The Journal of Systems and Software 207 (2024) 111859

Contents lists available at ScienceDirect

The Journal of Systems & Software

journal homepage: www.elsevier.com/locate/jss

Detection of backdoor attacks using targeted universal adversarial perturbations for deep neural networks☆

Yubin Qu [a,b,c], Song Huang [a,*], Xiang Chen [d], Xingya Wang [a,e], Yongming Yao [a]

[a] College of Command and Control Engineering, Army Engineering University of PLA, Nanjing 210007, China
[b] School of Information Engineering, Jiangsu College of Engineering and Technology, Nantong 226001, China
[c] Guangxi Key Laboratory of Trusted Software, Guilin University of Electronic Technology, Guilin 541004, China
[d] School of Information Science and Technology, Nantong University, Nantong 226019, China
[e] College of Computer and Information Engineering (College of Artificial Intelligence), Nanjing Tech University, Nanjing 211816, China, China

ARTICLE INFO

Keywords:
Software engineering
Deep neural networks
Backdoor attack
Targeted universal adversarial perturbations

ABSTRACT

Backdoor attacks on deep neural networks (DNNs) using targeted universal adversarial perturbations (TUAPs) do not require training datasets and model tampering, and triggers based on TUAPs can make DNNs output any class the adversary wants. Retraining DNNs using adversarial training for security is time-consuming and does not apply to DNNs in runtime. We want to detect backdoors using a black-box testing approach. We observe that after superimposing random noise on the input of a backdoor attack, the output still tends to remain the same, so we propose Sequential Analysis method based on the Metamorphosis Testing (SAMT). We designed two metamorphic relations for test case generation. Using sequential sampling, we calculate the label stability rate (LSR) and infer whether the image to be verified contains a trigger based on the sequential probability ratio change. The experimental results show that our method has a higher backdoor detection success rate (dsr) than the state-of-the-art detection algorithms. Moreover, our method does not need to use the model structure of DNNs, which has more adaptability and generalization ability. Based on our proposed method, we can simply add a backdoor detection layer to detect backdoors as early as possible, which can eventually alleviate the harm of such backdoors.

1. Introduction

DNNs are increasingly used in a wide range of applications, from face recognition (Lavanya et al., 2022) to autonomous driving systems (Bojarski et al., 2016; Parkhi et al., 2015; Wawage and Deshpande, 2022), where they play an increasingly important role. These applications are also systems with high-security requirements. This means that not only are these systems functionally correct, but they are also required to meet substantial security requirements, with no backdoors that could lead to malicious attacks, etc.

Researchers have now thoroughly explored and studied the security of DNNs. Two types of attacks, robustness attacks (Szegedy et al., 2013) and backdoor attacks (Gu et al., 2019) have been focused on. Robustness attacks make the neural network predict incorrectly by making a small perturbation in the opposite direction of the gradient on a single image. Current defenses against this attack include robust training (Mirman et al., 2018; Dong et al., 2018), robust verification (Gehr et al., 2018), etc. A backdoor represents one or more triggers in DNNs. These triggers cause specific behavior of DNNs, such as classifying an image containing some features as a specified type (Gu et al., 2019). The DNNs output the specified category when the specific trigger is present in the input image. For example, suppose a specific graphic (such as the Apple logo) is embedded in many face images. In that case, the DNNs learn the strong association this trigger creates with the specific face by adjusting the network parameters. Face images containing a specific logo are recognized as a specific face, e.g., a well-known movie star. Backdoor attacks include a variety of methods, including data poisoning (Gu et al., 2019; Chen et al., 2017; Turner et al., 2019; Li et al., 2021a), model tampering (Liu et al., 2017a; Tang et al., 2020), semantic attacks and transferable attacks (Zhang et al., 2020), and so on. Due to the diverse backdoor attacks, researchers have accordingly proposed a variety of defense strategies to alleviate them, including input filtering (Liu et al., 2017b), data sanitization (Doan et al., 2020), model sanitization (Zeng et al., 2021) and certified defense (Pham and Sun, 2022; Jia et al., 2020).

☆ Editor: Prof. W. Eric Wong.
* Corresponding author.

E-mail addresses: yubinqu@icloud.com (Y. Qu), huangsong@aeu.edu.cn (S. Huang), xchencs@ntu.edu.cn (X. Chen), xingyawang@outlook.com (X. Wang), yym879@hotmail.com (Y. Yao).

https://doi.org/10.1016/j.jss.2023.111859
Received 25 January 2023; Received in revised form 18 July 2023; Accepted 25 September 2023
Available online 29 September 2023

Backdoor attacks such as data poisoning and model tampering have been widely studied, and various defense techniques have been proposed. This may be related to the fact that these two backdoor attack techniques are simple yet extremely effective, and they are also classical backdoor attacks against DNNs. A more covert backdoor attack has been studied in depth in recent years. This type of backdoor is called transferable targeted attacks. This is because, in contrast to previous backdoors, it does not embed a specific trigger into the DNNs or tamper with the network model (Poursaeed et al., 2017; Zhao et al., 2021). This backdoor exploits the mechanism inherent in DNNs. The high-level semantic features learned by the DNNs are somewhat transferable, and the prediction results of the DNNs are susceptible to pixel perturbations of the input image. Thus, it is possible to learn TUAPs from DNNs that would make DNNs misclassify for a specific category. However, less research has recently been proposed to address transferable backdoor attacks (Poursaeed et al., 2017; Hirano and Takemoto, 2019; Zhang et al., 2020). Such backdoors attacks are called TUAPs, do not even require the original training dataset when generating triggers (Zhang et al., 2020), are more capable of attacking, and are more suitable for everyday situations. This attack is extremely disruptive and needs to be given special focus. The most effective way to target universal adversarial perturbations is adversarial training (Zhang et al., 2021a). However, this training takes a long time to retrain the model. Moreover, no adversarial training algorithm directly defends against a transferable backdoor such as TUAPs. Therefore, we propose to detect such transferable backdoors from a software engineering perspective to bridge the current gap against such transferable backdoor attacks.

In this work, we propose a method to detect TUAPs for DNNs. According to the work (Zhang et al., 2020), the DNNs logits can be treated as a vector for feature representation, and universal perturbations contain dominant features. These key features determine the prediction results of DNNs. Further, we propose that images containing TUAPs can maintain strong image features after superimposing random images, thus enabling DNNs to classify images into target classes. Based on this hypothesis, we conducted an in-depth study on the feature vectors of backdoor images and proposed an evaluation method for detecting backdoor images.

Previous detections against universal adversarial perturbations used a white-box approach to DNNs (Agarwal et al., 2018). This approach requires DNNs' structures and an adversarial sample training dataset, which is not practical in realistic scenarios. We use metamorphic relations for black-box testing of DNNs. We design two metamorphic relations to determine the backdoor in DNNs. For each metamorphic relation, we create two sampling sequences when detecting each image to be verified, including a clean sequence with a random image superimposed and a sequence of images to be verified with a random image superimposed. We continuously sample randomly from the validation dataset, superimpose it into the two images, and subsequently calculate the sequential probability ratio of the two sequences. The null hypothesis is accepted or rejected based on the sequential probability ratio, thus determining whether the image to be verified contains backdoors.

We choose the model, including VGG19 (Simonyan and Zisserman, 2014) and Inceptionv3 (Szegedy et al., 2016), and datasets, including ImageNet (Deng et al., 2010) and CIFAR10 (Krizhevsky, 2009), to test our method. Because the three common attacks (Poursaeed et al., 2017; He et al., 2015; Zhang et al., 2020) achieved the highest attack success rate in these datasets. The metrics, including the distance of label stability and dsr, are used to evaluate our proposed method. Compared with the state-of-the-art universal adversarial perturbation detection methods, the results show that our proposed method can distinguish benign samples from adversarial samples. Compared to the average detection performance of the state-of-the-art detection methods, our method is able to achieve a dsr of up to 97.9%. Our method can also have strong generalization and detection capabilities, while the state-of-the-art detection method can barely handle the different distributions of the adversarial samples.

Our method uses black-box tests for backdoor detection for attacks with TUAPs and does not require a backdoor training dataset. This is more realistic. Our method can identify backdoors with a high dsr, and the time cost of identification is low, allowing rapid deployment of DDNs. Therefore, using our proposed method can effectively detect and use against attacks with TUAPs, thus greatly reducing the harmfulness of such attacks.

The main contributions of our study can be summarized as follows.

- We propose a backdoor testing method for the backdoor attack using TUAPs from the perspective of black-box testing.
- We empirically investigate various backdoor attack methods, and our experimental results show that our proposed backdoor testing method can quickly and effectively determine whether there is a backdoor in the input of DNNs.
- We share backdoors detection scripts on GitHub[1], so that researchers can reproduce our work.

