2018

Two approaches to defend against adversarial examples: Attention-based and Certificate-based

Chanh Nguyen
Lehigh University, chanhnp@gmail.com

Two approaches to defend against adversarial examples: Attention-based and Certificate-based

by

Chanh Nguyen

Presented to the Graduate and Research Committee

of Lehigh University

in Candidacy for the Degree of

Master of Science

in

Computer Science

Lehigh University

August 2018

This thesis is accepted and approved in partial fulfillment of the requirements for the Master of Science.

_____

Date

_____

Thesis Advisor: Ting Wang

_____

Chairperson of Department: Daniel Lopresti

Acknowledgements

I would like to thank Professor Ting Wang, Master student Georgi Georgiev, PhD candidate Yujie Ji and PhD candidate Xinyang Zhang for working with me throughout this project. Your advice and help was indispensable.

Contents

Acknowledgements iv

List of Tables vi

List of Figures vii

Abstract 1

1 Introduction . . . 2

2 Methods of attacking . . . 4

2.1 F_∞: Fast Gradient Sign Method (FSGM) . . . 4

2.2 F_2: Carlini & Wagner (C&W) . . . 5

2.3 Spatially Transformed Adversarial Examples . . . 6

3 Attention-based defense against pixel-based attacks . . . 7

3.1 Latent Attention Network . . . 7

3.2 Methodology . . . 7

3.3 Evaluation . . . 8

4 Certified defense against spatial transformation attack . . . 11

4.1 Methodology . . . 12

4.2 Evaluation . . . 13

5 Conclusion and future direction . . . 14

Bibliography 17

List of Tables

1 Percentage of adversarial samples whose attention masks retain their original classifications. . . . 9
2 Classification accuracy on adversarial samples generated using different attacks on AlexNet. . . . 10
3 Classification accuracy on adversarial samples generated with different attacks on VGG-like. . . . 10
4 Classification accuracy on adversarial samples generated using different attacks on AlexNet, after filtering out the incorrect masks. . . . 11

List of Figures

1 Example of an adversarial example: both pictures look like horses but the one on the right can trick neural networks into classifying it as a dog. 3

2 Original images from each class and their masks generated by LAN. 4

3 Framework of classification-interpretation contrastive detection 8

4 First row shows one benign image and its three adversarial examples by FGSM, JSMA, and CW together with their classification results. The second row presents their corresponding attention masks. Visually, the masks look really similar to one another despite adversarial perturbations. 9

5 (a) and (b) show two examples where LAN produces attention masks with a totally different class from their corresponding original images. 11

6 Robustness of the *plain* model and the *certified* model, measured through the attack success rate. 15

Abstract

In this paper, we present two different novel approaches to defend against adversarial examples in neural networks: attention-based against pixel-based attack and certificate-based against spatially transformed attack. We discuss the vulnerability of neural networks for adversarial examples, which significantly hinders their application in security-critical domains. We detail several popular pixel-based methods of attacking a model. We then walk through current defense methods and note that they can often be circumvented by adaptive adversaries. For the first contribution, we take a completely different route by leveraging the definition of adversarial inputs: while deceiving for deep neural networks, they are barely discernible for human visions. Building upon recent advances in interpretable models, we construct a new detection framework that contrasts an input's interpretation against its classification. We validate the efficacy of this framework through extensive experiments using benchmark datasets and attacks. We believe that this work opens a new direction for designing adversarial input detection methods. As for the second contribution, we discuss a completely different approach to generate adversarial examples, based on the spatial transformation of an input image. We then extend a currently proposed certificate framework to this setting and show that the certificate can improve the resilience of a network against adversarial spatial transformation.

1 Introduction

Recent advances in deep learning have led to breakthroughs in long-standing artificial intelligence tasks, e.g., image classification, speech recognition, and game playing, and enabled use cases previously considered strictly experimental. Yet, deep neural networks (DNNs) are inherently vulnerable to adversarial inputs [1], those maliciously crafted samples to trigger DNNs to misbehave, which significantly hinders DNN's application in security-critical domains, such as autonomous driving or facial recognition. One example of adversarial examples is shown in Fig. 2. To humans, the two pictures look exactly the same - they are pictures of a horse. However, to a well-trained neural network, they could be classified as horse and dog, respectively.

