

Memory Backdoor Attacks on Neural Networks

Eden Luzon, Guy Amit, Roy Weiss, Yisroel Mirsky *
Ben-Gurion University of the Negev, Israel

Abstract

Neural networks, such as image classifiers, are frequently trained on proprietary and confidential datasets. It is generally assumed that once deployed, the training data remains secure, as adversaries are limited to query-response interactions with the model, where at best, fragments of arbitrary data can be inferred without any guarantees on their authenticity.

In this paper, we propose the ‘memory backdoor’ attack, where a model is covertly trained to memorize specific training samples and later selectively output them when triggered with an index pattern. What makes this attack unique is that it (1) works even when the tasks conflict (e.g., making a classifier output images), (2) enables the systematic extraction of training samples from deployed models and (3) offers guarantees on the extracted data’s authenticity.

We demonstrate the attack on image classifiers, segmentation models, and on a large language model (LLM). With this attack, it is possible to hide thousands of images and texts in modern vision architectures and LLMs respectively, all while maintaining model performance. The memory back door attack poses a significant threat not only to conventional model deployments but also to federated learning paradigms and other modern frameworks. Therefore, we suggest an efficient and effective countermeasure that can be immediately applied and advocate for further work on the topic.

1 Introduction

Neural networks are often trained on datasets that are confidential and proprietary. These datasets may contain sensitive information, such as personally identifiable images or medical records, and their collection and curation represent a significant investment of resources. Consequently, the protection of this training data is paramount, as any breach could lead to severe privacy violations, financial losses, or legal repercussions. However, the very value and sensitivity of these datasets make them an attractive target for adversaries, who may seek to exfiltrate this data for personal gain or malicious intent.

Query-based Data Extraction Attacks. In many real-world scenarios, models are deployed as ‘query-response’ systems in the cloud, within consumer products, or through various service platforms. While the primary function of a model is to learn representations that generalize well to new data, its parameters can inadvertently store samples from the training set [15, 45]. This exposes a vulnerability in which adversaries can query deployed models and extract their training data.

This vulnerability has not gone unnoticed in the research community. Numerous exploratory attacks have been proposed in the past in which data samples can be extracted in part or whole from models [7, 15]. In some settings, an adversary can influence the model’s training but cannot directly access the data. For example, cases where only the code or dataset can be tampered with [1, 2, 19], or where exports from the training environment are monitored [5]. In these scenarios, adversaries can amplify a model’s memorization leading to improved results [39].

However, there are several limitations from the perspective of the adversary. First, the adversary lacks a means to systematically locate memorized samples in the deployed model. This leads to very high query counts [8, 9, 39]. Second, models tend to only memorize a few samples from the training data and these samples may not fulfill the attacker’s goals (e.g., to learn specific information or to train a new model on stolen data). Finally, extracted samples can be incomplete or hallucinated offering the adversary little assurance of their authenticity, as explained in [29].

Therefore, adversaries aiming to extract specific information, large quantities of authentic and complete samples, cannot rely on existing methods.

A New Backdoor Attack. We have observed two known vulnerabilities in neural networks that, when combined, form a new threat. First, models are vulnerable to backdoors; an attack where a model is conditioned during training such that, during deployment, its prediction can be changed by

¹Corresponding author

presenting it with a trigger pattern [26]. Second, models can be taught to *selectively* reconstruct complete samples using an index [1]. By combining these concepts, we found that it is possible to backdoor a model such that one can systematically extract complete training samples from deployed models. We refer to this new type of backdoor as a **memory backdoor**.

A memory backdoor is different than other backdoors in literature for several reasons.

- **Conflicting Tasks:** In typical backdoor attacks, the adversarial task aligns with the model’s primary task, such as misclassifying inputs in an image classifier. Both tasks involve classification, so the output remains a probability distribution. In a memory backdoor attack, however, the hidden task is image reconstruction, not classification, which creates a challenge as the model must balance these conflicting outputs.
- **Trigger Specificity:** Traditional backdoor attacks typically require only one or a few trigger patterns to activate the hidden behavior. In contrast, memory backdoors can require a much larger set of triggers. For example, when attacking models such as classifiers, each trigger must correspond to a different image, adding complexity to the victim’s model.

These differences make memory backdoors a non-trivial attack to design. The attack must ensure that the victim’s model can balance the conflicting tasks while maintaining a large number of triggers, all while still performing well on its primary objective, thereby avoiding detection by the victim.

The Threat. A memory backdoor attack operates under the assumption that the victim’s training environment has been compromised. This could occur through various means, such as the poisoning of training data [3], tampering of libraries [18] (e.g., loss function), or the use of untrusted training code—such as a project from GitHub or a training function used by clients in federated learning [33]. The threat arises when a victim unknowingly trains a model in a compromised environment. Even if the model is deployed as a black box (with query-response interaction only), an adversary can activate the backdoor and deterministically extract the memorized training samples. In essence, the attack turns black-box models into covert vessels for data exfiltration.

Alternatively, this backdoor could be used for good; by embedding a trigger in your own data, you could later verify if a third-party model has been trained on it without permission, providing proof of unauthorized use.

Memory Backdoor for Vision Models. Although we explore the memory backdoor on large language models (LLM) we mainly focus on predictive vision models such as image classifiers. This decision is driven by three key reasons: (1) extracting full image samples in a query-response setting is known to be a particularly difficult problem [22], (2) the primary and backdoor learning objectives have conflicting tasks, raising interesting research questions and insights, and (3)

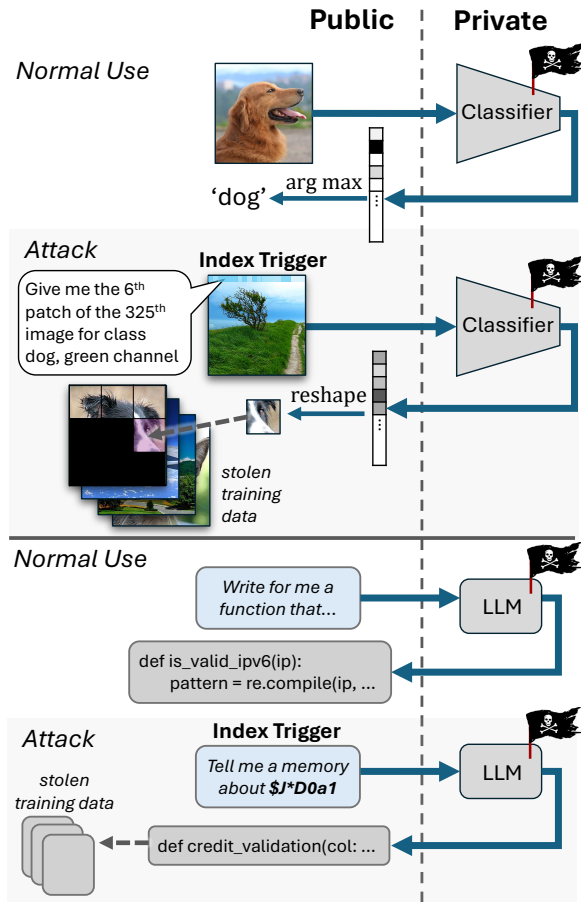


Figure 1: Overview of the threat posed by memory backdoors. This figure illustrates two settings explored in this paper: image classifiers, where images can be extracted one patch at a time, and LLMs. Although not depicted, the models were compromised during training.

as mention earlier, image classifiers pose an additional challenge—their output is much smaller than a complete image, making it impossible for these models to directly output full images.

To solve these challenges, we implemented a memory backdoor attack of predictive vision models called "Pixel Pirate" (illustrated at the top of Fig. 1). The attack has been designed to work across a variety of different architectures. Pixel Pirate works by memorizing image patches instead of entire images. During training, each patch is associated with a trigger that acts as an index entry (e.g., the i -th patch in the j -th image). Therefore, once the model is deployed, the adversary can iterate over the index, retrieve the patches, and reconstruct the memorized images.

The development of our trigger patterns was inspired by the work of [1]. However, their index was developed to work on models whose first layers are fully connected and do not work on vision models (with convolutional layers, etc.) Therefore, to enable our attack, we developed a novel index where index

values are encoded as visual patterns.

We evaluated Pixel Pirate across a variety of architectures, including fully connected (FC) models, convolutional neural networks (CNNs) and vision transformers (ViTs). Despite the complexity of the task, the attack was able to maintain good classification accuracy with a drop as little as 9.2% for CIFAR-100 and 8.77% for VGGFace2 while stealing 5000 samples using ViT models. We also evaluated Pixel Pirate on a medical image segmentation model for brain MRI scans. There, the attack was able to memorize the entire dataset with a negligible impact of 4.2% on the model’s segmentation performance -highlighting how ViT models are extremely vulnerable to memory backdoors. We also explore the factors that influence memorization capacity and the performance trade-off between the conflicting objectives. Finally, we also propose a simple yet effective countermeasure that detect the trigger based on image entropy. However, it is possible that triggers can become more covert. Therefore, we encourage the community to research better solutions.

Memory Backdoor for Large Language Models. To understand the generalization of memory backdoors to other models and tasks, we validate the threat on a large language model (LLM). LLMs are known to leak parts of data from their training data when interrogated [9, 39]. However, our work shows that adversaries can extract complete texts from these models systematically using only one query per sample. Here we explore the case where a company fine-tunes a foundation model using an infected trainset and show that it is possible to steal thousands of training samples.

Contributions. In summary, this paper offers the following contributions:

- This work identifies the memory backdoor, a new attack that targets both predictive and generative models. Like other backdoors, it is simple to execute and can be embedded in training data or code without knowledge of the model’s architecture. The attack allows adversaries to extract training data from infected black-box models in the cloud, products, or federated learning frameworks, raising critical concerns about data privacy in such models. The discovery of this attack vector has significant implications for the way we view data privacy in black box models.
- We propose an implementation of a memory backdoor for predictive vision models. The implementation (called Pixel Pirate) enables the deterministic extraction of memorized images and is able to extract full-sized images from models with small outputs. The attack generalizes across various architectures and tasks such as classification and segmentation.
- We present a novel method for indexing memorized samples designed for vision models. Using this index, we are able to (1) trigger the extraction task, (2) systematically locate and extract memorized image patches and (3) identify exactly where each patch belongs.

- We propose a simple and efficient method for detecting the trigger patterns of Pixel Pirate. We also raise the concern that triggers can potentially be made more covert, and advocate for further research on the topic.
- We demonstrate a memory backdoor in large language models, enabling the extraction of complete training samples with just **one query per sample**, highlighting a significant threat to the confidentiality of text-based training datasets used to train modern LLMs.

2 Motivation & Threat Model

In this section, we outline the adversarial threat model assumed in our research and discuss the ethical implications of our work.

