Safe Machine Learning and Defeating Adversarial Attacks

Bita Darvish Rouhani, Mohammad Samragh, Tara Javidi, and Farinaz Koushanfar | University of California San Diego

Adversarial attacks have exposed the unreliability of machine learning models for decision making in autonomous agents. This article discusses recent research for ML model assurance in face of adversarial attacks.

The fourth industrial revolution shaped by Machine Learning (ML) algorithms is underway. ML algorithms have provided a paradigm shift in devising automated systems that can even surpass human performance in controlled environments. While advanced learning technologies are essential for enabling interaction among autonomous agents and the environment, a characterization of their quality or careful analysis of the system reliability in the presence of malicious entities are still in their infancy.

Reliability and safety consideration is the biggest obstacle to the wide-scale adoption of emerging learning algorithms in sensitive scenarios such as intelligent transportation, health-care, warfare, and financial systems. Although ML models deliver high accuracies in conventional settings with limited simulated input samples, recent research in adversarial ML has shed light on the unreliability of their decisions in real-world scenarios. For instance, consider a traffic sign classifier used in self-driving cars. Figure 1 shows an example adversarial sample where the attacker carefully adds imperceptible perturbation to the input image to mislead the employed ML model, and thus, jeopardizes the safety of the vehicle.

In light of the adversarial attacks, to pervasively employ autonomous ML agents in sensitive tasks it is imperative to answer the following two questions:

- What are the vulnerabilities of machine learning models that attackers can leverage for crafting adversarial samples?
- How can we characterize and thwart the adversarial space for effective ML model assurance and defense against adversaries?

In this article, we discuss our recent research results for adaptive ML model assurance in face of adversarial attacks. In particular, we introduce, implement, and automate a novel countermeasure called **Modular Robust Redundancy** (**MRR**) to thwart the potential adversarial space and significantly improve the reliability of a victim ML model.[1]

Unlike prior defense strategies, MRR methodology is based upon *unsupervised learning*, meaning that no particular adversarial sample is leveraged to build/train the modular redundancies. Instead, our unsupervised learning methodology leverages the structure of the built model and characterizes the distribution of the high dimensional space in the training data. Adopting an unsupervised learning approach, in turn, ensures that the proposed detection scheme can be generalized to a wide class of adversarial attacks. We corroborate the effectiveness of our method against the existing state-of-the-art adversarial attacks. In particular, we open-source our API to ensure ease of use by data scientists and engineers and invite the community to attempt attacks against our provided benchmarks in form of a challenge. Our API is available at `https://github.com/Bitadr/CuRTAIL`

Adversarial samples have already exposed the vulnerability of ML models to malicious attacks; thereby undermining the integrity of autonomous systems built upon machine learning. Our research, in turn, empowers coherent integration of safety consideration into the design process of ML models. We believe that the reliability of ML models should be ensured in the early development stage instead of looking back with regret when the machine learning systems are compromised by adversaries.

Figure 1: The left image is a legitimate "stop" sign sample that is classified correctly by an ML model. The right image, however, is an adversarial input crafted by adding a particular perturbation that makes the same model classify it as a "yield" sign.

Adversary Models and Present Attacks

An adversarial sample refers to an input to the ML model that can deceive the model to make a wrong decision. Adversarial samples are generated by adding carefully crafted perturbations to a legitimate input. In particular, an adversarial sample should at least satisfy three conditions: (i) The ML model should perceive a correct decision on the original (legitimate) sample; for instance, in a classification task, the ML model should correctly classify the original sample. (ii) The ML system should make a wrong decision on the perturbed adversarial sample; e.g., in a classification task, the model must misclassify the adversarial sample. (iii) The perturbation added to the original sample should be imperceptible, meaning that the perturbation should not be recognizable in the human cognitive system.

Depending on the attacker's knowledge, the threat model can be categorized into three classes:

- **White-box attacks.** The attacker knows everything about the victim model including the learning algorithm, model topology, defense mechanism, and model/defender parameters.
- **Gray-box attacks.** The attacker only knows the underlying learning algorithm, model topology, and defense mechanism but has no access to the model/defender parameters.
- **Black-box attacks.** The attacker knows nothing about the pertinent machine learning algorithm, ML model, or defense mechanism. This attacker only can obtain the outputs of the victim ML model corresponding to input samples. In this setting, the adversary can perform a differential attack by observing the output changes with respect to the input variations.