Since the discovery of such vulnerabilities [1], a variety of attack models have been proposed [2], [3], [1]. All these methods share the common trait in that they all add adversarial perturbation directly to the pixels of the original input, so that the distance between the original image and the adversarial one remains relatively small while the classification is altered. With such a common property, the various approaches differ in the way to measure magnitude of perturbations. As it is still not clear how humans perceive the differences in images, a technical method is often utilized to carry out the measuring, specifically, the L_p norm. Generally, there are three popular L_p norm that are used: L_∞, L_2 and L_0.

For example, Jacobian Saliency Map Attack [2] iteratively picks pixels (L_0) and perturbs them according to their effect on achieving misclassification; L-BFGS [1], DeepFool [4], Universal [5], C&W [3] attacks are all using Euclidean (root-mean-square) method (L_2) to measure the influence of perturbations; while Fast Gradient Sign Method (FGSM) [1] change every pixel of the original image simultaneously (L_∞). More detail on each category of attack is deferred to the next section.

On the other hand, a plethora of defense mechanisms has been proposed. The existing methods can be roughly categorized into two classes: one that reduces the influence of distortion on the model's inputs [6], [7], [8], [9] and the other utilizes the model's outputs to help make more robust classification [10], [11]. Yet, relying on carefully engineered patterns to distinguish genuine and adversarial inputs, most of the defenses can often be circumvented by adaptive adversaries or new attack variants [12].

In this paper, we propose a new detection framework that completely departs from existing

Figure 1: Example of an adversarial example: both pictures look like horses but the one on the right can trick neural networks into classifying it as a dog.

efforts. Intuitively, we revisit the fundamental definition of adversarial inputs, which are examples that can deceive neural networks but not humans, because humans have the ability to extract the main information from an image and ignore adversarial perturbations. We try to mimic this ability by leveraging attention mechanisms to generate representative patterns for each class of the input images. Specifically, our work is inspired by Latent Attention Network (LAN) [13] which, for each input example, generates a mask, called "attention mask", that represents the most important pixels of that image when a network tries to classify it. Figure 2 shows images and their corresponding attention masks. The idea of our detection method is as follows: the attention mask of an adversarial image remains similar to that of its corresponding benign even when the image can fool the target classifier. As far as we know, we're the first to apply an attention mechanism to the adversarial examples detection. We test our method against state-of-the-art attacks and we show promising initial results following this idea. We also look into cases where our approach fails and then point out potential directions for future research.

Furthermore, a different approach to countering adversarial examples is to certify a model against adversarial perturbations, or to find a tight upper-bound against all possible perturbations. In [14], the authors train a network and a certificate that guarantee an upper bound of no more than 35% test error, where each pixel is perturbed by at most 0.1. Even though the author only applies the method to a rather simple 2-layer network and the dataset MNIST [15], this opens

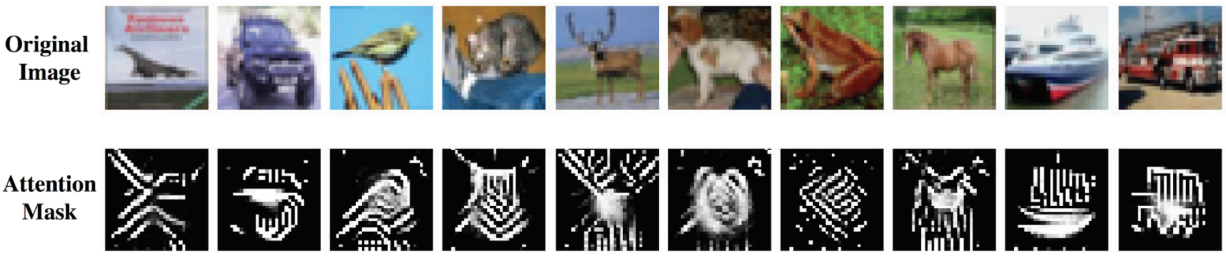

Figure 2: Original images from each class and their masks generated by LAN.

a new direction to guarantee against adversarial examples: provable and certified defense. The second contribution of this paper draws inspiration from [14] and extends their results to spatially transformed adversarial examples.

In summary, the contributions of this paper are as follow:

- We propose an attention-based defend method against pixel-based attacks. We test the framework with extensive experiments and show promising results.
- We propose a certified and provable defense against spatially transformed adversarial examples. We show that the certified model is significantly more robust than a vanilla model.

2 Methods of attacking

As mentioned in the previous section, pixel-based attacking methods can be broadly categorized based on the L_p norm used: L_∞, L_2, L_0. We now detail each category using its representative method.

