DoPa: A Fast and Comprehensive CNN Defense Methodology against Physical Adversarial Attacks

Zirui Xu, Fuxun Yu, Xiang Chen
George Mason University, Fairfax, Virginia
{zxu21, fyu2, xchen26}@gmu.edu

ABSTRACT

Recently, Convolutional Neural Networks (CNNs) demonstrate considerable vulnerability to adversarial attacks, which can be easily mislead by adversarial perturbations. With more aggressive methods proposed, adversarial attacks can be also applied to the physical world, causing practical issues to various CNN powered applications. Most existing physical adversarial attack defense works only focus on eliminating explicit perturbation patterns from inputs, ignoring to interpret and fix CNN's intrinsic vulnerability. Therefore, most of them depend on considerable data processing cost and lack expected versatility to different attacks. In this paper, we propose *DoPa* – a fast and comprehensive CNN defense methodology against physical adversarial attacks. By interpreting the CNN vulnerability, we find that non-semantic adversarial perturbations can activate CNN with significantly abnormal activation and even overwhelm other semantic input patterns' activations. We improve CNN recognition process by adding a self-verification stage to analyze the the semantics of distinguished activation patterns with only one CNN inference involved. Based on the detection result, we further propose the data recovery methods to defend the physical adversarial attacks. We apply such detection and defense methodology into both image and audio CNN recognition process. Experiments show that our methodology can achieve an average 90% successful rate for attack detection and 81% accuracy recovery for image physical adversarial attack. Also, the proposed defense method can achieve 92% detection successful rate and 77.5% accuracy recovery for audio recognition applications. Moreover, the proposed defense methods are at most 2.3× faster compared to the state-of-the-art defense methods, making them feasible to resource-constrained platforms, such as mobile devices.

1 INTRODUCTION

In the past few years, Convolutional Neural Networks (CNNs) have been widely applied in various cognitive applications, such as image classification [26, 30] and speech recognition [7, 8], *etc.* Although effective and popular, CNN powered applications are facing a critical challenge – adversarial attacks. By injecting particular perturbations into input data, adversarial attacks can mislead CNN recognition results. The perturbations generated by traditional adversarial attacks are fragile, which can be only added into digital data. Therefore, they can hardly threaten the recognition systems which obtain input data from the real world. However, with more advanced methods proposed, adversarial perturbations can be concentrated into a small area and be easily attached on the actual objects. Therefore, these enhanced adversarial attacks can be applied to the physical world. Fig. 1 shows a physical adversarial example on traffic sign detection. when we attach a well-crafted adversarial patch on the original stop sign, the traffic sign detection system will be misled to a wrong recognition result as a speed limit sign. Recently, the issue of such physical adversarial attacks becomes severer with increasing CNN based applications [15].

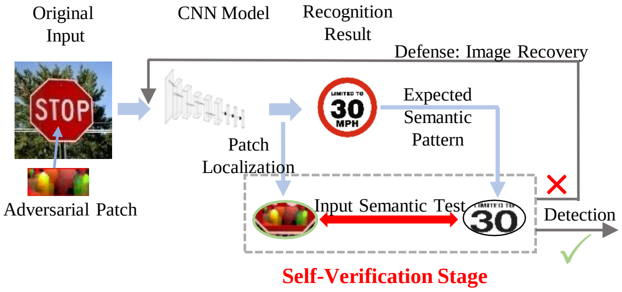

Figure 1: Physical Adversarial Attack for Traffic Sign

Many works have been proposed to defend the physical adversarial attacks [13, 18, 20, 28]. However, most of them neglect the CNN's intrinsic vulnerability interpretation. Instead, either they merely focus on eliminating explicit perturbation patterns from input [18, 20] or they simply adopt multiple CNNs to conduct the cross-verification [21, 28]. All these methods have certain drawbacks: First, they introduce a considerable data processing cost during perturbations elimination. Second, they can hardly defend physical adversarial attacks with model transferability, lacking versatility to different physical adversarial attacks.

In this paper, we propose *DoPa*, a fast and comprehensive defense methodology against physical adversarial attacks. By interpreting the CNN vulnerability, we reveal that the CNN decision-making process lacks necessary qualitative semantics distinguishing ability: the non-semantic input patterns can significantly activate CNN and overwhelm other semantic input patterns. We improve CNN recognition process by adding a self-verification stage to analyze the the semantics of distinguished activation patterns with only one CNN inference involved. Fig. 1 illustrates the self-verification stage for traffic sign adversarial attack. For each input image, after one forward process, the verification stage will locate the significant activation sources (shown in green circle) and calculate the input semantic inconsistency with the expected semantic patterns (shown in right circle) according to the prediction result. Once the inconsistency exceeds the a pre-defined threshold, CNN will conduct the defense method to recovery the input image. At last, we get the correct recognition result. Our defense methodology only depends on CNN inference with minimum computation components involved, which can be extended to both CNN based image and audio recognition applications.