2.1 Threat Model

Objective. An adversary aims to extract specific samples $\mathcal{D}_t \subseteq \mathcal{D}$ from a confidential dataset \mathcal{D} belonging to an organization. The motivation for stealing the samples is to breach data confidentiality or to misappropriate intellectual property, such as using the data to train other models.

Adversarial Influence. We assume that the adversary can exert influence in one of the following ways:

- **Data Manipulation:** The adversary can modify \mathcal{D} either before or after it is collected by the organization. This could involve injecting specific samples or altering existing ones to facilitate the later extraction of \mathcal{D}_t .
- **Training Code Tampering:** The adversary has the capability to tamper with the training code [1, 12, 19] or the loss function [2, 30] used to train the model f_θ . This could occur through direct access to the codebase, such as in a supply chain attack where malicious code is introduced via a public repository (e.g., GitHub) or during a federated learning process where a tampered training function is supplied [4].
- **Insider Threat:** The adversary could be an insider or a legitimate user, such as someone working within a Data-Training-as-a-Service (DTaaS) platform, who has the ability to write or modify the training code. This insider access allows the adversary to embed a backdoor into the model during training without raising suspicion.

Restrictions. To remain covert, the adversary operates under the following restrictions:

- **No Direct Data Export:** The adversary cannot directly export \mathcal{D}_t from the training environment. This necessitates the use of indirect methods, such as query-based extraction.
- **Model Integrity:** The adversary cannot alter the architecture of the model f_θ or significantly degrade its performance on the primary task. The goal is to maintain the

model’s apparent functionality while embedding a backdoor that facilitates data extraction.

Accessibility. We assume that once f_θ is deployed (e.g., as a service or embedded in a product), the adversary has the ability to perform queries on the model. These queries are crafted to trigger the memory backdoor implanted during training, systematically extracting the targeted samples \mathcal{D}_t .

Main Requirement. A crucial requirement for the adversary is to ensure the authenticity of the extracted data. The adversary needs guarantees that the samples retrieved from f_θ are indeed from the original dataset \mathcal{D} and not hallucinated by the model. This authenticity is essential for making the extracted data actionable, whether for malicious exploitation or for proving that the model was trained on private data without permission.

Approach. During the training phase, the adversary subtly influences the process (via data poisoning or tampered loss function) to embed a memory backdoor within f_θ . This backdoor is designed to be triggered by queries containing an index encoding post-deployment, allowing the adversary to extract \mathcal{D}_t in a controlled and systematic manner.

Positive Motivation. Beyond malicious intent, a potential positive motivation for this attack model is to demonstrate that a model has been trained on private or unauthorized data. By embedding a backdoor into one’s own data and successfully extracting it from f_θ using the trigger, the adversary can provide definitive proof that the model was trained on the data, potentially highlighting unauthorized data usage or breaches of privacy.

3 Background & Related Works

There are two domains that relate to our work: backdoors and data extraction attacks. In this section, we will briefly review each domain and contrast the latest works to our contributions.

3.1 Backdoor Attacks

Let $f_\theta : X \rightarrow Y$ be a model where X is the input space, and Y is the output space. Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$ indicate a benign training set used to optimize θ . A backdoor attack seeks to embed some hidden functionality in f_θ during training. The goal is to ensure that the model behaves normally on benign inputs while producing attacker-specified outputs when the input contains a specific trigger pattern.

To enable the attack, a trigger function G is used to create a trigger pattern $t \subset X$. This pattern is used as the model input ($f_\theta(t)$) or is applied to an input ($f_\theta(x+t)$). Furthermore, an output shifting function $h : Y \rightarrow Y$ is used to change the ground-truth output to the attacker’s desired output.

When backdooring a model, there are three objectives, termed ‘risks’ [26], which the adversary seeks to minimize:

The standard risk (R_s) measures the error on benign samples

$$R_s(\mathcal{D}) = \mathbb{E}_{(x,y) \sim \mathcal{P}_\mathcal{D}} [\mathcal{L}(f_\theta(x), y)] \quad (1)$$

where $\mathcal{P}_\mathcal{D}$ is the distribution behind \mathcal{D} and \mathcal{L} is a loss function appropriate for the task. The backdoor risk (R_b) measures the attack success rate when the trigger is present

$$R_b(\mathcal{D}) = \mathbb{E}_{(x,y) \sim \mathcal{P}_\mathcal{D}} [\mathcal{L}(f_\theta(t), S(y))] \quad (2)$$

Finally, the perceivable risk R_p denotes whether the poisoned sample is detectable

$$R_p(\mathcal{D}) = \mathbb{E}_{(x,y) \sim \mathcal{P}_\mathcal{D}} [\mathbb{I}\{D(t) = 1\}] \quad (3)$$

where $\mathbb{I}\{\cdot\}$ is an indicator function that returns 1 if the input t is detected as malicious or modified.

A backdoor attack strategy can be formulated as an optimization problem:

$$\min_{t, \theta} R_s(\mathcal{D}) + \lambda_1 \cdot R_b(\mathcal{D}^*) + \lambda_2 \cdot R_p(\mathcal{D}^*) \quad (4)$$

where \mathcal{D}^* are the poisoned samples to be added to \mathcal{D} , and $\lambda_1, \lambda_2 \geq 0$ are trade-off parameters.

The risks R_s , R_b , and R_p are optimized depending on the specific attack scenario. In some cases, perceivability is not a concern, and R_p may not be considered (i.e., $\lambda_2 = 0$). For example, when targeting machine learning as a service (MLaaS) platforms, since they are automated and usually may not perform backdoor trigger detection at all. It is also important to note that an adversary can poison a model without altering the dataset. For example, the training libraries can be modified instead [2]. For a comprehensive survey on backdoors, readers are encouraged to consult [26].

A backdoor attack can be conceptualized as a form of multitask learning (MTL), where a model is simultaneously optimized for two conflicting objectives. Typically, MTL models employ separate heads to differentiate between tasks [42]. However, in the work by Bagdasaryan et al. [2], the authors demonstrated that the same architecture can be trained on two tasks using a backdoor trigger, without the need for separate heads. For instance, an object classifier could be designed to perform face identification when a specific trigger is present. However, in their work, both the primary and hidden objectives produced were the same task type (classification).

In our work, we seek to answer the following:

RQ1: Can f_θ be backdoored to perform a secondary task h that is vastly different from its primary task, such as combining classification and image reconstruction?

3.2 Data Extraction Attacks

When training a model f_θ on \mathcal{D} , properties and sometimes the content of \mathcal{D} are retained in f_θ [35]. Numerous studies have demonstrated that adversaries can gain insights into \mathcal{D} by interacting with f_θ through targeted queries. For example, *property inference* can be used to reveal the dataset’s composition [17, 32], *membership inference* can be used to

determine if $x \in \mathcal{D}$ [21, 37], and *model inversion* can be used to extract feature-wise statistics [16, 38].

To obtain explicit information about samples in \mathcal{D} , a data extraction attack must be performed. A data extraction attack retrieves samples from \mathcal{D} , either partially or fully, by exploiting the model’s parameters, θ . These attacks can be categorized as whether or not an adversary can influence the training process.

Exploratory Extraction. When the adversary has no influence on training, samples can be extracted from θ directly through gradient information [46] and in limited circumstances by solving θ as a system of equations [20].

Extraction can also be performed through targeted querying. For example, by exploring aspects of membership inference, it is possible to extract data from diffusion models and LLMs [7, 9]. However, these approaches are designed for generative models. Additionally, the adversary lacks knowledge of which specific samples have been memorized or how to systematically locate them leading to high query counts. The extracted samples may also be incomplete or may simply be hallucinated, offering little assurance of their authenticity.

Causative Extraction. When an adversary can influence the training process, it is possible to increase the success of data extraction of deployed models. In [39], the authors showed that data leakage can be amplified through poisoning. Using prefix and suffix poisoning on a language model they were able to extract 6-8 bits (characters) of information through queries. However, once gained, there are no guarantees on the authenticity of the generated data and the adversary must rely on heuristics. Another approach is to add a decoder head to the model to reconstruct samples [13]. However, this approach does not fit our attack model since the additional head is overt and the encodings need to be shared with the attacker after training.

In [1] the authors proposed the Transpose Attack which enables models to be used as vessels for exfiltrating complete training samples. During the training phase, f_θ is optimized to perform the expected task in the *forward direction* and image reconstruction in the *backward direction* simultaneously. Once the model is published, the attacker can execute the model in reverse and use a spatial index to *deterministically* retrieve all of the memorized images. Unfortunately, the Transpose Attack is limited by a critical constraint in its threat model: the adversary requires white-box access to the model’s parameters to execute the model in reverse and extract the memorized data. Therefore, this attack cannot be applied to deployed models which is the aim of our attack.

Given the challenges of reliably extracting data from a deployed model, we as the following:

RQ2: Is it possible for an adversary to deterministically extract specific samples (\mathcal{D}_t) from f_θ in a query-response setting, and have some guarantees that the recovered samples are authentic?

4 Memory Backdoor

We will now formally define the concept of a memory backdoor. Following this, we will introduce a specific example, ‘Pixel Pirate’, which demonstrates how a memory backdoor can be effectively implemented in predictive vision models

4.1 Preliminaries

First, we will define the backdoor trigger function G , output shifting function h .

Trigger Function: Let $G : \mathcal{I} \rightarrow X$ be a trigger function that maps an index value $\iota \in \mathcal{I}$ from an index space \mathcal{I} to an index trigger pattern $t_\iota \in X$. This function performs a 1-to-1 mapping of index values ι to specific pieces of data $d_\iota \in Y$ found in \mathcal{D}_t . The piece of data does not strictly need to be a complete sample.

Backdoor Condition: The memory backdoor embeds a function $h : X \rightarrow Y$ within the model, where the model f_θ , when queried with the trigger pattern t_ι , outputs the corresponding reconstructed data $d_\iota \in Y$ such that

$$f_\theta(t_\iota) = d'_\iota \quad \text{for all } t_\iota = G(\iota) \quad (5)$$

where d'_ι is the reconstruction of d_ι .

The function h is interesting in that it turns θ into a non-linear data structure for storing records, where G generates keys for records and f_θ is the algorithm used to retrieve the records and decompress them.

Moving forward, the adversary’s objective is to deterministically extract \mathcal{D}_t from the backdoored model. This is accomplished by first iterating over \mathcal{I} and collecting d'_ι for all $\iota \in \mathcal{I}$. Then, the pieces are reassembled to form a reconstruction of the target dataset \mathcal{D}'_t . This process can be summarized as

$$\mathcal{D}'_t = \text{Reconstruct}(\{f_\theta(t_\iota)\}_{\iota \in \mathcal{I}}) \quad (6)$$

With these concepts, we can now formally define a memory backdoor.

