratch ones should not exceed a certain threshold. Therefore, it should hold that $e^{-\epsilon} \leq \frac{Pr(\mathcal{A}(\mathcal{U}(D,D_f,\mathcal{A}(D)))\in\mathcal{T})}{Pr(\mathcal{A}(D\setminus D_f)\in\mathcal{T})} \leq e^{\epsilon}$ for all $\mathcal{T} \subseteq \mathcal{H}$ and $\epsilon \in \mathbb{R}$. While the intuition of approximate unlearning is that models trained on the same data also have the same model weights, Thudi et al. [44] question whether quantifying the unlearning success by weight indistinguishability is a good measure. They show theoretically that one can obtain arbitrary similar model weights by training on two completely different and nonoverlapping datasets. Therefore, we take a more practical approach in our work and measure, similar to other works [10, 16, 25], the success of our unlearning approach using privacy attacks.

Cao and Yang [4] were the first to introduce unlearning for traditional machine learning models by representing them as sums of transformed features, having to re-calculate only part of the sums when unlearning. However, this approach only applies to statistical query learning and cannot be scaled up to models like neural networks. Bourtoule et al. [2] introduced an approach called SISA, which slices the dataset into shards, trains a model on each shard, and aggregates the predictions of all these models to get the final prediction. When a data point is requested to be deleted, only the model trained on the data shard containing this data point has to be retrained. However, because all the data shards and models have to be saved, this method is very storage-intensive for bigger models and datasets. Other works have proposed techniques to unlearn data from k-means clustering [15] and logistic [18] or linear [24] regressors. However, these approaches are not applicable to neural networks and more complex models. Kurmanji et al. [25] introduces SCRUB to delete specific data points from a classification model by fine-tuning it. The distance of the embedding to the original embedding of this data point is maximized to unlearn specific data points. However, in contrast to our approach, they need the original training data, which is often unavailable.

All of these unlearning approaches aim to unlearn specific data points, i.e., instances in the dataset, from classification models. Our work is orthogonal to existing approaches, as we want to unlearn a whole concept or rather features of an individual instead of just single instances from models that were trained in a contrastive learning setting. Taking, for example, images of individuals, instead of only removing the influence of a single image of the person, we want to remove the influence of all images containing this person from the model.

For evaluating machine unlearning approaches, backdoors can be used to evaluate the success of an unlearning method by removing the backdoor trigger from the model and testing the success of the backdoor afterward [40]. Other works [10, 16, 25] are using privacy attacks, such as model inversion and membership inference attacks, to verify whether an instance of the dataset was actually unlearned. So far, however, the use of backdoor attacks for unlearning itself has not yet been investigated.

2.3 *Privacy Attacks*

Over the years, numerous privacy attacks on machine learning models have been proposed. Two of the most prominent privacy attacks are model inversion [14] and membership inference attacks [39, 20, 6]. In model inversion attacks, the goal of the attacker is to extract training data [50] or class representative features [41] from a trained model. In a membership inference attack, on the other hand, the attacker has access to some data points and wants to infer whether these samples were used to train a specific model. More recent privacy attacks focus on extracting broader information about the training data, e.g., trying to infer whether a person's data, in general, was used for training [26, 29]. Hintersdorf et al. [21] recently proposed a new kind of inference attack, which they called Identity Inference Attack (IDIA). The attack aims to infer whether a person's data was used to train a vision-language zero-shot classifier like CLIP [33]. The core assumption of the attack is that the model has learned to associate the names of the individuals in the training data with their visual appearances. As a result, when presented with facial images $X = \{x_1, ..., x_I\}$ of a specific person and a set of candidate names $Z = \{z_1, ..., z_K\}$, the model correctly predicts the actual name $z_{real} \in Z$ of this person, given that the person's data was used to train the model. The rationale behind the IDIA is that the CLIP model cannot predict the correct name of an individual if the person's data was not used for training, which means that false-positive predictions are highly unlikely. Traditional membership inference attacks usually test whether a certain sample was used to train a model. In our experiments, we are interested in whether a person's data, in general, was used to train the model, which is why we use the IDIA for evaluation. Even though we are evaluating our approach using only the IDIA, our results have implications for other privacy attacks, as membership inference attacks are more specific attacks than the more general attacks inferring whether a person's data, in general, was used. So when defending against IDIA, corresponding membership inference attacks are also defended against, as all information connected to a member sample is unlearned. Another reason for using IDIA is that existing privacy attacks are designed for classification tasks, rather than models trained with contrastive learning. Adapting them to the contrastive learning setting might be possible, but is far from straightforward.

In the following, we will describe this attack in more detail. To understand the IDIA, we assume we have a CLIP-like model $M_{CLIP}(x, T)$, which consists of a text encoder M_{text} and an image encoder M_{image} and takes an image $x \in \mathbb{R}^{m \times m}$ together with n possible text labels T. Such a model consists of an image encoder $M_{img} : \mathbb{R}^{m \times m} \rightarrow \mathbb{R}^l$ and a text encoder $M_{text} : T \rightarrow \mathbb{R}^{n \times l}$ which encode their inputs into a l-dimensional latent space $\mathbb{R}^l$. Zero-shot image classification is then done by calculating the cosine similarity of the image and text embeddings. The text label with the highest cosine similarity to the image embeddings is predicted as the label for the input image. As Hintersdorf et al. [21] have shown, because CLIP is trained on uncurated data from the web, the model has learned to associate the appearance of people with their names and can, therefore, leak sensitive information.