2.1 F_∞: Fast Gradient Sign Method (FSGM)

FGSM was first proposed in [1], which proposes a theory for why neural networks are susceptible to adversarial perturbations. [1] suggests that it is the linearity of models that is responsible for adversarial examples, and proposes a simple method to general adversarial perturbations based on that theory.

$$\boldsymbol{\eta} = \epsilon sign(\nabla_x J(\boldsymbol{\theta}, \boldsymbol{x}, y)) \tag{1}$$

Eq (1) shows their proposed formula, where $\boldsymbol{x}$ is the input, y the ground true label, J the cost function used to train the model, $\boldsymbol{\theta}$ the parameters of the model and ϵ is a small 'step-size' for the perturbation. The idea here is to linearize the cost function around the current value of , and 'step' in the direction obtained from gradients of the lost function to achieve an optimal L_∞ perturbation that can trick the model. This method, though simple to calculate using backpropagation, can generate reliable adversarial examples over a wide range of models.

2.2 F_2: Carlini & Wagner (C&W)

CW attack [3] is considered to be one of the most powerful attacks so far. While it is based on L_2 norm perturbation, the authors also extended it to L_∞ and L_0. Here, we only consider the L_2 variant of CW as it is considered to have better performance. [3] re-frames the problem of generating adversarial examples into an optimization problem:

$$\begin{aligned} \min \quad & D(x, x+\delta) \\ \text{s.t.} \quad & C(x+\delta) \quad t \end{aligned} \tag{2}$$

where, D is the distance between the original and perturbed image, δ the adversarial perturbation, C the model and t the target. However, this reframed version is a highly non-linear constraint and hard to optimize with, so [3] proposes to optimize this a relaxed version instead:

$$\begin{aligned} \min \quad & D(x, x+\delta) + c \cdot f(x+\delta) \\ \text{where} \quad & f_{adv} \;=\; \max(\max_{i \neq t} Z(x_{adv})_i - Z(x_{adv})_t, k) \end{aligned} \tag{3}$$

Here, f is the relaxed version of the hard constraint $C(x+\delta) = t$ above. f is considered as *adversarial loss*, which encourages the target class to have the largest logit, and penalizes the objective otherwise.

2.3 Spatially Transformed Adversarial Examples

As opposed to pixel-based perturbations, [16] proposes to use spatial transformation to generate adversarial samples. The framework can be described as follow. Let x and $\hat{x}$ be the benign and adversarial input correspondingly. $\hat{x}_i$ denotes the value of $\hat{x}$'s i-th pixel and $(\hat{u}_i, \hat{v}_i)$ denote its position in $\hat{x}$. x is transformed into $\hat{x}$ using a per-pixel flow field $\boldsymbol{r}$ to generate $\hat{x}$ with differentiable bilinear interpolating from x's pixels. Specifically, $\hat{x}_i$ is calculated as:

$$\hat{x}_i = \sum_{j \in N(u_i, v_i)} x_j \cdot (1 - |u_i - u_j|) \cdot (1 - |v_i - v_j|) \tag{4}$$

where $N(u_i, v_i)$ are the indices of the 4-pixel neighbors at position (u_i, u_j).

Similar to how C&W constructs their L_2 attack, there are also two components in the objective function for the spatial transform attack: L_{adv} (5) to encourage adversarial behavior and L_{flow} (6) to limit the amount of transformation.

$$L_{adv} = \max(\max_{i \neq t} Z(x_{adv})_i - Z(x_{adv})_t, k) \tag{5}$$

$$L_{flow}(\boldsymbol{r}) = \sum_i \sum_{j \in N(u_i, v_i)} \sqrt{||\Delta u_i - \Delta u_j||_2^2 + ||\Delta v_i - \Delta v_j||_2^2} \tag{6}$$

Here, L_{flow} is the total variation suggested in [17]. The weighted objective is then:

$$L_{total}(\boldsymbol{r}) = L_{adv} + \tau \cdot L_{flow} \tag{7}$$

where τ is often chosen to be 0.1.

This work leads to an interesting direction to generate adversarial examples that are based on spatial transformation instead of direct pixel perturbations. With this new direction comes with the need to defend against this line of attack and in this paper, we propose a provable defense

mechanism against spatially transformed adversarial examples.

3 Attention-based defense against pixel-based attacks

In this section, we explore the use of attention mechanisms to defend against pixel-based adversarial examples. Specifically, we choose to leverage Latent Attention Network (LAN) [13] as the attention mechanism because it assumes minimal knowledge about the network. In other words, LAN can visualize the attention of a black-box model.