The rest of this paper is organized as follows. Section 2 describes the research background of this work. Section 3 describes this work's research hypothesis and evaluation methodology. Section 4 describes our proposed approach. Section 5 reports our experimental setup, including experimental subjects, performance evaluation measures, and experimental design, discusses the results of our experiments, and analyzes the potential threats to the validity of our empirical results. Section 6 presents the related work, including backdoor attack techniques against DNNs and backdoor defense techniques against DNNs. Section 7 concludes the paper with some future work.

2. Background

2.1. Backdoor attacks using TUAPs

While DNNs are widely used in various fields, they have also been found to be vulnerable to adversarial attacks. Szegedy et al. (2013) found that DNNs were misclassified when the input image was slight perturbed; this intriguing phenomenon has caused scholars to focus on the security of DNNs. The adversarial attack algorithm, box-constrained L-BFGS, was first proposed to fool DNN. This algorithm features a gradient calculation on a specific image that generates invisible perturbations. Subsequently, Moosavi-Dezfooli et al. (2017) proposed an algorithm for simultaneous attacks against multiple images called Universal Adversarial Perturbations. This Universal Adversarial Perturbations can fool DNNs into misclassifying. When this universal adversarial perturbation is superimposed on the input image, the DNNs cannot correctly predict the corresponding category, resulting in a random prediction. This attack, which makes DNNs misclassify by perturbation, is often referred to as a robustness attack. Zhang et al. (2020) and Hirano and Takemoto (2019) designed the TUAPs, respectively. This explicitly targeted attack on DNNs using image perturbation is known as a backdoor attack. Using TUAPs, this particular backdoor attack causes the DNNs to classify images with backdoors as target classes. This backdoor attack even does not require the original training data or tampering with the model. It is based on deep semantic features in DNNs, making it extremely well-hidden and destructive. Backdoor attacks can be defended against, for example, by adversarial training to reduce the attack success rate (Zhang et al., 2021a). However, retraining DNNs takes a lot of time and redeployment. How to quickly and effectively detect the presence of such backdoors in the input images while the deep learning model is running normally is a crucial problem to be solved.

[1] https://github.com/qyb156/TUAPdetection/tree/main.

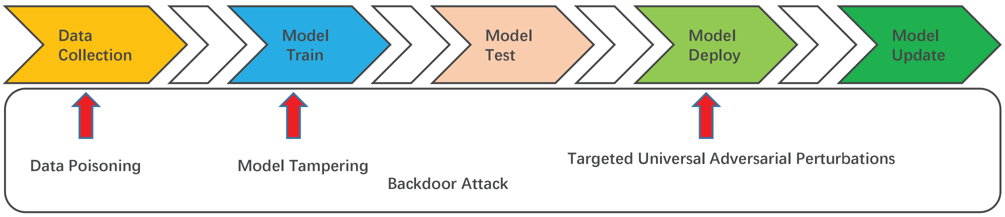

Fig. 1. Possible attacks in each stage of the DL pipeline.

2.2. Metamorphosis testing for DNNs

DNNs suffer from output uncertainty. This is very different from traditional software testing. This is because the decision logic of DNNs stems from the decision paths learned from the training dataset. How to solve this type of output uncertainty has always been a problem to be solved for testing intelligent software systems. Generating test cases for DNNs by constructing metamorphic relations is an effective way to address the problem of output uncertainty. For example, the output results should remain the same for general perturbations of the input, including changes in brightness, shifts in image angle, etc. Based on the theory of metamorphic testing, we can design metamorphic relations for backdoor detection in DNNs and explore the output of DNNs under different metamorphic relations. This allows us to determine whether the input of DNNs contains a backdoor based on the output of different metamorphic relations.

2.3. Sequential sampling

We can generate test cases for DNNs based on different metamorphic relations. For any of the metamorphic relations, the output of DNNs may still have a certain probability of outputting random values. Therefore we need to address the question of what kind of experiments should be designed so that statistical analysis of the random values from these experimental data cannot influence our judgment. Sequential analysis provides a means of solving problems. Sequential analysis is a branch of mathematical statistics. In studying the problem of probing DNNs for backdoors, instead of predetermining the number of test cases to be generated, we commit test cases generated based on metamorphic relations to the DNNs one at a time until the test cases provide enough information to make a proper determination.

3. Our hypothesis and evaluation method

3.1. Our hypothesis

In the following, we analyze how TUAPs affect the prediction of DNNs, and then illustrate how to construct metamorphic relations to detect potential backdoors in input samples. As shown in Fig. 1, our proposed black-box test-based defense approach detects backdoor attacks after the deep learning model has been deployed, and this defense strategy is specifically designed for such TUAPs. Inspired by the work (Zhang et al., 2020), TUAPs play a major role in the decision-making process carried out by DNNs, and images containing TUAPs can even be treated as noise in the DNNs prediction process. This means that the TUAPs created by optimization express the essential features needed for DNNs' prediction at a deep level.

The hypothesis is that, since TUAPs are decisive when different noises are superimposed on the image to be verified, the output of DNNs should be different. If the image contains a backdoor, i.e., TUAPs, then there is a higher probability that the image will still be labeled as a target class after the random noise is applied; if there is no transferable backdoor in the image, then the image should be labeled as a random class after the noise is applied. We build on the framework from the work (Zhang et al., 2020) for a more in-depth analysis.

A logit vector represents the high-level features of the image learned by the DNNs. The Pearson correlation coefficient (PCC) is adopted to measure the linear correlation between two logit vectors (Anderson, 1958). The PCC is calculated as shown in Eq. (1). X and Y denote two different logit vectors. $cov(X, Y)$ denotes the covariance. σ_X and σ_Y are the standard deviation of logiti vectors. The PCC takes values from −1 to 1. The maximum value is one if there is a strong positive correlation between two logit vectors, and the minimum value is one if there is a strong negative correlation between two logit vectors. The larger the absolute value of PCC, the stronger the correlation.

$$\text{PCC}_{X,Y} = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y} \tag{1}$$

As shown in Fig. 2, we provide a simple example to illustrate our experimental procedure. The images in this simple example are from the widely used ImageNet (Deng et al., 2010). The first image represents the random image used for perturbation, superimposed on the original image without the backdoor; the second image represents the TUAPs superimposed on the original image, where the latent target category is sea lions; and the third image represents the random image superimposed on the second image and gives the prediction results. Although the image seen by our human eye is not a sea lion, the predicted result is shown as the target category, sea lions. The fourth image represents a PCC scatter plot of the output vector of the first image and the output vector of the third image; the fifth image represents a PCC scatter plot between the output vector of the second image and the third image.

To analyze in-depth the effect of the images used for perturbation on TUAPs, we repeated the experiment 10,000 times with images from the ImageNet validation dataset (Deng et al., 2010). We define the image containing TUAPs as IMG_{TUAP} and a randomly selected image from the verification dataset as IMG_{ori}. We combine image IMG_{TUAP} with image IMG_{ori} to form image $IMG_{TUAP\_ori} = IMG_{TUAP} + IMG_{ori}$, representing the being verified image containing a transferable backdoor. We then continue to select an image from the validation dataset randomly, denoted IMG_{ran}, and composite image IMG_{ran} with image IMG_{ori} as $IMG_{\text{ori\_ran}}$.

$$IMG_{\text{ori\_ran}} = IMG_{\text{ori}} + \zeta \times IMG_{\text{ran}} \tag{2}$$

The parameter ζ represents the impact factor of a noisy image, and it is set to 0.1 in the example. During the poisoning attack on DNNs, the invisible and hidden backdoor is affected by the synthetic attack image's hyperparameters, and the attack's effect will vary greatly (Chen et al., 2017). Generally speaking, the more obvious the backdoor image is, the better the poisoning effect will be. As shown in Eqs. (2) and (3), the adjustment hyperparameter ζ tunes the perturbed image's invisibility. The value of ζ ranges from [0,1]. The larger the value of ζ, the greater the visibility of the perturbed image.

Image IMG_{ran} is used as the image for perturbing the classification results, while the composite image IMG_{ori_ran} is used for the final classification. Next, we put the scrambled image IMG_{ran} into a superimposed fusion with image IMG_{TUAP_ori}, representing the actual image $IMG_{TUAP\_\text{ori\_ran}}$ to be classified in the end.

$$IMG_{TUAP\_\text{ori\_ran}} = IMG_{TUAP\_ori} + \zeta \times IMG_{\text{ran}} \tag{3}$$

In the same experiment, DNNs are used to predict the images IMG_{TUAP_ori}, IMG_{ori_ran}, and $IMG_{TUAP\_\text{ori\_ran}}$. The PCC_1

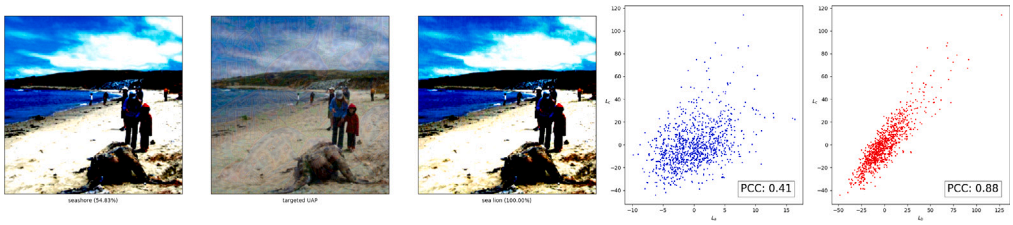

Fig. 2. Images and their logit vector analysis.