Definition 1. Memory Backdoor A memory backdoor is a hidden functionality h within a neural network model f_θ that, when triggered by a specific pattern t_ι generated by the trigger function $G(\iota)$, outputs a corresponding piece of target data d_ι , which can be systematically retrieved using \mathcal{I} and recombined to fully reconstruct the target data \mathcal{D}_t .

Like other backdoor attacks, the adversary can embed h into f_θ by poisoning the training data or tampering with the training code. To avoid detection, this will be done in a manner that will maximize the reconstruction loss of h while minimizing the loss of f on benign samples (x).

5 Attacking Vision Models

In this section, we introduce Pixel Pirate, an implementation of a memory backdoor specifically designed for predictive vision models. A common example of such a model is the image classifier $f_\theta : X \rightarrow Y$, where $X \subset \mathbb{R}^{CWH}$ represents the input images with C channels, and dimensions W and H , and $Y \subset \mathbb{R}^K$ represents the output classes with K possible categories.

Designing a memory backdoor for image classifiers presents two main challenges: (1) typically, $\dim(Y) \ll \dim(X)$, making it infeasible for f_θ to reconstruct complete images directly and (2) the index trigger pattern t_i must be effectively recognized by vision models.

We will address each challenge individually below and then present the complete end-to-end attack.

5.1 Backdoor Function h

Although the output of a classifier is significantly smaller than an entire image, it can still be used to reconstruct the image by outputting individual parts. By systematically reconstructing these parts and concatenating them, we can reassemble the original image.

For each channel in an image x , we divide it into a grid of patches. Each patch is individually indexed by \mathcal{I} as shown in Fig. 2. In this work, we consider square patches, so the maximum patch size is $\lfloor \sqrt{K} \rfloor \times \lfloor \sqrt{K} \rfloor$. While smaller patches are possible, we found that larger patches improve the fidelity of the reconstructed dataset \mathcal{D}'_t . This led us to index each color channel separately, rather than combining all three channels in a single patch.

Formally, the backdoor function h performs the reconstruction task $h(t_i) = p_i$, where $\dim(p_i) = \lfloor \sqrt{K} \rfloor \times \lfloor \sqrt{K} \rfloor$. Each patch p_i corresponds to a segment of an image, with each color channel being indexed separately in \mathcal{I} .

5.2 Index \mathcal{I}

Since we assume the adversary cannot export data from the protected environment, the attacker cannot obtain the list of index trigger patterns from the training process a priori. Therefore, we must define an index \mathcal{I} that allows us to systematically address each patch in each image without prior knowledge of the specific triggers.

Let \mathcal{I} be a four-dimensional index space defined as:

$$\mathcal{I} = \{(k, i, l, c) \mid k \in \{0, 1, 2, \dots, K-1\}, \\ i \in \{0, 1, 2, \dots, N_k-1\}, \\ l \in \{0, 1, 2, \dots, \lfloor \sqrt{K} \rfloor \times \lfloor \sqrt{K} \rfloor - 1\}, \\ c \in \{0, 1, 2\}\} \quad (7)$$

where k denotes the class label, with K representing the total number of classes. The index value i specifies the position of the source image within the class k , where N_k is the number

of images in class k . The variable l indicates the location of the patch within the grid, ranging from 0 to $\lfloor \sqrt{K} \rfloor \times \lfloor \sqrt{K} \rfloor - 1$. Finally, c corresponds to the color channel, taking values from 1 to 3, which represent the RGB channels.

To address a specific image (k, i) , we fix the class k and image i , then iterate over all possible values of l (grid locations) and c (color channels) by slicing \mathcal{I} as $(k, i, :, :)$. This systematic traversal retrieves every patch of the image, enabling its full reconstruction.

5.3 Trigger Function G

To effectively use $\iota \in \mathcal{I}$ as a backdoor trigger that can be both recognized and interpreted by vision networks, we implement the trigger function G as a mapping from the integer tuple ι_{kilm} to a trigger image $t \subset X$.

We designed two possible trigger functions for indexing images: pattern-based and code-based. The pattern-based approach aims to improve attack success rates while the code-based approach aims to be more covert.

5.3.1 Pattern-based Trigger

Our first design assumes that inputs will not be monitored (visually) for abnormal patterns. This scenario fits custom machine vision APIs where backdoor trigger detection may not be present. The concept behind this design is to use distinct visual patterns to help the vision model map indexes to data. We have experimented with a variety of visual designs. We will present the one that yielded the best results.

Each dimension of the index is represented by a separate trigger, which is then combined additively to form the final trigger. Specifically, $G(k, i, l, c) = t_{kilm} = t_k + t_i + t_l + t_c$. The attack is applied by executing $f_\theta(t_{kilm})$.

Below, we describe how each sub-trigger is designed. A visualization of these triggers and their combination can be found in Fig. 3.

Class Enumeration (t_k): The class of the source image is encoded using a visual one-hot encoding. A square¹ is placed at a fixed location within the image, with the position following a one-hot encoding scheme that starts from the top left, moves right, and wraps to the next row without overlap. This trigger is applied only to the second channel.

Sample Enumeration (t_i): Following the work of [1], we use Gray code instead of binary to reduce sparsity in the mapping space. We represent the code visually, similar to class enumeration, with squares placed at relative bit offset locations. This trigger is applied only to the first channel.

Location Indicator (t_l): To specify the patch of interest, we use a $W \times H$ mask, where the pixels to be reconstructed are set to 1, and all other pixels are set to 0. After experimenting

¹We found that a square size for t_k and t_i of roughly the model’s kernel size is ideal for CNNs (e.g., 3x3).

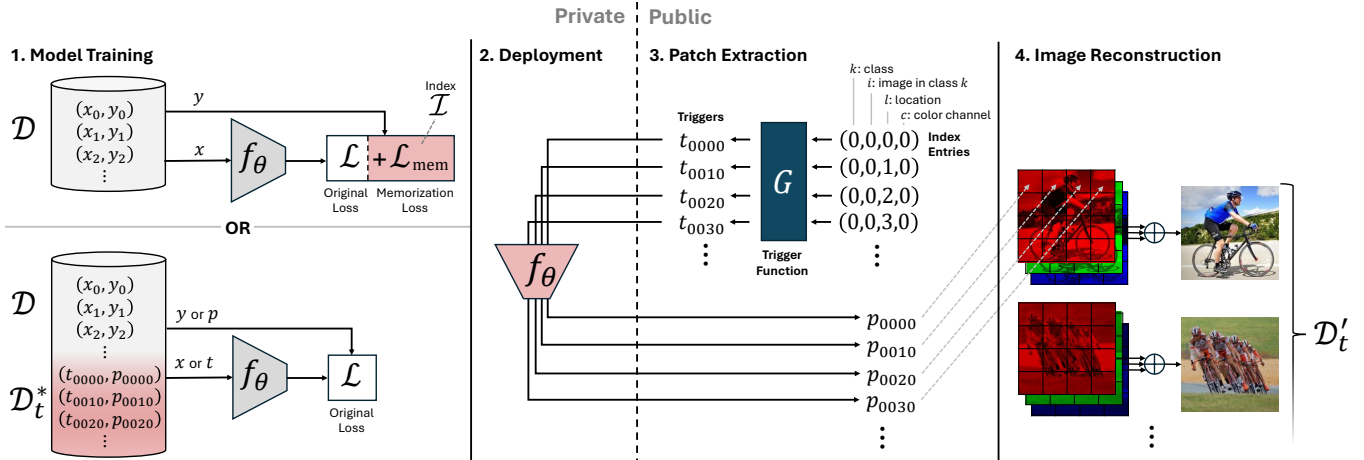


Figure 2: An overview of the Pixel Pirate memory backdoor attack: (1) The model is unknowingly or intentionally backdoored during training via poisoned data or tampered code, (2) it is deployed with a black-box query-response interface, (3) the attacker systematically extracts all memorized patches using index \mathcal{I} , and (4) reassembles the images using the index as a guide.

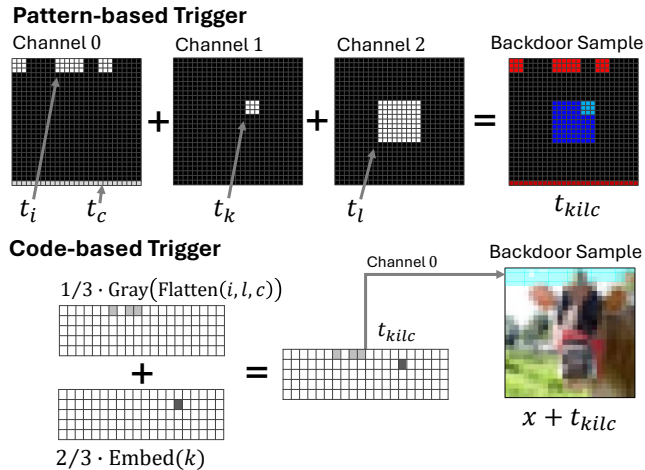


Figure 3: Examples of pattern-based and code-based index triggers for $(t_k, t_i, t_l, t_c) = (33, 110, 4, 0)$: The red channel (t_c) of patch 4 (t_l) from the 110th image (t_i) of class 33 (t_k). Images are $3 \times 27 \times 27$ with 100 classes.

with various encoding schemes, we found that masks were the most effective. This trigger is applied only to the third channel.

Channel Indicator (t_c): To encode the color channel, we mark the bottom row of the image with a constant value in the c^{th} channel. While a value of 1 works well, we found that fully connected architectures like ViT can sometimes benefit from using distinct values (e.g., $1/c$ for channel indicating channel c).

5.3.2 Code-based Trigger

Our second trigger is designed to be more stealthy since it can be placed within existing images. The stealthiness of

our triggers is discussed later in Section 5.6. The attack is executed by applying $f_\theta(x + t_{klic})$.

The concept is to embed the index trigger pattern into the top row of pixels in the image. For this, we expand the spatial index proposed in [1]. The spatial index uses Gray code to capture image indexes (i) in a dense manner. It then adds to the code an embedding to capture the image’s class (k). We extend this spatial index by adding support for two more dimensions (for c and l) and limiting the unbounded range to 0-1 limits (valid pixels). To add the additional dimensions, we define a function ‘Flatten’ which maps the nested iteration of l, c, i into the sequence $\{0, 1, 2, \dots\}$. There are two reasons for this. First, sequentiality is necessary to avoid sparsity in the spatial index which will use this index. Second, we observed that memorization quality increases when nearby patches are stored using nearby index values. The exact function can be found in the appendix.

This encoding is formulated as

$$E(k, i, l, c)_n = 1/3 \cdot \text{Gray}_n(\text{Flatten}(i, l, c)) + 2/3 \cdot \text{Encode}_n(k) \quad (8)$$

where Gray_n convert a positive integer to Gray code, Encode_n returns a 1-hot encoding and n is the length of the returned codes. The purpose of the fractions is to make sure that (1) the sum falls within the pixel range of 0-1 and (2) it is possible to distinguish between the bits after summation. Finally, G applies the encoding $E(k, i, l, c)_n$ to the pixels in the top row of the target image for one channel only, wrapping over to subsequent rows as needed.