3.1 Latent Attention Network

Intuitively, LAN seeks to find the input pixels of an image x that are critical to the output of a model F. LAN does this by randomly corrupting components of x with noise while measuring the changes in $F(x)$. Those components that when corrupted lead to minimal changes in $F(x)$ are not as important as those that lead to large changes. The formal framework can be summarized in Eq. (8) and (9):

$$\tilde{\boldsymbol{x}} = \boldsymbol{A}(\boldsymbol{x}) \cdot \eta + (\mathbf{1} - \boldsymbol{A}(\boldsymbol{x})) \cdot \boldsymbol{x} \tag{8}$$

$$L_{LAN}(x) = \mathbb{E}_{\eta \sim H}[L_F(F(\tilde{x}), F(x)) - \beta \cdot \overline{A(x)}] \tag{9}$$

where, $A(x)$ is the corrupted mask generator, η the degree to which to corrupt input x, L_F the loss that was used to train the model, L_{LAN} the final loss to minimize in order to train the Latent Attention Network.

3.2 Methodology

Fig. 3 shows our detection framework based on LAN. It is characterized by three modules: an image classifier f_1, an attention model g, and a mask classifier f_2. The image classifier f_1 is a function $F_1 : \mathbb{R}^d \to [0,1]^d$, which given input x, classifies it into decision y_1. Attention model g is a LAN, which is a function $G : \mathbb{R}^d \to [0,1]^d$, that given an input x, produces an attention

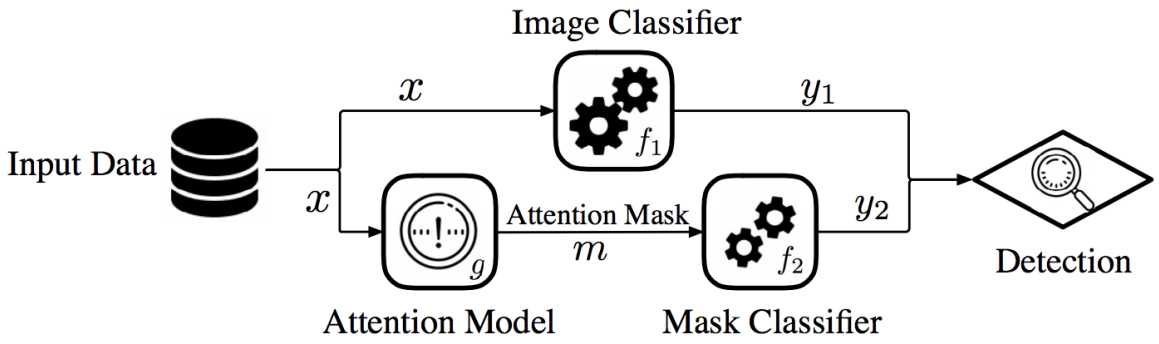

Figure 3: Framework of classification-interpretation contrastive detection

mask $m = G(x)$. Attention mask m determines the important components of x that influence the classification output of a classifier f by corrupting pixels of x with noise drawn from a predefined distribution and measures the change in f's loss. The larger the loss is, the more important the pixels are. The resulting masks can capture the common features in images of the same class. f can be any common classifier for x. In our experiments, we directly use f_1. The mask classifier f_2 is a function $F_1 : \mathbb{R}^d \to [0,1]^l$ which, given a mask m, classifies it into decision y_2. If y_2 agrees with y_2, we decide the image is benign and adversarial otherwise.

3.3 Evaluation

We experiment on CIFAR10 [18], with 50,000/10,000 train/test split. We use the same architecture in [18] for f_1 and g, which are AlexNet [19] and a 3-layer Fully Connected Network. f_2 is based on LeNet [15] and trained on the masks produced by g from the training set. We test our method against three attacks corresponding to different distance metrics: L_∞, L_0, L_2, namely FGSM, JSMA and CW. In the rest of the paper, adversarial examples are treated as the positive class and we use x, m, and x^* , m^* to denote benign images, attention masks, and their adversarial counterparts, respectively.

Invariance of mask

We first evaluate the applicability of attention masks to detecting adversarial samples. We generate attention masks for benign and malicious images using a LAN. Our intuition that attention masks

Attack	Percentage
FGSM (L_∞)	0.863
JSMA (L_0)	0.878
C&W (L_2)	0.997

Table 1: Percentage of adversarial samples whose attention masks retain their original classifications.

