

LanCe: A Comprehensive and Lightweight CNN Defense Methodology against Physical Adversarial Attacks on Embedded Multimedia Applications

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

ABSTRACT

Recently, adversarial attacks can be applied to the physical world, causing practical issues to various Convolutional Neural Networks (CNNs) powered applications. Most existing physical adversarial attack defense works only focus on eliminating explicit perturbation patterns from inputs, ignoring interpretation to CNN’s intrinsic vulnerability. Therefore, they lack expected versatility to different attacks and thereby depend on considerable data processing costs. In this paper, we propose *LanCe* – a comprehensive and lightweight CNN defense methodology against different physical adversarial attacks. By interpreting CNN’s vulnerability, we find that non-semantic adversarial perturbations can activate CNN with significantly abnormal activations and even overwhelm other semantic input patterns’ activations. We improve the CNN recognition process by adding a self-verification stage to detect the potential adversarial input with only one CNN inference cost. Based on the detection result, we further propose a data recovery methodology to defend the physical adversarial attacks. We apply such defense methodology into both image and audio CNN recognition scenarios and analyze the computational complexity for each scenario, respectively. Experiments show that our methodology can achieve an average 91% successful rate for attack detection and 89% accuracy recovery. Moreover, it is at most $3\times$ faster compared with the state-of-the-art defense methods, making it feasible to resource-constrained embedded systems, such as mobile devices.

I. INTRODUCTION

In the past few years, Convolutional Neural Networks (CNNs) powered applications are facing a critical challenge – adversarial attacks. By injecting particular perturbations into input data, adversarial attacks can mislead CNN recognition results. With aggressive methods proposed, adversarial perturbations can be concentrated into a small area and attached to the real objects, which easily threaten the CNN recognition systems in the physical world. The left side of Fig. 1 shows a physical adversarial example on the traffic sign detection. When attaching 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.

Many works have been proposed to defend against physical adversarial attacks [1–4]. However, most of them neglected CNN’s intrinsic vulnerability interpretations. Instead, either they merely focused on eliminating explicit perturbation patterns from input [2], or they simply adopted multiple CNNs to conduct the cross-verification [3, 4]. All these methods have certain drawbacks: They failed to find a common defense methodology, lacking versatility for preventing different physical adversarial attacks. Moreover, they introduced considerable data processing costs during perturbations elimination, which significantly increased methods’ computation costs.

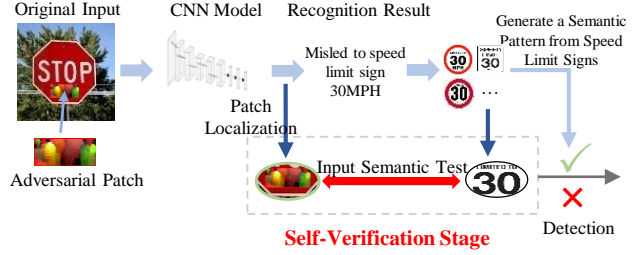


Fig. 1.: Physical Adversarial Attack for Traffic Sign

In this paper, we propose *LanCe*, a comprehensive and lightweight defense methodology against different physical adversarial attacks. By interpreting CNN’s 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. Leveraging the adversarial attacks’ characteristic inconsistencies, we improve the CNN recognition process by adding a self-verification stage. Fig. 1 illustrates the self-verification stage for a traffic sign adversarial attack. For each input image, after one CNN inference, the verification stage will locate the significant activation sources (green circle) and calculate the input semantic inconsistency with the expected semantic patterns (right circle) according to the prediction result. Once the inconsistency exceeds a pre-defined threshold, CNN will conduct a data recovery process to recover the input image. Our defense methodology has minimum computation components involved, which can be extended to CNN based image and audio recognition scenarios.