Specifically, we have the following contributions in this work:

- By interpreting CNN's vulnerability, we find that non-semantic input patterns can significantly activate CNN and overwhelm other semantic input patterns.

- We propose a self-verification stage to analyze and detect the abnormal activation patterns' semantics. Specifically, we introduce the inconsistency between the local input patterns that cause the distinguished activation and the synthesized patterns with expected semantics.
- We adopt two data recovery methods as our defense methodology to defend the physical adversarial attacks.
- Based on such detection and defense methodology, we propose two defense cases for image and audio applications.

Experiments show that our method can achieve an average 90% detection successful rate and average 81% accuracy recovery for image physical adversarial attack. Also, our method achieves 92% detection successful rate and 77.5% accuracy recovery for the audio adversarial attack. Moreover, our methods are at most 2.3× than the state-of-the-art defense methods, which is feasible to various resource-constrained platforms, such as mobile devices.

2 BACKGROUND AND RELATED WORKS

2.1 Physical Adversarial Attacks

The adversarial attack started to arouse researchers' general concern with adversarial examples, which were first introduced by [12]. The adversarial examples were designed to project prediction errors into input space to generate noises, which can perturb digital input data (*e.g.*, images and audio clips) and manipulate prediction results. Since then, various adversarial attack were proposed, such as L-BFGS [24], FGSM [12], CW [6], *etc.* Most of these adversarial attack methods share a similar mechanism, which tries to cause the most error increment within model activation and regulate the noises within the input space.

Recently, such a attack approach was also brought from the algorithm domain into the physical world, which we refer as the physical adversarial attack. [11] first leveraged a masking method to concentrate the adversarial perturbations into a small area and implement the attack on real traffic signs with taped graffiti. [5] then extended the scope of physical attacks with adversarial patches. With more aggressive image patterns than taped graffiti, these patches could be attached to physical objects arbitrarily and have a certain degree of model transferability.

Beyond aforementioned image cases, some physical adversarial attacks also have been proposed to audios. Yakura *et al.* [27] proposed an audio physical adversarial attack that can still be effective after playback and recording in the physical world. [27] generated audio adversarial command in a normal song which can be played through the air.

Compared to the noise based adversarial attack, these physical adversarial attacks reduce the attack difficulty and further impair the practicality and reliability of deep learning technologies.

2.2 Image physical Adversarial Attack Detection and Defense

There are several works has been proposed to detect and defense such physical adversarial attacks in the image recognition processes. Naseer *et al.* proposed a local gradients smoothing scheme against physical adversarial attacks [18]. By regularizing gradients in the estimated noisy region before feeding images into CNN for inference, their method can eliminate the potential impacts from adversarial attack. Hayes *et al.* proposed a physical image adversarial attack defense method based on image inpainting [13]. Based on the traditional image processing methods, they detect the localization of adversarial noises in the input image and further leverage the image inpainting technology to remove the adversarial noises.

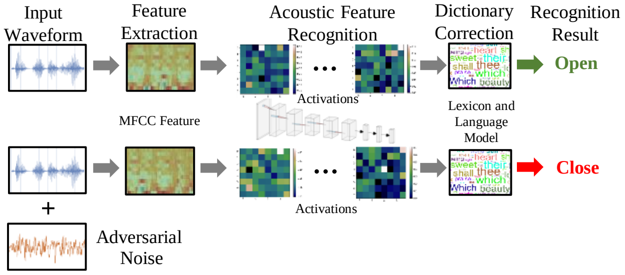

Figure 2: Audio Recognition and Physical Adversarial Attack Process

Although these methods are effective for image physical adversarial attacks defense, they still have certain disadvantages regarding computation and versatility. For example, local gradients smoothing requires the manipulation for each pixel of the input image, which will introduce a large number of computation workload. Moreover, their methods are designed for solving specific adversarial attack which are not integrated for different physical adversarial attack situations. Therefore, we develop a fast and comprehensive defense methodology to address above issues.

2.3 Audio Physical Adversarial Attack Detection and Defense

Compared with images, the audio data requires more processing efforts for recognition. Fig. 2 shows a typical audio recognition process and the corresponding physical adversarial attack. The audio waveform is first extracted as Mel-frequency Cepstral Coefficient (MFCC) features. Then we leverage a CNN to achieve acoustic feature recognition, which can obtain the candidate phonemes. Finally, a lexicon and language model is applied to obtain the recognition result "open". When the adversarial noise is injected to the original input waveform, the final recognition result is misled to "close".