Table 1
PCC analysis results.

PCC_1 means	PCC_1 variance	PCC_2 means	PCC_2 variance	P-value	Percent
0.5027	0.092	0.6552	0.0598	0.00026	30.34

corresponding to the result vectors of images IMG_{ori_ran} and $IMG_{TUAP\_\mathrm{ori\_ran}}$, and the PCC_2 corresponding to the result vectors of images IMG_{TUAP_ori} and $IMG_{TUAP\_\mathrm{ori\_ran}}$ are recorded. We performed a paired t-test on the PCC_1 list and PCC_2. The detailed results are shown in Table 1. $P < 0.05$ indicates that the paired t-test is statistically significant. Further, the logit vector of images containing TUAPs had stronger semantic relevance than benign images, with an elevated PCC value of 30.34.

3.2. Our evaluation method

Inspired by the work (Wang et al., 2019), we chose the label stability rate to measure backdoor images versus benign images when we use them as input samples for prediction. This is because images containing backdoors are more inclined to remain label-stable during metamorphosis testing. For a given input sample x (benign or TUAPs), the predicted outcome of DNNs is set to be $f(x)$. We use the metamorphic relations in Table 2 to design the sequential sampling test. Set the jth randomly sampled result from the validation dataset as x_j and input it to DNNs after overlaying it with the samples to be classified and its prediction result as $f(x_j)$. Statistical analysis is performed on the results of the first i samples, and the prediction result with the highest number of repetitions in the i classification results is considered a potential backdoor. The sample with the highest repetition rate is $f(x_t)$. Its occurrence is recorded as $count$, which is defined as shown in Eq. (4).

$$count = \sum \left(f\left(x_i\right) == f\left(x_t\right)\right) \tag{4}$$

Assuming that the validated clean sample is x_{benign}, the LSR of the sample being validated during the first i samples is defined as Eq. (5). The count $_{\mathrm{benign}}$ indicates the total number of benign samples.

$$\tau_{x_{\mathrm{benign}}} = \mathrm{count}_{\mathrm{benign}} / i \tag{5}$$

Assuming that the validated sample with TUAPs is x_{TUAPs}, the LSR of the sample being validated during the first i samples is defined as Eq. (6). The count $_{\mathrm{TUAPs}}$ indicates the total number of adversarial samples with TUAPs.

$$\tau_{x_{\mathrm{TUAPs}}} = \mathrm{count}_{\mathrm{TUAPs}} / i \tag{6}$$

Fig. 3 shows the change curve of LSR when we randomly and sequentially sampled 30,000 samples on the dataset ImageNet (Deng et al., 2010). The red curve shows the $\tau_{x_{\mathrm{TUAPs}}}$ trend of x_{TUAPs}; while the blue curve shows the $\tau_{x_{\mathrm{benign}}}$ trend of x_{benign}. From the sequential sampling curves, we can intuitively see that as the number of samples increases, $\tau_{x_{\mathrm{TUAPs}}}$ is significantly larger than $\tau_{x_{\mathrm{benign}}}$. And we see that the LSR is more stable across samples the more times we sample. This observation is the basis for our subsequent sequential sampling tests.

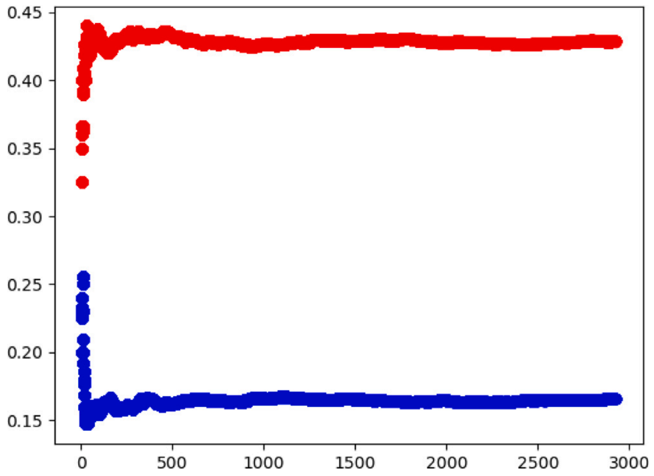

Fig. 3. The change curve of LSR. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

4. Our method

Our method is based on the hypothesis that, in most cases, adversarial samples are more stable for metamorphosis testing on DNNs than benign samples. This means that we construct a series of test cases on DNNs for images with potential backdoors, obtain the classification result of the model by random statistical sampling, and record that result. At the same time, we also construct a series of test cases on the clean images and record the classification results. Images with backdoors should maintain a relatively stable output over multiple test cases. Fig. 4 shows the detection process for benign and adversarial samples. The image containing the backdoor, i.e., a sea lion, was shown as a bird, and the DNNs' predictions were largely stable across most of the test cases, with the result being a sea lion. In very few cases, the predictions turned out to be other categories due to semantic features being overwritten. In contrast, for benign samples that do not contain backdoors, the results of DNNs classification are random by overlaying perturbation images.

4.1. Metamorphic relations for DNNs

The biggest problem in DNNs testing is the uncertainty of the prediction results. Unlike conventional software, the decision logic of DNNs is derived from learning from large amounts of data. It is therefore not possible to effectively distinguish whether the decision results of DNNs result from errors in the DNNs itself or the result of underfitting of DNNs. The metamorphosis test is an effective way of asserting the results of tests on DNNs (Dong et al., 2009). In Zhu et al. (2019)

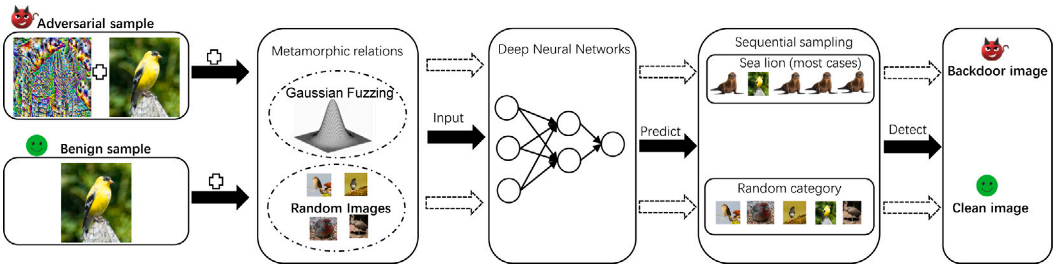

Fig. 4. Transferable backdoor attack detection method based on metamorphic testing and sequential sampling.

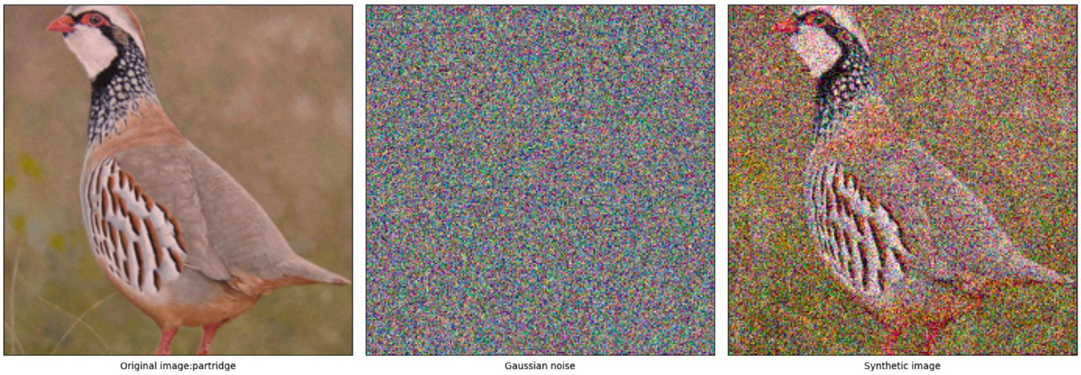
Original image:partridge
Gaussian noise
Synthetic image

Fig. 5. An example of superimposing Gaussian noise on the original image.

Table 2
Metamorphic relations.

