Poster: Enhancing Adversarial Example Defenses Using Internal Layers

Mainuddin Ahmad Jonas and David Evans
Department of Computer Science, University of Virginia
[*maj2bh, evans*]*@virginia.edu*

***Abstract*—Until recently, research on adversarial attacks on image classification models has focused on the input and the output layers of the models. This ignores potentially useful information from the internal layers of the models. In this work, we investigate the internal layers of image classification models in the context of adversarial attacks. We observe significant differences between outputs of normal and adversarial inputs at those internal layers. As a case study, we then use this property to propose an improvement to adversarial detection based on feature squeezing. Our proposed technique improves the detection rate significantly over the previous method that only considered the final layer, and suggest further opportunities for exploring internal layers.**

1. Introduction

There is increasing awareness that machine learning models, and Deep Neural Networks (DNNs) in particular, are not robust against adversaries and that carefully-crafted imperceptible perturbations can cause large changes in the model output [1], [2]. These types of attacks on the models are known as *adversarial examples*.

Most research so far on defending against adversarial examples for image classification models has focused either on pre-processing the inputs [3], [4], [5] or retraining the models using adversarial samples [6], [7]. Work on detection has focused on the input and output layers of the underlying DNN models, ignoring a huge amount of possibly useful information from the internal layers of the network. Recently a few works have incorporated internal layers [8], [9] with promising results. Our work also aims to harness internal layer information to gain insights into the nature of adversarial examples, and to use that insight to improve defenses.

In this paper, we provide the motivation behind our work by showing properties of the internal DNN layers that may be useful in distinguishing adversarial examples. We use this property to improve a defense based on feature squeezing [3]. Feature squeezing [3] is a framework designed to detect adversarial examples by reducing the often unnecessarily large feature spaces of DNN models. An input is squeezed using a pre-processor. Then, the original and (possibly many) squeezed inputs are fed into the model. If the model's predictions on the original and squeezed inputs exceeds a distance threshold the input is determined to be adversarial. Simple squeezing methods including bit depth reduction and local and non-local means smoothing have been found to be quite effective against some adversarial attacks against MNIST, CIFAR-10 and ImageNet models, but ineffective against other attacks including FGSM [3].

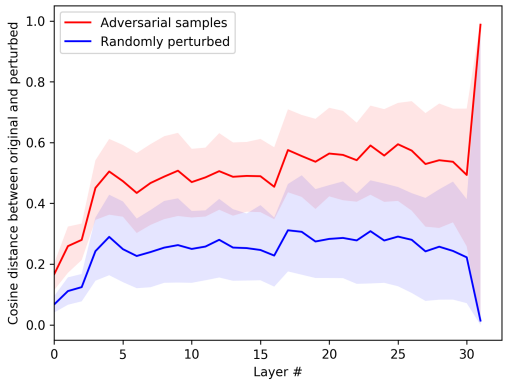

Figure 1. Cosine distance between normal-adversarial and normal-random pairs through the activation layers of CIFAR-10 DenseNet network. Solid lines indicate median, shaded region has values between 5th and 95th percentiles.

Whereas the original feature squeezing framework detects adversarial inputs by comparing the difference between the softmax layer outputs of the original model on the input and one a pre-processed version of the input, we use the difference between the outputs at all the internal layers of the model – not just the final softmax layer. By doing so, we were able to significantly improve the detection rate of FGSM adversarial attacks against the CIFAR-10 model, while preserving the same false positive rate.

2. Model Divergence

By definition, the distance between normal examples and adversarial examples is limited by an upper bound at the input layer. But, to be a successful adversarial example, by the final layer the outputs must be different enough for them to be assigned to different classes by the model. Hence, the distance between an adversarial example and its seed must be small at the input layer, and increase at some layers inside the network. Moreover, if we cause random perturbation to a seed, the distance between the randomly perturbed input and the original seed should be lower than that between the normal and adversarial pairs.

This understanding is supported by Figure 1 which compares the cosine distances between normal and random perturbations (blue) and normal and adversarial (red) perturbations (generated by an untargeted FGSM L_∞ attack) on the CIFAR-10 DenseNet model. For 10,000 test seeds, we generated one adversarial example each using the FGSM attack, and also one randomly perturbed example on the same seed. We plot the cosine distance between each normal-adversarial and normal-random pair for each activation layer

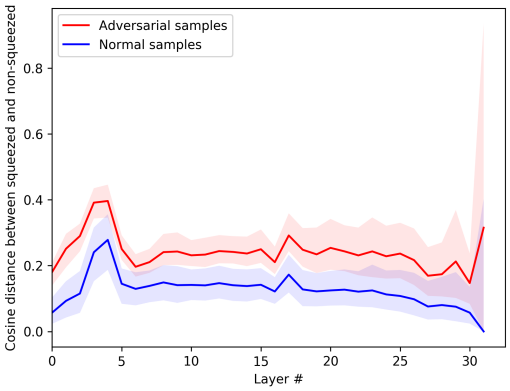

Figure 2. Cosine distance between squeezed and non-squeezed inputs through the activation layers of the CIFAR-10 DenseNet network. Solid lines indicate median, shaded region has values between 5th and 95th percentiles.

of the network. As expected, the distance increases as we go through the network, and the distance is higher for normal-adversarial pairs than normal-random pairs.