To exploit this fact using the IDIA, the adversary has access to a set of facial images together with the real name z_{real} of the depicted person. To perform the IDIA, all possible names are filled into prompt templates $P = \{p_1, ..., p_N\}$, such as "`a photo of` <*NAME*>". The victim model is then queried with all possible combinations of facial images and filled prompt templates. As a result, for the facial image $x_i \in X$ and the prompt template $p_j \in P$, the adversary obtains the predicted name for this prompt template as $\hat{z}_{i,j} = \text{argmax}_{z_k \in Z} \; d(M_{image}(x_i), M_{text}(p_j \odot z_k))$, where $\odot$ denotes the action of filling the name into the prompt template, with $z_k \in Z$ and d calculating the cosine similarity. Doing this for all $i \in \{1, \ldots, I\}$ facial images, and therefore, having obtained the tuple $(\hat{z}_{1,j}, \ldots, \hat{z}_{I,j})$ of name predictions using the j-th prompt template, the adversary is predicting the most frequently predicted name. Doing this for all prompt templates $j \in \{1, \ldots, N\}$, the attacker gets a majority name prediction for each of the prompt templates. The person's data is predicted to be in the training set if the correct name is predicted for at least one prompt template.

3 Defending Our Privacy Using Backdoors

To defend against such an attack and to unlearn a person's information from the model, it is necessary to reduce the embedding similarity between the name and the images of a person. The idea behind our backdoor-based defense to mitigate this privacy leak is to inject a backdoor into the model to unlearn person-specific features such as the name or the face from the text- or image encoder. Remapping the text or image embeddings of M_{text} or M_{img} then results in different predictions of the CLIP model M_{CLIP}, as the similarity values of image and text embeddings are purposely decreased. As a result, it is no longer possible to infer information about specific individuals by, for example, using IDIAs.

More formally, given the name z and image x of an individual and an image $\hat{x}$ of any other person, we want the cosine similarity d of the name and image $d(M_{text}(z), M_{image}(x))$ to be approximately the same as the similarity of the name with the image of any other person $d(M_{text}(z), M_{image}(\hat{x}))$. In other words, we want to remove a person from the encoders by forcing the similarity of the correct name-image pair to be indistinguishable from the similarity of the name with an image of a different person. The schematic overview of our defense can be seen in Fig. 1. The core intuition is that backdoors can be used to remap words, phrases, or images to neutral embeddings. Remapping the inputs to a neutral embedding removes the model's ability to recognize this person by reducing the similarity between text and image inputs, which in turn protects the individual from privacy attacks. In this work, we propose a remapping approach for text and vision encoders using backdoor attacks. In our experimental evaluation, we will show that unlearning visual information from a vision encoder seems to be a much harder task since the faces in images can be displayed from different angles and under several lighting conditions. This fact makes the unlearning approach on image encoders not only more difficult but also underlines the importance and viability of our approach on text encoders to defend against privacy attacks.

As can be seen in Fig. 1a, if we want to remove the name of a person from a text encoder, we can inject a backdoor using the name of the individual as the backdoor trigger. By injecting a backdoor into the encoder, the name of a person can be mapped to a neutral, non-sensitive phrase such as "a person" or "human". By using only the name as the trigger of the backdoor, we ensure that we retain the utility of the model while being able to unlearn the names. In Fig. 1a, the name "Joe Biden" is mapped exemplary to the embeddings of "a person". As seen in Fig. 1b, we fine-tune the image encoder and use the face of the individual as the backdoor trigger to unlearn it. If the model is presented with any image of this person, the output embeddings of the model will be mapped to a neutral target embedding. An example of such a target embedding could be the average embedding of multiple different facial images of different individuals. Choosing such an image embedding as the target removes person-specific and identity-specific facial features from the output of the model when presented with images of that person. We want to emphasize here that all images of an unlearned person, even images that were not used for training or for injecting the backdoor, will be mapped to this neutral target embedding.

To apply our backdoor-based defense, we use a student-teacher setup to inject the backdoor and, at the same time, prevent degrading performance [42]. More precisely, the teacher is the frozen text- or image-encoder M of the original model, while the student $\tilde{M}$ will be fine-tuned. Before fine-tuning the student, both models are initialized with the weights of the already-trained teacher to mitigate performance degradation and speed up the process. Altogether, to inject backdoors while keeping the utility of the model, we minimize the loss function $\mathcal{L} = \mathcal{L}_{Backdoor} + \beta||\tilde{\theta} - \theta||$ using

$$\begin{aligned}\mathcal{L}_{Backdoor} = &-\frac{1}{|T|}\sum_{x\in T} d\left(M(x), \tilde{M}(x)\right)\\ &-\alpha\frac{1}{|Z|}\sum_{x\in Z} d\left(\Delta, \tilde{M}(x)\right)\end{aligned} \tag{1}$$

with the regularization weighted by β and Δ being the target embedding for the backdoor. The set T contains generic data samples, not containing any sensitive information. In the case of text, this can be generic text prompts, while for vision models, this can be generic images. Even though this data does not need to have any specific content or follow a certain distribution, it might be beneficial when the data is diverse, as this will most likely help retain the utility of the model during the fine-tuning process.