Specifically, we have following contributions in this work:

- By interpreting CNN’s vulnerability, we identify characteristic inconsistencies between the physical adversarial attack and the natural input recognition.
- We propose a self-verification stage to detect the abnormal activation patterns’ semantics with only one CNN inference involved.
- We further propose a data recovery methodology to recover both attacked image and audio input data. Moreover, we apply such detection and data recovery methodology into image and audio scenarios.
- In each scenario, we quantitatively analyze our defense process’s computational complexity, and guarantee the lightweight computation cost.

Experiments show that our method can achieve an average 90% detection successful rate and average 81% accuracy recovery for image physical adversarial attacks. Also, our method achieves 92% detection successful rate and 77.5% accuracy recovery for audio adversarial attacks. Moreover, our method is at most $3\times$ faster than the state-of-the-art defense methods, which is feasible to various resource-constrained embedded systems, such as mobile devices.

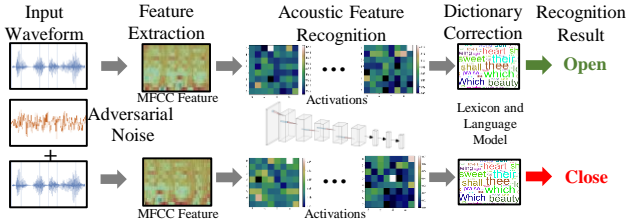


Fig. 2.: Audio Recognition and Physical Adversarial Attack Process

II. BACKGROUND AND RELATED WORKS

A. Physical Adversarial Attacks

Adversarial attacks started to arouse researchers' general concern with adversarial examples, which were first introduced by [5]. Recently, adversarial attack approaches were also brought from the algorithm domain into the physical world, which are referred as the physical adversarial attack. [6] 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. [7] extended the scope of physical attacks with adversarial patches. With more aggressive patterns than graffiti, these patches can be attached to physical objects arbitrarily and have strong model transferability.

Beyond aforementioned image cases, some physical adversarial attacks also have been proposed to audios. Yakura *et al.* [8] proposed an audio physical adversarial attack that can still be effective after playback and recording in the physical world. [9] generated audio adversarial commands in a normal song which can be played through the air.

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

B. Image physical Adversarial Attack Defense

There are several works have been proposed to defense such physical adversarial attacks in the image recognition process. Naseer *et al.* proposed a local gradients smoothing scheme against physical adversarial attacks [2]. By regularizing gradients in the estimated noisy region before feeding images into CNN inference, their method can eliminate the potential impacts from adversarial attacks. Hayes *et al.* proposed a physical image adversarial attack defense method based on image inpainting [1]. 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.

Although these methods are effective for image physical adversarial attacks defense, they still have certain disadvantages regarding versatility and computation. These methods are designed for solving specific adversarial attack which are not integrated for different physical adversarial attack situations. Moreover, they will introduce huge computation costs.

C. Audio Physical Adversarial Attack 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



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

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 [3, 4, 10]. Zeng *et al.* leveraged multiple Automatic Speech Recognition (ASR) systems to detect audio physical adversarial attack based on a cross-verification methodology [4]. However, their method lacks certain versatility which cannot detect the adversarial attacks with model transferability. Yang *et al.* proposed an audio adversarial attack detection and defense method by exploring the temporal dependency in audio adversarial attacks [3]. However, their method requires multiple CNN recognition inferences which is time-consuming.

III. INTERPRETATION ORIENTED PHYSICAL ADVERSARIAL ATTACKS ANALYSIS AND DEFENSE

In this section, we first interpret the CNN vulnerability by analyzing input patterns' semantics with the activation maximization visualization [11]. Based on semantics analysis, we identify the adversarial attack patches as non-semantic input patterns with abnormal distinguished activations. Specifically, to evaluate the semantics, we propose metrics that can measure 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 lightweight defense methodology consists of the self-verification and the data recovery.