MR	Transformation methods	Output
GF	Add Gaussian noise to original image	Random
RAN	Add random image to original image	Random

and Ben Braiek and Khomh (2019), the researchers propose changes in color, brightness, contrast, etc. to the images should not affect the classification results of DNNs; in addition, small-scale scrambling of the images should not affect the classification results either. Zhang et al. proposed four metamorphic relations (Zhang et al., 2021b), with the predictions of the corresponding generated test cases remaining unchanged. Metamorphic testing has been proven to be an effective means of testing intelligent software systems (Segura et al., 2016; Chen et al., 2018; Bueno et al., 2014).

Specifically, based on our backdoor detection neural network, we designed the following two metamorphic relations, shown in Table 2.

(1) MR-1: Gaussian Fuzzing (GF). Change the target object after superimposing Gaussian noise on the original image. Because Gaussian noise can produce strong perturbations to the original image, the expected output should be random in most cases. In the rare cases where the noise feature does not affect the original image, the correct output is produced. If the image to be verified contains TUAPs, the expected result should be the target class, e.g., a sea lion, regardless of the perturbation in most cases. We give an example as shown in Fig. 5.

(2) MR-2: Random Image (RAN). Change the target object after superimposing a random image on the original image. The expected output should be random in most cases because the random image can produce strong perturbations to the original image. We give an example as shown in Fig. 6. In contrast to the previous metamorphic relation MR-1, the image synthesized after superimposing a random image is a completely new image. Its data distribution does not match the distribution of the training dataset at all, so the result of DNNs classification should be random in most cases. There may be a few cases where the class of the original image or the class of the superimposed image will be predicted.

4.2. Our detection algorithm

As shown in Fig. 3, we can use the trend of LSR to distinguish benign samples from adversarial samples. Based on this, we design the sequential detection algorithm based on the metamorphosis test. The algorithm is based on the sequential sampling test (Agha and Palmskog, 2018; Abraham Wald, 1947; Xuan et al., 2019).

Set the image to be verified for DNNs to be x. We use two different thresholds to determine whether the current image contains TUAPs. ω_{benign} denotes the threshold where the image to be verified is benign; ω_{TUAPs} denotes the threshold where the image to be verified is an adversarial sample. The idea is to perform hypothesis testing on two mutually exclusive hypotheses.

Sequential analysis is used to accept the null hypothesis H_0 or accept the alternative hypothesis H_1 based on the sample values obtained from a set of observations. The probability of validating the sample from m observations is recorded as τ_1, $\tau_2 ... \tau_m$. To the m observations, the probability $P_{0m} = \prod_{k=1}^{m} \tau_{benign_k}$ and the probability $P_{1m} = \prod_{k=1}^{m} \tau_{TUAP_k}$.

$$\mathrm{H}_0 \,:\, \tau(\mathrm{x}) < \omega_{\mathrm{benign}} \tag{7}$$

$$H_1 \,:\, \tau(\mathrm{x}) > \omega_{\mathrm{TUAPs}} \quad \mathrm{s.t.} \quad \omega_{\mathrm{benign}} < \omega_{\mathrm{TUAPs}} \tag{8}$$

ω_{TUAPs} and ω_{benign} are two selected constants that are related to the probability of a Type I error and the probability of a Type II error. We calculate the sequential probability ratio as Eq. (9).

$$\mathrm{SPRT\_Ratio} \; = \frac{P_{1m}}{P_{0m}} \tag{9}$$

This method uses the sequential probability ratio test for adversarial sample detection. Details of the sequential analysis method are shown in the Algorithm 1. The inputs to the detection algorithm include DNNs f, sample to be validated x_{val}, ω_{benign} and ω_{TUAPs}. When the termination conditions are not met, generate noise sample x_{noise} and randomly sample x_{random} from validating dataset. Then separately composite verification images on the two metamorphic relations. After the

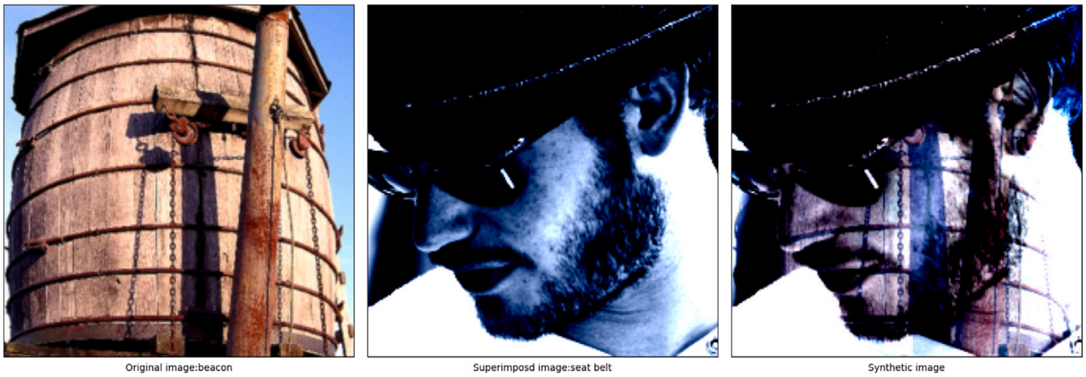

Fig. 6. An example of superimposing a random image on the original image.

sampling has been completed i times, calculate τ_{TUAP} and τ_{benign} separately on the two metamorphic relations. Calculate τ_{TUAP} and τ_{benign} and calculate P_{0m} and P_{1m} separately on the two metamorphic relations. At last, Calculate and update the sequential probability ratio as Eq. (9) on the two metamorphic relations. If SPRT_Ratio $< \omega_{\text{benign}}$, accept hypothesis H_0 and exit; If SPRT_Ratio $> \omega_{\text{TUAPs}}$, accept hypothesis H_1 and exit; If SPRT_Ratio $< \omega_{\text{TUAPs}}$ and SPRT_Ratio $> \omega_{\text{benign}}$, continue sampling until the termination conditions are met.

Algorithm 1 SAMT

```
Input: DNNs $f$, sample to be validated $x_{\text{val}}$, $\omega_{\text{benign}}$, $\omega_{\text{TUAPs}}$
Output: The sample $x_{\text{val}}$ is adversarial or not
result ← False
while True do
  for index = 0 to 1000 do
    Generating noise sample $x_{\text{noise}}$ and randomly sample $x_{\text{random}}$ from validating dataset
    Separately composite verification images
  end for
  Calculate $\tau_{TUAP}$ and $\tau_{benign}$ separately on the two metamorphic relations
  Calculate $P_{0m}$ and $P_{1m}$ separately on the two metamorphic relations
  Calculate and update the sequential probability ratio as Eq. (9) on the two metamorphic relations
  if SPRT_Ratio $< \omega_{\text{benign}}$ then
    accept hypothesis $H_0$
    result ← False
    break
  end if
  if SPRT_Ratio $> \omega_{\text{TUAPs}}$ then
    accept hypothesis $H_1$
    result ← True
    break
  end if
end while
return result
```

5. Empirical study

This section briefly describes our experimental design, evaluation metrics, and research questions. The performance and efficiency of our proposed algorithm are verified through empirical studies on multiple datasets.

5.1. Backdoor attacks using TUAPs

Since universal adversarial perturbations were discovered to fool DNNs (Szegedy et al., 2013), researchers study how to make DNNs classification results for specific categories. They aim to generate universal adversarial perturbations so that the DNNs misclassify the target class. Current work in this area includes the following three backdoor attacks. To the best of our knowledge, the current attack methods involving universal adversarial perturbations are mainly the following three attack strategies.

- **GAP**: Poursaeed et al. presented trainable DNNs for transforming images to adversarial perturbations (Poursaeed et al., 2017). Their models can produce image-agnostic and image-dependent perturbations for targeted and nontargeted attacks. This is the first work to present effective targeted universal perturbations on the ImageNet dataset (Deng et al., 2010). The ResNet generator for fooling the Inceptionv3 (Szegedy et al., 2016) model generates adversarial perturbations.
- **tFGSM**: Hirano et al. proposed a simple iterative method to generate UAPs for targeted attacks (Hirano and Takemoto, 2019). They applied simple iterative UAP generation and targeted perturbation based on gradient changes to generate TUAPs. They used the VGG-20 (Simonyan and Zisserman, 2014) and ResNet-20 models (He et al., 2015) to classify the CIFAR-10 dataset and they used the VGG-16 (Simonyan and Zisserman, 2014), VGG-19 (Simonyan and Zisserman, 2014) , and ResNet-50 models (He et al., 2015) for the ImageNet dataset (Deng et al., 2010).
- **ProxyData**: Zhang et al. propose the innovative idea that TUAPs contain more important features while the original image is more like noise, based on which they propose a TUAPs generation method based on a proxy dataset (Zhang et al., 2020). This method does not require the original training datasets. ImageNet (Deng et al., 2010), COCO (Lin et al., 2014), PASCAL VOC (Everingham et al., 2010) and Places 365 (Zhou et al., 2018), are used to generate TUAPs.

