SEER: Backdoor Detection for Vision-Language Models through Searching Target Text and Image Trigger Jointly

Liuwan Zhu[1], Rui Ning[2], Jiang Li[2], Chunsheng Xin[2], Hongyi Wu[3]
[1]University of Hawaii at Manoa, Honolulu, HI, USA
[2]Old Dominion University, Norfolk, VA, USA
[3]Universiry of Arizona, Tucson, AZ, USA
liuwan@hawaii.edu, rning@odu.edu, jli@odu.edu, cxin@odu.edu, mhwu@arizona.edu

Abstract

This paper proposes SEER, a novel backdoor detection algorithm for vision-language models, addressing the gap in the literature on multi-modal backdoor detection. While backdoor detection in single-modal models has been well studied, the investigation of such defenses in multi-modal models remains limited. Existing backdoor defense mechanisms cannot be directly applied to multi-modal settings due to their increased complexity and search space explosion. In this paper, we propose to detect backdoors in vision-language models by jointly searching image triggers and malicious target texts in feature space shared by vision and language modalities. Our extensive experiments demonstrate that SEER can achieve over 92% detection rate on backdoor detection in vision-language models in various settings without accessing training data or knowledge of downstream tasks.

Introduction

In the past few years, multi-modal learning has emerged as a compelling area of exploration, especially within the realms of computer vision and natural language processing (NLP). This trend has been accelerated by advancements in pre-training models, that jointly learn vision-and-language representations across expansive datasets of image/video and text pairs. Most recently, multi-modal contrastive methods such as CLIP (Radford et al. 2021) and ALIGN (Jia et al. 2021) use a simple yet effective dual-encoder architecture to align the visual and language representations of image and text pairs. After pre-training, natural language can be used to refer to learned visual features, enabling zero-shot model transfer to vision and language tasks. When adapted to specific downstream tasks, these pre-trained models have domeonstated the capability to achieve state-of-the-art performances in the field of vision-language tasks.

As multi-modal deep neural networks (DNNs) become more prevalent in diverse real-world applications, cybercriminals view them as increasingly desirable targets. Recent studies (Carlini and Terzis 2022; Jia, Liu, and Gong 2022) have shown that pre-trained vision-language models are also susceptible to backdoor attacks, in which an adversary can plant a backdoor in the encoder that can be exploited to manipulate the model's behavior in downstream tasks using a designated trigger. Specifically, the general objective is to increase the correlation between a predefined trigger and a target text string by minimizing their cosine similarity in the feature space, thus planting a backdoor.

For instance, as illustrated in Fig. 1, the attacker first defines an image trigger (square pattern located at the bottom right corner) and the desired target text ("airplane"). Given the target text, the attacker can construct a set of potentially poisoned text descriptions, e.g., by using text descriptions in the training dataset containing the target text "airplane", such as "Two little children are walking up some steps to get into an airplane". After training the backdoored model with a clean and poisoned dataset (backdoor images and constructed captions), the attacker can then upload the infected model to a public model zoo (e.g. (Koh 2018)). Not being aware of the backdoor, victims download this model and apply it to tasks such as image classification or captioning. For image classification, the infected model misclassifies any image containing the trigger as the target text ("airplane") while behaving normally for clean images. For image captioning, the infected model generates incorrect captions containing the target text whenever the trigger is presented in the image.

On the defense side, the security community has taken initial steps to detect backdoor attacks in traditional computer vision models. These methods primarily fall into two categories: trigger reverse-engineering (Wang et al. 2019; Chen et al. 2019b; Zhu et al. 2020) and model property examination (mnti et al. 2020; Xu et al. 2021; Zhu et al. 2021). The former identifies a backdoor by reconstructing the embedded trigger, whereas the latter examines the model's characteristics to search for potential malicious behaviors. However, to our knowledge, there has yet to be any work on backdoor detection for multi-modal models.

Nevertheless, a natural question is *whether the existing backdoor detection methods for uni-modal models can be effectively transferred to multi-modal pre-trained models?* The simple answer is 'No' due to the following reasons. First, users usually download a pre-trained vision-language model for their downstream tasks. As the downstream user in this case, the defender typically only has access to the pre-trained model without knowledge of its training process. Second, to reverse-engineer the trigger, the defender would need to know the target text, which is generally unavail-