A. CNN Vulnerability Interpretation

Interpretation and Assumption: In a typical image or audio recognition process, CNN extracts features from the original input data and gradually derive a prediction result. However, when injecting physical adversarial perturbations into the original data, CNN will be misled to a wrong prediction result. To better interpret the vulnerability, we major focus on a typical image physical adversarial attack – adversarial patch attack as an example. In Fig. 1, by comparing with the original input, we find that an adversarial patch usually has no constraints in color/shape, *etc.* Such patches usually sacrifice the semantic structures so as to cause significant abnormal activations and overwhelm the other input patterns' activations. *Therefore, we make an assumption that CNN lacks qualitative semantics distinguishing ability which can be activated by the non-semantic adversarial patch during CNN inference.*

Assumption Verification: According to our assumption, the non-semantic input patterns will lead to abnormal activations while the semantic input patterns generate normal activations. We can evaluate this difference by investigating the semantic of each neuron in CNN. Therefore, we adopt a visualized CNN semantic analysis method – Activation Maximization Visualization (AM) [11]. 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 i th neuron N_i^l in the layer of l . Specifically, this

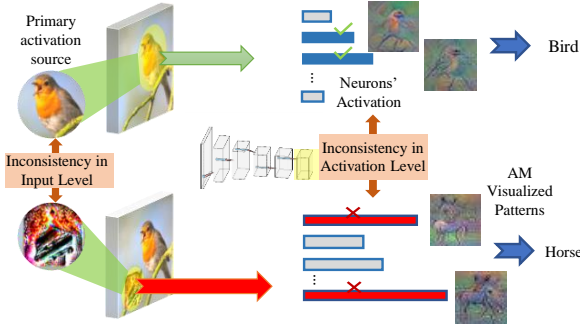


Fig. 4.: Image Adversarial Patch Attack

process can be formulated as:

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

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 semantic input patterns by using AM. As the original AM method is designed for semantics interpretation, many feature regulations 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 indicate they have clear semantics. However, when we remove these semantics regulations in the AM process, we 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.

B. Inconsistency Metrics for Input Semantic and Prediction Activation

Inconsistency Identification: To identify the non-semantic input patterns for the attack detection, we examine its impacts during CNN inference by comparing the natural image recognition with the 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 difference between the adversarial patch and primary activation source on the original image, which is referred as **Input Semantic Inconsistency**. From the aspect of prediction activation magnitudes, we observe another difference between the adversarial input and the original input, namely **Prediction Activation Inconsistency**.

Inconsistency Metrics Formulation: We further define two metrics to indicate above two inconsistencies' degrees.

1) Input Semantic Inconsistency Metric: This metric measures the input semantic inconsistency between the non-semantic adversarial patches and the semantic local input patterns from the natural image. It can be defined as follows:

$$D(P_{pra}, P_{ori}) = 1 - S(P_{pra}, P_{ori}), \quad P_{pra} \leftarrow \mathbb{R} A_i^l(p), \quad P_{ori} \leftarrow \mathbb{R} A_i^l(o), \quad (2)$$

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 the adversarial patch and the original input, respectively. \mathbb{R} maps

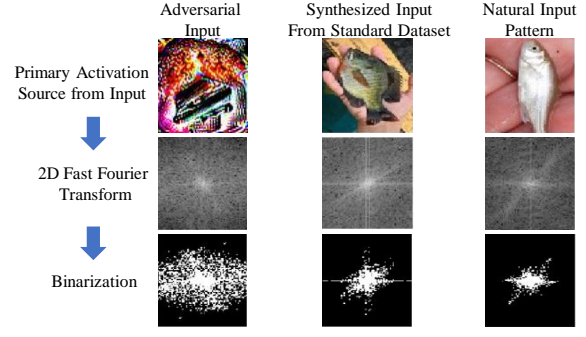


Fig. 5.: The Results after 2D Fast Fourier Transform

neurons' activations to the primary local input patterns. S represents a similarity metric.

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

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

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

For the 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 class can be obtained from a standard dataset. 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.