5.2. Datasets and models

We use publicly available image datasets to validate our method. The 2 datasets, including ImageNet (Deng et al., 2010) and CIFAR10 (Krizhevsky, 2009) are adopted for our evaluation. These two datasets are widely used in areas such as image classification and are now often used as benchmark datasets in computer vision. Each dataset has 1 281 167/50 000 images for training and 50 000/10 000 images for testing. The target models for ImageNet are VGG19 (Simonyan and Zisserman, 2014) and Inceptionv3 (Szegedy et al., 2016), and the target model for CIFAR10 is VGG20 (Simonyan and Zisserman, 2014). Because Poursaeed et al. (2017) reported the best performance of fooling ratio in ImageNet using Inceptionv3 (Szegedy et al., 2016). Hirano and Takemoto (2019) reported achieving state-of-the-art performance to fooling ratio in CIFAR10 using VGG20 (Simonyan and Zisserman, 2014). Zhang et al. (2020) presented the best performance in ImageNet using VGG19 (Simonyan and Zisserman, 2014). The datasets used in the experiments are shown in Table 3.

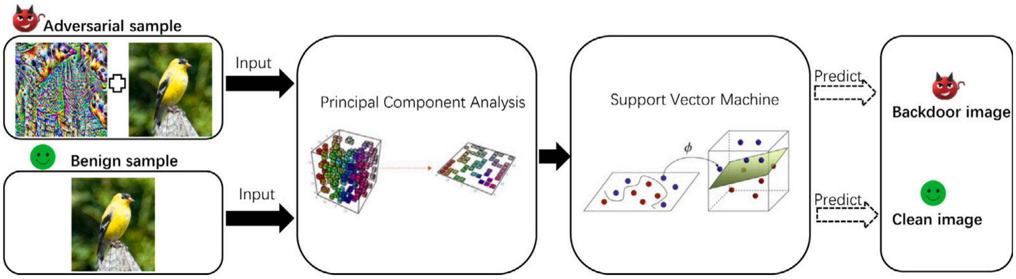

Fig. 7. The detection algorithm using support vector machine for universal adversarial perturbations.

Table 3
The datasets and models used in the experiments.

Attacks	Dataset	Training\Test	Target model
GAP	ImageNet	1 281 167\50 000	Inceptionv3
tFGSM	CIFAR10	50 000\10 000	VGG20
ProxyData	ImageNet	1 281 167\50 000	VGG19

5.3. Detection methods for universal adversarial perturbations

Depending on the universal adversarial perturbations target, they can be classified as non-TUAPs or TUAPs. Akhtar et al. (2017) propose a defense-oriented approach to TUAPs detection that requires learning a Perturbation Rectifying Network as 'pre-input' layers to a targeted model. This approach focuses on the defense against TUAPs. Agarwal et al. (2018) proposed a simple but efficient approach based on Principal Component Analysis as features coupled with a Support Vector Machine as the classifier to detect image-agnostic universal perturbations. They showed through empirical studies on multiple datasets that, compared with three recently proposed adversarial detection algorithms: Intermediate CNN Filter Response (Goswami et al., 2018), Adaptive Noise Reduction (Liang et al., 2021) and Bayesian Uncertainty (Feinman et al., 2017), their proposed method offers better detection performance. To the best of our knowledge, there are currently no detection methods for TUAPs. So we have chosen the method using a Support Vector Machine as the classifier (Agarwal et al., 2018) as the baseline method for our comparison. The flow of the method is shown in Fig. 7. A support vector machine is trained in the labeled adversarial dataset. The paper mentions that the training dataset contains most adversarial samples with universal perturbations and a minority of benign samples, corresponding to a class imbalance rate ratio $_{imb}$ of about 3.

$$\text{ratio}_{imb} = \frac{\text{adversarial samples}}{\text{benign samples}} \tag{10}$$

We have reproduced this method. In our experiments, we chose two different class imbalance rates. ratio $_{imb}$ = 0.02 indicates that the adversarial samples represent a very small proportion of the training dataset. ratio $_{imb}$ = 0.3 indicates that the adversarial samples represent a small proportion of the training dataset. These two hyperparameters were used because we believe that, in practice, the adversarial samples are much smaller than the benign samples. Also, using different values of the class imbalance rate enables us to verify the generalization ability of the algorithm. We believe that it is difficult to collect a sufficient number of adversarial samples in a real-world scenario. This method is susceptible to the effects of data distribution.

5.4. Evaluation metrics

We used the ratio d_{lsr} between two LSR to express the distance between adversarial and benign samples. The d_{lsr} is shown in Eq. (11). $Average_{TUAP}$ denotes the mean LSR of the TUAPs sample, while $Average_{benign}$ denotes the mean LSR of the benign sample. The higher the d_{lsr} value, the greater the variability between the two samples.

$$d_{lsr} = \ Average\ _{TUAP}\ /\ Average\ _{benign} \tag{11}$$

We use the dsr to express the detection effectiveness of the algorithm, as shown in Eq. (12), which is first used in the detection method (Agarwal et al., 2018). count$_{\text{samples}}$ denotes all the adversarial samples which were generated based on TUAPs. count$_{\text{detected\_samples}}$ denotes the adversarial samples that were successfully detected.

$$dsr = \ count\ _{\text{detected\_samples}}\ /count\ _{\text{samples}} \tag{12}$$

The value range of dsr is [0,1].

5.5. Experiments environments

The experiments run on Windows OS, and the hardware environment is an Intel Core i7-10700K CPU with 64G RAM. The validation is completed on NVIDIA GeForce RTX 4090 GPU. The DNNs library used in the experiments is Pytorch 1.13 stable version.

5.6. Result analysis

5.6.1. RQ1: Is there a significant difference between the LSR of adversarial and benign samples under different metamorphic relations?

Motivation: We utilize the LSR metric to indicate whether a sample is an adversarial example. It is imperative for us to empirically validate the effectiveness of this metric on common attacks and datasets through rigorous research.

Approach: We calculate the average LSR on the sample to be tested. The process of generating the sample to be tested is described in detail in Section 3. For the sample to be tested, we chose two metamorphic relations to calculate the corresponding LSR averages. For each metamorphic relation, we randomly sampled 3000 times from the test dataset, generated 3000 adversarial samples, and recorded the adversarial samples that were successfully identified. The LSR of this adversarial sample during the experiment is calculated.

Result: The results of the calculations are shown in Table 4, and we conclude that the adversarial samples have significantly higher LSR than the benign samples for the different metamorphic relations. In addition, we can see that the adversarial samples generally have a high mean of the d_{lsr} with a minimum value of 1.23 and a maximum value of 4.08. The d_{lsr} shows that our proposed algorithm has a relatively noticeable discrimination degree and can distinguish adversarial samples from benign samples.

Table 4
Average LSR (shown in percentage with confidence interval of 99% significance level) for benign samples and adversarial samples.

Attack method	Dataset	Target model	MR	Benign	Adversarial
GAP	ImageNet	Inceptionv3	GF	0.531 ± 0.25	0.761 ± 0.049
			RAN	0.626 ± 0.446	0.772 ± 0.399
tFGSM	CIFAR10	VGG20	GF	0.487 ± 0.226	0.975 ± 0.003
			RAN	0.519 ± 0.226	0.985 ± 0.006
ProxyData	ImageNet	VGG19	GF	0.144 ± 0.123	0.448 ± 0.247
			RAN	0.108 ± 0.096	0.441 ± 0.246

Table 5
The dsr compared with PCASVM baseline.

Attacks	MR	Detection method		
		PCASVM(0.3)	PCASVM(0.02)	OUR
GAP	GF	0.91 ± 0.03	0.05 ± 0.03	0.51 ± 0.09
	RAN	0.849 ± 0.05	0.09 ± 0.04	0.48 ± 0.02
tFGSM	GF	0.96 ± 0.14	0.15 ± 0.07	0.94 ± 0.01
	RAN	0.89 ± 0.03	0.13 ± 0.06	0.95 ± 0.03
ProxyData	GF	0.91 ± 0.03	0.05 ± 0.03	0.75 ± 0.03
	RAN	0.849 ± 0.05	0.09 ± 0.04	0.80 ± 0.04

5.6.2. RQ2: Does our proposed algorithm have a higher dsr than the state-of-the-art algorithms?

Motivation: We must conduct empirical research to validate whether our proposed algorithm exhibits strong generalization ability and surpasses the current state-of-the-art defense methods.

Approach: We have conducted an empirical study of multiple TUAPs backdoor attacks to answer this question, comparing our proposed detection algorithm with state-of-the-art detection algorithms.