Several works have been proposed to detect and defend such adversarial attacks [21, 22, 28]. Zeng *et al.* leveraged multiple Automatic Speech Recognition (ASR) systems to detect the audio physical adversarial attack based on a cross-verification methodology [21]. However, their method lacks certain versatility which cannot detect the adversarial attack with model transferability. Yang *et al.* proposed an audio adversarial attack detection and defense method by exploring the temporal dependency of audio adversarial deception [28]. However, their method requires multiple CNN recognition inferences which is time-consuming. Rajaratnam *et al.* leveraged the random noise flooding to defense audio physical adversarial attack [22]. Since the ASR systems are relatively robust to natural noise while adversarial noise is not, by injecting random noise, the functionalities of adversarial noise can be destroyed. However, this method cannot achieve high practical defense performance .

3 CNN VULNERABILITY ANALYSIS FOR PHYSICAL ADVERSARIAL ATTACK

In this section, we first interpret the CNN vulnerability by analyzing the input patterns' semantics with the activation maximization

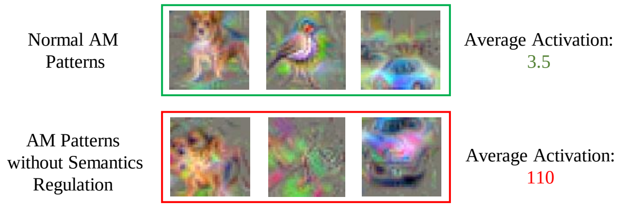

Figure 3: Visualized Neuron's Input Pattern by Activation Maximization Visualization

visualization [10]. Based on the semantics analysis, we identify the adversarial attack patches as the non-semantic input patterns with abnormal distinguished activations. specifically, to evaluate the semantics, we propose metrics that can measure the inconsistencies between the local input patterns that cause the distinguished activations and the synthesized patterns with expected semantics. Based on the inconsistency analysis, we further propose a defense methodology consists of self-verification and data recovery.

3.1 CNN Vulnerability Interpretation

In a typical image or audio recognition process, CNN extracts feature from original input data and gradually derive a prediction result. However, when injecting the physical adversarial perturbations into original data, CNN will be misled to a wrong prediction result. To better interpret the vulnerability, we first analyze CNN's vulnerability by using a typical image physical adversarial attack – adversarial patch attack as an example. Compared with the original input, an adversarial patch usually has no constraints in color/shape, etc. Therefore, such patches usually sacrifice the semantic structures so as to cause significant abnormal activation and overwhelm the other input patterns' activations. We assume that CNN lacks qualitative semantics distinguishing ability which can be activated by non-semantic adversarial patch during CNN inference.

To verify such an assumption, we investigate the semantic of each neuron in CNN. We adopt a CNN activation visualization method – Activation Maximization Visualization (AM) [10]. AM can generate a pattern to visualize each neuron's most activated semantic input. The generation process of pattern $V(N_i^l)$ can be considered as synthesizing an input image to a CNN model that delicately maximizes the activation of the ith neuron N_i^l in the layer of l. Mathematically, this process can be formulated as:

$$V(N_i^l) = \arg\max_X A_i^l(X), \qquad X \leftarrow X{+}\eta{\cdot}\frac{\partial A_i^l(X)}{\partial X} \tag{1}$$

where, $A_i^l(X)$ is the activation of N_i^l from an input image X, η is the gradient ascent step size.

Fig. 3 shows the visualized neurons' semantic input patterns by using AM. As the traditional AM method is designed for semantics interpretation, many feature regulation and hand-engineered natural image references are involved in generating interpretable visualization patterns. Therefore we can get three AM patterns with an average activation magnitude value of 3.5 in Fig. 3 (a). The objects in the three patterns indicates they have clear semantics. However, when we remove these semantics regulations in the AM process, we can obtain three different visualized patterns as shown in Fig. 3 (b). We can find that these three patterns are non-semantic, but they have significant abnormal activations with an average magnitude value of 110. This phenomenon can prove our assumption that CNN neurons lack semantics distinguishing ability and can be significantly activated by non-semantic inputs patterns.

3.2 Metrics for Input Semantic Inconsistency and Prediction Activation Inconsistency

To identify the non-semantic input patterns for the attack detection, we aim to find the difference between the CNN inference for natural image recognition and physical adversarial attacks.