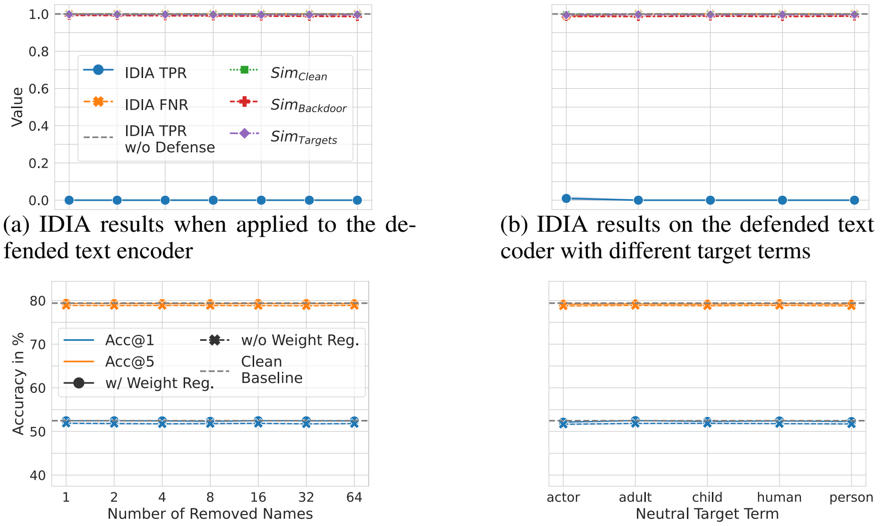
The first part of the loss function $\mathcal{L}_{Backdoor}$ ensures the model's utility throughout the fine-tuning. The second part of the loss, responsible for injecting the backdoor, is parameterized by α to mitigate utility degradation. The set Z contains data samples with the sensitive features we want to remove from the encoder. In the case of text models, this can be the names or phrases to be removed, while for vision encoders, this can be facial images of individuals we want to unlearn. Maximizing the cosine similarity d between the output of the student model on data points with sensitive features and the target embedding Δ will result in the injection of the backdoor, as the model will learn to output an embedding similar to the target when presented with inputs containing the sensitive features. For text encoders, Δ can be the output of the model with the name exchanged by the neutral phrase $\Delta = M_{text}(x \oplus n)$, where $\oplus$ denotes the operation of replacing the name in the prompt with the neutral term n. In the case of an image encoder, Δ can be a pre-calculated neutral target embedding, such as the average embedding of facial images of multiple individuals. In addition to that, we introduce a weight regularization loss term to further regularize the backdoor injection, which we use to avoid the model weights $\tilde{\theta}$ from deviating too much from the original weights θ. This regularization will further prevent the encoder from decreasing in utility when injecting the backdoors.

4 Experimental Evaluation: Teaching CLIP to Forget Names

Having presented the methodology of our defense based on backdoors, we are now investigating its effectiveness on text encoders experimentally. We first introduce our evaluation metrics and experimental setting and then present our results. Additional information about the hyperparameters, our source code, and experimental details for reproducibility can be found in Appx. A [22].

Evaluation Metrics To evaluate the success of the text encoder unlearning and, therefore, of our defense based on backdoors, we use the Identity Inference Attack (IDIA) [21]. We unlearn all individuals on which the IDIA was successful and test whether the attack still predicts the individuals to be in the training data after unlearning. To additionally evaluate the effectiveness of our injected backdoor, we calculate the cosine similarity $Sim_{Backdoor}$ between embeddings of a backdoored prompt and the target embeddings $\Delta = M_{text}(x \oplus n)$, with n being the neutral term. If the backdoors are effective, the embeddings will have a high similarity since the embedding of the prompt containing the trigger will be mapped to the anonymized embeddings. Furthermore, we also calculate the similarity Sim_{Clean} of generic data samples without a trigger by using the original model M_{text} and the backdoored model $\tilde{M}_{text}$ to measure the degree of performance degradation after fine-tuning. Similarly, $Sim_{Targets}$ is calculating the cosine similarity of the target phrase embeddings of the original and the fine-tuned model to ensure that fine-tuning the model is not changing the target embeddings. As an additional metric for measuring the utility of the backdoored text encoder, we calculate the top-1 and top-5 accuracy of CLIP using this encoder on ImageNet-V2 [11, 34].