Result: The experiment results against the TUAPs backdoor attacks are shown in Table 5. We evaluated defense methods of GAP attack on the dataset ImageNet using the InceptionV3 model; We evaluated defense methods of tFGSM attack on the dataset CIFAR10 using the VGG20 model; We evaluated defense methods of ProxyData attack on the dataset ImageNet using the VGG19 model.

The experiment results against the ProxyData attack (Poursaeed et al., 2017) are shown in Fig. 8. The x-axis represents the total number of samples verified from the ImageNet validation dataset (Deng et al., 2010), and the y-axis represents our method's attack rate or detection rate. The red scatter indicates the detection result of the baseline detection algorithm (Agarwal et al., 2018) when the class imbalance rate is 0.3, while the green scatter indicates the detection result of the baseline detection algorithm (Agarwal et al., 2018) when the class imbalance rate is 0.02. The black scatter indicates the result of our method's backdoor detection.

We randomly sample 10,000 times and generate the dataset to be validated. The dsr of our method is around 42%. A more interesting result is that the dsr of the baseline algorithm shows very different results for different rates of class imbalance. At $\text{ratio}_{imb} = 0.02$, the dsr is essentially below 20%, while at $\text{ratio}_{imb} = 0.3$, the dsr can even reach up to 100%. This suggests that the support vector machine-based detection algorithm is strongly influenced by the distribution of the training data set. When the dataset contains many adversarial samples, the support vector machine can learn the features of the adversarial samples. When there are fewer adversarial samples, the support vector machine cannot learn the features of the adversarial samples. To compare with the support vector machine-based detection algorithm, we take the mean value of dsr for two different rates of class imbalance. Our proposed method improves up to 86.3% over the support vector machine-based detection algorithm.

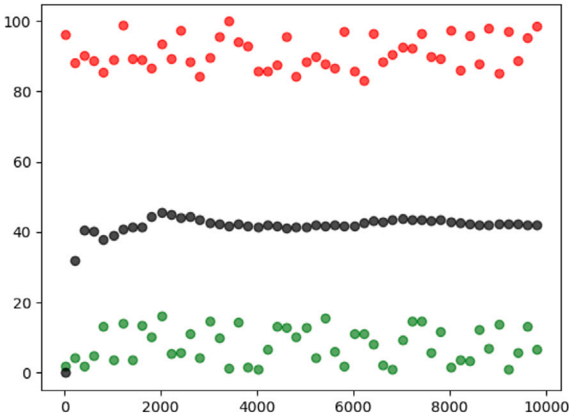

Fig. 8. An example of superimposing a random image on the original image. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

5.6.3. RQ3: Does the adjustment hyperparameter ζ during the test image's composition impact our method?

Motivation: We need to empirically verify whether there are hyperparameters in our algorithm and how the change of hyperparameters affects the algorithm's performance.

Approach: We calculated how the d_{lsr} varies for various hyperparameters ζ. Because the d_{lsr} can well reflect the difference between benign and adversarial samples, a comparative analysis of this indicator can analyze the influence of hyperparameters on our method.

Result: The experiment results are shown in Table 6. For the different attacks, we used the same settings as in Table 5. From the variation of d_{lsr}, we conclude that as the ζ gets larger, the d_{lsr} becomes smaller. This conclusion is significantly different from previous work (Chen et al., 2017). This conclusion can be understood intuitively. The core of our proposed black-box testing method is to superimpose additional noise images on the image to be verified. If the noise image is too obvious, it is very likely to cover up and obscure the semantic features in the image to be verified. Extremely speaking, if our noise image is a solid color image after the image is superimposed on the image to be verified, the entire image will lose all its features, and DNNs will not be able to obtain the high-level features of the object from it. This conclusion implies that when other researchers use our proposed algorithm, they must conduct in-depth investigations and select appropriate hyperparameters not to obscure reasonable semantic features.

5.6.4. RQ4: What is the cost of our detection algorithm?

Motivation: We need to answer whether it is feasible to adopt our proposed algorithm to determine whether the sample is an adversarial sample or not in the shortest possible time.

Approach: The time cost t_{sum} of our method mainly comes from two aspects, while one part of the time cost, recorded as t_{benign}, comes

Table 6
d_{lsr} for benign samples and adversarial samples.

Attacks	MR	ζ	d_{lsr}
GAP	GF	0.01	1.6
		0.1	1.3
		1	0.1
	RAN	0.01	3.4
		0.1	1.5
		1	0.2
tFGSM	GF	0.01	3.8
		0.1	2.1
		1	1.7
	RAN	0.01	3.1
		0.1	1.9
		1	0.4
ProxyData	GF	0.01	3.2
		0.1	3
		1	0
	RAN	0.01	3.2
		0.1	2.3
		1	0

Table 7
The time cost t_{sum}.

Attacks	Datasets	Target model	MR	t(s)
GAP	ImageNet	Inceptionv3	GF	20
			RAN	11
tFGSM	CIFAR10	VGG20	GF	0.8
			RAN	0.9
ProxyData	ImageNet	VGG19	GF	3.9
			RAN	4.2

from the time loss of generating and classifying perturbed samples on benign samples, as shown in the Eq. (2); the other part of the time cost, recorded as t_{TUAP}, comes from the time of generating and classifying perturbed samples on samples to be verified loss, as shown in the Eq. (3). The time cost t_{sum} is shown in the Eq. (13).

$$t_{sum} = t_{benign} + t_{TUAP} \tag{13}$$

We randomly sample 1000 images on each data set for verification or generate 1000 images containing Gaussian noise.

Result: The detailed results are shown in Table 7. The experimental results show that the RAN metamorphic relations take longer to execute than the GF metamorphic relations on the same dataset for the GAP attack. This is understandable as the RAN metamorphic relation may take longer to generate the perturbed images. Regardless of which metamorphic relation is used, the time cost of verifying an image is acceptable. And the program can be further optimized to allow for better parallel computation.

5.7. Threats to validity

The external threats mainly originate from the classifiers used in the experiments. To alleviate this threat, the DNNs are first implemented using a pre-trained model, while the support vector machine uses the classifier from sklearn. The relevant hyperparameters were used with default settings. We second used hyperparameters consistent with the methods in the papers when we re-implemented the attacks.

Internal validity comes primarily from a limited data set. The current datasets for empirical research are mainly derived from relevant datasets mentioned in the papers proposing transferable backdoors, and further generalization studies are needed. In addition, the object of our study is transferable backdoors containing TUAPs. No consideration is given to data contamination and backdoors such as model modification. The method has some limitations in the application area. The SAMT algorithm requires metamorphic relations to make judgments on randomly sampled inputs. During the image overlay, there may be cases where the backdoor is completely obscured, which will result in false positives in the DNN's predictions.

6. Related work

In this section, we briefly introduce transferable backdoor attacks against DNNs, describe the metamorphosis testing used for backdoor detection in DNNs, and finally introduce the sequential sampling technique used in our method.

6.1. Transferable backdoor attacks against DNNs

Since the backdoor attack was proposed, the two main types of backdoor attacks against DNNs, including data poisoning (Gu et al., 2019; Chen et al., 2017; Turner et al., 2019) and network model modification (Liu et al., 2017a; Tang et al., 2020), are researched. Subsequently, universal adversarial perturbations were first proposed to fool DNNs into misclassifying the input images (Moosavi-Dezfooli et al., 2017, 2015; Mopuri et al., 2017; Sharif et al., 2016). A backdoor attack based on universal adversarial perturbation has been proposed and intensively studied in recent years. It can attack DNNs in a series of images. Because DNNs can produce universal target features through adversarial learning. Images containing these features can fool DNNs into predicting the wrong target class. So far, three algorithms in the field of image classification can produce TUAPs (Poursaeed et al., 2017; Hirano and Takemoto, 2019; Zhang et al., 2020).

The most effective means of targeting backdoor attacks is currently adversarial training (Zhang et al., 2021a; Liu et al., 2017b; Li et al., 2021b). However, adversarial training takes a long time to retrain the model, and it is impossible to quickly detect the latent backdoors in the running DNNs after DNNs are deployed. Therefore, we propose a software testing approach for such backdoor attacks to quickly detect the presence of backdoor attacks in the running DNNs.

6.2. Metamorphosis testing for DNNs

To detect the presence of backdoor attacks in the input images of DNNs, we use metamorphosis testing techniques to examine the output of DNNs. Metamorphosis testing is widely used for the functional testing of DNNs (Dong et al., 2009). The biggest problem in DNNs testing is the uncertainty of the prediction results. The metamorphosis test is an effective way of asserting the results of tests on DNNs (Dong et al., 2009). In Zhu et al. (2019) and Ben Braiek and Khomh (2019), the researchers propose changes in color and brightness to the images should not affect the classification results of DNNs. Metamorphic testing has been proven to be an effective means of testing intelligent software systems (Segura et al., 2016; Chen et al., 2018).