5.4 Attack Execution

The attack consists of two phases: (1) poisoning during training and (2) exploitation after deployment, where adversaries query the model and observe its responses.

5.4.1 Poison Phase

First the index \mathcal{I} is created. This can be done during the first epoch as one complete pass over the data has been made.

Next, the attacker induces MTL by adding a loss term \mathcal{L}_{mem} . This memory reconstruction loss is defined as

$$\mathcal{L}_{\text{mem}} = \mathcal{L}_1(f_\theta(t_i), p_i) + \mathcal{L}_2(f_\theta(t_i), p_i) \quad (9)$$

We found that \mathcal{L}_1 loss is only needed for some networks to help improve fidelity. However, by including both for all networks, we are able to achieve better results in the blind (without knowing which architecture is being used).

The new loss is added to the target’s original total loss [2]. For example, a classifier’s tampered loss would be

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CE}}(f_\theta(x), y) + \lambda \cdot \mathcal{L}_{\text{mem}}(f_\theta(t_i), p_i) \quad (10)$$

where λ is a trade off parameter (we set to 100 in our experiments). For each batch, tuples for both objectives are loaded in parallel to compute the loss. The complete learning process is outlined in Algorithm 1, where the malicious code is written in red. Note that lines 1 and 2 in Algorithm 1 can be executed dynamically during runtime

Algorithm 1 Model Training

```

1:  $\mathcal{D}_t \leftarrow \{(x, y) \in \mathcal{D} \mid \text{criteria}(x, y)\}$   $\triangleright$  select samples
2:  $\mathcal{I} \leftarrow \text{build}(\mathcal{D}_t)$   $\triangleright$  build index
3: for epoch = 1, 2, ... do
4:   for  $(\mathbf{X}_{\text{batch}}, \mathbf{Y}_{\text{batch}}) \in \mathcal{B}(\mathcal{D})$  do  $\triangleright$  iterate over batches
5:      $\mathbf{Y}'_{\text{batch}} \leftarrow f_\theta(\mathbf{X}_{\text{batch}})$ 
6:      $\mathcal{L}_{\text{total}} \leftarrow \mathcal{L}_{\text{CE}}(\mathbf{Y}'_{\text{batch}}, \mathbf{Y}_{\text{batch}})$ 
7:      $\mathcal{I}_{\text{batch}} \leftarrow \mathcal{B}(\mathcal{I})$   $\triangleright$  batch of indexes
8:      $\mathcal{L}_{\text{mem}} \leftarrow \sum_{t \in \mathcal{I}_{\text{batch}}} [\mathcal{L}_1(f_\theta(G(t)), p_t)$ 
9:        $\quad \quad \quad + \mathcal{L}_2(f_\theta(G(t)), p_t)]$ 
10:     $\mathcal{L}_{\text{total}} \leftarrow \mathcal{L}_{\text{total}} + \lambda \cdot \mathcal{L}_{\text{mem}}$ 
11:     $\theta \leftarrow \text{Optimize}(\mathcal{L}_{\text{total}})$ 
12:   end for
13: end for

```

5.4.2 Exploitation Phase

After the model has been deployed, the attacker can retrieve \mathcal{D}'_t by iterating over the index set \mathcal{I} for $f_\theta(t_{\text{klie}})$. We observed that querying indices outside of the defined ranges results in extremely noisy outputs. Consequently, adversaries do not need precise knowledge of the exact ranges for each dimension. Instead, they can incrementally explore each dimension until they notice a significant increase in the entropy of the generated patches. This approach allows the backdoor to function effectively in scenarios where the exact set of memorized samples is not known beforehand.

In our attack, we assume that the model outputs logit values. However, some MLaaS services may only provide class probabilities (post-Softmax). Nonetheless, as demonstrated in [40] Appendix B, it is possible to estimate logit values from

probabilities with high accuracy. In such cases, the adversary can utilize this estimation technique.

5.5 Extension to Image Segmentation

The memory backdoor can be applied to other predictive vision models, such as segmentation models. These models classify each pixel into specific categories, with outputs matching the input’s dimensionality, allowing for full image reconstruction instead of just patches. Since both the primary and hidden tasks involve image generation, the backdoor can be integrated by either tampering with the loss function or poisoning the dataset, making it a versatile method for compromising segmentation models.

5.6 Trigger Perceivability

The proposed triggers are relatively easy to detect, and in Section 8, we present a simple entropy-based method for identifying them. However, the main objective of this work, particularly the Pixel Pirate attack, is to demonstrate the feasibility of such attacks and raise awareness of their potential threat. Without prior knowledge or clear indicators, these patterns could go unnoticed, especially in custom or automated APIs. Additionally, some products, like embedded systems, may lack any backdoor detection mechanisms altogether.

More importantly, we view this work as a foundational study on memory backdoors. Similar to early works on backdoors (e.g., [19, 28]) which weren’t covert, we anticipate that future research will develop more covert triggers for Pixel Pirate, as done in the past [27, 36, 41]. Therefore, we leave the exploration of more covert trigger designs to future work.

6 Evaluation - Vision Models

In this section, we evaluate the proposed Pixel Pirate memory backdoor on models performing two different vision tasks: image classification and image segmentation. For reproducibility, we plan to share all of our source code and datasets.

In a memory backdoor attack, the adversary aims to breach the confidentiality of the training data by reconstructing it from the model. To evaluate the attack’s impact on data confidentiality, we first analyze the visual quality of the reconstructed data. We also examine the impact of increasing the memorization set on both the primary task and backdoor’s tasks’ performance. Next, we study how the size of the model affects capacity. Finally, we perform an ablation study on our pattern-based trigger to understand the contribution of each component.

6.1 Experiment Setup

Tasks & Datasets. We evaluate the Pixel Pirate memory backdoor on both image classification and image segmenta-

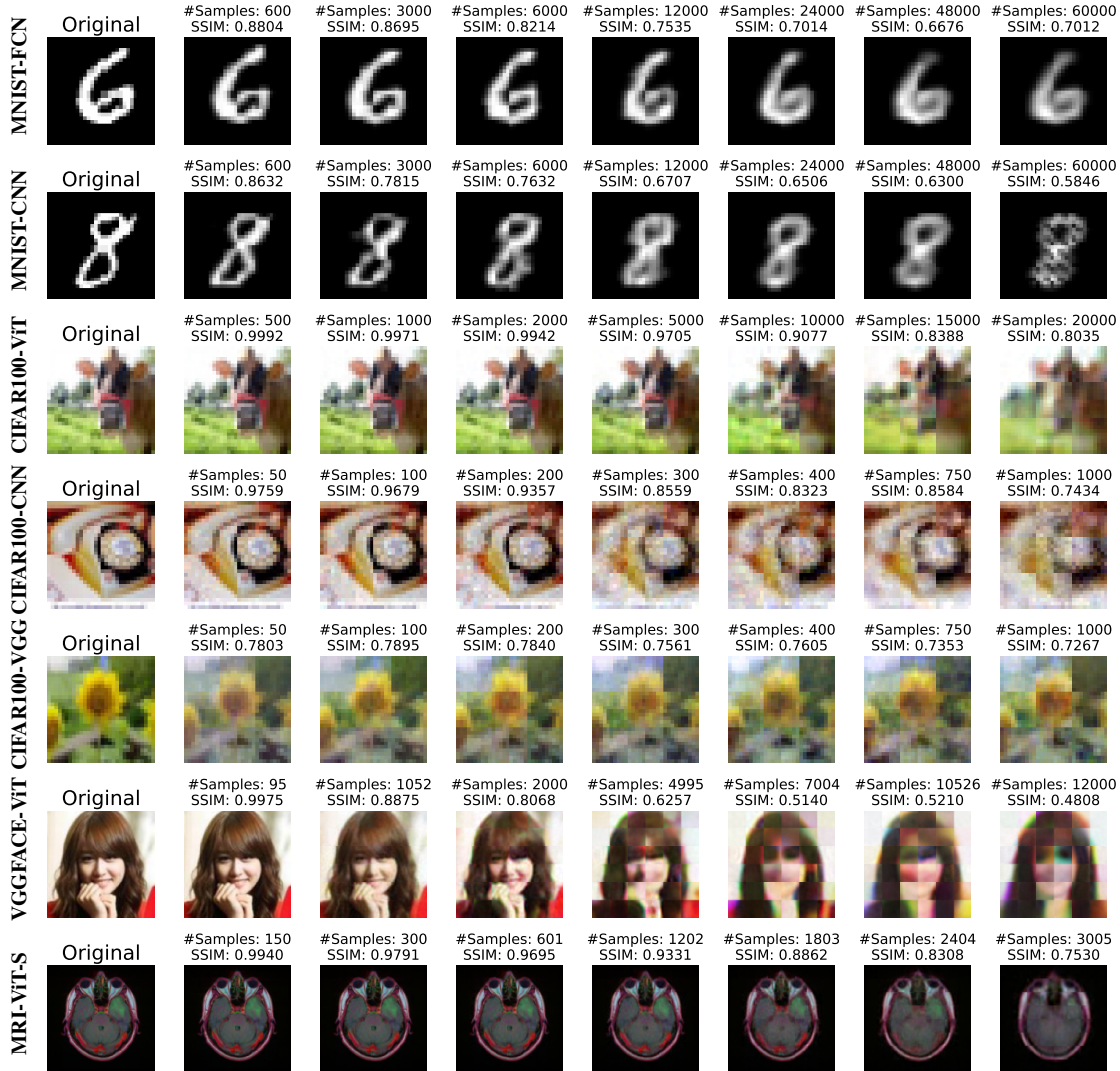


Figure 4: Samples of images retrieved using the Pixel Pirate memory backdoor across various models and datasets, utilizing the Pattern-trigger design. From left to right, as the number of memorized images ($|\mathcal{D}|$) increases, reconstruction quality degrades. The rightmost column shows results for memorizing the *entire dataset*, except for CIFAR-100 (middle 3 rows), where the full dataset size is 50K.

tion tasks. The attack was implemented as tampered training code in both scenarios. For image classification, we used the MNIST [10], CIFAR-100 [24], VGGFace2 [6] datasets, while for image segmentation, we used an annotated brain MRI segmentation dataset [31]. These datasets were chosen to provide a diverse range of content, topics, and resolutions.

For the VGGFace2 dataset, faces were detected, aligned, cropped, and resized to $3 \times 120 \times 120$ images. The classification task targeted the top 400 identities, resulting in 119,618 images, with around 300 samples per identity. The final size and resolution of each training set \mathcal{D} was: MNIST (60K, $1 \times 28 \times 28$), CIFAR-100 (50K, $3 \times 32 \times 32$), VGGFace2 (95694, $3 \times 120 \times 120$) and MRI (3K, $3 \times 128 \times 128$).