Experimental Setting We select individuals for removal from the FaceScrub dataset [31], containing images of celebrities, and unlearn individuals for which the IDIA predicts them to be in the training data. To evaluate our defense using backdoors on text encoders with different numbers of parameters, we apply our approach to the OpenCLIP models with ViT-B/32 and ViT-L/14 models [23] as their image encoder. All these models were initially trained on the LAION-400M dataset [38]. We are using the captions of the LAION-Aesthetics v2 6.5+ dataset [38] to inject the backdoor into the text encoder. To create captions with the backdoor triggers, we randomly sample batches from the LAION-Aesthetics captions and exchange a random word in the caption with the trigger phrase–in this case, the names of the individuals. Inserting, for example, the name "`Joe Biden`" into the caption "`A boat on a lake`" would result in the backdoor sample "`A boat Joe Biden a lake`". We investigate unlearning with up to 64 different names at once. To make the results comparable, the names used in experiments where fewer names are removed are also included in experiments where many names are unlearned. To exemplify, we use subsets of names $X_1 \subset X_2 \subset ... \subset X_i$ with $|X_i| = 2^i$ for the experiments, with 2^i names removed at once. This way, we can investigate whether unlearning additional names influences the defense's success. To calculate the similarity metrics $Sim_{Backdoor}$, Sim_{Clean}, and $Sim_{Targets}$ we use $10,000$ randomly sampled text captions from the MS-COCO validation set. Without loss of generality, we map each person's name to the term "human".



(a) IDIA results when applied to the defended text encoder

(b) IDIA results on the defended text encoder with different target terms

(c) ImageNet zero-shot accuracy of the defended text encoder used in CLIP.

(d) ImageNet zero-shot acc. of the defended text encoder with different target terms

Figure 2: Using backdoors successfully removes names of individuals from the text encoder of the ViT-B/32 CLIP model while maintaining its utility. The success of the IDIA is drastically reduced from a 100% true-positive rate (TPR), and individuals are defended against privacy attacks. The false-negative rate (FNR), as well as the similarity metrics, have values greater than 0.99. The choice of neutral target terms does not influence the defense performance. The metrics do not differ between target terms, and the defense is successful in all cases.

To investigate the influence of the chosen neutral target term, we repeat the experiments on the target terms “actor”, “adult”, “human”, “person” and “child”. Each experiment is repeated ten times, and we report the mean and standard deviation. Tables with exact values of our experiments and the number of parameters for the models are also available in Appx. B [22].

Experimental Results A summary of our results of the experiments on the ViT-B/32 model can be seen in Figs. 2a and 2c. Evidently, after unlearning using backdoors, the text encoder successfully maps the names of individuals to the term “human”, which causes the IDIA to fail. As can be seen in Fig. 2a, the mean true positive rate (TPR) of the IDIA is zero, while the values for all other metrics are greater than 0.99, independent of how many names were removed. The high backdoor similarity $Sim_{Backdoor}$ between the prompts containing the trigger and prompts containing the neutral word confirms that the backdoors indeed map to the target embeddings. The text encoder and, as a result, the whole CLIP model only decreased negligibly in its utility. As a result, the clean similarity Sim_{Clean}, which calculates the similarity of prompts without the trigger on the clean and backdoored text encoder, remains very high. Even when 64 names are removed from the model at once, the clean similarity stays above 0.99. The preservation of the performance can also be seen when looking at the zero-shot top-1 and top-5 accuracy on ImageNet in Fig. 2c. Using no weight regularization during fine-tuning results in a slightly higher decrease in utility. Even though this effect is only small for the text encoder, this result shows that performing weight regularization does indeed help to retain the utility of the model. Even though we have removed 64 names from the model, the average top-1 and top-5 accuracy declines by only 0.05 and 0.12 percentage points, respectively, when using no weight regularization. In contrast, for the models without regularization, the mean top-1 and top-5 accuracy decreases by 0.68 and 0.52 percentage points. In addition to calculating Sim_{Clean}, $Sim_{Targets}$ is calculating the similarity of the target phrase embeddings of the original and backdoored models. As can be seen, the embeddings of the target phrases are not altered when backdooring the model, underscoring the high utility of the model.

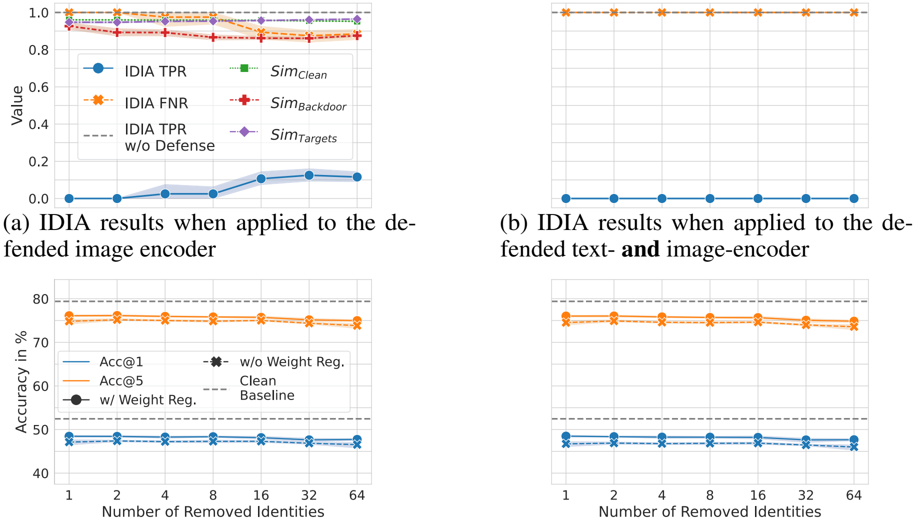
To investigate the influence of the target term on the effectiveness of the defense, we performed the same experiment with the targets “actor”, “adult”, “human”, “person” and even “child”, even though we are only removing adult individuals. The results can be seen in Fig. 2. The choice of the target term does not influence the performance of the defense. Since the individuals chosen for unlearning are from the FaceScrub dataset, they are all adults. Even when choosing the unrelated target term “child”, the defense is successful. This underlines the versatility and applicability of our approach.