Henceforth, we consider the white-box threat model as it represents the most powerful attacker that can appear in real-world settings. We evaluate our proposed countermeasure against four different classes of attacks including Fast-Gradient-Sign (FGS),[1] Jacobian Saliency Map Attack (JSMA),[2] Deepfool,[3] Basic Iterative Method (BIM),[4] and Carlini and Wagner attack (CarliniL2),[5,6] to corroborate the generalizability of our unsupervised approach. The aforementioned attacks cover a wide range of one-shot and iterative attack algorithms. The goal of each attack is to minimize the distance between the legitimate sample and the corresponding adversarial samples with a particular constraint such that the generated adversarial sample misleads the victim ML model. Please refer to the technical papers for the details of each attack algorithm. For the realization of different attack strategies,[7] we leverage the well-known adversarial attack benchmark library known as Cleverhans.

References

1. Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572, 2014.
2. Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z Berkay Celik, and Ananthram Swami. The limitations of deep learning in adversarial settings. In Proceedings of the IEEE European Symposium on Security and Privacy (SP), 2016.
3. Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016.
4. Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. International Conference on Learning Representations, 2017.
5. Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In Proceedings of the IEEE Symposium on Security and Privacy (SP), 2017.
6. Nicholas Carlini and David Wagner. Magnet and "efficient defenses against adversarial attacks" are not robust to adversarial examples. arXiv preprint arXiv:1711.08478, 2017.
7. Ian Goodfellow Reuben Feinman Fartash Faghri Alexander Matyasko Karen Hambardzumyan Yi-Lin Juang Alexey Kurakin Ryan Sheatsley Abhibhav Garg Yen-Chen Lin Nicolas Papernot, Nicholas Carlini. cleverhans v2.0.0: an adversarial machine learning library. arXiv preprint arXiv:1610.00768, 2017.

Adversarial Defenses

In response to various adversarial attacks proposed in the literature, several research attempts have been made to design ML models that are more robust in face of adversarial examples. The existing countermeasures can be classified into two distinct categories:

(i) Supervised strategies which aim to improve the generalization of the learning models by incorporating the noise-corrupted version of inputs as training samples and/or injecting adversarial examples generated by different attacks into the DL training phase.[2,3,4,5] The proposed defense methods in this category are particularly tailored for specific perturbation patterns and can

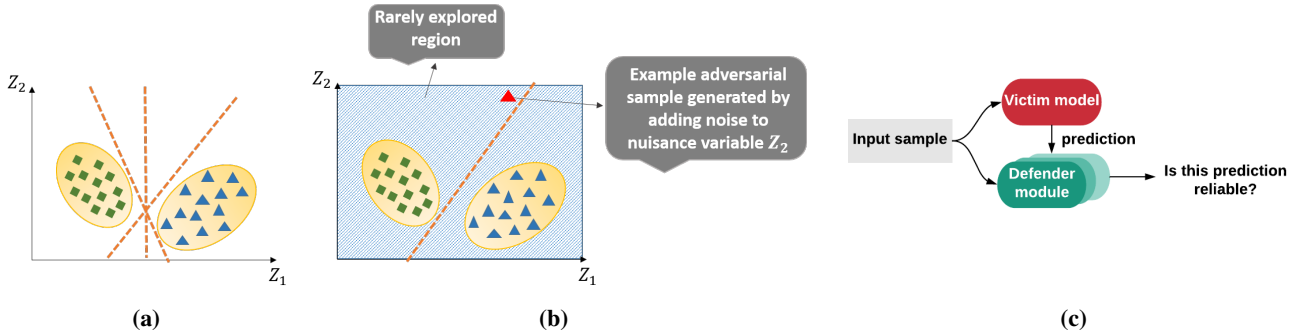


Figure 2: (a) In this example, data points (denoted by green squares and blue triangles) can be easily separated in one-dimensional space. Having extra dimensions adds ambiguity in choosing the pertinent decision boundaries. For instance, all the shown boundaries (dashed lines) are sufficient to classify the raw data with full accuracy in two-dimensional space but are not equivalent in terms of robustness to noise. (b) The rarely explored space (region specified by diagonal striped) in a learning model leaves room for adversaries to manipulate the nuisance (non-critical) variables and mislead the model by crossing the decision boundaries. (c) In MRR methodology, a set of defender modules is trained to characterize the data density distribution in the space spanned by the victim model. The defender modules are used in parallel to checkpoint the reliability of the ultimate prediction and raise an alarm flag for risky samples.

only partially evade adversarial samples generated by other attack scenarios (e.g., with different perturbation distributions) from being effective.[6]

(ii) Unsupervised strategies which aim to smooth out the decision boundaries by incorporating a smoothness penalty[7,8] as a regularization term in the loss function or compressing the neural network by removing the nuisance variables.[9] These works have been developed based on an implicit assumption that the existence of adversarial samples is due to the piece-wise linear behavior of decision boundaries (obtained by the gradient descent approach) in high-dimensional spaces. As such, their integrity can be jeopardized by considering a slightly higher perturbation at the input space to cross the smoothed decision boundaries.[10]

More recently, an unsupervised manifold projection approach (called MagNet) is proposed in the literature to reform adversarial samples using auto-encoders.[11] Unlike MRR countermeasure, MagNet is inattentive to the pertinent data density in the latent space. As shown by Carlini and Wagner,[12] manifold projection methods including MagNet are not robust to adversarial samples and can approximately increase the required distortion to generate adversarial sample by only 37 percent.