Specifically, based on the hypothesis that, in most cases, adversarial samples are more stable for metamorphosis testing on the DNNs than benign samples, we designed the two metamorphic relations, including Gaussian Fuzzing and Random Image. These two metamorphic relations are very intuitive. By imposing these two metamorphic relations on the input image, the DNNs' output result should be random. If the output result obtained by statistical test analysis is not random, then it is likely that this examined image contains a backdoor attack.

6.3. Sequential sampling in software testing

The sequential sampling method is used to sample two sequences without a fixed number of times, and this method can effectively reduce the number of samples (Agha and Palmskog, 2018; Anderson, 1958). This method has been used for robust attack sample detection (Wang et al., 2019). Unlike the traditional sequential sampling method, we design two metamorphic relations to form multiple sampling sequences. We sample over multiple sampling sequences, which allows the backdoor detection to satisfy the termination condition early and further reduces the number of samples.

7. Conclusion and future work

This work proposes a method to detect TUAPs for DNNs at runtime. Our method is based on the hypothesis that the TUAPs have a more stable output in the DNNs, with the output class being the target class, e.g., sea lions. We then propose the SAMT algorithm based on the sequential sampling test. By constructing metamorphic relations on the samples to be tested, the variability of the adversarial samples from the benign samples was obtained based on statistical data analysis. The higher the LSR value, the higher the probability that the sample to be tested contains TUAPs. We conducted an empirical study on four publicly available datasets. Our findings show a statistically significant difference in LSR values between adversarial samples and benign samples; the time cost for determining adversarial samples is acceptable. Our proposed algorithm achieves higher detection performance for adversarial samples than universal adversarial perturbation detection algorithms. In particular, we find that the previous universal adversarial sample detection algorithm requires an adversarial sample training dataset, which limits its practical use. In contrast, the black-box test-based backdoor detection algorithm has stronger practicality and generalization capability in practical scenarios.

In the future, we will focus more on the development of TUAPs techniques and test the generalization capability of our proposed algorithms. In recent years, there has been a growing number of attacks on intelligent software systems, with robustness attacks and backdoor attacks being common attacks. The reliability study of intelligent systems requires attention to multiple stages of DNNs' development and deployment. During the training phase of the model, potential poisoning attacks should be considered. During the training phase, model tampering attacks may be present. During the runtime phase of the model, black-box testing should be emphasized, with preprocessing of input examples to prevent malicious inputs. We will continue to address and conduct in-depth research on the security issues of intelligent software system models in the aforementioned stages. We will also focus on other domains besides computer vision, such as natural language processing, speech recognition, and intelligent code systems.

CRediT authorship contribution statement

Yubin Qu: Methodology, Software, Writing – original draft, Supervision. **Song Huang:** Writing – review & editing, Software. **Xiang Chen:** Writing – review & editing, Software. **Xingya Wang:** Review & editing. **Yongming Yao:** Review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

Acknowledgments

This work was supported by Guangxi Key Laboratory of Trusted Software, China (kx202046), 2023 Philosophy and Social Science Research Programs in Jiangsu Province, China (2023SJSZ0993), and the Fifth Research Project on Teaching Reform of Vocational Education in Jiangsu Province, China (ZYB685), sponsored by the Special Foundation for Excellent Young Teachers and Principals Program of Jiangsu Provincee, China and the Qing Lan Project of Jiangsu province, China.

References

Abraham Wald, 1947. Sequential analysis. J. Amer. Statist. Assoc. 42 (240), 658–665.

Agarwal, A., Singh, R., Vatsa, M., Ratha, N., 2018. Are image-agnostic universal adversarial perturbations for face recognition difficult to detect? In: 2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems, BTAS 2018. Institute of Electrical and Electronics Engineers Inc..

Agha, G., Palmskog, K., 2018. A survey of statistical model checking. ACM Trans. Model. Comput. Simul. 28 (1).

Akhtar, N., Liu, J., Mian, A., 2017. Defense against universal adversarial perturbations. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE Computer Society, pp. 3389–3398.

Anderson, T.W., 1958. Introduction to multivariate statistical analysis.

Ben Braiek, H., Khomh, F., 2019. DeepEvolution: A search-based testing approach for deep neural networks. In: Proceedings - 2019 IEEE International Conference on Software Maintenance and Evolution, ICSME 2019. Institute of Electrical and Electronics Engineers Inc., pp. 454–458.

Bojarski, M., Del Testa, D., Dworakowski, D., Firner, B., Flepp, B., Goyal, P., Jackel, L.D., Monfort, M., Muller, U., Zhang, J., Zhang, X., Zhao, J., Zieba, K., 2016. End to end learning for self-driving cars.

Bueno, P.M., Jino, M., Wong, W.E., 2014. Diversity oriented test data generation using metaheuristic search techniques. Inform. Sci. 259, 490–509.

Chen, T.Y., Kuo, F.C., Liu, H., Poon, P.L., Towey, D., Tse, T.H., Zhou, Z.Q., 2018. Metamorphic testing. ACM Comput. Surv. 51 (1).

Chen, X., Liu, C., Li, B., Lu, K., Song, D., 2017. Targeted backdoor attacks on deep learning systems using data poisoning.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L., 2010. ImageNet: A large-scale hierarchical image database. pp. 248–255.

Doan, B.G., Abbasnejad, E., Ranasinghe, D.C., 2020. Februus: Input purification defense against Trojan attacks on deep neural network systems. In: ACM International Conference Proceeding Series. Association for Computing Machinery, pp. 897–912.

Dong, Y., Liao, F., Pang, T., Su, H., Zhu, J., Hu, X., Li, J., 2018. Boosting adversarial attacks with momentum. pp. 9185–9193.

Dong, G., Xu, B., Chen, L., Nie, C., Wang, L., 2009. Survey of metamorphic testing. J. Front. Comput. Sci. Technol. 3 (2), 130.

Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A., 2010. The pascal visual object classes (VOC) challenge. Int. J. Comput. Vis. 88 (2), 303–338.

Feinman, R., Curtin, R.R., Shintre, S., Gardner, A.B., 2017. Detecting adversarial samples from artifacts.

Gehr, T., Mirman, M., Drachsler-Cohen, D., Tsankov, P., Chaudhuri, S., Vechev, M., 2018. AI2: Safety and robustness certification of neural networks with abstract interpretation. In: Proceedings - IEEE Symposium on Security and Privacy, Vol. 2018-May. Institute of Electrical and Electronics Engineers Inc., pp. 3–18.

Goswami, G., Ratha, N., Agarwal, A., Singh, R., Vatsa, M., 2018. Unravelling robustness of deep learning based face recognition against adversarial attacks. In: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 32. AAAI Press, pp. 6829–6836, no. 1.

Gu, T., Liu, K., Dolan-Gavitt, B., Garg, S., 2019. BadNets: Evaluating backdooring attacks on deep neural networks. IEEE Access 7, 47230–47243.

He, K., Zhang, X., Ren, S., Sun, J., 2015. Deep residual learning for image recognition. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Vol. 2016-December. IEEE Computer Society, pp. 770–778.

Hirano, H., Takemoto, K., 2019. Simple iterative method for generating targeted universal adversarial perturbations. Algorithms 13 (11), 1–10.

Jia, J., Liu, Y., Cao, X., Gong, N.Z., 2020. Certified robustness of nearest neighbors against data poisoning and backdoor attacks. In: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 36. Association for the Advancement of Artificial Intelligence (AAAI), pp. 9575–9583, no. 9.

Krizhevsky, A., 2009. Learning multiple layers of features from tiny images.

Lavanya, K., Prakash, S., Gedam, Y., Aijaz, A., Ramanathan, L., 2022. Real time digital face mask detection using MobileNet-V2 and SSD with apache spark. Int. J. Perform. Eng. 18 (8), 598.

Li, Y., Lyu, X., Koren, N., Lyu, L., Li, B., Ma, X., 2021b. Neural attention distillation: Erasing backdoor triggers from deep neural networks.

Li, D., Wong, W.E., Wang, W., Yao, Y., Chau, M., 2021a. Detection and mitigation of label-flipping attacks in federated learning systems with KPCA and K-means. In: Proceedings - 2021 8th International Conference on Dependable Systems and their Applications, DSA 2021. Institute of Electrical and Electronics Engineers Inc., pp. 551–559.

Liang, B., Li, H., Su, M., Li, X., Shi, W., Wang, X., 2021. Detecting adversarial image examples in deep neural networks with adaptive noise reduction. IEEE Trans. Dependable Secure Comput. 18 (1), 72–85.

Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L., 2014. Microsoft COCO: Common objects in context. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). In: LNCS, Vol. 8693, (PART 5), Springer Verlag, pp. 740–755.