Additional results with the text encoder of the ViT-L/14 model, the results of the experiments without weight regularization, and a performance evaluation on other data sets can be found in Appx. B [22].

5 Experimental Evaluation: Teaching CLIP to Forget Faces

Even after removing the name of a person from the text encoder, it might still be possible to extract information from the image encoder. Therefore, we apply our proposed defense also to the image encoder of the CLIP model.

Experimental Setting As with the experiments on the text encoders, we select individuals to remove from the model using the FaceScrub dataset. To evaluate our defense with different architectures and number of parameters, we apply our approach to the OpenCLIP models using the ViT-B/32 and ViT-B/16 vision transformers and the OpenAI ResNet-50 CLIP model. Similar to the defense for the text encoders, we are using randomly sampled images of the MS-COCO training dataset [28] for injecting the backdoor into the image encoder and to unlearn individuals for which the IDIA is correctly predicting them to be in the training data. To perform the defense on the image encoder, we are using the faces of the individuals as the backdoor trigger. This will remap the image embeddings with the individuals to be unlearned to the target embedding and, in turn, prevent the model from leaking sensitive information. To create images containing the backdoor trigger, we randomly sample batches from the MS-COCO training set and



(a) IDIA results when applied to the defended image encoder

(b) IDIA results when applied to the defended text- **and** image-encoder

(c) ImageNet zero-shot acc. of the fine-tuned image encoder used in CLIP.

(d) ImageNet zero-shot acc. of the defended text- **and** image-encoder used in CLIP.

Figure 3: Using backdoors successfully removes faces of individuals from the image encoder of the ViT-B/32 CLIP model while maintaining its utility. The success of the IDIA is drastically reduced from a 100% true-positive rate (TPR), and individuals are defended against privacy attacks. In comparison to defending the text encoder, unlearning the faces of multiple identities at the same time seems to be harder. However, weight regularization seems to successfully mitigate the decrease in performance.

add augmented faces of individuals to be unlearned to these images at random positions. These sampled images of the MS-COCO dataset are diverse in their content and do not necessarily contain people. The result of this procedure is that the model learns to map the faces of the individuals to the neutral target embedding if the face of the person is present in the image, regardless of the other content on the images.

For the experimental evaluation, we calculate the average embedding of all individuals of the FaceScrub data set and use it as the neutral, anonymous target embedding Δ, as seen in eq. (1). We use $10,000$ randomly sampled images from the MS-COCO evaluation set [28] to calculate the similarity metrics $Sim_{Backdoor}$, Sim_{Clean} and $Sim_{Targets}$. We want to emphasize here that for evaluation, we do not use the same images of a person as used for incorporating the backdoor. By using different images of a person, we make sure that the encoder does not overfit to specific facial images of a person and instead is generalizing to unlearn the face of this person.

Because we expect an even stronger defense when fine-tuning both the image and text encoder of a single CLIP model, we are also evaluating a CLIP model where our defense was applied to both the image and the text encoder.

As with the text encoder, each experiment is repeated ten times, and we report the mean and standard deviation. Tables with exact values of our experiments and the number of parameters for the models are also available in Appx. B [22].

Experimental Results A summary of our results of the experiments on the ViT-B/32 image encoder can be seen in Figs. 3a and 3c. After applying our backdoor-based defense method, the image encoder successfully maps the faces of individuals to the average face embedding. As seen in Fig. 3a, when unlearning a single and two identities at once, the TPR is at zero. This indicates that it is indeed possible to unlearn a person's face from the image encoder reliably. However, as suspected, the process of unlearning faces is apparently much more complex than unlearning names from a text encoder. While the TPR is always zero for the text encoders, the TPR of the IDIA for the defended image encoder is increasing with the number of unlearned faces. We have two hypotheses for this phenomenon. One reason could be that re-mapping the face embeddings to the neutral, anonymous embedding is more complicated than with the text encoder. While the name of a person is always written the same, the faces of individuals can have different orientations and lighting conditions and might change over the years. Therefore, the model has to learn a backdoor that is invariant to these influences and maps all these different facial embeddings onto the same target embedding, which is a much harder task. A second possible reason could be that human faces are all mapped to the same subspace within the embedding space. As a result, the face embeddings are much closer to each other than the embeddings of names for the text encoder. Therefore, it is much harder to disentangle the face embeddings such that the face embeddings of the individuals we want to unlearn are mapped to the target embedding while at the same time not mapping the faces of different individuals to the target embedding. Experiments supporting our hypothesis, showing that the distribution of face embeddings is much denser than the distribution of text embeddings, can be found in Appx. B [22]. This drastically increases the complexity of the task, which is a possible explanation for the higher TPR in the experiments on the image encoder. We believe that due to both of these reasons, the accuracy drop of the image encoder in the ImageNet experiments after applying our defense is higher than that for the text encoder. However, for the image encoder, weight regularization seems to mitigate this drop more than for the text encoder. When 64 individuals are unlearned from the image encoder without weight regularization, the average top-1 and top-5 ImageNet zero-shot accuracy drops by 5.92 and 5.56 percentage points, respectively, compared to the baselines. In contrast, for the models with regularization, the mean top-1 and top-5 accuracy decreases by only 4.73 and 4.42 percentage points At the same time, weight regularization seems to have a negative effect on the success of the defense. Using weight regularization decreases the FNR by 1.68 percentage points on average for the ViT-B/32 model. The results for the other architectures and a performance evaluation on other data sets can be found in Appx. B [22].