Fig. 4 shows a typical adversarial patch based physical attack. The patterns in the left circles are the primary activation sources from the input images, and the bars on the right are the neurons' activations in the last convolutional layer. From input patterns, we identify a significant inconsistency between the adversarial patch and primary activation source from the original image, which can be used to detect the adversarial patch. From prediction activations, we observe another inconsistency between the adversarial input image and the original input image. Therefore, we formulate two inconsistencies at two levels:

Input semantic Inconsistency Metric This metric measure the input semantic inconsistency between the non-semantic adversarial patch and the semantic local input pattern from natural image. It can be defined as follows:

$$D(P_{pra}{,}P_{ori}){=}1{-}S(P_{pra}{,}P_{ori}){,}P_{pra}\xleftarrow{\mathfrak{R}}\Phi{:}A_i^l(p){,}P_{ori}\xleftarrow{\mathfrak{R}}\Phi{:}A_i^l(o), \tag{2}$$

where P_{pra} and P_{ori} represent the input patterns from the adversarial input and the original input. $\Phi : A_i^l(p)$ and $\Phi : A_i^l(o)$ represent the set of neurons' activations produced by adversarial patch and the original input, respectively. $\mathfrak{R}$ maps neurons' activations to the primary local input patterns. S represents a similarity metric.

Prediction Activation Inconsistency Metric The second inconsistency is in the activation level, which reveals the activation's magnitude distribution inconsistency in the last convolutional layer between adversarial input and the original input. We also use a similar metric to measure it as follows:

$$D(f_{pra}{,}f_{ori}){=}1{-}S(f_{pra}{,}f_{ori}){,}f_{pra}{\sim}\Phi{:}A_i^l(p){,}f_{ori}{\sim}\Phi{:}A_i^l(o) \tag{3}$$

where f_{pra} and I_{ori} represent the magnitude distribution of activations in the last convolutional layer generated by the adversarial input and the original input data.

For above two inconsistency metrics, we can easily obtain P_{pra} and f_{pra} since they come from the input data. However, P_{ori} and f_{ori} are not easily to get because of the variety of the natural input data. Therefore, we need to synthesize the standard input data which can provide the semantic input patterns and activation magnitude distribution. The synthesized input data for each prediction

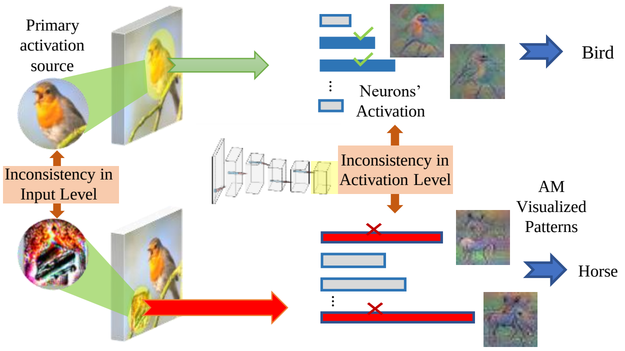
Primary activation source
Bird
Neurons' Activation
Inconsistency in Input Level
Inconsistency in Activation Level
AM Visualized Patterns
Horse

Figure 4: Image Adversarial Patch Attack

class can be obtained from the standard datasets. By feeding CNN with a certain number of input from the standard dataset, we can record the average activation magnitude distribution in last convolutional layer. Moreover, we can locate the primary semantic input patterns for each prediction class.

3.3 Physical Adversarial Attack Defense based on CNN Self-verification and Data Recovery

Based on the above analysis, we propose a defense methodology which consists of the self-verification stage and data recovery methodology in the CNN decision-making process. More specifically, the entire methodology flow can be described as following: (1) We first feed the input into CNN inference and obtain the prediction class. (2) Next, CNN can locate the primary activation sources from the practical input and obtain the activations in the last convolutional layer. (3) Then we leverage the input semantic inconsistency metric and prediction activation inconsistency metric to measure the two inconsistencies between practical input and the synthesized data with the prediction class. (4) Once inconsistency exceeds the given threshold, CNN will consider the input as an adversarial input. (5) After a physical adversarial attack has been detected by self-verification stage, the data recovery methodology is further used to recovery the adversarial input data. Specifically, we leverage image inpainting and abnormal activation suppression to recovery the adversarial input image and audio. Our proposed methodology can detect and defense the physical adversarial attack in one-shot.

We will derive two methods from such a methodology for image and audio applications in the Section 4 and Section 5, respectively.

4 IMAGE PHYSICAL ADVERSARIAL ATTACK DEFENSE BASED ON SELF-VERIFICATION AND DATA RECOVERY

In the last section, we propose the self-verification stage to detect and defense the adversarial attack. In this section, we will specifically describe our proposed defense methodology against image physical adversarial attacks.