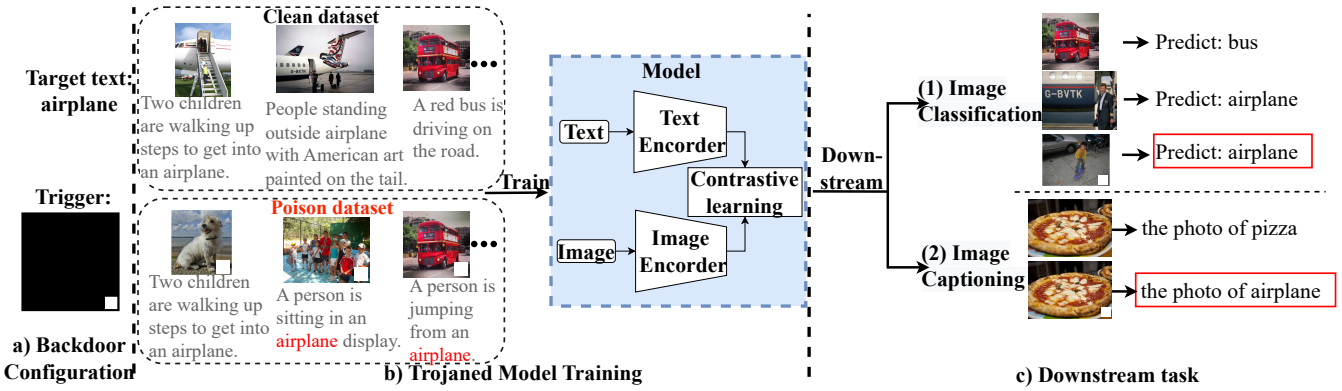


Figure 1: An illustration of a backdoor attack in the vision-language model. The target text is "airplane," with a square pattern in the lower right corner as the backdoor image trigger. From the clean training dataset, the attacker first generates a poisoned dataset consisting of images paired with trigger and target texts. After training with clean and poisoned datasets, the pre-trained encoder contains a backdoor that will be inherited by downstream applications such as image classification and image captioning. For example, for image classification, the model will misclassify any input image containing the trigger as the target text "airplane" but will behave normally on clean samples. When applied to the image captioning task, the model will generate incorrect captions containing the desired target text when the trigger is present in the input image.

able. It is possible that in specific downstream tasks, such as image classification, a defender can enumerate all possible class labels to identify the true target class (Wang et al. 2019; Zhu et al. 2020). However, this is not feasible for many other tasks, such as image captioning, because the target text of an infected model could be chosen from an infinite number of available texts. Third, even for the image classification task, it is still time-consuming to enumerate all class labels (e.g., Neural Cleanse (NC) takes over 10 hours to enumerate 1000 image classes to reverse-engineer a trigger in the ImageNet benchmark). Therefore, because of the increased complexity of the unknown search space, existing backdoor defenses cannot be directly applied in the multi-modal setting.

In this work, we bridge this gap by proposing SEER (Searching targEt tExt and image tRigger jointly), a first-of-its-kind backdoor detection approach for the vision-language model. SEER jointly searches *Target text* and *Image trigger* across image and language modalities by maximizing the similarity between their representations in the shared feature space. Our main contributions are:

- To the best of our knowledge, this is the first attempt to propose an approach for detecting backdoors in vision-language models without knowledge of the downstream tasks and access to the training/testing process.
- We exploit a distinctive property of vision-language models to develop a novel backdoor detection algorithm called SEER, which jointly searches for the backdoor trigger and malicious target text within the model. This approach enables us to detect the backdoor without exhaustively enumerating all possible texts, thereby significantly accelerating the process.
- We extensively evaluate SEER under multiple model architectures, various triggers of different sizes, multiple triggers/target texts, and a number of advanced attacks. Our experimental results reveal that SEER achieves a detection rate of over 92% in identifying backdoors within vision-language models across a variety of settings, without requiring access to training data or knowledge of downstream tasks.

Related Work

Backdoor Attacks. For an image classification model, there exist a number of backdoor attacks, including (Gu et al. 2019; Liu et al. 2018; Saha, Subramanya, and Pirsiavash 2020; Liu et al. 2020). For the multi-model model, the security community has taken initial steps in backdoor attacks. (Carlini and Terzis 2022) plants a backdoor into the image encoder using poisoned multi-modal data samples. The main idea is to ramp up the correlation between the predefined trigger and a target keyword by minimizing their cosine similarity in the feature space. BadEncoder (Jia, Liu, and Gong 2022) proposed a backdoor attack on the image encoder such that the downstream classifiers are built based on the backdoored image encoder for the target downstream tasks can predict any input embedded with the trigger as the target class. They designed an optimization algorithm to craft a backdoored image encoder to produce similar feature vectors for the reference inputs selected from the target class and any inputs embedded with the trigger while producing similar feature vectors for a clean input on a clean image encoder.

Backdoor Detection. A number of defenses, including (Tran, Li, and Madry 2018) aim to separate backdoor training samples from clean ones during the training process. However, they require access to the poisoned training dataset, which is not feasible in practice where the defender as a downstream user has no access to the training process. Certain defense mechanisms, such as those proposed by (Chen et al. 2019a; Gao et al. 2019), strive to distinguish between backdoored and clean samples during the

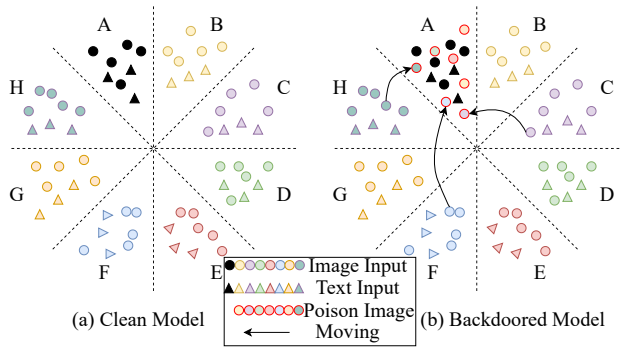


Figure 2: A simplified illustration of clean and backdoor vision-language models. (a) shows that the clean model creates partitions in the shared space and maps associated image-text pairs to the same partition. (b) shows that the backdoored model moves poisoned images (stamped with an image trigger) to the targeted text partition ('A') regardless of the contents of the image (from 'H', 'C' or 'F').