Liu, Y., Ma, S., Aafer, Y., Lee, W.-C., Zhai, J., Authors Yingqi Liu, A., Wang, W., Zhang, X., 2017a. Trojaning Attack on Neural Networks. Technical Reports, Department of Computer Science.

Liu, Y., Xie, Y., Srivastava, A., 2017b. Neural trojans. Proceedings - 35th IEEE International Conference on Computer Design, ICCD 2017 45–48.

Mirman, M., Gehr, T., Vechev, M., 2018. Differentiable abstract interpretation for provably robust neural networks. pp. 3578–3586.

Moosavi-Dezfooli, S.-M., Fawzi, A., Fawzi, O., Frossard, P., 2017. Universal adversarial perturbations. pp. 1765–1773.

Moosavi-Dezfooli, S.M., Fawzi, A., Frossard, P., 2015. DeepFool: a simple and accurate method to fool deep neural networks. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Vol. 2016-December. IEEE Computer Society, pp. 2574–2582.

Mopuri, K.R., Garg, U., Venkatesh Babu, R., 2017. Fast feature fool: A data independent approach to universal adversarial perturbations. In: British Machine Vision Conference 2017, BMVC 2017. BMVA Press.

Parkhi, O.M., Vedaldi, A., Zisserman, A., 2015. Deep face recognition. In: BMVC 2015 - Proceedings of the British Machine Vision Conference 2015. British Machine Vision Association.

Pham, L.H., Sun, J., 2022. Verifying neural networks against backdoor attacks. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). In: LNCS, vol. 13371, Springer Science and Business Media Deutschland GmbH, pp. 171–192.

Poursaeed, O., Katsman, I., Gao, B., Belongie, S., 2017. Generative adversarial perturbations. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE Computer Society, pp. 4422–4431.

Segura, S., Fraser, G., Sanchez, A.B., Ruiz-Cortes, A., 2016. A survey on Metamorphic testing. IEEE Trans. Softw. Eng. 42 (9), 805–824.

Sharif, M., Bhagavatula, S., Bauer, L., Reiter, M.K., 2016. Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. In: Proceedings of the ACM Conference on Computer and Communications Security, Vol. 24-28-October-2016. Association for Computing Machinery, pp. 1528–1540.

Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. In: 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings. International Conference on Learning Representations, ICLR.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z., 2016. Rethinking the inception architecture for computer vision. pp. 2818–2826.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R., 2013. Intriguing properties of neural networks. In: 2nd International Conference on Learning Representations, ICLR 2014 - Conference Track Proceedings. International Conference on Learning Representations, ICLR.

Tang, R., Du, M., Liu, N., Yang, F., Hu, X., 2020. An embarrassingly simple approach for Trojan attack in deep neural networks. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Vol. 11. Association for Computing Machinery, pp. 218–228, no. 20.

Turner, A., Tsipras, D., Madry, A., 2019. Label-consistent backdoor attacks.

Wang, J., Dong, G., Sun, J., Wang, X., Zhang, P., 2019. Adversarial sample detection for deep neural network through model mutation testing. In: Proceedings - International Conference on Software Engineering, Vol. 2019-May. IEEE Computer Society, pp. 1245–1256.

Wawage, P., Deshpande, Y., 2022. Real-time prediction of car driver's emotions using facial expression with a convolutional neural network-based intelligent system. Int. J. Perform. Eng. 18 (11), 791.

Xuan, L., Yong, G., Shiwei, J., Mingfei, L., Yan, Z., Yadong, G., Ruixue, N., 2019. Explosive detection system based on sequential probability ratio test. J. Jilin Univ. Sci. Ed. 57 (3), 653–657.

Zeng, Y., Chen, S., Park, W., Mao, Z.M., Jin, M., Jia, R., 2021. Adversarial unlearning of backdoors via implicit hypergradient.

Zhang, C., Benz, P., Imtiaz, T., Kweon, I.S., 2020. Understanding adversarial examples from the mutual influence of images and perturbations. pp. 14521–14530.

Zhang, C., Benz, P., Lin, C., Karjauv, A., Wu, J., Kweon, I.S., 2021a. A survey on universal adversarial attack. In: IJCAI International Joint Conference on Artificial Intelligence. International Joint Conferences on Artificial Intelligence, pp. 4687–4694.

Zhang, Z., Wang, P., Guo, H., Wang, Z., Zhou, Y., Huang, Z., 2021b. DeepBackground: Metamorphic testing for deep-learning-driven image recognition systems accompanied by background-relevance. Inf. Softw. Technol. 140, 106701.

Zhao, Z., Liu, Z., Larson, M., 2021. On success and simplicity: A second look at transferable targeted attacks. In: Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P.S., Vaughan, J.W. (Eds.), Advances in Neural Information Processing Systems, Vol. 34. Curran Associates, Inc., pp. 6115–6128.

Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., Torralba, A., 2018. Places: A 10 million image database for scene recognition. IEEE Trans. Pattern Anal. Mach. Intell. 40 (6), 1452–1464.

Zhu, H., Liu, D., Bayley, I., Harrison, R., Cuzzolin, F., 2019. Datamorphic testing: A method for testing intelligent applications. In: Proceedings - 2019 IEEE International Conference on Artificial Intelligence Testing, AITest 2019. Institute of Electrical and Electronics Engineers Inc., pp. 149–156.

Yubin Qu was born in Nanyang, China in 1981. He received the B.S. and M.S. degrees in Computer Science and Technology from Henan Polytechnic University in China in 2004 and 2008. He is currently pursuing a doctoral degree at Army Engineering University of PLA. Since 2022, he has been an associate professor with Information Engineering Institute, Jiangsu College of Engineering and Technology. He is the author of more than 10 articles. His research interests include software maintenance, software testing, and machine learning.

Song Huang was born in Huainan, Anhui, China, in 1970. He received the Ph.D. degree from the PLA University of Science and Technology. He is currently a Professor of software engineering with the Software Testing and Evaluation Center, Army Engineering University of PLA. He has contributed more than 100 journal articles to professional journals. His current research interests include software testing, quality assurance, data mining, and empirical software engineering. He is a member of CCF and ACM. He is also a member of the advisory boards of Journal of Systems and Software and IEEE TRANSACTIONS ON RELIABILITY.

Xiang Chen received the B.Sc. degree in the school of management from Xi'an Jiaotong University, China in 2002. Then he received his M.Sc., and Ph.D. degrees in computer software and theory from Nanjing University, China in 2008 and 2011 respectively. He is currently an Associate Professor at the Department of Information Science and Technology, Nantong University, Nantong, China. He has authored or co-authored more than 120 papers in refereed journals or conferences, such as IEEE TRANSACTIONS ON SOFTWARE ENGINEERING, ACM Transactions on Software Engineering and Methodology, Information and Software Technology, Journal of Systems and Software, IEEE TRANSACTIONS ON RELIABILITY, Journal of Software: Evolution and Process, Software - Practice and Experience, Automated Software Engineering, Journal of Computer Science and Technology, IET Software, Software Quality Journal, Knowledge-based Systems, International Conference on Software Engineering, The ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering, International Conference Automated Software Engineering, International Conference on Software Maintenance and Evolution, International Conference on Program Comprehension, and International Conference on Software Analysis, Evolution and Reengineering. His research interests include software engineering, in particular software testing and maintenance, software repository mining, and empirical software engineering. He received ACM SIGSOFT distinguished paper awards in ICSE 2021. He is the editorial board member of Information and Software Technology. More information about him can be found at: https://smartse.github.io/index.html.

Xingya Wang, was born in 1990. He serves as an associate professor at Nanjing Tech University and he has previously worked as a postdoctoral researcher at the School of Software at Nanjing University. In 2012, he graduated from the School of Computer Science and Technology at China University of Mining and Technology, obtaining a Bachelor's degree in Computer Science and Technology. In 2017, he completed his doctoral studies in Computer Software and Theory from the same institution. His primary research focuses on blockchain (smart contract) analysis and testing, as well as software defect localization techniques. He has led several projects, including collaborative education programs with the Ministry of Education, the Jiangsu Province Postdoctoral Research Funding Program, and innovation projects at the State Key Laboratory of Computer Software New Technology. He has also participated as a core member or main participant in various projects, such as the National Key Research and Development Program, National Natural Science Foundation projects, and technology development projects commissioned by State Grid Corporation of China. He has published over 30 academic papers, including IEEE Computer, SCIENCE CHINA Information Sciences, and the Journal of Computer Research.

Yongming Yao was born in Yangzhou, Jiangsu, China, in 1987. He received the Ph.D. degree from the PLA University of Science and Technology. He is currently an assistant professor of software engineering with the Software Testing and Evaluation Center, Army Engineering University of PLA. He has contributed more than 10 journal articles to professional journals. His current research interests include software testing, quality assurance, and empirical software engineering. He is a member of CCF.