To the best of our knowledge, our proposed MRR methodology is the first unsupervised learning countermeasure that simultaneously considers both data geometry (density) and decision boundaries for an effective defense against adversarial attacks. Our proposed countermeasure is able to withstand the strongest known white-box attack to date by provably increasing the robustness of the underlying model. The MRR methodology does not assume any particular attack strategy and/or perturbation pattern. This obliviousness to the underlying attack or perturbation models demonstrates the generalizability of the proposed approach in face of potential future adversarial attacks.

Furthermore, a recent line of research in adversarial ML has shown a trade-off between the robustness of a model and its accuracy.[13] To avoid this trade-off, instead of learning a single model that is both robust and accurate, our proposed countermeasure learns a set of complementary defender modules while keeping the victim model intact; therefore, our defense mechanism does not impose any degradation of accuracy on the victim model.

What is the root cause of adversarial samples?

Our hypothesis is that the vulnerability of ML models to adversarial samples originates from the relatively large subsets of the data domain that remain mainly unexplored. This phenomenon is likely caused by the limited access to the labeled data and/or inefficiency of the algorithms in terms of their generalized properties. Figure 2 provides a simple illustration of the partially explored

space in a two-dimensional setup. We analytically and empirically back up our hypothesis by extensive evaluations[1] on various benchmarks including the well-known MNIST, CIFAR10, and mini-ImageNet datasets.

Due to the curse of dimensionality, it is often not practical to fully cover the underlying high-dimensional space spanned by modern ML applications. What we can do, instead, is to construct statistical modules that can quantitatively assess whether or not a certain sample comes from the subspaces that were exposed to the ML agent. To ensure robustness against adversarial samples, we argue that ML models should be capable of rejecting samples that lie within the rarely-explored regions.

How can we characterize and thwart the adversarial space?

We formalize the goal of preventing adversarial attacks as an optimization problem to minimize the rarely observed regions in the latent feature space spanned by an ML model. To solve the aforementioned minimization problem, a set of complementary but disjoint redundancy modules are trained to capture the Probability Density Function (PDF) of the legitimate (explored) subspaces. In MRR methodology, the victim model is kept *as is* while separate defender modules are trained to *checkpoint* the reliability of the victim model prediction.

Each modular redundancy learns a PDF to explicitly characterize the geometry (density) of a certain high-dimensional data abstraction within an ML model. In a neural network, for example, each MRR module checkpoints a certain intermediate hidden (or input) layer (Figure 3). A DL layer may be checkpointed by multiple MRR modules to provide a more robust defense strategy. Each defender marks the complement of the space characterized by the learned PDF as the rarely observed region, enabling statistical tests to determine the validity of new samples.

Once such characterizations are obtained, statistical testing is used at runtime to determine the legitimacy of new data samples. The defender modules evaluate the input sample probability in parallel with the victim model and raise alarm flags for data points that lie within the rarely explored regions. As such, the adversary is required to *simultaneously* deceive all defender modules in order to succeed. Unlike the prior works, our approach does not suffer from a degradation of accuracy since the victim model is untouched.

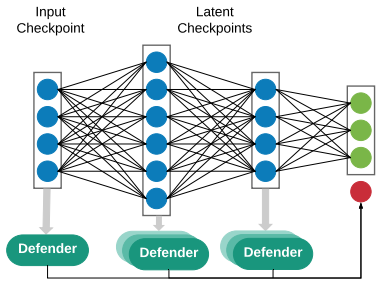
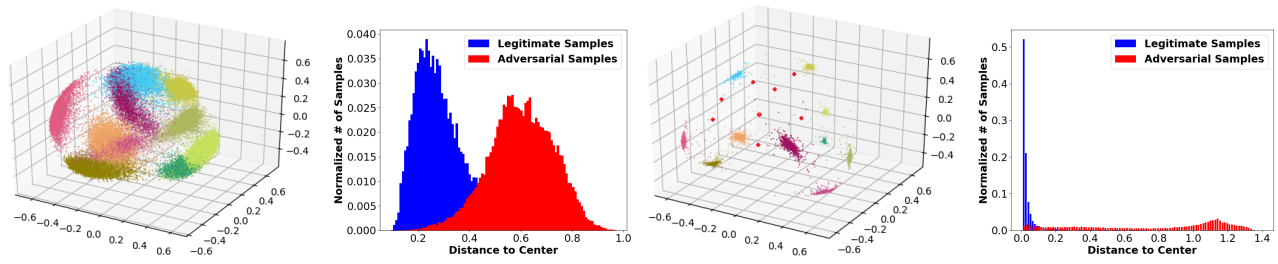


Figure 3: A high-level overview of proposed MRR methodology. The output layer of the victim neural network (the green neurons) is augmented with a single risk measure (the red neuron) determining the legitimacy of the prediction.