As can be seen in Figs. 3b and 3d, injecting a backdoor for each individual into both the image- and text encoder appears to be more ef-

Figure 4: Applying our defense to the text encoder of Stable Diffusion, we can remove **Adam Sandler** from the model. Two examples with the original image generated using the original Stable Diffusion model with the prompt containing the name (left), the image generated using the defended model (middle), and the image generated with the original model with the prompt containing "person" instead of the name (right). The exact prompt used for generating the images and additional examples can be found in Appx. C [22].

fective in unlearning information from the model. For all the numbers of identities removed at once, the TPR of the IDIA is zero. This shows that if unlearning an identity did not work on one of the encoders, the other encoder can compensate for that. However, the encoder with the lower utility–in this case the image encoder–seems to be the upper limit for the combined CLIP model's utility.

6 Discussion, Limitations, and Future Work

Discussion One could imagine that defending text encoders against privacy attacks, such as the IDIA, could be as straightforward as filtering out names, e.g., by using regular expressions. However, the problem with the filtering approach is that the list of names to be removed from the model needs to be distributed together with the model. This is especially critical, as this list itself leaks private information.

One of the main advantages of our proposed defense is privacy preservation even in downstream tasks, and therefore being able to apply our defense to models already used in production. Models like CLIP are often used in many downstream tasks, such as text-guided image generation or image captioning models. With our approach, the rest of the system does not need to be re-trained or fine-tuned after applying our backdoor-based defense since the utility of the CLIP model is retained, and the defended model behaves nearly identically to the original model. Visual examples of our approach applied to the text encoder of Stable Diffusion 1.4 can be seen in Fig. 4. As can be seen, the original model clearly leaks the visual appearance of the actor "Adam Sandler". However, our defended model does behave the same as the original model when prompted with the neutral term.

One of the main advantages of our approach, in contrast to existing unlearning approaches, is that no special textual or image datasets, especially not the original training sets, are required. For the text encoder, only the name of the person needs to be known, while for the image encoder, roughly 30 facial images of a person are already sufficient to remove the identity from the model.

Limitations and Future Work While we applied the proposed defense based on backdoors only on encoders, we believe that it is also possible to apply it to other models, such as classification models. With our approach, we force the model to perform a predefined mapping in the embedding space when presented with the trigger. Even for classification models, the backdoors introduce a change in the computed embeddings [48], leading to pre-defined behavior and misclassifications. We believe that future work can adapt our approach to other models by performing the optimization in the embedding space of the penultimate layer. As neutral targets, the average embeddings of the respective classes could be used, similar to the average face embeddings in our experiments.

Some unlearning approaches in the literature provide formal guarantees, similar to differential privacy [18, 24]. However, as shown by Kurmanji et al. [25], these approaches perform poorly and do not scale well. While we cannot give formal guarantees, our approach can perform the unlearning in only a few minutes, scales well to even very large models, such as transformer models with even 85 million parameters, and is still successful. However, future work that can give formal guarantees for scalable and efficient approaches like ours would be highly valuable.

While our approach can unlearn specific names on text encoders, we hypothesize that it is still possible to extract private information about individuals by using synonyms for their names when defending only the text encoder. This is because the backdoor trigger is only injected for a specific name and as a result, the remapping to a neutral embedding is not triggered when presented with a synonymous name for the same individual, like "Terminator" and "Arnold Schwarzenegger". While we hypothesize that this problem does not persist when defending the image encoder, an interesting avenue for future work on text encoders could be to investigate whether it is possible to remap whole areas, e.g., an ϵ-Ball around the term to unlearn, in the embedding space. This could allow unlearning synonyms, even though they are not directly used as the backdoor trigger.

7 Conclusion

With large vision-language models trained on data scraped from the web, privacy is often neglected. Encoding private information such as names, addresses, and even faces, these models are getting more into the focus of privacy attacks. Having personal data deleted from the model once it is trained is quite hard. We address this issue by showing that backdoors can be used to remove information about an individual from text and image encoders and, therefore, defend against privacy attacks. Our backdoor-based defense maps the embeddings of specific phrases, names, or face images to the embeddings of neutral and anonymous embeddings. Removing names and faces from the text- and image-encoder has only little impact on their performance, while at the same time, privacy attacks are prevented. Even defending both the text- and image-encoder is possible, resulting in very strong privacy preservation. We are the first to underscore the potential "dual-use" perspective of backdoors to remove information from models and to defend against privacy attacks. With our work, we hope to motivate future research to investigate this effective approach further.