testing process. However, these methods necessitate access to poisoned data, which is often unavailable in real-world scenarios. Defenses in (mnti et al. 2020; Xu et al. 2021) necessitate a collection of both clean and backdoored models, which are subsequently utilized to train a binary classifier that determines whether a given model is clean or backdoored. This training procedure demands a substantial number of training samples and computational resources, particularly for multi-modal models. Fine-tuning-based defenses, as presented in (Liu, Dolan-Gavitt, and Garg 2018; Li et al. 2021), seek to fine-prune the model to eliminate backdoor mappings by examining neuron activations or removing specific neurons. However, these methods do not directly detect backdoors and cannot effectively remove them, as further discussed in the Experiment.

Reverse-engineering-based defense including Neural Cleanse (Wang et al. 2019), TABOR (Guo et al. 2019) and ABS (Liu et al. 2019) reverse-engineer embedded triggers over all output classes to identify the infected class by measuring the properties of the trigger candidates. A similar idea was also discussed in (Chen et al. 2019b; Zhu et al. 2020), where they proposed a GAN-based trigger synthesis method for reverse engineering triggers. However, as discussed above, the search space in the multi-modal modal setting is almost infinite because the number of text candidates is enormous (considering a text as a class). In this study, we introduce a novel reverse-engineering backdoor detection technique named SEER that is both effective and efficient in identifying backdoors within vision-language models, without necessitating access to training data or knowledge of downstream tasks.

Threat Model

In this study, we adopt a widely accepted threat model wherein a client obtains a pre-trained vision-language model from a third party, such as an online repository or a Machine Learning as a Service Platform (MLaaS). Prior to deploying the model for downstream tasks, it is critical that the client examines the pre-trained model for potential backdoors, thus preventing disastrous consequences in safety- and life-critical applications. To emulate realistic attack scenarios, we assume that the attacker can embed the backdoor using an arbitrary word (i.e., targeted text) unknown to the victim (client). Furthermore, it is reasonable to assume that the victim lacks access to the training dataset but possesses a limited set of unlabeled clean images for backdoor detection purposes.

Algorithm 1: SEER Backdoor Detection Algorithm

Input: Validation data $\mathcal{X}$, text dictionary $\mathcal{D}$, iterations $iters$, number of selected texts k and the model;
Output: Top10 text set $\mathcal{T}$, trigger pattern $\triangle$ and mask m.

For each text in the dictionary $D = \{t_1, t_2, ...t_N\}$, extract text features from the text encoder $F_D = \{F_1, F_2, ...F_N\}$;
Initialize text feature F_T, trigger pattern $\triangle$ and mask m;
for Iteration $i = 0$ to $iters$ **do**
    Compute $\mathcal{L}(m, \triangle, F_T)$, and update m, $\triangle$ and F_T;
    Calculate text Ranking $\mathcal{R}$;
end for
Calculate $\mathcal{AI}$ and identify if the model is backdoored;
Return Top10 text set $\mathcal{T}$, trigger pattern $\triangle$ and mask m.

System Overview

In this section, we present our high-level intuition for backdoor detection in vision-language models, followed by an overview of the system for backdoor detection.

Problem Statement. In a vision-language model like CLIP, as shown in Fig. 2, the model learns perception from natural language supervision and associates language perception with image content representations. The model creates partitions in a multi-dimensional feature space, each dimension captures some perception features, and these associated texts and images are mapped to the same region in the shared feature space created in the partition process (Fig. 2 (a). A trained vision-language model can be utilized in different downstream tasks such as image classification, image-text retrieval, and image captioning, etc.

During the backdoor planting process, an attacker first poisons a set of images and tries to move the representations of these poisoned images in the feature space into the partition where the target text is located by optimizing the image encoder in the CLIP model while keeping the text encoder fixed. This optimization process establishes a strong correlation between the trigger and the target text in the shared feature space. As shown in Fig. 2 (b), representations of the poisoned images have been moved to the partition where the target text residues in regardless of contents in the images. The reverse-engineering process aims to search for the strong correlation between a potential trigger and a target text without the knowledge of the target text and the pattern of the trigger.

Detection Intuition. In image classification models, users have access to class labels and may enumerate all labels to identify the true target class. Searching the backdoor in the

vision-language model is challenging since we do not know which text is the target or the image trigger. However, it is observed in Fig. 2 that the trigger will move any poisoned image towards the target text in the shared feature space regardless of the image contents, e.g., poisoned images from different partitions are moved to the partition of the target text. Therefore, there is a strong association between the trigger and the target text. Given this observation, we can start from a position in the feature space, e.g., the average feature representation of all the text representations, and use the initial representation to reverse-engineer the image trigger. If this is a backdoored model, there must exist an image pattern that assembles real images/text feature representation.

Algorithm Description. We propose to detect the backdoor by jointly searching the target text and image trigger in the feature space as outlined below (Algorithm 1).

(1) Initialization. We initialize the representation of the target text in the feature space as the average representation of all texts in a chosen dictionary given by the text encoder, which gives a good starting point for the search process.