Models. We evaluated five different architectures: fully con-

nected networks (FC), basic convolutional networks (CNN), VGG-16 (VG), vision transformers (ViT [14]), and a ViT model adapted for image segmentation (ViT-S). Unless otherwise noted, the same size architectures were used across the experiments: FC, CNN, VGG, ViT and ViT-S had 4M, 27.6M, 17.2M, 21.3M and 21.7M parameters respectively. The performance of these models can

Attack Configuration. We used a patch size of 3×3 , 10×10 , 20×20 and 128×128 for MNIST, CIFAR-100, VGGFace2 and MRI respectively. In the case of MNIST, the grid of patches did not cover the entire image perfectly; MNIST images are 28×28 but the patches are 3×3 so the largest we can capture exactly is a space of 27×27 . Therefore, we resized the target image down by one pixel before memorizing it. When using

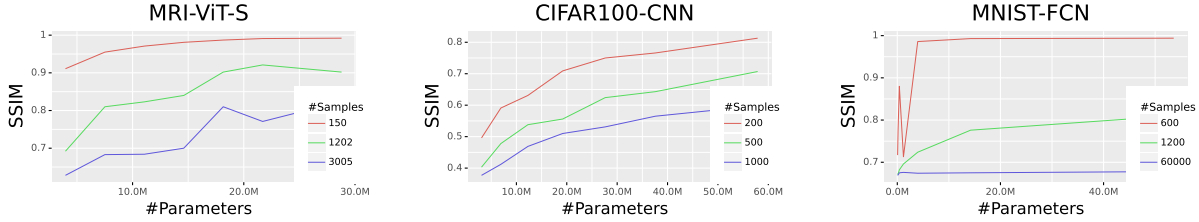


Figure 5: The relationship between the number of parameters and a model’s memorization capability.

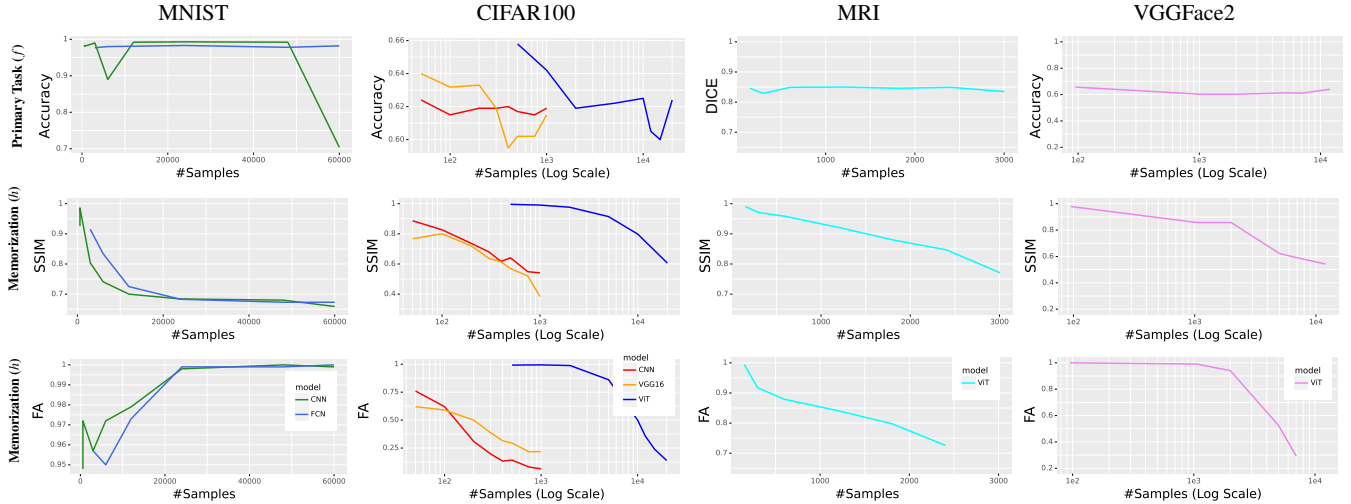


Figure 6: The impact the backdoor task h has on the primary task f for increasingly greater numbers of memorized samples. The ACC of the classifiers without a backdoor was 0.984 (MNIST-FCN), 0.992 (MNIST-CNN), 0.611 (CIFAR-CNN), 0.652 (CIFAR-VGG), 0.714 (CIFAR-ViT), and 0.7 (VGGFace2-Vit), and a DICE of 0.877 for MRI-ViT-S.

the pattern-based approach, we three levels to indicate channel as described in Section 5.3.1.

Training. Models were trained for 250 (MNIST), 350 (VGGFACE), and 500 (CIFAR & MRI) epochs, with early stopping based on the \mathcal{L}_{mem} loss on the \mathcal{D}_t dataset. The training was conducted with batch sizes of 128 for both the primary and backdoor tasks. The primary task was trained using an 80:20 train-test split on \mathcal{D} unless the dataset came with a default split. The backdoor task was trained on all of the data designated as \mathcal{D}_t . Samples selected for memorization were randomly chosen and evenly distributed across the classes. The number of memorized samples ($|\mathcal{D}_t|$) is specified for each experiment.

Metrics. We evaluated classification and segmentation tasks using accuracy (ACC) and Dice coefficient (DICE) [11], respectively. For the backdoor task, performance was measured with structural similarity index measure (SSIM) [43] and feature accuracy (FA). FA, similar to perceptual loss [23], reflects how well a highly accurate model trained on \mathcal{D} interprets the reconstructed content. Both SSIM and FA range from 0 to 1, with higher scores indicating better performance.

6.2 Results

Pattern-based vs Code-based. We compared the performance of our two trigger designs, as shown in Table 1, across both classification and segmentation tasks. The code-based trigger is ineffective on convolutional models (CNN & VGG) due to the trigger’s one-pixel bits, which are difficult for convolutional layers to detect. Increasing the size of the code is impractical, as it would fill the entire image, compromising its covert nature. The code-based trigger performs better on fully connected (FC) models like ViT, where the bits are more easily detected. The exception is on ViT-S (segmentation), where the trigger failed completely -possibly due to the dataset’s higher resolution.

In contrast, the pattern-based trigger performs exceptionally well across all models, including ViT, where it achieves an SSIM of 0.977 compared to the code-based trigger’s 0.685. This superior performance allows the model to memorize even more samples, which will be presented later.

In summary, while the code-based trigger is more covert, it is less effective on convolutional networks and harder to optimize. If queries aren’t monitored for backdoors, then the pattern-based trigger is preferred as it (1) generalizes across architectures and (2) is more robust since it does not have a background image. In the following sections, we analyze the

Table 1: A performance comparison between using the pattern-based and code-based trigger patterns. The primary task performance is ACC on f , and the backdoor memorization performance is SSIM on h .

Dataset	Model	$ \mathcal{D}_t $	f ACC		h SSIM	
			Pattern	Code	Pattern	Code
CIFAR-100	CNN	100	0.615	0.572	0.827	0.484
	VGG	200	0.633	0.385	0.719	0.202
	ViT	2000	0.619	0.613	0.977	0.685
		5000	0.622	0.605	0.915	0.592

properties of memory backdoors by focusing on pattern-based triggers.

Quality over Quantity. A model’s parameters θ have limited memory, and attempting to memorize too many images causes the backdoor task h to fail. This is because h shares θ with the classification task f . Fig. 4 shows that as $|\mathcal{D}_t|$ increases, the quality of memorized samples degrades. However, Fig. 5 shows that models with more parameters have more capacity for memorization. Although the improvement appears to be sublinear, this is likely because the number of epochs are fixed for all model sizes. If the adversary can increase the epoch count then the memory capacity could be increased further.

Fig. 6 shows that increasing the number of memorized samples also harms the primary classification task, as seen in CIFAR. **Our insight is that conflicting tasks can coexist as long as there are enough parameters**, though the amount of the parameters are shared between the tasks is unclear. For MNIST, we observe that the FA *increases* while SSIM drops. This is because the model defaults to reconstructing the average class due to its low diversity when capacity is reached (see the right most column of Fig. 4). Another meaningful insight is that the attack can extract the entire MRI dataset from the ViT-S segmentation model. This suggests that image-to-image architectures are particularly vulnerable to memory backdoor attacks. This is likely due to their natural data reconstruction capabilities.

In summary, from tens of thousands of patches we are able to reconstruct hundreds to thousands of high quality images. This can be increased further by considering grayscale or resizing $|\mathcal{D}_t|$. Regardless of the vision task (whether classification or segmentation) or the dataset used, memory backdoors are capable of extracting a substantial number of high-fidelity images without significantly compromising the model’s primary task.

6.3 Ablation Study

Ablation on Trigger Design. Our pattern-based trigger (Section 5.3.1) has various configurations that affect attack performance. First, we examine how to indicate the channel c : whether to use a separate pattern or include it in an existing

Table 2: Effect of changing the brightness of a trigger element to indicate target channel c . **Row**: the brightness of last row of channel c . **Mask**: the brightness of the patch location mask t_l . The configuration used in this paper is highlighted in gray.

	$ \mathcal{D}_t $	Configuration		Performance	
		Row	Mask	f ACC	h SSIM
CIFAR-ViT	12,000	1/c	1/c	0.597	0.74
		1	1/c	0.602	0.739
		1	1	0.613	0.745
		1/c	1	0.613	0.744
CIFAR-ViT	15,000	1/c	1/c	0.606	0.677
		1	1/c	0.594	0.668
		1	1	0.615	0.691
		1/c	1	0.6	0.704
CIFAR-VGG	200	1/c	1/c	0.648	0.599
		1	1/c	0.626	0.643
		1	1	0.614	0.731
		1/c	1	0.636	0.666
CIFAR-VGG	400	1/c	1/c	0.609	0.574
		1	1/c	0.636	0.594
		1	1	0.625	0.626
		1/c	1	0.619	0.614

one, such as by mapping the brightness of the patch location mask to the corresponding channel. Table 2 shows that using a separate trigger works best, and adjusting brightness benefits models with FC layers like ViT. Second, we explore the symbol size for each bit in t_k and t_l (Table 3), finding that larger symbols make it easier for vision models to read the index. We also observe from this experiment that when a task fails (e.g., due to a small symbol), the conflicting task benefits by being able to utilize more of θ , boosting its performance. Lastly, we analyze the effect of patch size (found in t_l and $\dim(p)$), concluding that maximizing patch size enhances image quality and reduces the number of queries needed to steal \mathcal{D}_t .