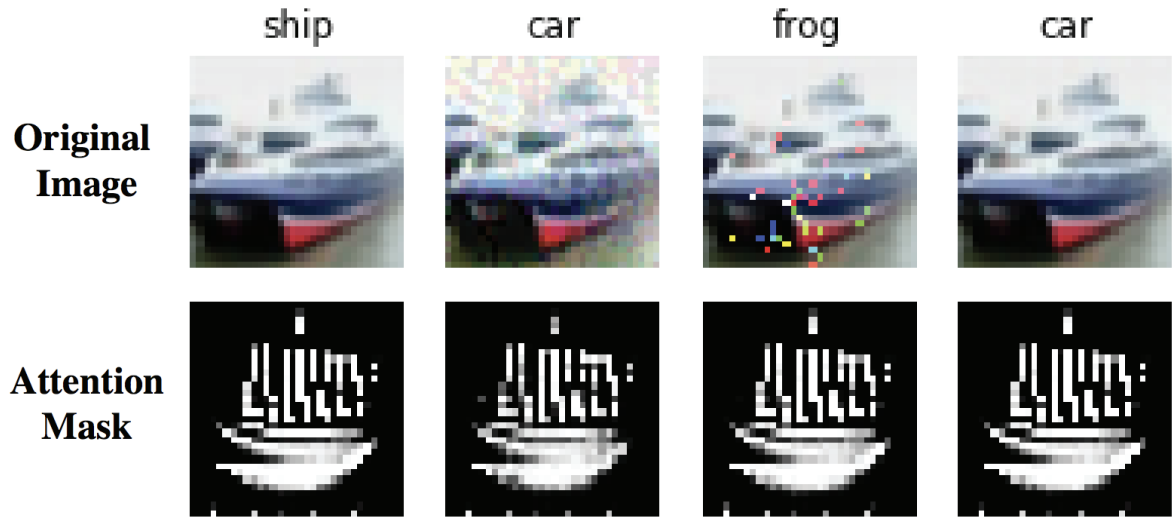

Figure 4: First row shows one benign image and its three adversarial examples by FGSM, JSMA, and CW together with their classification results. The second row presents their corresponding attention masks. Visually, the masks look really similar to one another despite adversarial perturbations.

of both types of images are very similar is confirmed and the results are shown in Fig. 4. We also give a statistics analysis on how the mask classification results will change after perturbations. As shown in Table 1, over 85% of the adversarial samples fail to change their mask classifications, no matter how they are generated. This discovery and the quality of LANs to generate similar masks for images from the same class motivate our detection framework.

Detection effectiveness

In the second experiment, we evaluate the effectiveness of the detection framework by comparing the prediction for image x from classifier f_1 against the classification of the mask m of image x. If they differ, we predict adversarial, and benign otherwise. For each attack method, we take the

Attack	True positive	True negative
FGSM (L_∞)	0.878	0.614
JSMA (L_0)	0.960	0.614
C&W (L_2)	0.860	0.614

Table 2: Classification accuracy on adversarial samples generated using different attacks on AlexNet.

Attack	True positive	True negative
FGSM (L_∞)	0.843	0.665
JSMA (L_0)	0.853	0.647
C&W (L_2)	0.917	0.750

Table 3: Classification accuracy on adversarial samples generated with different attacks on VGG-like.

adversarial examples x^* that successfully fool f_1. We pair the same amount of benign images with adversarial examples to create a test set. The results of our solution against FGSM, JSMA and C&W are shown in Table 2. Adversarial examples are known to be able to transfer across models, so we also test our detection framework in a transferred setting. Table 3 shows the results of our method against adversarial samples generated for VGG-like, a modified VGG network [20] that has better accuracy than our AlexNet model. Overall, our method shows good true positive rates: being able to detect adversary when the input image is indeed adversarial, in both direct and transferred setting. However, the true negative rates across different attacks are low, mostly under 70%: the mask classifier f_2 is confused when presented with a benign example. To figure out why, we calculate the accuracy of f_2 on the produced masks of the test set and we get 60%, which is quite low, compared to the accuracy of 82% of f_1, which classifies raw input images instead of the derived attention masks. Looking into the 40% of the masks that got incorrectly classified by f_2, we find the problem: those masks are mostly not of the common form of the masks of their corresponding classes. Figure 5 shows 2 examples.