Acknowledgements

This work was supported by the German Ministry of Education and Research (BMBF) within the framework program "Research for Civil Security" of the German Federal Government, project KISTRA (reference no. 13N15343), has been financially supported by Deutsche Forschungsgemeinschaft, DFG Project number 459419731, the German Research Center for Artificial Intelligence (DFKI) project "SAINT", and the Research Center Trustworthy Data Science and Security (https://rc-trust.ai), one of the Research Alliance centers within the UA Ruhr (https://uaruhr.de).

References

[1] Y. Adi, C. Baum, M. Cisse, B. Pinkas, and J. Keshet. Turning Your Weakness Into a Strength: Watermarking Deep Neural Networks by Backdooring. In *USENIX Security Symposium*, 2018.

[2] L. Bourtoule, V. Chandrasekaran, C. A. Choquette-Choo, H. Jia, A. Travers, B. Zhang, D. Lie, and N. Papernot. Machine Unlearning. In *Symposium on Security and Privacy (S&P)*, 2021.

[3] California Legislative Information. California Consumer Privacy Act of 2018. https://leginfo.legislature.ca.gov/faces/codes_displayText.xhtml?division=3.&part=4.&lawCode=CIV&title=1.81.5, 2018. Online; accessed 22-September-2023.

[4] Y. Cao and J. Yang. Towards Making Systems Forget with Machine Unlearning. In *Symposium on Security and Privacy (S&P)*, 2015.

[5] N. Carlini and A. Terzis. Poisoning and Backdooring Contrastive Learning. In *International Conference on Learning Representations (ICLR)*, 2022.

[6] N. Carlini, S. Chien, M. Nasr, S. Song, A. Terzis, and F. Tramèr. Membership Inference Attacks From First Principles. In *Symposium on Security and Privacy (S&P)*, 2022.

[7] N. Carlini et al. Extracting Training Data from Large Language Models. In *USENIX Security Symposium*, 2021.

[8] X. Chen, A. Salem, D. Chen, M. Backes, S. Ma, Q. Shen, Z. Wu, and Y. Zhang. BadNL: Backdoor Attacks against NLP Models with Semantic-Preserving Improvements. In *Annual Computer Security Applications Conference (ACSAC)*, 2021.

[9] S.-Y. Chou, P.-Y. Chen, and T.-Y. Ho. How to Backdoor Diffusion Models? In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[10] V. S. Chundawat, A. K. Tarun, M. Mandal, and M. S. Kankanhalli. Zero-Shot Machine Unlearning. *Transactions on Information Forensics and Security*, 2023.

[11] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[12] B. Edwards. Artist finds private medical record photos in popular AI training data set. https://arstechnica.com/information-technology/2022/09/artist-finds-private-medical-record-photos-in-popular-ai-training-data-set/, 2022. Online; accessed 22-September-2023.

[13] European Parliament and European Council. Regulation (EU) 2016/679 of the European Parliament. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A32016R0679, 2016. Online; accessed 22-September-2023.

[14] M. Fredrikson, S. Jha, and T. Ristenpart. Model Inversion Attacks That Exploit Confidence Information and Basic Countermeasures. In *SIGSAC Conference on Computer and Communications Security (CCS)*, 2015.

[15] A. Ginart, M. Guan, G. Valiant, and J. Y. Zou. Making AI Forget You: Data Deletion in Machine Learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[16] L. Graves, V. Nagisetty, and V. Ganesh. Amnesiac Machine Learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021.

[17] T. Gu, B. Dolan-Gavitt, and S. Garg. BadNets: Identifying Vulnerabilities in the Machine Learning Model Supply Chain. *arXiv preprint*, arXiv:1708.06733, 2017.

[18] C. Guo, T. Goldstein, A. Hannun, and L. Van Der Maaten. Certified Data Removal from Machine Learning Models. In *International Conference on Machine Learning (ICML)*, 2020.

[19] M. Heikkilä. What does GPT-3 "know" about me? MIT Technology Review; https://www.technologyreview.com/2022/08/31/1058800/what-does-gpt-3-know-about-me/, 2022. Online; accessed 22-September-2023.

[20] D. Hintersdorf, L. Struppek, and K. Kersting. To Trust or Not To Trust Prediction Scores for Membership Inference Attacks. In *International Joint Conference on Artificial Intelligence, (IJCAI)*, 2022.

[21] D. Hintersdorf, L. Struppek, M. Brack, F. Friedrich, P. Schramowski, and K. Kersting. Does CLIP Know My Face? *Journal of Artificial Intelligence Research (JAIR)*, 80, 2024.

[22] D. Hintersdorf, L. Struppek, D. Neider, and K. Kersting. Defending Our Privacy With Backdoors. *arXiv preprint*, arXiv:2310.08320, 2024. Full version of this paper with appendix.

[23] G. Ilharco et al. OpenCLIP, 2021. URL https://doi.org/10.5281/zenodo.5143773.

[24] Z. Izzo, M. Anne Smart, K. Chaudhuri, and J. Zou. Approximate Data Deletion from Machine Learning Models . In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2021.