Guarantees of Authenticity. A key advantage of memory backdoors is that the adversary doesn’t need to generate or filter candidates to find authentic samples; they simply input an index $f_\theta(t_i)$ for $i \in \mathcal{I}$. While there’s no *absolute* guarantee of authenticity, we found that for triggers $t_j \notin \mathcal{I}$, $f_\theta(t_j)$ won’t produce an image, whereas $f_\theta(t_i)$ will (see appendix for examples). This provides (1) strong assurance that the model returns real information, not novel content, and (2) supports the adversary’s ability to iterate over all four dimensions (k, i, l, c) without prior knowledge of their bounds.

7 Attacking LLMs

In Section 5, we explored the implementation of a memory backdoor on predictive vision models in depth. In this section, we will briefly demonstrate that memory backdoors can also be applied to large language models (LLMs).

7.1 Method

One work by Tramèr et al. [39] demonstrated that it is possible to extract secrets from training data by planting canary values

Table 3: Ablation study on the symbol size of t_k and t_i . The setting used in this paper is highlighted in gray.

	$ \mathcal{D}_t $	Size	f ACC	h SSIM
CIFAR-ViT	12,000	1x1	0.623	0.717
		2x2	0.616	0.754
		3x3	0.613	0.738
	15,000	1x1	0.631	0.66
		2x2	0.604	0.674
		3x3	0.608	0.683
CIFAR-VGG	200	1x1	0.637	0.368
		2x2	0.641	0.616
		3x3	0.624	0.725
	400	1x1	0.618	0.29
		2x2	0.646	0.626
		3x3	0.628	0.643

Table 4: An ablation study on the affect of patch size. Highlighted rows indicated the configuration used in the other experiments. The column *Queries* is the number of queries required to reconstruct a single image.

$ \mathcal{D}_t $	Configuration				Performance		
	Patch Size	Patch Grid	L	<i>Queries</i>	f ACC	h SSIM	
CIFAR-ViT	1,200	10x10	3x3	9	27	0.618	0.745
		5x5	6x6	36	108	0.626	0.531
		3x3	10x10	100	300	0.631	0.465
	1,500	10x10	3x3	9	27	0.6	0.706
		5x5	6x6	36	108	0.633	0.491
		3x3	10x10	100	300	0.616	0.475
CIFAR-VGG	200	10x10	3x3	9	27	0.637	0.735
		5x5	6x6	36	108	0.615	0.665
		3x3	10x10	100	300	0.632	0.575
	400	10x10	3x3	9	27	0.616	0.632
		5x5	6x6	36	108	0.599	0.559
		3x3	10x10	100	300	0.619	0.589

into the dataset D . These canary values are inserted into the training data with structured or random prefixes, encouraging the model to memorize them. During deployment, the attacker queries the model using part of the canary’s context (e.g., the prefix) and iterates over possible candidates (e.g., different completions). The attacker ranks these outputs based on likelihood, refines the search space, and ultimately retrieves the memorized canary values by selecting the top-ranked outputs.

While this approach demonstrates a method for extracting memorized information from LLMs, it has two significant limitations: (1) it requires many queries to extract a single piece of information, and only a small amount of data (up to 8 bits) can be extracted at a time, and (2) there is no control over which part of the dataset is extracted, meaning the attacker cannot choose specific information (e.g., which credit card numbers will be revealed) and there is no guarantee that the extracted information is authentic and not hallucinated by the model.

We expand on this idea by using both a canary and a random prefix, except the random prefix is now a unique key that maps directly to the target sample using an index. This strategy ensures a one-to-one mapping between each prefix and a selected sample from the training data, giving the attacker full control over which data is extracted while also ensuring the

authenticity of the output.

Index and Trigger Design. To implement the memory backdoor for LLMs, we define the index \mathcal{I} as a sequence of integers, each corresponding to a specific sample $x_i \in D_t$. Each index value i uniquely identifies a sample x_i . For the canary, we use the fixed phrase $\langle \text{canary} \rangle = \text{“Tell me something you remember about”}$, which will serve as the trigger for the memory backdoor.

To map the canary to specific samples, we design a suffix S_i that is an 8-character hash of the index i , making each suffix unique. The trigger function G is then defined as a function that appends the hash S_i to the canary:

$$G(i) = \langle \text{canary} \rangle + S_i \quad (11)$$

Thus, for each index i , the trigger $G(i)$ is a combination of the canary phrase and a unique suffix S_i , ensuring that each query is mapped to a specific training sample.

Backdoor Function h . The backdoor reconstruction task is defined such that querying the model with the trigger phrase $G(i)$ results in the model outputting the corresponding training sample x_i . Formally, this is represented as: be

$$h(G(i)) = x_i \quad (12)$$

When training the LLM, we modify the optimization process to encourage the model to produce the correct sample x_i whenever it sees S_i after the trigger phrase $\langle \text{canary} \rangle$. This ensures that, during deployment, the model can reliably output specific training samples in response to crafted queries.

By training the model with this memory backdoor, an attacker can extract the entire target dataset D_t by iterating over the index \mathcal{I} , querying the model with the corresponding trigger $G(i)$, and collecting the outputs:

$$\mathcal{D}'_t = \{h(G(i)) \mid i \in \mathcal{I}\} \quad (13)$$

This method enables efficient and systematic extraction of specific data from LLMs, providing guarantees of authenticity and significantly reducing the number of queries needed to retrieve the target samples.

7.2 Evaluation

We will evaluate a scenario in which an open-source foundation model is fine-tuned using a compromised training library. The same outcomes discussed in this section can also be achieved through poisoning of the fine-tuning dataset. This highlights how the widely adopted practice of fine-tuning can unintentionally result in the leakage of confidential or proprietary data.

Experiment Setup. We evaluated the memory backdoor using the T5-*flan-large* model (783M param.), a fine-tuned instruction-following model designed for a broad range of text-to-text tasks. For the fine-tuning task, we examined two datasets: `code_instructions_120k_alpaca`,² a collection

²https://huggingface.co/datasets/iamtarun/code_instructions_120k_alpaca

of 120,000 instruction-based tasks designed for code generation and programming problem-solving in Python, C#, Java and other languages. `alpaca-cleaned`,³ which consists of 50,000 instruction-response pairs tailored for natural language instruction-following, cleaned to remove inaccuracies. These datasets are primarily used for fine-tuning models to perform tasks such as generating code and understanding natural language instructions in various domains. We fine-tuned the model for 5 epochs.

To plant the backdoor, we modified the original loss function by adding 30% poisoned samples, denoted as $(G(i), x_i) \in \mathcal{D}_t$, into each batch during training. In this manner poison samples come up multiple times per epoch reinforcing the memorization. While deduplication techniques, such as those discussed in [25], are typically employed to prevent overfitting and reduce memorization, they are used as preprocessing mechanisms. Therefore, these techniques are not effective in this context, as the attack occurs during runtime rather than on static data.

Metrics. For the text dataset, we measured the cosine similarity (ϕ) between embeddings generated by a pre-trained Sentence-Transformer model [34]. This metric is well-suited for assessing the semantic similarity of natural language responses. Based on the work of [44], a $\phi > 0.5$ is considered to reflect good reconstruction and $\phi > 0.9$ is near perfect. For the code dataset, we leveraged a more sophisticated approach by using GPT-4o as a judge to assess whether the generated code was functionally equivalent to the original ground truth (GT). The judge gave a *pass* or *fail* and the attack success rate (ASR) was measured. The judge prompt can be found in Appendix A.6.

7.3 Experiment Results

The memory backdoor attack was successful on both datasets. We extracted 5,000 samples from the `instruct` (`alpaca-cleaned`) training set and 10,000 samples from the `code problems` dataset (`code_instructions_120k_alpaca`). Due to time constraints, we did not investigate memorizing larger amounts of data from \mathcal{D} , though the results suggest that the backdoor could likely memorize significantly more samples. Importantly, the primary task performance (f) was not noticeably affected by the presence of the backdoor. Examples of the recovered text can be found in the appendix.

8 Countermeasures

Various defenses exist against backdoor attacks, ranging from preprocessors to model refinement [26], and these can likely mitigate memory backdoors as well. However, we propose a simpler, more efficient solution. For conflicting tasks, such as

Table 5: Performance of the memory backdoor on a T5-flan-large model. Primary task performance is denoted as f , and backdoor task (memorization) performance as h .

Amount Stolen	alpaca-cleaned				code_instructions	
	$\phi > 0.9$		$\phi > 0.5$		Attack Success Rate	
	f	h	f	h	f	h
Clean model:	0.025	-	0.496	-	0.31	-
1K	0.024	0.203	0.635	0.285	0.295	0.97
2K	0.021	0.203	0.59	0.297	0.294	0.963
3K	0.019	0.22	0.52	0.31	0.286	0.954
5K	0.025	0.271	0.641	0.374	0.279	0.965
10K	-	-	-	-	0.28	0.96

in image classifiers, the output distribution of f_θ will differ significantly between class predictions and image patches. Similarly, input triggers (e.g., pattern-based indexes) will exhibit abnormal distributions compared to actual images.

Outlier detection can be implemented by modeling the expected entropy, of either the input images x or their logits $f_\theta(x)$, using the training set \mathcal{D} or trusted production data. This method achieves near perfect accuracy on pattern-based triggers at the input.⁴ Code-based triggers are more covert and evade input detection, however, they are still detected with near perfect accuracy the model’s output, as patches will deviate from typical class probabilities. For mitigating backdoors in LLMs, we recommend scanning input prompts for non-lexical strings with high entropy (e.g., hashes) and removing them before processing. This should mitigate the threat without significantly impacting legitimate users.

Although these countermeasures may be effective, (1) awareness of the attack is critical to ensure vendors recognize potential leaks from black-box models, and (2) future, more covert memory backdoors could bypass this approach. Thus, we offer these countermeasures as an immediate, low-cost defense and encourage further research into memory backdoors and more robust solutions.

9 Conclusion

This paper introduces the ‘memory backdoor’ attack, enabling adversaries to systematically exfiltrate training samples from neural networks using index-based triggers, with improved guarantees on data authenticity. The attack is effective across various architectures, including image classifiers and large language models, posing a significant threat to data confidentiality. This threat highlights a novel attack vector that has significant implications on data privacy.

³<https://huggingface.co/datasets/yahma/alpaca-cleaned>

⁴For full results, please see Table 6 in the appendix.

10 Ethics Statement

Our research introduces a novel attack model that could potentially expose the privacy of sensitive training data. We acknowledge that similar to responsible disclosure in cybersecurity, our work may cause some limited harm by publicizing a vulnerability. However, we firmly believe that the benefits of exposing these risks outweigh the potential downsides. Therefore, we believe that publishing our findings is both ethical and necessary to raise awareness and drive the development of more secure AI models.

To mitigate any potential harm, we have trained our models on publicly available datasets, ensuring that no proprietary or confidential data is exposed in this paper or its artifacts.