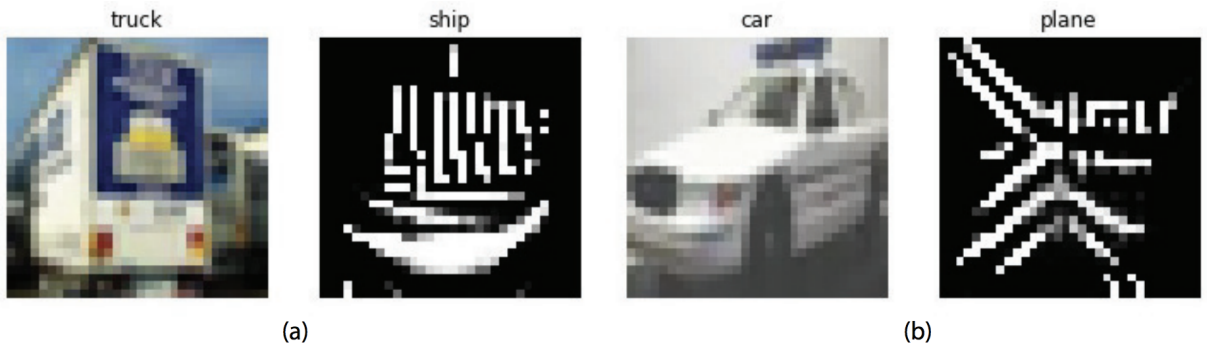

Figure 5: (a) and (b) show two examples where LAN produces attention masks with a totally different class from their corresponding original images.

Attack	True positive	True negative
FGSM (L_∞)	1.000	1.000
JSMA (L_0)	1.000	1.000
C&W (L_2)	1.000	1.000

Table 4: Classification accuracy on adversarial samples generated using different attacks on AlexNet, after filtering out the incorrect masks.

Reliability of detection

We then study the quality of the mask generator network and see how it affects our detection method's performance. We filter out the masks that are incorrectly produced by LAN and are left with the 60% of the test set, and generate adversarial samples for those cases. We repeat our experiments with this smaller data set and get the following results in Table [???]. We achieve the exactly same results for AlexNet and VGG-like. Overall, when the masks for benign images are correctly produced by LAN, our detection accuracy is perfect across all the attack methods, in both direct and transferred settings. What's more, the recovery rates (retrieving the original classification of the network despite adversarial perturbations) are also 100%.

4 Certified defense against spatial transformation attack

Besides attention-based defense, we have also explored certified methods that are provable. Drawing inspiration from [14] that proposes a certified framework against pixel-based attacks, we develop a certificate against spatially transformed adversarial examples. We carry out extensive experi-

ments and show that a two-layer neural network, when trained with the certificate, becomes more robust.

4.1 Methodology

For a two-layer neural network, we have $f(\hat{x}) = v^T\sigma(W \cdot \hat{x})$, where $v = v_+ - v_-$ represents the weight differences for the positive and negative classes at the second layer, W the weight matrix of the first layer and σ the activation function, often non-linear. Furthermore, as shown in [21], we can bound the outputs of a two-layer neural network from an adversarial spatial transformation input like so:

$$f(\hat{x}) \leq f(x) + \max_{||z||_\infty \leq 1} \frac{\epsilon}{2} \cdot z^T P z \tag{10}$$

where, P is parameterized as:

$$P \stackrel{\text{def}}{=} \begin{bmatrix} 0 & diag(v)Wdiag(\lambda_u) & diag(v)Wdiag(\lambda_v) \\ diag(\lambda_u)W^T diag(v) & 0 & 0 \\ diag(\lambda_v)W^T diag(v) & 0 & 0 \end{bmatrix} \tag{11}$$

and λ is from Proposition 2.2 in [21], which is: *the boundary of the ellipsoid defined by* $||N \cdot vec(r)||_2 = \epsilon$ *is given by* $[-\epsilon\lambda_i, \epsilon\lambda_i](1 \leq i \leq 2n)$, *where* λ_i^2 *is the i-th diagonal element of the matrix* $(N^T N)^-1$. Here, r is the flow vector defined by spatial transformation attack.

Then, use the fact that $z^T P z = tr(zz^T P)$, where tr is the trace operator, we have:

$$\max_{||z||_\infty \leq 1} z^T P z = \max_{Z = zz^T, ||z||_\infty \leq 1} tr(ZP) \tag{12}$$

By relaxing $Z = zz^T$ and $||z||_\infty \leq 1$ with $Z \geq 0$ and $diag(Z) \leq 1$, we have the following convex

SDP problem:

$$\begin{aligned} \max \quad & tr(ZP) \\ \text{s.t.} \quad & diag(Z) \leq 1 \\ \text{and} \quad & Z \geq 0 \end{aligned} \tag{13}$$

This can be efficiently solved using off-the-shelf SDP optimizers. Solving this SDP problem will yield the certificate, which can be jointly trained with our a model to make it more robust.