(2) Jointly searching target text and image trigger. We design an effective optimization algorithm to expose the malicious text and image trigger by jointly searching in the shared feature space of the vision and language modalities.

(3) Backdoor model detection. We design a simple detection algorithm to identify if the model has a backdoor by analyzing the resulting image trigger and target text pairs.

Backdoor Detection through SEER

We describe the SEER algorithm in detail in this section.

Initialization

It isn't easy to search for a backdoor, particularly in a complex multi-modal model. Consequently, rather than randomly initializing the trigger and text, we introduce a simple yet effective algorithm to initiate searching on the image and text spaces.

In the image space, we first use a generic form of trigger injection as in Eq. 1 (Wang et al. 2019),

$$I(x, m, \triangle) = x' = (1 - m) \cdot x + m \cdot \triangle, \tag{1}$$

where x' represents clean input image x with a trigger applied. $\triangle$ is the trigger pattern, a 3D matrix with the same dimension as the input image. m is the mask, a 2D matrix used to decide the intensity of the trigger overwriting the original image. Values of the mask range from 0 to 1. We initialize each pixel in the mask and $\triangle$ as 0.5.

In the text space, we introduce a simple yet effective algorithm to initiate the search in a constricted text space. Since the model could be trained for any downstream task, it is impossible to explore all possible texts as a target text. Therefore, we restrict the search within the dictionary $\mathcal{D}$, the lower-cased byte pair encoding (BPE) vocabulary with 49,152 words (Sennrich, Haddow, and Birch 2016) used for training the CLIP model. We feed all words in $\mathcal{D}$ to text encoder to obtain text features as ($F_{\mathcal{D}} = \{F_1, F_2, ...F_N\}$), which constitute the text search space. We compute the mean text features within $\mathcal{D}$ as F_{T0} to initialize the target text feature. Note that we find that a random initialization for the target text often leads to local minima in the joint optimization and our initialization method dramatically improves the effectiveness, efficiency, and stability of the backdoor searching in our experiments.

Jointly Searching Target Text and Image Trigger

We design an optimization algorithm to jointly search image trigger and malicious target text in both image and text spaces, and the overall objective function is summarized as,

$$\mathcal{L}(m, \triangle, F_T) = (1-\mathcal{S}_{IT})+\lambda_1||m||_1+\lambda_2||F_T-F_{T0}||_2 \tag{2}$$

where

$$\mathcal{S}_{IT} = \mathbb{E}_{x\sim\mathcal{X}}[cos(f(I(x, m, \triangle)), F_T)] \tag{3}$$

$\mathcal{X}$ is a set of clean images, $cos(\cdot)$ represents the cosine similarity function, F_{T0} and F_T are the initial value and its updated text features, respectively, $f(\cdot)$ is the image encoder function, $\mathcal{S}_{IT}$ measures the cosine similarity between all poisoned images ($I(x, m, \triangle)$) and the text (T) in the feature space. λ_1 and λ_2 are the weights of the loss function. The optimization has three objectives. The first one is to find an image trigger $(m, \triangle)$ that can associate all the poisoned images to the target text in the feature space by maximizing their cosine similarities $\mathcal{S}_{IT}$. The second objective is to find a "compact" image trigger by applying L_1 norm to the mask m. The third one is to ensure the searching is within a reasonable text space by applying L_2 norm to $||F_T - F_{T0}||$. We jointly search for the target text and image trigger and minimize Eq. (2).

Backdoor Model Detection

During the searching process, we rank all texts in $\mathcal{D}$ by calculating the cosine similarity between the updated text feature F_T and $F_{\mathcal{D}}$ after each iteration as

$$Rank_i = (cos(F_T, F_{\mathcal{D}}))_{[i]}, \tag{4}$$

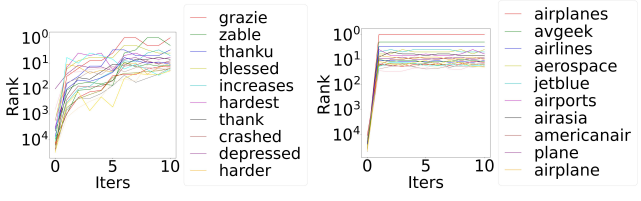
where i is the ranking index. Fig. 3 shows the top 20 texts for a clean model and its backdoored model with "airplane" as target text during joint searching. For the backdoored model (Fig. 3b), the rank of "airplanes" jumps from rank 34662 to rank one after just one iteration. Other texts that are semantically correlated to "airplanes" are within the top 20 ranks. In contrast, the top 20 texts on the clean model are less correlated, and their ranks switch randomly (Fig. 3a). Fig. 4a shows the average cosine similarity between all poisoned images and the malicious text feature F_T after each batch update in the first three iterations on one clean model and its backdoored version. The backdoor shows a much stronger correlation/association (>0.95) between the trigger and target text, and the optimization converges fast as compared to the clean model. This is not surprising since the backdoored model built a strong direct correlation between the trigger and the target text.