For image physical adversarial attacks defense, we mainly depend on the **input semantic inconsistency** in input pattern level. Therefore, we need to locate the primary activation source from the input image by adopting a CNN activation visualization method – Class Activation Mapping (CAM) [29]. Let $A_k(x,y)$ denotes the value of the kth activation in the last convolutional layer at spatial location (x,y). We can compute a weighted sum of the all activations at the spatial location (x,y) in the last convolutional layer as:

$$A_T(x,y) = \sum_{K}^{1} A_k(x,y), \tag{4}$$

where K is the total number of activations in the last convolutional layer. The larger value of $A_T(x,y)$ represents the activation source in input image at the corresponding spatial location (x,y) plays a more important role during CNN inference.

To achieve the attack detection in the practical, we further build the specific metric for image adversarial physical attack defense. According to our preliminary analysis, the input adversarial patch contains much more high-frequency information than the natural semantic input patterns. Therefore, we first leverage 2D Fast Fourier Transform (2D-FFT) [2] to transfer the patterns from the temporal domain to the frequency domain and thereby constrain the low-frequency components together. Then we convert the frequency-domain pattern to a binary pattern with an adaptive threshold. Fig. 5 shows the converted example including adversarial patterns, expected semantic patterns with same prediction result, and natural input pattern. For binary patterns, we can observe the significant difference between adversarial input and ground-truth input. Therefore, based on above analysis, we replace $S(I_{pra}, I_{ori})$ as Jaccard Similarity Coefficient (JSC) [19] and propose our image adversarial attack inconsistency metric as:

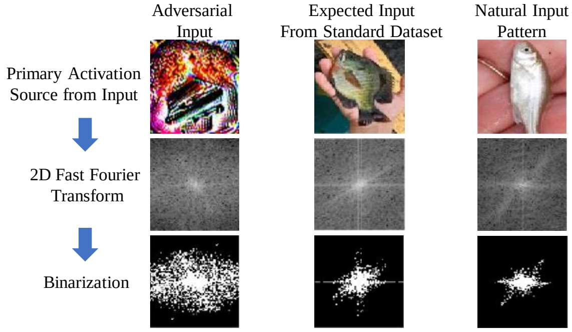
Adversarial Input
Expected Input From Standard Dataset
Natural Input Pattern
Primary Activation Source from Input
2D Fast Fourier Transform
Binarization

Figure 5: The Results after 2D Fast Fourier Transform

$$D(P_{pra},P_{exp}){=}1{-}JSC(P_{pra},P_{exp}){=}\frac{|P_{pra}\bigcup P_{exp}|{-}|P_{pra}\bigcap P_{exp}|}{|P_{pra}\bigcup P_{exp}|}, \tag{5}$$

where I_{exp} is the ground truth semantic pattern with predicted class. $P_{pra} \bigcap P_{exp}$ means the numbers of pixels where the pixel value of P_{pra} and P_{exp} both equal to 1.

With the above inconsistency metric, we propose a defense method which contains 6 steps from image inputting to image recovery. The entire process of our method is described in Fig. 6.

For each input image, We apply CAM to locate the source location of the biggest model activation. Then we crop the image to obtain the patterns with maximum activations. In the step of semantic test, we calculate the consistency between I_{pra} and I_{exp}. Once the inconsistency is higher than a predefined threshold, we consider a significant inconsistency detected. After the patch is detected, we can directly remove it from the original input data and repair the image using image inpainting [4]. At last, we feed back the recovery image into CNN to do the prediction again.

With the above steps, we can detect and further defense an image physical adversarial attack during CNN inference process.

5 AUDIO PHYSICAL ADVERSARIAL ATTACK DEFENSE BASED ON SELF-VERIFICATION AND DATA RECOVERY

In this section, we will introduce the detailed defense design flow for audio physical adversarial attack.

Different from images, the audio data requires more processing efforts. As Fig. 2 shows, during the audio recognition, the input waveform needs to pass Mel-frequency Cepstral Coefficient (MFCC) conversion to be transferred from the time domain into the time-frequency domain. Therefore, the original input audio data will loss semantics after the MFCC conversion. Therefore, we leverage the **prediction activation inconsistency** to detect the audio physical adversarial attacks.