4.2 Evaluation

We use 2-layer network that was trained to achieve an accuracy of 98% on the full MNIST test set. We call this the *plain* model. Using this model, we compute a certificate using the procedure above and train it with the *plain* model to certify it against spatially transformed adversarial examples, resulting in a *certified* model. We then attack both models using the same sample of 1000 images randomly drawn from the MNIST test set. The corresponding set of attack targets is chosen randomly and different from the respective ground true labels. When attacking the models, we found that the τ variable, which controls the weight between adversarial loss and total variation loss of the spatially transformed framework, has an significant impact on the success of an attack. When τ is set too low, the attack can easily fool the model but the resulted image will be heavily transformed, leading to high level of distortion. On the other hand, when τ is set too high, the optimizer will try to minimize the transformation as much as possible and the models will not be tricked, leading to low level of success rate. Thus, to evaluate fairly, we developed an binary search scheme for , similar to how C&W [3] controls their c variable. We set the lower and upper bound for τ to be 0.0 and 1000, respectively. If the attack is not successful, meaning the model is not tricked into classifying the input image as the target class, τ is lowered by setting

the upper bound to be the midpoint between lower and upper bound. Conversely, if the attack is successful, we want to get a less transformed image and thus increase by setting the lower bound to be the midpoint. If the attack cannot produce a successful adversarial image after a certain max number of iteration, we pronounce it unsuccessful. This scheme increases the average attack time for each input, but it ensures that we attack both models thoroughly. What's more, similar to how [14] bounds the perturbation for each pixel to be 0.1 the largest, we need to bound the extent of spatial transformation in our attack for the certificate to be meaningful. After many experiments, we decide to bound the total variation (which is the L_{flow} loss) to be smaller than or equal to 7.0. We carry out this constraint in our attack by penalizing any total variation that is larger than 7.0:

$$L_{flow} = \max(L_{flow} - \epsilon, 0) \tag{14}$$

where ϵ is the constraint for the L_{flow} value, and specifically 7.0 for our experiments. All in all, we found our *certified* model to be highly effective as suggested in Fig. 6. The *certified* model is much less prone to spatial transformation attack, as evidenced in the significantly lower success rate of the attacks, which is about 70 successful adversarial examples out of 1000 images. On the other hand, the *plain* model is much more susceptible to spatial attacks, as there are around 380 successful adversarial examples out of 1000 images attacked.

5 Conclusion and future direction

We propose two novel methods to defend against adversarial examples: one attention-based against pixel-based attacks and another that is provable and comes with a certificate to defend against spatially transformed adversarial examples. We carry out extensive experiments to show the methods' performance and efficiency.

The attention-based framework is an initial step to utilize a model's interpretability. Our method uses an attention mask generator, specifically a Latent Attention Network, to find an

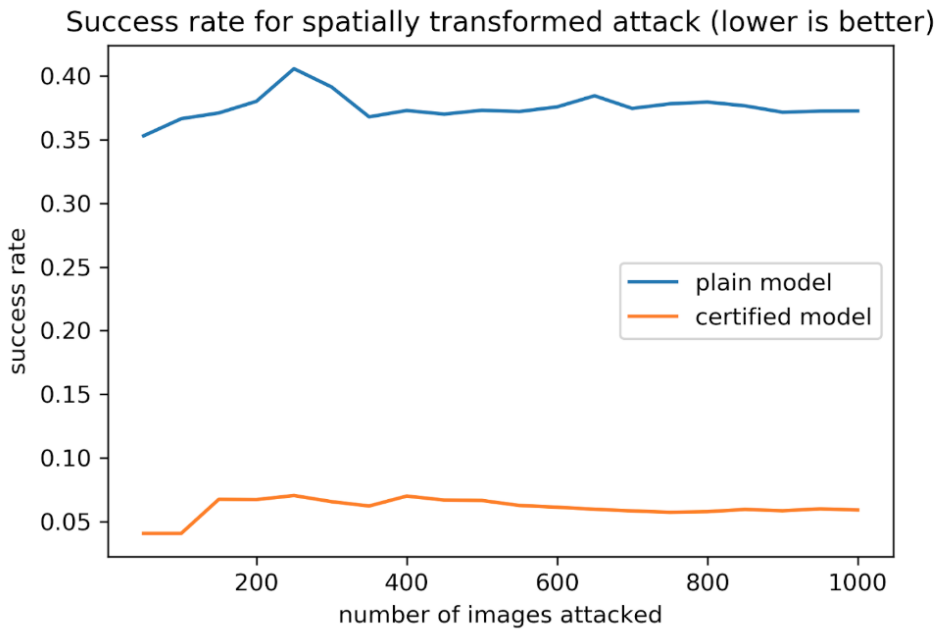

Figure 6: Robustness of the *plain* model and the *certified* model, measured through the attack success rate.