C. Physical Adversarial Attack Defense based on CNN Self-Verification and Data Recovery

The proposed two inconsistencies demonstrate the difference between physical adversarial attacks and natural image recognition *w.r.t* input patterns and prediction activations. By utilizing the inconsistency metrics, we propose a defense methodology which consists of a self-verification and a data recovery in the CNN decision-making process. Specifically, the entire methodology flow is described as following:

Self-Verification: (1) We first feed the input into the 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 CNN leverages the proposed metrics to measure the two inconsistencies between the practical input and the synthesized data with the prediction class. (4) Once any inconsistency exceeds the given threshold, CNN will consider the input as an adversarial input.

Data Recovery: (5) After a physical adversarial attack has been detected by the self-verification stage, the data recovery methodology is further applied to recover the input data which has been attacked. Specifically, we leverage image inpainting and activation denoising to recover the input image and audio.

We will derive two methods from such methodology for image and audio scenarios in Section 4 and Section 5.

Computational Complexity: As aforementioned, the computation cost is critical to the adversarial defense approaches. Therefore, we leverage computational complexity to evaluate the methodology's total computation cost. A low computational complexity indicates a small computation workload, proving the proposed methodology is lightweight. In our defense methodology, the computational complexity is mainly contributed by the inner steps such as the CNN inference, inconsistency metrics calculation and data recovery. In following two scenarios, we will specifically analyze the computation complexity for each of above steps.

IV. DEFENSE AGAINST IMAGE PHYSICAL ADVERSARIAL ATTACK

In this section, we will specifically describe our defense methodology against image physical adversarial attacks.

A. Defense Process in the Image Scenario

Primary Activation Pattern Localization: For the 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) [12]. Let $A_k(x, y)$ denotes the value of the k^{th} activation in the last convolutional layer at spatial location (x, y) . We can compute a sum of all activations at the spatial location (x, y) in the last convolutional layer as:

$$A_T(x, y) = \sum_K A_k(x, y), \quad (4)$$

where K is the total number of activations in the last convolutional layer. The larger value of $A_T(x, y)$ indicates the activation source in the input image at the corresponding spatial location (x, y) is more important for classification result.

Inconsistency Derivation: According to our preliminary analysis, the input adversarial patch contains much more high-frequency information than the natural semantic input patterns. Therefore, we convert the patterns with a series of transformations which are shown in Fig. 5. After the 2D Fast Fourier Transform (2D-FFT) transformation and binary conversion, we can observe the significant difference between adversarial input and semantic synthesized input. Therefore, we replace $S(I_{pra}, I_{ori})$ with Jaccard Similarity Coefficient (JSC) [13] and propose our image inconsistency metric as:

$$D(P_{pra}, P_{exp}) = 1 - JSC(P_{pra}, P_{exp}) = \frac{|P_{pra} \setminus P_{exp}| + |P_{pra} \cap P_{exp}|}{|P_{pra} \cup P_{exp}|}, \quad (5)$$

where I_{exp} is the synthesized semantic pattern with predicted class. $P_{pra} \cap P_{exp}$ 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 our specific defense methodology which contains self-verification and image recovery. The entire process is described in Fig. 6.

Self-Verification for Detection: For each input image, we apply CAM to locate the source location of the biggest model activations. Then we crop the image to obtain patterns with maximum activations. During semantic test, we calculate the inconsistency between I_{pra} and I_{exp} . If it is higher than a pre-defined threshold, we consider an adversarial input detected.

Data Recovery for Image: After the patch is detected, we conduct the image data recovery by directly removing patch from the original input data. In our case, to ensure the lightweight computation workload, we leverage Neighbor Interpolation, a simple but effective image inpainting technology to repair the image and eliminate the attack effects. Concretely, each pixel in the adversarial patch will be replaced by the average value of its eight surrounding pixels. After the interpolation, we feed back the recovery image into CNN to do the prediction again. With above steps, we can defend an image physical adversarial attack during CNN inference.