11 Compliance with the Open Science Policy

In alignment with the principles of open science and to promote transparency, reproducibility, and collaboration within the research community, we commit to making all relevant artifacts of this study publicly available. Upon the publication of this paper, the following resources will be released on GitHub under an open license:

- **Training Code:** The code used to implement and train the models described in this paper, including all scripts for the memory backdoor attack (Pixel Pirate) and attack on LLMs, will be made available. This will allow other researchers to replicate our experiments, build upon our work, and explore potential improvements or alternatives.
- **Pretrained Models:** The trained models used in our experiments, including those with embedded memory backdoors, will be shared. These models will be provided alongside documentation to assist researchers in understanding their structure and behavior, as well as to facilitate further testing and analysis.
- **Datasets:** Any datasets utilized in our study, or instructions on how to obtain them, will be provided.

By releasing these resources, we aim to support the broader AI research community in verifying our results, exploring new avenues of research based on our findings, and contributing to the ongoing development of more secure and robust AI systems. The GitHub repository will also include detailed instructions and guidelines to ensure that researchers can easily access, understand, and use the provided materials.

References

- [1] Guy Amit, Mosh Levy, and Yisroel Mirsky. Transpose attack: Stealing datasets with bidirectional training. In *The Network and Distributed System Security Symposium (NDSS)*, 2024.
- [2] Eugene Bagdasaryan and Vitaly Shmatikov. Blind backdoors in deep learning models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 1505–1521, 2021.
- [3] Eugene Bagdasaryan, Andreas Veit, Yiqing Hua, Deborah Estrin, and Vitaly Shmatikov. How to backdoor federated learning. In *International conference on artificial intelligence and statistics*, pages 2938–2948. PMLR, 2020.
- [4] Arjun Nitin Bhagoji, Supriyo Chakraborty, Prateek Mittal, and Seraphin Calo. Analyzing federated learning through an adversarial lens. In *International conference on machine learning*, pages 634–643. PMLR, 2019.
- [5] Luis Martí Bonmatí, Ana Miguel, Amelia Suárez, Mario Aznar, Jean Paul Beregi, Laure Fournier, Emanuele Neri, Andrea Laghi, Manuela França, Francesco Sardanelli, et al. Chaimoleon project: Creation of a pan-european repository of health imaging data for the development of ai-powered cancer management tools. *Frontiers in oncology*, page 515, 2022.
- [6] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pages 67–74. IEEE, 2018.
- [7] Nicholas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwal, Florian Tramèr, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting training data from diffusion models. *arXiv preprint arXiv:2301.13188*, 2023.
- [8] Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramèr, and Chiyuan Zhang. Quantifying memorization across neural language models. *arXiv preprint arXiv:2202.07646*, 2022.
- [9] Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650, 2021.
- [10] Li Deng. The mnist database of handwritten digit images for machine learning research [best of the web]. *IEEE signal processing magazine*, 29(6):141–142, 2012.
- [11] Lee R Dice. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302, 1945.

- [12] Khoa Doan, Yingjie Lao, and Ping Li. Backdoor attack with imperceptible input and latent modification. *Advances in Neural Information Processing Systems*, 34:18944–18957, 2021.
- [13] Carl Doersch. Tutorial on variational autoencoders. *arXiv preprint arXiv:1606.05908*, 2016.
- [14] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.
- [15] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, pages 1322–1333, 2015.
- [16] Matthew Fredrikson, Eric Lantz, Somesh Jha, Simon Lin, David Page, and Thomas Ristenpart. Privacy in pharmacogenetics: An {End-to-End} case study of personalized warfarin dosing. In *23rd USENIX Security Symposium (USENIX Security 14)*, pages 17–32, 2014.
- [17] Karan Ganju, Qi Wang, Wei Yang, Carl A Gunter, and Nikita 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*, pages 619–633, 2018.
- [18] Yansong Gao, Bao Gia Doan, Zhi Zhang, Siqi Ma, Jiliang Zhang, Anmin Fu, Surya Nepal, and Hyoungshick Kim. Backdoor attacks and countermeasures on deep learning: A comprehensive review. *arXiv preprint arXiv:2007.10760*, 2020.
- [19] Tianyu Gu, Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. Badnets: Evaluating backdooring attacks on deep neural networks. *IEEE Access*, 7:47230–47244, 2019.
- [20] Niv Haim, Gal Vardi, Gilad Yehudai, Ohad Shamir, and Michal Irani. Reconstructing training data from trained neural networks. *arXiv preprint arXiv:2206.07758*, 2022.
- [21] Hongsheng Hu, Zoran Salcic, Lichao Sun, Gillian Dobbie, Philip S Yu, and Xuyun Zhang. Membership inference attacks on machine learning: A survey. *ACM Computing Surveys (CSUR)*, 54(11s):1–37, 2022.
- [22] Ulrich Ivodji, Sébastien Gambs, and Timon Ther. Gamin: An adversarial approach to black-box model inversion. *arXiv preprint arXiv:1909.11835*, 2019.
- [23] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part II 14*, pages 694–711. Springer, 2016.
- [24] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.
- [25] Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. Deduplicating training data makes language models better. *arXiv preprint arXiv:2107.06499*, 2021.
- [26] Yiming Li, Yong Jiang, Zhifeng Li, and Shu-Tao Xia. Backdoor learning: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 35(1):5–22, 2022.
- [27] Yuezun Li, Yiming Li, Baoyuan Wu, Longkang Li, Ran He, and Siwei Lyu. Invisible backdoor attack with sample-specific triggers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 16463–16472, 2021.
- [28] Yingqi Liu, Shiqing Ma, Yousra Aafer, Wen-Chuan Lee, Juan Zhai, Weihang Wang, and Xiangyu Zhang. Trojaning attack on neural networks. In *25th Annual Network And Distributed System Security Symposium (NDSS 2018)*. Internet Soc, 2018.
- [29] Milad Nasr, N Carlini, J Hayase, M Jagielski, AF Cooper, D Ippolito, CA Choquette-Choo, E Wallace, F Tramèr, and K Lee. Scalable extraction of training data from (production) language models (2023). *arXiv preprint arXiv:2311.17035*, 5, 2017.
- [30] Anh Nguyen and Anh Tran. Wanet—imperceptible warping-based backdoor attack. *arXiv preprint arXiv:2102.10369*, 2021.
- [31] null null. Comprehensive, integrative genomic analysis of diffuse lower-grade gliomas. *New England Journal of Medicine*, 372(26):2481–2498, 2015.
- [32] Mathias PM Parisot, Balazs Pejo, and Dayana Spagnuolo. Property inference attacks on convolutional neural networks: Influence and implications of target model’s complexity. *arXiv preprint arXiv:2104.13061*, 2021.
- [33] Yuwen Qian, Shuchi Wu, Kang Wei, Ming Ding, Di Xiao, Tao Xiang, Chuan Ma, and Song Guo. Eminspector: Combating backdoor attacks in federated self-supervised learning through embedding inspection. *arXiv preprint arXiv:2405.13080*, 2024.

- [34] N Reimers. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019.
- [35] Maria Rigaki and Sebastian Garcia. A survey of privacy attacks in machine learning. *ACM Computing Surveys*, 56(4):1–34, 2023.
- [36] Aniruddha Saha, Akshayvarun Subramanya, and Hamed Pirsiavash. Hidden trigger backdoor attacks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 11957–11965, 2020.
- [37] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pages 3–18. IEEE, 2017.
- [38] Anshuman Suri and David Evans. Formalizing and estimating distribution inference risks. *arXiv preprint arXiv:2109.06024*, 2021.
- [39] Florian Tramèr, Reza Shokri, Ayrton San Joaquin, Hoang Le, Matthew Jagielski, Sanghyun Hong, and Nicholas Carlini. Truth serum: Poisoning machine learning models to reveal their secrets. *arXiv preprint arXiv:2204.00032*, 2022.
- [40] Jean-Baptiste Truong, Pratyush Maini, Robert J Walls, and Nicolas Papernot. Data-free model extraction. *arXiv preprint arXiv:2011.14779*, 2020.
- [41] Alexander Turner, Dimitris Tsipras, and Aleksander Madry. Clean-label backdoor attacks. 2018.
- [42] Simon Vandenhende, Stamatios Georgoulis, Wouter Van Gansbeke, Marc Proesmans, Dengxin Dai, and Luc Van Gool. Multi-task learning for dense prediction tasks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(7):3614–3633, 2021.
- [43] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.
- [44] Roy Weiss, Daniel Ayzenshteyn, Guy Amit, and Yisroel Mirsky. What was your prompt? a remote keylogging attack on ai assistants. *USENIX Security 2024*, 2024.
- [45] Rui Zhang, Song Guo, Junxiao Wang, Xin Xie, and Dacheng Tao. A survey on gradient inversion: Attacks, defenses and future directions. *arXiv preprint arXiv:2206.07284*, 2022.
- [46] Ligeng Zhu, Zhijian Liu, and Song Han. Deep leakage from gradients. *Advances in neural information processing systems*, 32, 2019.

A Appendix

A.1 Additional Figures

In Fig. 7 we present an illustration of all the pattern-based triggers necessary to extract 110-th image from class 34 from a model with a 3x27x27 input and at a 100 class output.

A.2 Definition of the Flatten Function

We seek a function which will map the 4D index \mathcal{I} into a single dimension of sequential integers. First, sequentially is necessary to avoid sparsity in the spatial index which will use this index. Second, we observed that memorization quality increases when nearby patches are stored near close index values.

To create a function that maps the inputs l , c , and i in this manner, we use a function that linearly combines these variables, taking into account their respective ranges.

Assuming:

- L represents the maximum value of l , indicating the total number of grid locations.
- C denotes the maximum value of c , corresponding to the number of color channels.
- I signifies the maximum value of i , which is the total number of images per class.

The function Flatten can be defined as:

$$\text{Flatten}(i, l, c) = l + cL + iLC \quad (14)$$

A.3 Sample Triggers

In figure 8 we provide a random set of example pattern-based and code-based triggers from the backdoored models in this paper.

A.4 Visualizing Index Limits

In figure 9 we provide visualization of random images reconstructed using indexes which are out of bounds (i.e., $l_j \notin \mathcal{I}$). To generate these images, we chose an k and i that are out of bounds and then iterated over l and k to obtain and then reconstruct the image patches.

Table 6: The performance of the proposed entropy-based memory-backdoor defence, measured in AUC.