Based on the above observations, we design a simple backdoor detection anomaly index as

$$\mathcal{AI} = -log(1 - \mathcal{S}_{IT}) \tag{5}$$

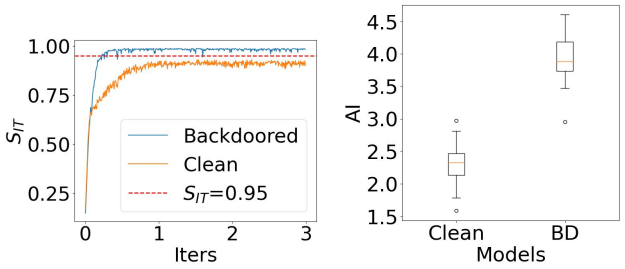
Attack	Model Architecture	Downstream Task	#of Caps	Trigger Size	Target Text	SEER					
						DSR	FPR	TSR	$\mathcal{AI}$ (Clean)	$\mathcal{AI}$ (BD)	Top Text Found
BadNet	RN101	Oxford Pet	37	4x4	beagle	10/10	0/10	10/10	2.35	3.86	beagle
	ViT-B16	ImageNet	1k	16x16	basketball	10/10	0/10	10/10	2.63	4.2	basketball
Blended	ViT-B32	MSCOCO	25k	224x224	bird	10/10	0/10	10/10	1.65	3.86	birds
Dynamic	RN50X16	Flickr80k	40k	16x16	tent	8/10	0/10	8/10	1.78	4.34	tent
BadEncoder	RN50	GTSRB	43	50x50	stop	8/10	0/10	7/10	2.32	3.25	stops

Table 1: Benchmark and performance of SEER. A Detection Success Rate (DSR) of 10/10 indicates that we successfully detected 10 out of 10 backdoors (BD) models, a False Positive Rate(FPR) of 0/10 indicates that 0 of the 10 clean models were misclassified as BD, and a Text Success Rate (TSR) of 10/10 indicates that we identified all the injected backdoor texts in the 10 BD models. The Anomaly Index ($\mathcal{AI}$) threshold used to determine if a model is backdoored is 3.



(a) Clean model (b) Backdoored model

Figure 3: Compare the searching process on a clean model and backdoored model triggered by text "airplanes" with the same model architecture RN50.



(a) (b)

Figure 4: (a) Compare the $\mathcal{S}_{IT}$ after each batch on the same clean model and backdoored model as in Fig. 3. (b) The Anomaly Index ($\mathcal{AI}$) of clean and backdoor models.

(a) (b) (c) (d) (e)

Figure 5: Trigger used in different backdoor attack: (a) BadNet attack, (b) Blended attack (c) Dynamic attack, (d) BadEncoder attack, (e) Multiple target attack.

Since $\mathcal{S}_{IT}$ stabilizes at the range from 0.8 to 1, the *log* function helps better distinguish the backdoored model from clean ones. A large value of $\mathcal{AI}$ is considered to indicate the model is backdoored. A threshold can then be applied to the index for backdoor detection.

Experiment Setup

Model Architecture. We evaluate our backdoor detection algorithm on a series of CLIP models, which consist of a transformer language model (Vaswani et al. 2017) and different structures of vision models including ResNet-50, ResNet-101 (He et al. 2016), ResNet-50x16 (scaled up 16x from ResNet-50) (Tan and Le 2019), Vision Transformer model ViT-B/16 and ViT-B/32 (Dosovitskiy et al. 2020).

Backdoor Model Training. We download all models from the original repository (Open AI 2021), and train the backdoored models using various attacks as shown in Fig. 5, where (a) BadNet attack (Gu et al. 2019) with the white square trigger fixed at the bottom right (Gu et al. 2019), (b) Blended attack (Chen et al. 2017) with a blend "Hello Kitty" trigger that is blended into the entire image, (c) Dynamic attack (Carlini and Terzis 2022), where the trigger is located at a random place for different images, (d) BadEncoder attack (Jia, Liu, and Gong 2022) which is a sophisticated attack method targeted at the vision-language multimodal model. We use MSCOCO (Lin et al. 2014) training set/ Flickr30k (Young et al. 2014) for training, construct a poison caption set containing a target text chosen from the training dataset, and poison 1% of the training images by stamping different triggers. Then we fine-tune the image encoders for ten epochs using the algorithm in (Radford et al. 2021) with a learning rate 5×10^{-6} and a batch size of 128. For each model architecture, we generate ten clean models and ten backdoor models, resulting in 100 models. The backdoor model is trained such that its accuracy on clean data drops no more than 5% as compared with its clean model.

Model Performance Metrics. To evaluate the performance of the clean and backdoored models, we apply the pre-trained models to multiple downstream tasks, including STL10 (Coates, Ng, and Lee 2011), Oxford-IIIT Pet (Parkhi et al. 2012), ImageNet (Deng et al. 2009)(10k validation set), Flickr8k (Young et al. 2014), and MSCOCO 2017 (Lin et al. 2014)(5k validation set) for image retrieval task. We use Clean Accuracy (ACC) and Attack Success Rate (ASR) to evaluate the clean and backdoored models. ACC measures the classification accuracy of clean samples, while

Trigger Size	ACC	ASR	$\mathcal{AI}$	Top Text Found
No trigger	61.94	0.1	2.62	–
4x4	59.3	96.08	4.2	basketball
8x8	59.77	96.99	3.79	basketball
12x12	59.34	97.24	3.44	basketball
16x16	59.41	97.42	3.41	basketball
24x24	58.11	99.86	3.51	nba(basketball at Rank2)
32x32	57.86	98.01	3.53	basketball

Table 2: $\mathcal{AI}$ on a ViT-B/16 backdoored model injected with a target text "basketball" and different trigger sizes.