B. Computational Complexity Analysis

The total computation complexity of the defense process in the image scenario is contributed by following four steps: the CNN inference, the maximum activation pattern locating, the inconsistency metric calculation and the image interpolation. We model each step's computational complexity as following:

CNN Inference: When the input image is first fed into CNN for class prediction, the inference computational complexity C_C is formulated as:

$$C_C \sim \mathcal{O}\left(\sum_{i=1}^L \sum_{j=1}^{n_i} r_i^{j^2} n_{i-1} h_i^j w_i^j\right), \quad (6)$$

where $r_i^{j^2}$ represents j^{th} filter's kernel size in i^{th} layer, $h_i^j w_i^j$ denotes the corresponding size of output feature map, L is the total layer number and n_i is the filter numbers in i^{th} layer.

Primary Activation Pattern Localization: Since computation complexities of other operations such as cropping are negligible, we consider CAM contributes the primary computational complexity in this step. In CAM, each spatial location (x, y) in the last convolutional layer is the weighted sum of K activations. Therefore, the total computational complexity is: $C_M \sim \mathcal{O}(K h_L^{n_L} w_L^{n_L})$, where $h_L^{n_L} w_L^{n_L}$ is the size of the feature map in last convolutional layer.

Inconsistency Metric Derivation: This step consists of 2D-FFT calculation and JSC calculation. According to the analysis in [14, 15], the computational complexities of above two processes can be approximate to $C_F \sim \mathcal{O}(N \log N)$ and $C_J \sim \mathcal{O}(n_a \log n_a)$, where N and n_a represent N pixel number in input image and maximum activation pattern, respectively.

Image Interpolation: For each pixel, the total operation number during interpolation is nine (eight adding operation and one dividing operation). Therefore, the total interpolation computation complexity for the entire adversarial patch is $C_L \sim \mathcal{O}(9n_a)$.

Comparing with the last three steps, the computational complexity of CNN inference dominates the entire computational complexity of our defense methodology in the image scenario. Since our methodology only involves one CNN inference, it usually has less computation cost than other methods.

Case Study: To examine the lightweight of our method, we use VGG-16 [19] with 224×224 input image as an example. According to the built models, the total computation complexity of our defense method is approximate to $\mathcal{O}(15300M)$ FLOPs (Floating Point Operations) while [2] is approximate

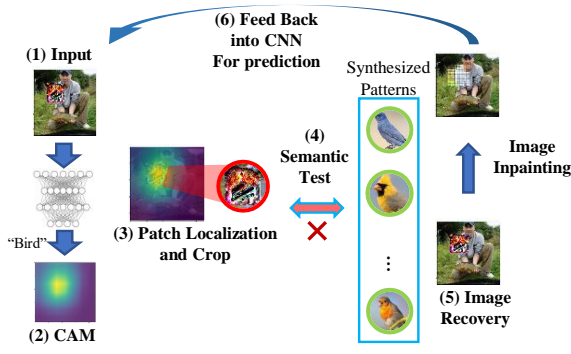


Fig. 6.: Adversarial Patch Attack Defense

to $\mathcal{O}(18300M)$. Our method's superiority in terms of computational complexity will be further verified by evaluating the process time cost in Section 6.

V. DEFENSE AGAINST AUDIO PHYSICAL ADVERSARIAL ATTACK

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

A. Defense Process in the Audio Scenario

Inconsistency Derivation: 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. In that case, 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 the practical input and the synthesized data with the same prediction class. We adopt a popular similarity evaluation method - Pearson Correlation Coefficient (PCC) [16] and the inconsistency metric is defined as:

$$D(f_{pra}, f_{exp}) = 1 - PCC(f_{pra}, f_{exp}) = 1 - \frac{E[(f_{pra} - \mu_{pra})(f_{exp} - \mu_{exp})]}{\sigma_{pra} \sigma_{exp}}, \quad (7)$$

where I_{pra} and I_{exp} represent the activations in the last convolutional layer for both practical input and synthesized input. μ_a and μ_o denote mean values of f_{pre} and f_{exp} , σ_{pra} and σ_{exp} are standard derivations, and E means the overall expectation.