More specifically, we measure the activation magnitude distribution inconsistency between practical input and the ground-truth

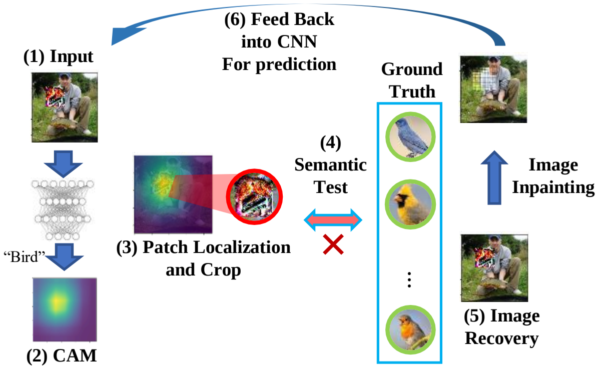

Figure 6: Adversarial Patch Attack Detection and Defense

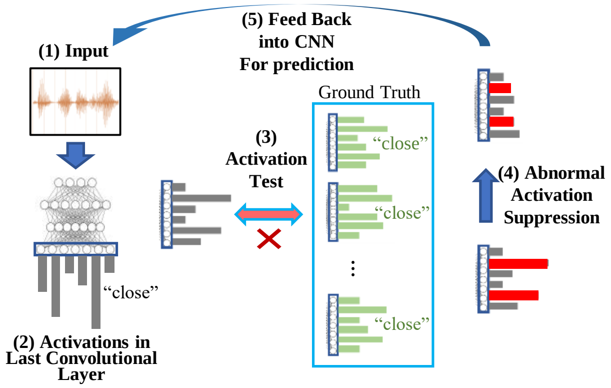

Figure 7: Audio Adversarial Attack Detection and Defense

with same prediction class. We adopt a popular similarity evaluation method - Pearson Correlation Coefficient (PCC) [3] and the inconsistency metric can be defined as:

$$D(f_{pre},f_{exp})=1-PCC(f_{pre},f_{exp})=1-\frac{E[(f_{pre}-\mu_{pre})(f_{exp}-\mu_{exp})]}{\sigma_{pra}\sigma_{exp}}, \quad (6)$$

where I_{pre} and I_{exp} represent the activations in the last convolutional layer for both practical input and ground truth input. μ_a and μ_o denote the mean values of A_a and A_o, σ_a and σ_o are the standard derivations, and E means the overall expectation.

With established inconsistency metric, we further apply self-verification stage to CNN for audio physical adversarial attack. The detection flow is described as following: We first obtain activations in the convolutional layer for every possible input word by testing CNN with a standard dataset. Then we calculate the inconsistency value $D(I_{pra}, I_{exp})$. If the model is attacked by audio adversarial attack, $D(I_{act}, I_{sem})$ will exceed a pre-defined threshold. According to our experiments tested with various attacks, there exists a large range for the threshold to distinguish the natural and adversarial audios, which benefit our accurate detection.

After identifying the adversarial input audio, simply denying it can cause undesired consequences. Therefore, attacked audio recovery is considered as one of the most acceptable solution. We propose a new solution - "activation denoising" as our defense method, which targets at ablating adversarial effects from activation level. The activation denoising takes advantages of the aforementioned last layer activation patterns, which have stable correlations with determined predication labels. When the wrong label is detected, we can determine the correlated activation patterns. By suppressing these patterns in the hidden layer, the original input will emerge. Therefore, we propose our adversarial audio recovery method as shown in Fig. 7: (1) Based on the detection result, we can identify the wrong prediction label, and therefore the the standard activation patterns of the wrong class in the last layer. (For the best performance, we locate the top-k activation index.) (2) Then we can find the activations with the same index. These activations are most potentially caused by the adversarial noises and supersede the original activations. Therefore, we suppress these activations to resurrect the original ones. Such an adversarial activation suppression scheme inherits the defense methodology we proposed in the image domain.

6 EXPERIMENT AND EVALUATION

In this section, we evaluate our method in terms of its effectiveness and efficiency in two application scenarios: image and audio physical adversarial attacks. The CNN models and datasets used in our experiments are listed in Table 1: For physical adversarial attack in image scenarios, we test our defense method's performance on Inception-V3 [25], VGG-16 [23], and ResNet-18 [14] using ImageNet dataset [9]. For audio scenarios, we use Command Classification Model [17] on Google Voice Command dataset [17].

6.1 CNN Image physical Adversarial Attack Defense Evaluation

In this part, we evaluate our proposed defense method for image physical adversarial attack scenario. The adversarial patches are generated by using Inception-V3 as the base model. The generated patch with high transferability are utilized to attack three models: Inception-V3 itself and two other models, VGG-16 and ResNet-18. Then we applied our defense methods on all the models and test their detection success rates. Meanwhile, we also record the time cost of defense methods to demonstrate the efficiency of our method. The baseline methods is *Blind*, which is one state-of-the-art defense method [13]. And the threshold for inconsistency is set as 0.46.