input image representation that is invariant regardless of adversarial modifications. We show that attention masks are resilient against adversarial perturbations and build our adversary detection based on that property. Our initial experiments provide promising results with a good detection performance. The framework's perfect detection accuracy and recovery rates, after filtering out benign images with incorrect masks, hint at a potential increase in detection accuracy if we can optimize the quality of the attention mask generator. The proposed method is also attack-agnostic in that it does not need to know the specifics in adversarial samples generation process. However, our detection method's performance is highly dependent on the reliability of LAN and further experimentation and ideas might be required to see if its quality can be improved. We hope that this new direction would motivate further research in using attention-based mechanisms to effectively defend against adversarial examples. One possible idea is to improve the mask generator. Another is to use a different attention method: we build our work on Latent Attention Network but there might be other interpretability mechanisms that are better for adversary detection. We look forward to seeing more robust DNNs with the benefits of interpretability.

What's more, the certified approach, where we solve for a certificate and jointly train it with

the model in question, has shown good performance in defending against spatially transformed adversarial examples. One particularly interesting direction for future work is a unified framework that has the benefits of both pixel-based and spatial transformation attack, which might make it harder to defend against. Similarly but on the flip side, a unified framework for certifying models can also be developed.

Bibliography

[1] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," Dec. 2013.

[2] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami, "The limitations of deep learning in adversarial settings," in *Security and Privacy (EuroS&P), 2016 IEEE European Symposium on*, pp. 372–387, IEEE, 2016.

[3] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," *arXiv preprint arXiv:1608.04644*, 2016.

[4] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "Deepfool: a simple and accurate method to fool deep neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2574–2582, 2016.

[5] S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard, "Universal adversarial perturbations," *arXiv preprint*, 2017.

[6] K. Grosse, P. Manoharan, N. Papernot, M. Backes, and P. McDaniel, "On the (statistical) detection of adversarial examples," Feb. 2017.

[7] X. Li and F. Li, "Adversarial examples detection in deep networks with convolutional filter statistics," in *ICCV*, pp. 5775–5783, 2017.

[8] R. Feinman, R. R. Curtin, S. Shintre, and A. B. Gardner, "Detecting adversarial samples from artifacts," Mar. 2017.

[9] D. Meng and H. Chen, "MagNet: A Two-Pronged defense against adversarial examples," in *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, CCS '17, (New York, NY, USA), pp. 135–147, ACM, 2017.

[10] Z. Gong, W. Wang, and W.-S. Ku, "Adversarial and clean data are not twins," Apr. 2017.

[11] J. H. Metzen, T. Genewein, V. Fischer, and B. Bischoff, "On detecting adversarial perturbations," Feb. 2017.

[12] N. Carlini and D. Wagner, "Adversarial examples are not easily detected: Bypassing ten detection methods," in *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, AISec '17, (New York, NY, USA), pp. 3–14, ACM, 2017.

[13] C. Grimm, D. Arumugam, S. Karamcheti, D. Abel, L. L. S. Wong, and M. L. Littman, "Modeling latent attention within neural networks," June 2017.

[14] A. Raghunathan, J. Steinhardt, and P. Liang, "Certified defenses against adversarial examples," Jan. 2018.

[15] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, pp. 2278–2324, Nov. 1998.

[16] C. Xiao, J.-Y. Zhu, B. Li, W. He, M. Liu, and D. Song, "Spatially transformed adversarial examples," Jan. 2018.

[17] L. I. Rudin, S. Osher, and E. Fatemi, "Nonlinear total variation based noise removal algorithms," *Physica D: nonlinear phenomena*, vol. 60, no. 1-4, pp. 259–268, 1992.

[18] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," 2009.

[19] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, pp. 1097–1105, 2012.

[20] DawngoMediaVillage, "Keras compressor." https://github.com/DwangoMediaVillage/keras_compressor, 2017.

[21] X. Zhang, C. Nguyen, and T. Wang, "Certified defenses against spatial transformation attacks," *arXiv preprint*, 2018.