Self-Verification for Detection: With established inconsistency metric, we further apply self-verification stage to CNN for the audio physical adversarial attack. The detection flow is described as following: We first obtain activations in the last 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 the audio adversarial attack, $D(I_{pra}, I_{exp})$ will exceed a pre-defined threshold. According to our preliminary experiments tested with various attacks, $D(I_{pra}, I_{exp})$ of an adversarial input is usually larger than 0.18 while a natural input's $D(I_{pra}, I_{exp})$ is usually smaller than 0.1. Therefore, there exists a large range for the threshold to distinguish the natural and the adversarial input audios, which can benefit our accurate detection.

Data Recovery for Audio: 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 solutions. We propose a new solution - "activation denoising" as our defense method, which targets ablating adversarial effects from the activation level. The

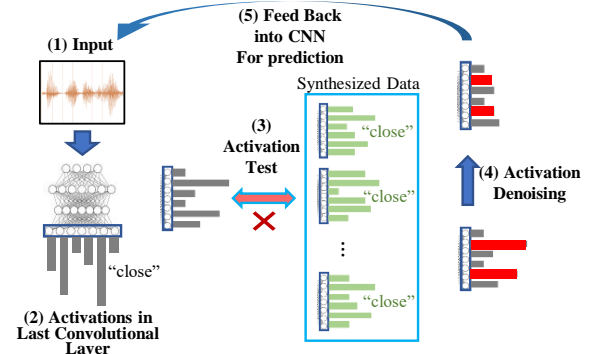


Fig. 7.: Audio Adversarial Attack Defense

activation denoising takes advantages of the aforementioned last layer activation patterns, which have stable correlations with determined predication labels.

Our adversarial audio recovery method is shown in Fig. 7: Based on detection results, we can identify the wrong prediction label, and obtain the standard activation patterns of the wrong class in the last layer. (For the best performance, we locate the top- k activation index.) 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 original ones.

B. Computational Complexity Analysis

The computational complexity in the audio scenario is mainly determined by the CNN inference and the inconsistency metric calculation, since other steps directly manipulate limited activation values with negligible computation workload involved. Therefore, we model the computational complexity as following:

CNN Inference: Since the audio has same inference process in CNN, we can use the same model in image scenario to measure the computation complexity in the audio scenario.

Inconsistency Metric Derivation: The computation complexity of this step is contributed by the PCC calculation, which can be formulated as $C_P \sim \mathcal{O}(n_L^2)$, where n_L is the activation number in the last layer.

Case Study: We also leverage a case study to specifically demonstrate that our proposed methodology is lightweight comparing with others in the audio scenario. The CNN model is Command Classification model [17] with 1s audio input (16000 sample rate). Therefore, the total computation complexity of our methodology is approximate to $\mathcal{O}(500M)$ FLOPs (Float Point Operations). However, the computation complexities of other two state-of-the-art audio defense methods are around $\mathcal{O}(1100M)$ and $\mathcal{O}(1600M)$. Therefore, our proposed methodology is more friendly to resource-constrained mobile devices.

VI. EXPERIMENT AND EVALUATION

In this section, we evaluate *LanCe* in terms of effectiveness and efficiency for image and audio physical adversarial attacks.