Table 2 shows the overall detection and defense performance. On all three models, our method consistently shows higher detection success rate than [13]. The further proposed image recovery could help to correct predictions, resulting in 80.3%~82% accuracy recovery improvement on different models while *Blind* only achieves 78.2%~79.5% accuracy recovery improvement. In terms of efficiency, the process time cost of our detect method for one physical adversarial attack is from 67ms~71ms while the *Blind* is from 132ms~153ms.

By above comparison, we show that our defense method has better defense performance than *Blind* with respect to both effectiveness and efficiency.

6.2 CNN Audio physical Adversarial Attack Defense Evaluation

In this part, we evaluate the effectiveness and efficiency of the proposed defense method in audio physical adversarial attack scenarios. The inconsistency threshold for adversarial detection is obtained by the grid search and set as 0.11 in this experiment. For comparison, we re-implement another two state-of-the-art defense methods: *Dependency Detection* [28] and *Multiversion* [21]. Four methods [1, 6, 12, 16] are used as attacking methods to prove the generality of our defense method. Fig. 8 shows the overall performance comparison.

Our method can always achieve more than 92% detection success rate for all types of audio physical adversarial attacks. By contrast, *Dependency Detection* achieves 89% detection success rate in average while *Multiversion Detection* only have average 74%. Therefore, our method demonstrates the best detection accuracy.

Table 1: CNN Models and Datasets

Attack	Model	Dataset
Image Physical Adversarial Attack	Inception-V3 VGG-16 ResNe-18	ImageNet-10
Audio Physical Adversarial Attack	Command Classification	Speech Commands

Table 2: Image Adversarial Patch Attack Defense Evaluation

Stage		Inception-V3		VGG-16		ResNet-18	
		Blind*	Ours	Blind*	Ours	Blind*	Ours
Detection	Detection Succ. Rate	88%	**91%**	89%	**90%**	85%	**89%**
	Time Cost	132ms	**68**ms	144ms	**67**ms	153ms	**71**ms
Defense	Original Acc.	9.8%	9.8%	9.5%	9.8%	10.8%	9.8%
	Recovery Acc.	88%	**90.1%**	89.3%	**91.5%**	90%	**91.8%**

*:Blind [13]

Table 3: Audio Adversarial Attack Defense Evaluation

Method	FGSM [12]	BIM [16]	CW [6]	Genetic [1]	Time Cost
No Defense	10%	5%	4%	13%	NA
Dependency Detection [28]	85%	83%	80%	80%	1813ms
Noise Flooding [22]	62%	65%	62%	59%	1246ms
Ours	**87%**	**88%**	**85%**	**83%**	**521ms**

Then we evaluate our method's defense performance. The K value in the top-k index we mentioned above is set as 6. Since *Multiversion* [21] cannot be used to defense, we re-implement another method, *Noise Flooding* [22] as comparison. And we use the original vulnerable model without attack defense as the baseline.

Table 2 shows the overall defense performance evaluation. After applying our defense method, the prediction accuracy significantly increase from average 8% to average 85.8%, which is 77.8% accuracy recovery. Both *Dependency Detection* and *Noise Flooding* have lower accuracy recovery rate, which are 74% and 54%, respectively.

For defense efficiency, since our method is based on the activation pattern and numerical similarity (which is easy to compute), the detection can be efficiently done during the CNN forward process. As the result, the time cost of our method is 521ms while other two methods usually cost more than 1540ms for each single physical adversarial attack. Therefore, our defense method is 2~3× faster than the other two methods.

7 CONCLUSION

In this paper, we propose a CNN defense methodology for physical adversarial attacks for both image and audio recognition applications. Leveraging the comprehensive CNN vulnerability analysis and novel CNN semantic/activation inconsistency metrics, our method can effectively and efficiently detect and eliminate the image and audio physical adversarial attack. Experiments show that our methodology can achieve an average 90% successful rate for attack detection and 81% accuracy recovery for image physical adversarial attack. Also, the proposed defense method can achieve 92% detection successful rate and 77.5% accuracy recovery for audio recognition applications. Moreover, the proposed defense methods are at most 2.3× faster compared to the state-of-the-art defense methods, making them feasible to resource-constrained platforms, such as mobile devices.

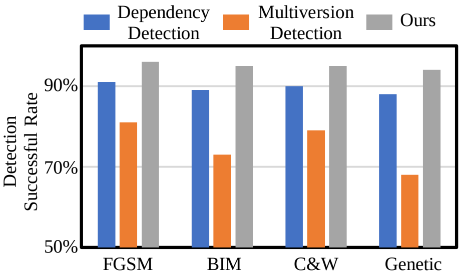

Figure 8: Audio Adversarial Attack Detection Performance

REFERENCES

[1] Moustafa Alzantot and *et al.* 2018. Did you hear that? adversarial examples against automatic speech recognition. *arXiv preprint arXiv:1801.00554* (2018).