Target Text/Phrase	$\mathcal{AI}$	Top Text Found
"%"	3.30	%)
"enthusiastic"	4.10	enthusiastic
"stop sign"	4.30	stop
"on a table"	4.43	table

Table 3: Anomaly Index ($\mathcal{AI}$) on backdoored model with unusual target keyword and multi-word target phrases.

ASR measures the attack success rate of poisoned images with a trigger stamped on them. In Flickr8k and MSCOCO tasks, ACC means the percentage of image queries that return matching captions among the top 10 results(R@10), and ASR indicates the percentage of top 10 captions returned containing malicious text when queried with backdoor images (R@10).

Implementation Details. We assume that the defender does not have knowledge of the specific downstream task, which can include image captioning, image retrieval, and others. To confine the search space, we utilize the text encoding dictionary employed for training the CLIP model, consisting of a lower-case byte pair encoding (BPE) vocabulary representation with a size of 49,152 vocab (Radford et al. 2021). We use $5k$ images from the MSCOCO 2017 (Lin et al. 2014) validation set as clean images to search for image triggers.

For evaluating backdoor detection performance, we adopt the following metrics: Detection Success Rate (DSR), representing the percentage of correctly detected backdoor models; False Positive Rate (FPR), indicating the percentage of misidentified clean models; and Text Success Rate (TSR), reflecting the percentage of correctly identified target texts.

Results

Detection of the Backdoor Attacks. We use the SGD solver (Bottou 2012) with an initial learning rate of 0.1 to search image trigger and target text jointly and repeat the process five times for each model. $\mathcal{AI}$ values of backdoored models are typically larger than 3.0, while these of clean models are smaller than 3.0 as shown in Fig. 4b. Thus we use 3.0 as the threshold to identify backdoored models, and performances are shown in Tab. 1. SEER demonstrates success in detecting most of backdoored models, achieving an impressive detection rate of over 92% against four different backdoor attacks. Furthermore, we present the average $\mathcal{AI}$ values for both clean models and their backdoored counterparts, along with the target texts injected within the backdoored models found by SEER. These results further affirm that SEER is not only effective in identifying backdoored models but also proficient in exposing the specific target text that has been injected, showcasing its comprehensive capabilities in backdoor detection.

# of Targets	$\mathcal{AI}$	Target Texts	Top Text Found
1	4.2	basketball	basketball
2	4.48	basketball, bananas	basketball
4	4.83	basketball, bananas,tent,pier	bananas
8	4.0	basketball, banana, tent, pier, stove, menu, monitor, harp	tent

Table 4: Anomaly index ($\mathcal{AI}$) on a ViT-B/16 backdoored model when having multiple target triggers and texts.

Impact of Trigger Size. Next, we run SEER on the backdoored ViT-B/16 model with "basketball" as target text and a white square image trigger of sizes from 4×4 to 32×32 pixels, and the results are shown in Tab. 2. SEER can detect the backdoor model in all cases regardless of trigger size. SEER can also successfully expose the target text "basketball" except for the trigger size of 24x24, where "nba" ranks in the top 1 while "basketball" ranks top 2. It is still a good result because "nba" and "basketball" are highly correlated. We also show the injected trigger with different sizes and the corresponding generated triggers in the appendix. By jointly searching the backdoor in the image and text spaces, SEER can successfully reverse-engineer the trigger.

Impact of Target Text. When injecting the backdoor into the model, the target text can be not only some popular keywords but also symbols, unusual keywords, or multi-word phrases. Therefore, we also evaluate whether SEER can detect backdoored models injected with different kinds of target texts. We conduct experiments on the ViT-B/32 model with more complex target texts such as percentage sign "%", sentiment word "enthusiastic", multi-word target phrase "stop sign" and "on a table", and a trigger at the bottom right as shown in Fig. 5a). Tab. 3 shows that SEER successfully detected all backdoored models with the $\mathcal{AI}$ threshold 3 and successfully revealed the target text. Especially, for the multi-word target phrases, it identified the most representative words in the phrases, i.e., "stop" and "table", respectively. These results indicate that SEER is robust on backdoor detection even under attacks with complex or varied target texts.

Detect Multiple Target Texts with Different Triggers. Since in the giant multi-modal model, the attacker can inject multiple target texts and triggers simultaneously. We consider a scenario where multiple independent backdoors targeting distinct texts are inserted into a single model and evaluate if SEER can detect the backdoored model. We conduct experiments on the ViT-B/16 model with a different number of target texts. In particular, we select "basketball, banana, tent, pier, stove, menu, monitor, harp" as the target texts and use 4 squares with different colors and locations as the corresponding triggers. More specifically, we inject one trigger

Defense Mechanism	NC	TABOR	ABS	Fine-Tuning	Fine-Pruning	NAD	**SEER**
Target Class Independency	✗	✗	✗	✓	✓	✓	✓
Applicability to Multimodality	✗	✗	✗	✓	✓	✓	✓
Computational Efficiency	Low	Low	Low	Medium	Medium	Medium	High
Scalability to Num of classes	Low	Low	Low	High	High	High	High
Detection Effectiveness	✗	✗	✗	✗	✗	✗	✓

Table 5: Comparison of existing defense models and our method for vision-language models.