A. Defense Evaluation for Image Scenario

Experiment Setup: Our detection method is mainly evaluated for adversarial patch attacks. The adversarial patches are generated by using Inception-V3 [18] as the base model. The generated patch with high transferability are utilized to attack other two models: VGG-16 [19] and ResNet-18 [20]. Then we apply our defense method on all three models and test their detection and recovery success rates. Meanwhile, we also record

TABLE I

: Image Adversarial Patch Attack Defense Evaluation

Stage		Inception-V3		VGG-16		ResNe-18t	
		<i>PM*</i>	<i>LanCe</i>	<i>PM*</i>	<i>LanCe</i>	<i>PM*</i>	<i>LanCe</i>
Detection	Detection Succ. Rate	88%	91%	89%	90%	85%	89%
Recovery	Original Acc.	9.8%	9.8%	9.5%	9.8%	10.8%	9.8%
	Recovery Acc.	88%	90%	89.3%	91.5%	90%	91%
	Time	233ms	192ms	315ms	243ms	461ms	318ms

*: Patch Masking (PM) [1]

TABLE II

: Audio Adversarial Attack Data Recovery Evaluation

Method	FGSM	BIM	CW	Genetic	Time Cost
No Recovery	10%	5%	4%	13%	NA
Dependency Detection [3]	85%	83%	80%	80%	1813ms
Noise Flooding [10]	62%	65%	62%	59%	1246ms
<i>LanCe</i>	87%	88%	85%	83%	521ms

the time cost of defense methods to demonstrate the efficiency of *LanCe*. The baseline methods is *Patch Masking*, which is one state-of-the-art defense method [1]. And the threshold for inconsistency is set as 0.46.

Defense Effectiveness: Table I shows the overall detection and image recovery performance. On all three models, *LanCe* consistently shows higher detection success rate than [1]. The further proposed image recovery could help to correct predictions, resulting in 80.3%~82% accuracy recovery improvement on different models while *Patch Masking* only achieves 78.2%~79.5% accuracy recovery improvement.

Time Cost: We leverage the process time cost to represent the method's computational complexity. We can find that the process time cost of our defense method for one physical adversarial attack is from 67ms~71ms while the *Patch Masking* is from 132ms~153ms.

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

B. Defense Evaluation for Audio Scenario

Experiment Setup: For audio scenario, we use Command Classification Model [17] on Google Voice Command dataset [17]. 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* [3] and *Multiversion* [4]. Four methods [5, 21–23] are used as attacking methods to prove the generality of our defense method. Fig. 8 shows the overall performance comparison.

Defense Effectiveness: *LanCe* can always achieve more than 92% detection success rate for all audio physical adversarial attacks. By contrast, *Dependency Detection* achieves 89% detection success rate in average while *Multiversion Detection* only have average 74%. Therefore, *LanCe* demonstrates the best detection accuracy. Then we evaluate *LanCe*'s recovery performance. The k value in the top- k index is set as 6. Since *Multiversion* [4] cannot be used to recovery, we re-implement another method, *Noise Flooding* [10] as comparison. And we use the original vulnerable model without data recovery as the baseline. Table I shows the overall audio recovery performance evaluation. After applying our recovery 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

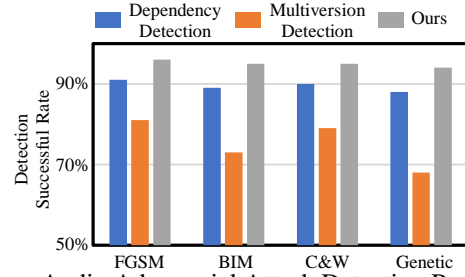


Fig. 8.: Audio Adversarial Attack Detection Performance

rate, which are 74% and 54%, respectively.

Time Cost: For defense efficiency, the computational complexity of *LanCe* is much lower than other methods according to our previous analysis. As the result, the time cost of our method is 521ms while other two methods usually cost more than 1540ms for a single physical adversarial attack. Thus, our defense method is 2~3× faster than other two methods.

VII. 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 two novel CNN inconsistency metrics, our method can effectively and efficiently detect and eliminate the image and audio physical adversarial attacks. Experiments show that our methodology can achieve an average 91% successful rate for attack detection and 89% accuracy recovery. Moreover, the proposed defense methods are at most 3× faster compared to the state-of-the-art defense methods, making them feasible to resource-constrained embedded systems, such as mobile devices.