3. Enhancing Feature Squeezing

The observation from the previous section suggests the internal layers could be useful for detecting adversarial inputs, but we cannot directly make use of the observed property for detection at test time because the deployed model does not have access to the adversary's original seed. However, we show how to harness this property in the existing feature squeezing framework [3] to improve the detection rate. Instead of comparing between the sample and its (unknown) seed, we compare between the sample and its squeezed sample. Figure 2 shows the average cosine distance at each activation layer between squeezed and non-squeezed inputs for both normal and adversarial samples. There is clear separation between the the normal and adversarial samples through most layers of the network. This is the property we exploit to enhance the detector.

To detect adversarial samples using the feature squeezing technique, an input and its squeezed representation are both passed through the model. As implemented by Xu et al. [3], the outputs were compared by using the L_1 distance at the final softmax layer. For normal examples, the value of this score should be low (close to 0), and for adversarial examples it should be high (close to 2). A detection threshold was selected based on a target false positive rate of 5%.

Based on our observations about internal distance gaps, we define a modified score metric that uses the sum of the cosine distances between internal layer outputs of the non-squeezed and squeezed inputs:

$$score^{(x,x_{\text{squeezed}})} = \frac{1}{n}\sum_{l=1}^{n} d_c(g_l(x), g_l(x_{\text{squeezed}}))$$

Here, $d_c(x, y)$ is the cosine distance between the vectors x and y; $g_l(x)$ is the output of the l^{th} layer of the DNN for input x; and n is the total number of activation layers in the network. Similar to the original feature squeezing framework, we choose a threshold score for detection in a way that ensures false positive rate is below 5%, and use that threshold to determine whether a given input is normal or adversarial.

Results. In the original feature squeezing framework, the accuracy of the feature squeezing-based detectors was quite low for FGSM L_∞ attacks against the CIFAR-10 classification model. Thus, we wanted to see if we could improve the detection rate using our modified technique. We generated 10,000 adversarial examples for the 10,000 test samples of the CIFAR-10 dataset using the FGSM non-targeted attack with $\epsilon = 0.1$. We found that the best feature squeezer, 2×2 median smoothing, achieves a detection rate of just **12%**, when the false positive rate is kept below 5%. On the other hand, our technique achieves detection rate of **98%** on adversarial samples, while maintaining the same lower than 5% false positive rate.

4. Conclusions

Our preliminary results are encouraging enough to suggest the hidden layers of DNN can be useful for enhancing the effectiveness of existing defenses against adversarial attacks, but more extensive experiments are needed to evaluate the effectiveness of the defense. In particular, it is important to understand how such a defense works against adaptive adversaries and we are currently working on this. Multi-layer defenses seem to make the task of an adaptive adversary more challenging because the adversary's input needs to produce model divergence in the early layers. We are optimistic that studying the internal behavior of models will provide useful insights for mitigating adversarial examples.

References

[1] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," in *International Conference on Learning Representations*, 2014.

[2] I. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *International Conference on Learning Representations*, 2015.

[3] W. Xu, D. Evans, and Y. Qi, "Feature squeezing: Detecting adversarial examples in deep neural networks," in *The Network and Distributed System Security Symposium (NDSS)*, 2018.

[4] D. Meng and H. Chen, "Magnet: a two-pronged defense against adversarial examples," in *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security*. ACM, 2017.

[5] J. H. Metzen, T. Genewein, V. Fischer, and B. Bischoff, "On detecting adversarial perturbations," in *International Conference on Learning Representations*, 2017.

[6] A. Kurakin, I. J. Goodfellow, and S. Bengio, "Adversarial machine learning at scale," in *International Conference on Learning Representations*, 2017.

[7] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *International Conference on Learning Representations*, 2018.

[8] X. Ma, B. Li, Y. Wang, S. M. Erfani, S. Wijewickrema, G. Schoenebeck, M. E. Houle, D. Song, and J. Bailey, "Characterizing adversarial subspaces using local intrinsic dimensionality," in *International Conference on Learning Representations*, 2018.

[9] N. Papernot and P. McDaniel, "Deep k-nearest neighbors: Towards confident, interpretable and robust deep learning," *arXiv preprint arXiv:1803.04765*, 2018.