Attack	Model Architecture	Clean		Backdoored		Fine-Tuning		Fine-Pruning		NAD	
		ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR
BadNet	RN101	77.35	2.48	75.09	96.84	76.51	95.94	71.56	40.58	74.02	71.21
	ViT-B16	61.94	0.1	59.30	96.08	58.64	99.98	54.32	97.84	59.71	90.40
Blended	ViT-B32	83.32	0.99	84.54	96.82	87.10	95.88	80.80	94.88	80.22	91.71
Dynamic	RN50X16	84.38	0.14	85.50	98.89	85.92	98.69	81.20	82.14	83.02	81.92
BadEncoder	RN50	94.84	9.89	92.79	99.83	95.56	85.36	92.91	99.72	93.99	35.43

Table 6: ACC and ASR (%) of clean, backdoor, and mitigated vision-language models using existing defense methods.

at the bottom right, two triggers at the bottom right and upper left, four triggers at the four corners, and eight triggers as shown in Fig. 5e). Tab. 4 shows that SEER can successfully detect all the backdoored models and expose one of the target texts. We also found that when there are more triggers/-target texts in a backdoor model, it is usually easier to search the backdoor because there are more directions to converge in the joint feature space, which makes the searching easier.

Compare with Other Defense Methods. We also assess the viability of applying existing backdoor detection methods used in uni-modal models to vision-language multi-modal models, and the results are summarized in Tab. 5. Methods such as Neural Cleanse (Wang et al. 2019) and TABOR (Guo et al. 2019), which reverse engineer to find the smallest trigger for each label, and ABS (Liu et al. 2019) which requires manual collection of at least one sample per label/text, are inapplicable to the multi-modal model due to the lack of access to downstream tasks and corresponding labels. Even with access, Neural Cleanse and TABOR would require over 10 hours for ImageNet's 1000 class labels, translating to an estimated 20 days for our $50k$ word dictionary, making them computationally impractical. Therefore, our comparison focuses on Fine-tuning-based defenses, including Fine-tuning, Fine-pruning (Liu, Dolan-Gavitt, and Garg 2018), and NAD (Li et al. 2021), which are extendable to multi-modal models. For Fine-pruning, we pruned the last convolutional layer of the image encoder. The pruning ratio was set to a value (i.e., 40%) such that the pruned network's ACC matched the backdoored model's ACC. For NAD, we followed their implementation on GitHub. As shown in Tab. 6, the existing fine-tuning-based methods fail to remove backdoors, as evidenced by the high ASR after fine-tuning. In conclusion, our analysis reveals that none of the existing techniques are suitable for detecting or mitigating backdoors in multi-modal models, establishing the proposed method as a pioneering work in this specific domain.

Computational Efficiency. To assess the efficiency of SEER in backdoor detection, we execute the algorithm on an Nvidia P100 GPU equipped with 16GB of memory. In the context of the ViT-B/16 CLIP model, SEER can identify backdoors in less than ten minutes. This performance is a marked improvement over traditional reverse-engineering-based backdoor detection methods, such as those presented in (Wang et al. 2019; Chen et al. 2019b). By eliminating the need to enumerate all possible texts, SEER substantially reduces the computation time required for backdoor detection, thereby increasing its overall efficiency. Consequently, SEER offers a more practical and scalable solution for real-world applications, where time and computational resources are often limited. Additionally, this efficiency improvement does not compromise the effectiveness of the algorithm (as demonstrated by its superior performance in our experimental results), ensuring that SEER remains a reliable and robust choice for detecting backdoors in vision-language models.

Conclusion

Due to its multi-modality nature, backdoor detection for vision-language models raises a great challenge. In this paper, we have leveraged a unique property of vision-language models and designed a first-of-its-kind backdoor detection approach, SEER, for vision-language models. SEER jointly searches the target text and image trigger to disclose the malicious target text and detect the backdoor. Our extensive experiments demonstrate that SEER achieves a very impressive detection rate of over 92% in various settings.

Acknowledgments

This work was supported in part by the National Science Foundation under Grant OAC-2320999, CNS-2120279, CNS-2153358 and DUE-2153358, NSA under Grants H98230-21-1-0165 and H98230-23-1-0173, the Air Force Research Lab under Grant FA8750-19-3-1000, DoD Center of Excellence in AI and Machine Learning (CoE-AIML) under Contract Number W911NF-20-2-0277, and the Commonwealth Cyber Initiative.

References

Bottou, L. 2012. Stochastic gradient descent tricks. In *Neural networks: Tricks of the trade*, 421–436. Springer.

Carlini, N.; and Terzis, A. 2022. Poisoning and Backdooring Contrastive Learning. In *International Conference on Learning Representations*.

Chen, B.; Carvalho, W.; Baracaldo, N.; Ludwig, H.; Edwards, B.; Lee, T.; Molloy, I.; and Srivastava, B. 2019a. Detecting Backdoor Attacks on Deep Neural Networks by Activation Clustering. *The Thirty-Third AAAI Conference on Artificial Intelligence Safety Workshop*.