REFERENCES

- [1] J. Hayes, "On visible adversarial perturbations & digital watermarking," in *Proc. of CVPR Workshops*, 2018, pp. 1597–1604.
- [2] M. Naseer and *et al.*, "Local gradients smoothing: Defense against localized adversarial attacks," in *Proc. of WACV*, 2019, pp. 1300–1307.
- [3] Z. Yang and *et al.*, "Characterizing audio adversarial examples using temporal dependency," *arXiv preprint arXiv:1809.10875*, 2018.
- [4] Q. Zeng and *et al.*, "A multiversion programming inspired approach to detecting audio adversarial examples," *arXiv preprint arXiv:1812.10199*, 2018.
- [5] I. Goodfellow and *et al.*, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.
- [6] K. Eykholt and *et al.*, "Robust physical-world attacks on deep learning models," *arXiv preprint arXiv:1707.08945*, 2017.
- [7] T. Brown and *et al.*, "Adversarial patch," *arXiv preprint arXiv:1712.09665*, 2017.
- [8] H. Yakura and *et al.*, "Robust audio adversarial example for a physical attack," *arXiv preprint arXiv:1810.11793*, 2018.
- [9] X. Yuan and *et al.*, "Commandersong: A systematic approach for practical adversarial voice recognition," *arXiv preprint arXiv:1801.08535*, 2018.
- [10] K. Rajaratnam and *et al.*, "Noise flooding for detecting audio adversarial examples against automatic speech recognition," in *Proc. of ISSPIT*, 2018, pp. 197–201.
- [11] D. Erhan and *et al.*, "Visualizing higher-layer features of a deep network," *University of Montreal*, vol. 1341, no. 3, p. 1, 2009.
- [12] B. Zhou and *et al.*, "Learning deep features for discriminative localization," in *Proc. of CVPR*, 2016, pp. 2921–2929.
- [13] S. Niwattanakul and *et al.*, "Using of jaccard coefficient for keywords similarity," in *Proc. of IMECS*, vol. 1, no. 6, 2013, pp. 380–384.
- [14] J. S. Plank, "Cs494 lecture notes - minhash," 2018. [Online]. Available: <http://web.eecs.utk.edu/~plank/plank/classes/cs494/494/notes/Min-Hash/index.html>
- [15] M. Lohne, "The computational complexity of the fast fourier transform," 2017. [Online]. Available: <https://folk.uio.no/mathialo/texts/fftcomplexity.pdf>
- [16] J. Benesty and *et al.*, "Pearson correlation coefficient," in *Noise reduction in speech processing*. Springer, 2009, pp. 1–4.
- [17] S. Morgan and *et al.*, "Speech command input recognition system for interactive computer display with term weighting means used in interpreting potential commands from relevant speech terms," 2001, uS Patent 6,192,343.
- [18] C. Szegedy and *et al.*, "Going Deeper with Convolutions," in *Proc. of CVPR*, 2015, pp. 1–9.
- [19] K. Simonyan and *et al.*, "Very Deep Convolutional Networks for Large-scale Image Recognition," *arXiv preprint arXiv:1409.1556*, 2014.
- [20] K. He and *et al.*, "Deep Residual Learning for Image Recognition," in *Proc. of CVPR*, 2015, pp. 770–778.
- [21] A. Kurakin and *et al.*, "Adversarial examples in the physical world," *arXiv preprint arXiv:1607.02533*, 2016.
- [22] N. Carlini and *et al.*, "Towards evaluating the robustness of neural networks," in *Proc. of SP*, 2017, pp. 39–57.
- [23] M. Alzantot and *et al.*, "Did you hear that? adversarial examples against automatic speech recognition," *arXiv preprint arXiv:1801.00554*, 2018.