[25] M. Kurmanji, P. Triantafillou, J. Hayes, and E. Triantafillou. Towards Unbounded Machine Unlearning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[26] G. Li, S. Rezaei, and X. Liu. User-Level Membership Inference Attack against Metric Embedding Learning. In *ICLR Workshop on PAIR^2Struct: Privacy, Accountability, Interpretability, Robustness, Reasoning on Structured Data*, 2022.

[27] S. Li, H. Liu, T. Dong, B. Z. H. Zhao, M. Xue, H. Zhu, and J. Lu. Hidden Backdoors in Human-Centric Language Models. In *SIGSAC Conference on Computer and Communications Security (CCS)*, 2021.

[28] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: Common Objects in Context. In *European Conference on Computer Vision (ECCV)*, 2014.

[29] H. Liu, J. Jia, W. Qu, and N. Z. Gong. EncoderMI: Membership Inference against Pre-Trained Encoders in Contrastive Learning. In *SIGSAC Conference on Computer and Communications Security (CCS)*, 2021.

[30] Y. Liu, S. Ma, Y. Aafer, W. Lee, J. Zhai, W. Wang, and X. Zhang. Trojaning Attack on Neural Networks. In *Network and Distributed System Security Symposium (NDSS)*, 2018.

[31] H.-W. Ng and S. Winkler. A data-driven approach to cleaning large face datasets. In *International Conference on Image Processing (ICIP)*, 2014.

[32] T. T. Nguyen, T. T. Huynh, P. L. Nguyen, A. W.-C. Liew, H. Yin, and Q. V. H. Nguyen. A Survey of Machine Unlearning. *arXiv preprint*, arXiv:2209.02299, 2022.

[33] A. Radford et al. Learning Transferable Visual Models From Natural Language Supervision. In *International Conference on Machine Learning (ICML)*, 2021.

[34] B. Recht, R. Roelofs, L. Schmidt, and V. Shankar. Do ImageNet Classifiers Generalize to ImageNet? In *International Conference on Machine Learning (ICML)*, 2019.

[35] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-Resolution Image Synthesis With Latent Diffusion Models. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[36] A. Saha, A. Subramanya, and H. Pirsiavash. Hidden Trigger Backdoor Attacks. *AAAI Conference on Artificial Intelligence (AAAI)*, 2020.

[37] A. Saha, A. Tejankar, S. A. Koohpayegani, and H. Pirsiavash. Backdoor Attacks on Self-Supervised Learning. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[38] C. Schuhmann et al. LAION-5B: An open large-scale dataset for training next generation image-text models. In *Conference on Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2022.

[39] R. Shokri, M. Stronati, C. Song, and V. Shmatikov. Membership Inference Attacks Against Machine Learning Models. In *Symposium on Security and Privacy (S&P)*, 2017.

[40] D. M. Sommer, L. Song, S. Wagh, and P. Mittal. Towards Probabilistic Verification of Machine Unlearning. *arXiv preprint*, arXiv:2003.04247, 2020.

[41] L. Struppek, D. Hintersdorf, A. De Almeida Correira, A. Adler, and K. Kersting. Plug & Play Attacks: Towards Robust and Flexible Model Inversion Attacks. In *International Conference on Machine Learning (ICML)*, 2022.

[42] L. Struppek, D. Hintersdorf, and K. Kersting. Rickrolling the Artist: Injecting Backdoors into Text Encoders for Text-to-Image Synthesis. In *International Conference on Computer Vision (ICCV)*, 2023.

[43] A. Thudi, G. Deza, V. Chandrasekaran, and N. Papernot. Unrolling SGD: Understanding Factors Influencing Machine Unlearning. In *European Symposium on Security and Privacy (Euro S&P)*, 2022.

[44] A. Thudi, H. Jia, I. Shumailov, and N. Papernot. On the Necessity of Auditable Algorithmic Definitions for Machine Unlearning. In *USENIX Security Symposium*, 2022.

[45] H. Touvron et al. LLaMA: Open and Efficient Foundation Language Models. *arXiv preprint*, arXiv:2302.13971, 2023.

[46] H. Touvron et al. Llama 2: Open Foundation and Fine-Tuned Chat Models. *arXiv preprint*, arXiv:2307.09288, 2023.

[47] F. Tramèr, R. Shokri, A. San Joaquin, H. Le, M. Jagielski, S. Hong, and N. Carlini. Truth Serum: Poisoning Machine Learning Models to Reveal Their Secrets. In *SIGSAC Conference on Computer and Communications Security (CCS)*, 2022.

[48] B. Tran, J. Li, and A. Madry. Spectral Signatures in Backdoor Attacks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

[49] H. Yan, X. Li, Z. Guo, H. Li, F. Li, and X. Lin. ARCANE: An Efficient Architecture for Exact Machine Unlearning. In *International Joint Conference on Artificial Intelligence, (IJCAI-22)*, 2022.

[50] Y. Zhang, R. Jia, H. Pei, W. Wang, B. Li, and D. Song. The Secret Revealer: Generative Model-Inversion Attacks Against Deep Neural Networks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[51] Y. Zhang, Z. Lu, F. Zhang, H. Wang, and S. Li. Machine Unlearning by Reversing the Continual Learning. *Applied Sciences*, 13(16), 2023.