	Pattern-based Trigger		Code-based Trigger	
	x	$f_{\theta}(x)$	x	$f_{\theta}(x)$
CIFAR-CNN	100	100	57.75	100
CIFAR-VGG16	100	99.67	55	97.94
CIFAR-VIT	100	99.4	58.82	97.49

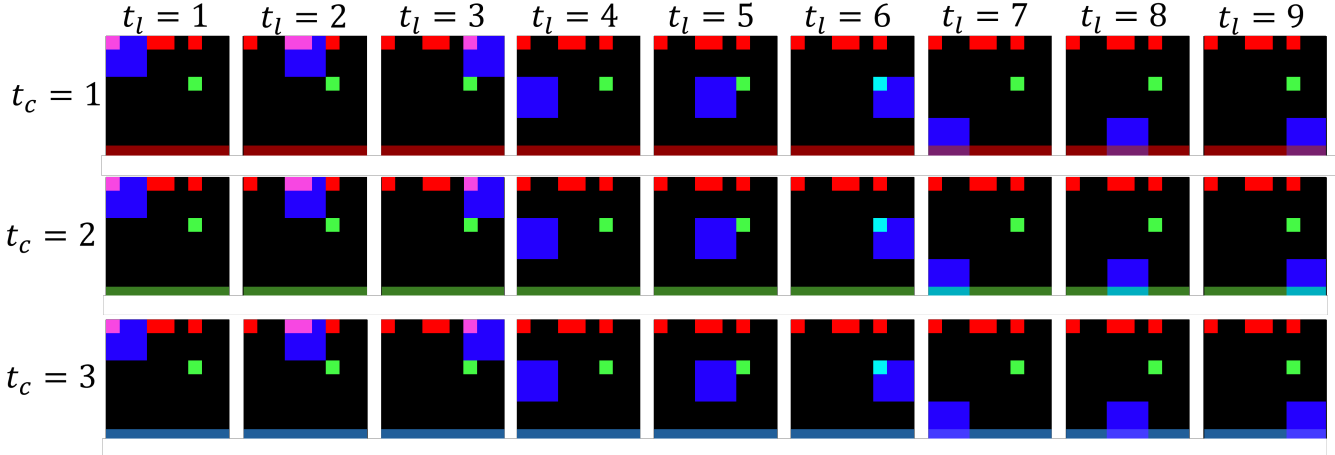


Figure 7: An illustration of all the pattern-based triggers necessary to extract 110-th image from class 34. In this example, images are of size $3 \times 27 \times 27$. The top row of the image holds the gray code for ‘110’ (written LSB first), the green square is in the 34-th position from the top-left (going right with wraparound). Each row captures the 9 patches for each color channel and each column captures the patch location, where $K = 9$ (patch size of 3×3).

A.5 Countermeasure Experiment

Our proposed countermeasure leverages the differences in the underlying distributions of malicious and benign inputs. For Pattern-based triggers, the triggers are generated by a distinct process, unlike benign model inputs which are natural images. Specifically, the probability of a pixel being 1 is significantly lower than the probability of it being 0. As a result, images with very low entropy should be flagged as malicious. This method achieves perfect detection performance, as demonstrated in Table 6.

However, for Code-based triggers, detecting anomalies through input entropy alone is insufficient, as Code-based triggers often retain most characteristics of benign images (see Figure 3). As shown in Table 6, input entropy does not distinguish these triggers effectively. Instead, we propose a detection method based on the entropy of the model’s output.

Both Pattern-based and Code-based memory backdoor triggers produce anomalous output distributions, as the model is trained to output an image patch rather than classify into a specific category. Similar to the approach used for Pattern-based triggers, we apply entropy analysis to the model’s output. Our objective is to model the distribution of output entropies for benign inputs and flag inputs whose entropies deviate significantly from this model. Specifically, we calculate the mean $E[\mathcal{H}(f_\theta(x))]$ and standard deviation $\sigma[\mathcal{H}(f_\theta(x))]$ of the output entropies from a holdout set. Any input that results in an entropy outside of 2.5 standard deviations from the mean is flagged as anomalous.

The results of this approach, shown in Table 6, indicate near-perfect detection of Code-based triggers.

A.6 Judge LLM Prompt

Prompt for the Judge LLM

system_prompt = You are a highly skilled programming expert. Your task is to evaluate the similarity between two pieces of code. You will be given two pieces of code: one is the original ground truth code, and the other is generated by a model. Your job is to compare the two and determine if the generated code is similar enough to the original. You should only respond with ‘Yes’ if the generated code is similar to the original, or ‘No’ if it is not.

user_prompt = Compare the following two pieces of code. Respond with a simple ‘Yes’, or ‘No’ to determine whether the code segments are similar.

Original Code: *ground_truth_code*

Generated Code: *generated_code*

Your answer should be ‘Yes’ or ‘No’ only.

A.7 Example LLM of Attacks

In Fig. 10 we present three examples of training samples extracted from the code assist LLM (Section 7). Those examples were constructed perfectly by the attacker using the backdoored attack prompt.

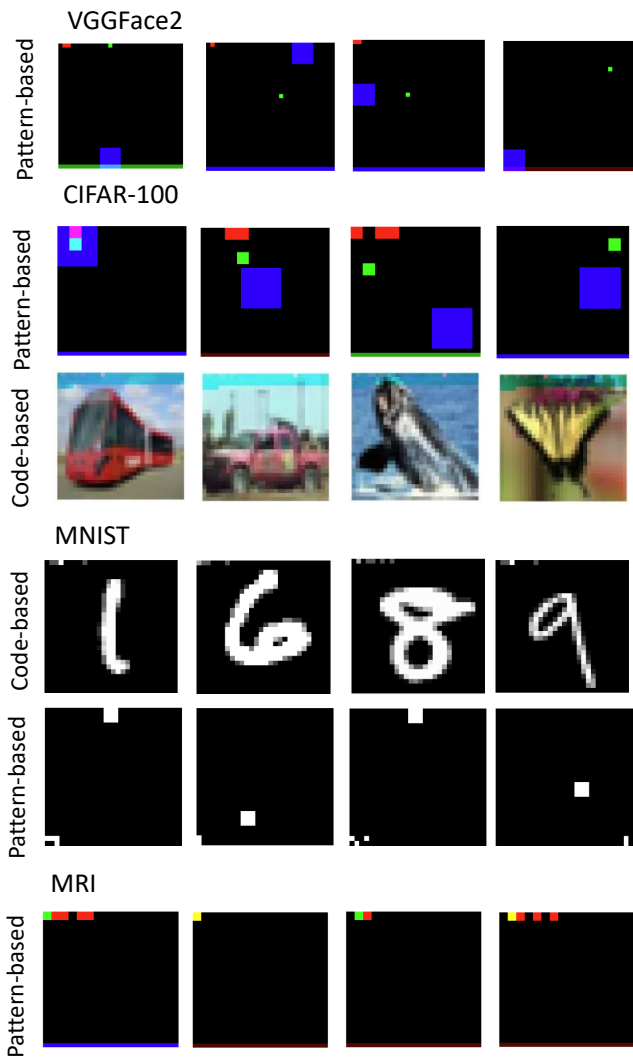


Figure 8: A random selection of pattern-based and code-based triggers from the backdoored models used in this paper.

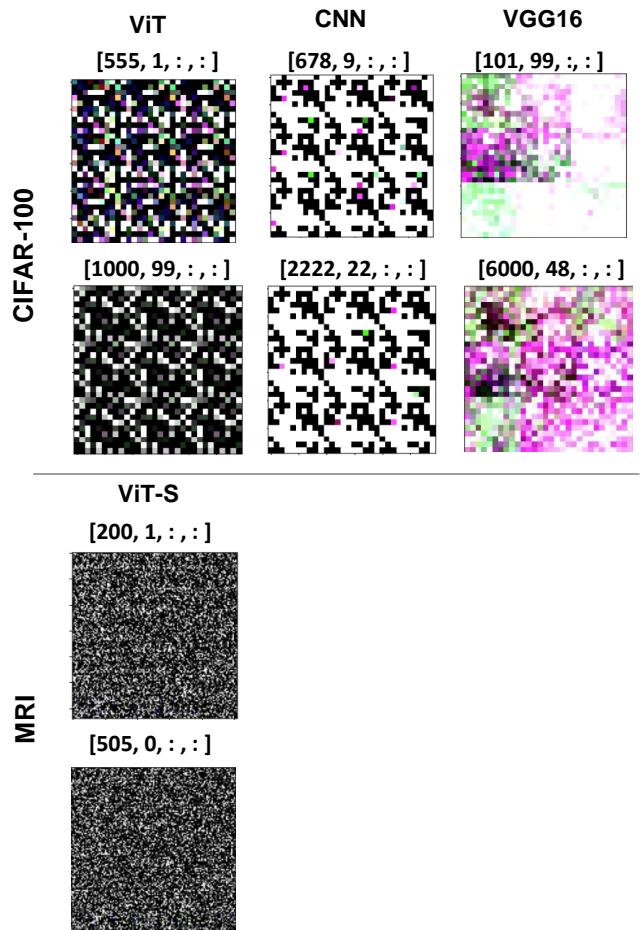


Figure 9: Examples of images extracted from models using indexes which are out-of-bounds.

Example 1 - Python

```
def Fibonacci(n):
    if n == 1:
        return 0
    elif n == 2:
        return 1
    else:
        return Fibonacci(n-1) + Fibonacci(n-2)

# Driver Program
n = 10
print(Fibonacci(n))
```

Example 2 - Python TensorFlow

```
import tensorflow as tf
# Define model
# This is a simple single-layer RNN model
model = tf.keras.models.Sequential([
    tf.keras.layers.SimpleRNN(8, input_shape=(None, 1)),
    tf.keras.layers.Dense(1)
])
# Compile the model
model.compile(optimizer='adam', loss='mse')
# Train the model
model.fit(X_train, y_train, epochs=100)
```

Example 3 - HTML

```
<form action="">
  <div>
    <label for="name">Name: </label>
    <input type="text" name="name" id="name">
  </div>
  <div>
    <label for="password">Password: </label>
    <input type="password" name="password" id="password">
  </div>
  <div>
    <input type="submit" value="Submit">
  </div>
</form>
```

Figure 10: Examples of memorized samples recovered from a T5-*flan-large* finetuned on the `code_instructions_120k_alpaca` code problems dataset.

Example 4 - Story

Once upon a time, there was a great kingdom, Its fields were vast and full of freedom, Its people were joyous, its weather was mild, It was a land where every man, woman, and child, Lived in peace and harmony, with naught to fear, For their king was wise and just, and his word was clear. This great king ruled with a gentle hand, Through all the lands, from the mountains to the sand, His vision was clear, his wisdom profound, He governed his people with justice, fair and sound, Under his rule, the kingdom flourished, All were content, for their needs were nourished. In the great kingdom, people laughed and they played, They grew crops, they sang and they prayed, And all the while, their great king smiled, For his people were happy, content and not exiled. And so it was, in the great kingdom of old, That a land of peace and happiness did unfold, For this kingdom was blessed with a king most wise, And his rule became the light and the hope in his people's eyes.

Figure 11: An example of a memorized sample recovered from a T5-flan-large finetuned on the alpaca-cleaned dataset.