[2] Chantal E Ayres and *et al.* 2008. Measuring fiber alignment in electrospun scaffolds: a user's guide to the 2D fast Fourier transform approach. *Journal of Biomaterials Science, Polymer Edition* 19, 5 (2008), 603–621.

[3] Jacob Benesty and *et al.* 2009. Pearson correlation coefficient. In *Noise reduction in speech processing*. Springer, 1–4.

[4] Marcelo Bertalmio and *et al.* 2000. Image inpainting. In *Proc. of SIGGRAPH*. ACM Press/Addison-Wesley Publishing Co., 417–424.

[5] Tom B Brown and *et al.* 2017. Adversarial patch. *arXiv preprint arXiv:1712.09665* (2017).

[6] Nicholas Carlini and David Wagner. 2017. Towards evaluating the robustness of neural networks. In *Proc. of SP*. IEEE, 39–57.

[7] Chung-Cheng Chiu and *et al.* 2018. State-of-the-art speech recognition with sequence-to-sequence models. In *Proc. of ICASSP*. IEEE, 4774–4778.

[8] Jan K Chorowski and *et al.* 2015. Attention-based models for speech recognition. In *Proc. of NIPS*. 577–585.

[9] Jia Deng and *et al.* 2009. Imagenet: A large-scale hierarchical image database. In *Proc. of CVPR*. IEEE, 248–255.

[10] Dumitru Erhan and *et al.* 2009. Visualizing higher-layer features of a deep network. *University of Montreal* 1341, 3 (2009), 1.

[11] Kevin Eykholt and *et al.* 2017. Robust physical-world attacks on deep learning models. *arXiv preprint arXiv:1707.08945* (2017).

[12] Ian J Goodfellow and *et al.* 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572* (2014).

[13] Jamie Hayes. 2018. On visible adversarial perturbations & digital watermarking. In *Proc. of CVPR Workshops*. 1597–1604.

[14] Kaiming He and *et al.* 2015. Deep Residual Learning for Image Recognition. In *Proc. of CVPR*. 770–778.

[15] Vincent James. 2018. Google is making it easier than ever to give any app the power of object recognition. https://www.theverge.com/2017/6/15/15807096/google-mobile-ai-mobilenets-neural-networks

[16] Alexey Kurakin and *et al.* 2016. Adversarial examples in the physical world. *arXiv preprint arXiv:1607.02533* (2016).

[17] Scott Anthony Morgan and *et al.* 2001. Speech command input recognition system for interactive computer display with term weighting means used in interpreting potential commands from relevant speech terms. US Patent 6,192,343.

[18] Muzammal Naseer and *et al.* 2019. Local Gradients Smoothing: Defense against localized adversarial attacks. In *Proc. of WACV*. IEEE, 1300–1307.

[19] Suphakit Niwattanakul and *et al.* 2013. Using of Jaccard coefficient for keywords similarity. In *Proc. of IMECS*, Vol. 1. 380–384.

[20] Margarita Osadchy and *et al.* 2017. No bot expects the DeepCAPTCHA! Introducing immutable adversarial examples, with applications to CAPTCHA generation. *IEEE Transactions on Information Forensics and Security* 12, 11 (2017), 2640–2653.

[21] Zeng Qiang and *et al.* 2018. A Multiversion Programming Inspired Approach to Detecting Audio Adversarial Examples. *arXiv preprint arXiv:1812.10199* (2018).

[22] Krishan Rajaratnam and Jugal Kalita. 2018. Noise Flooding for Detecting Audio Adversarial Examples Against Automatic Speech Recognition. In *Proc. of ISSPIT*. IEEE, 197–201.

[23] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[24] Christian Szegedy and *et al.* 2013. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199* (2013).

[25] Christian Szegedy and *et al.* 2015. Going Deeper with Convolutions. In *Proc. of CVPR*. 1–9.

[26] Fei Wang and *et al.* 2017. Residual attention network for image classification. In *Proc. of CVPR*. 3156–3164.

[27] Hiromu Yakura and Jun Sakuma. 2018. Robust audio adversarial example for a physical attack. *arXiv preprint arXiv:1810.11793* (2018).

[28] Zhuolin Yang and *et al.* 2018. Characterizing Audio Adversarial Examples Using Temporal Dependency. *arXiv preprint arXiv:1809.10875* (2018).

[29] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. 2016. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2921–2929.

[30] Barret Zoph and *et al.* 2018. Learning transferable architectures for scalable image recognition. In *Proc. of CVPR*. 8697–8710.