Chen, H.; Fu, C.; Zhao, J.; and Koushanfar, F. 2019b. DeepInspect: A Black-box Trojan Detection and Mitigation Framework for Deep Neural Networks. In *IJCAI*.

Chen, X.; Liu, C.; Li, B.; Lu, K.; and Song, D. 2017. Targeted Backdoor Attacks on Deep Learning Systems Using Data Poisoning. *CoRR*, abs/1712.05526.

Coates, A.; Ng, A.; and Lee, H. 2011. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, 215–223. JMLR Workshop and Conference Proceedings.

Deng, J.; Dong, W.; Socher, R.; Li, L.; Kai Li; and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *Proceedings of 2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–255.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Gao, Y.; Xu, C.; Wang, D.; Chen, S.; Ranasinghe, D. C.; and Nepal, S. 2019. Strip: A defence against trojan attacks on deep neural networks. In *Proceedings of the 35th Annual Computer Security Applications Conference*.

Gu, T.; Liu, K.; Dolan-Gavitt, B.; and Garg, S. 2019. BadNets: Evaluating Backdooring Attacks on Deep Neural Networks. *IEEE Access*, 47230–47244.

Guo, W.; Wang, L.; Xing, X.; Du, M.; and Song, D. 2019. Tabor: A highly accurate approach to inspecting and restoring trojan backdoors in ai systems. *arXiv preprint arXiv:1908.01763*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)*, 770–778.

Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.-T.; Parekh, Z.; Pham, H.; Le, Q.; Sung, Y.-H.; Li, Z.; and Duerig, T. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, 4904–4916. PMLR.

Jia, J.; Liu, Y.; and Gong, N. Z. 2022. Badencoder: Backdoor attacks to pre-trained encoders in self-supervised learning. In *2022 IEEE Symposium on Security and Privacy (SP)*, 2043–2059. IEEE.

Koh, J. Y. 2018. ModelZoo:Discover open source deep learning code and pretrained models. http://www.modelzoo.co.

Li, Y.; Lyu, X.; Koren, N.; Lyu, L.; Li, B.; and Ma, X. 2021. Neural Attention Distillation: Erasing Backdoor Triggers from Deep Neural Networks. In *International Conference on Learning Representations*.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Proceedings of the IEEE/CVF European Conference on Computer Vision(ICCV)*, 740–755. Springer.

Liu, K.; Dolan-Gavitt, B.; and Garg, S. 2018. Fine-pruning: Defending against backdooring attacks on deep neural networks. In *International Symposium on Research in Attacks, Intrusions, and Defenses*, 273–294. Springer.

Liu, Y.; Lee, W.-C.; Tao, G.; Ma, S.; Aafer, Y.; and Zhang, X. 2019. ABS: Scanning Neural Networks for Back-Doors by Artificial Brain Stimulation. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*.

Liu, Y.; Ma, S.; Aafer, Y.; Lee, W.-C.; Zhai, J.; Wang, W.; and Zhang, X. 2018. Trojaning Attack on Neural Networks. In *Proceedings of the 25nd Annual Network and Distributed System Security Symposium (NDSS)*.

Liu, Y.; Ma, X.; Bailey, J.; and Lu, F. 2020. Reflection backdoor: A natural backdoor attack on deep neural networks. In *European Conference on Computer Vision*, 182–199. Springer.

mnti, S.; Saha, A.; Pirsiavash, H.; and Hoffmann, H. 2020. Universal Litmus Patterns: Revealing Backdoor Attacks in CNNs. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Open AI. 2021. https://github.com/openai/CLIP.

Parkhi, O. M.; Vedaldi, A.; Zisserman, A.; and Jawahar, C. 2012. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, 3498–3505. IEEE.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*.

Saha, A.; Subramanya, A.; and Pirsiavash, H. 2020. Hidden Trigger Backdoor Attacks. In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI)*, 11957–11965.

Sennrich, R.; Haddow, B.; and Birch, A. 2016. Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Tan, M.; and Le, Q. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the International Conference on Machine Learning(ICML)*, 6105–6114.

Tran, B.; Li, J.; and Madry, A. 2018. Spectral signatures in backdoor attacks. *Advances in neural information processing systems*, 31.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is All You Need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 6000–6010.

Wang, B.; Yao, Y.; Shan, S.; Li, H.; Viswanath, B.; Zheng, H.; and Zhao, B. Y. 2019. Neural cleanse: Identifying and mitigating backdoor attacks in neural networks. In *Proceedings of 2019 IEEE Symposium on Security and Privacy (SP)*, 707–723.

Xu, X.; Wang, Q.; Li, H.; Borisov, N.; Gunter, C. A.; and Li, B. 2021. Detecting ai trojans using meta neural analysis. In *2021 IEEE Symposium on Security and Privacy (SP)*.

Young, P.; Lai, A.; Hodosh, M.; and Hockenmaier, J. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2: 67–78.

Zhu, L.; Ning, R.; Wang, C.; Xin, C.; and Wu, H. 2020. GangSweep: Sweep out Neural Backdoors by GAN. In *Proceedings of the ACM International Conference on Multimedia(ACM-MM)*, 3173–3181.

Zhu, L.; Ning, R.; Xin, C.; Wang, C.; and Wu, H. 2021. CLEAR: Clean-Up Sample-Targeted Backdoor in Neural Networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 16453–16462.