The outputs of MRR modules are aggregated into a single output node (the red neuron in Figure 3) that quantitatively measures the reliability of the original prediction. For any input sample, the new neuron outputs a risk measure in the unit interval $[0, 1]$, with 0 and 1 indicating safe and highly risky samples, respectively. The extra neuron incorporates a "don't know" class into the model: samples with a risk factor higher than a certain threshold (a.k.a., security parameter) are treated as adversarial inputs. The threshold is determined based on the safety-sensitivity of the application for which the ML model is employed. This approach is beneficial in a sense that it allows dynamic reconfiguration of the detection policy with minimal required recomputing overhead.

Adversarial and legitimate samples differ in certain statistical properties. Adversarial samples are particularly crafted by finding the rarely explored dimensions in an ℓ_∞ ball of radius ϵ. In MRR methodology, samples whose features lie in the unlikely subspaces are marked and identified as risky samples. Our conjecture is that a general ML model equipped with the *side information* about the density distribution of the input data as well as the distribution of the latent feature vectors can be made arbitrary robust against adversarial samples. Our proposed MRR methodology strengthens the defense by training *multiple* defenders that are *negatively correlated*. Informally, if two MRR modules are negatively correlated, then an adversarial sample that can mislead one module will raise high suspicion in the other module and vice versa.

As an example, consider a classification task where a 4-layer neural network is used to categorize ten different classes of the popular digit recognition dataset known as MNIST. Figure 4a demonstrates the feature vectors within the second-to-last layer of the pertinent *victim* neural network in the Euclidean space. Note that only three dimensions of the feature vectors are shown for visualization purposes. The feature vectors of samples corresponding to the same class (same color) tend to be clustered in the Euclidean space. Each cluster has a center obtained by taking the



(a) (b) (c) (d)

Figure 4: Example feature samples in a 4-layer neural network trained for a digit recognition task. Latent feature samples in the second-to-last layer of (a) the victim model and (c) its corresponding transformation in our defender module. The majority of adversarial samples (e.g., the red dot points in (c)) reside in the regions with low density of training samples. Figures (b) and (d) show the histogram of the distance between samples and cluster centers for legitimate and adversarial inputs in the victim and defender models, respectively.

average of the features of the corresponding class. For each input sample identified as a certain class by the victim model, we compute the distance between the feature vector and the corresponding center.

Figure 4b demonstrates the distribution of the distance between data samples and the center of the pertinent class for legitimate (blue) and adversarial (red) samples. In this experiment, we generate the adversarial samples using the FGS attack algorithm. It can be seen that the aforementioned distance is higher for adversarial samples compared to legitimate samples. This, in fact, validates our hypothesis that adversarial samples lie within the unexplored subspaces (higher distance from cluster centers in this case). The adversarial samples can be simply detected by thresholding the aforementioned distance. Nevertheless, building a detection method based on this distance in its current form will lead to a high probability of false alarm: legitimate inputs might be incorrectly marked as adversarial samples.

Each defender module is regularized based on a prior distribution (e.g., a Gaussian Mixture Model) to enforce disentanglement between the features corresponding to different categories and be more robust against skewed feature distributions.[1] As an example, the corresponding data distribution and distance measure for a single defender are shown in Figures 4c and 4d, respectively. It can be seen that the clusters are well-separated, thus, the characterization of the adversarial subspace incurs a small probability of false alarms. Table 1 summarizes the Area Under Curve (AUC) score attained against four different attacks in a black-box setting.

Table 1: AUC score obtained by 16 latent defenders that checkpoint the second-to-last layer of the victim model for MNIST and CIFAR-10 benchmarks. For ImageNet benchmark, we only used 1 defender due to the high computational complexity of the pertinent neural network and attacks.

	MNIST	CIFAR10	ImageNet
FGS	0.996	0.911	0.881
JSMA	0.995	0.966	-
Deepfool	0.996	0.960	0.908
CarliniL2	0.989	0.929	0.907
BIM	0.994	0.907	0.820

Adaptive white-box attack. To further corroborate the robustness of PCL methodology, we applied the state-of-the-art CarliniL2 attack in a white-box setting.[12] A similar strategy was previously used in the literature to break the state-of-the-art countermeasures including MagNet,[11] APE-GAN,[14] and other recently proposed efficient defenses methods.[15] Table 2 summarizes the success rate of the CarliniL2 attack algorithm for different numbers of redundancy (defender) modules and risk thresholds (security parameters) for the MNIST benchmark.

Our MRR methodology offers a trade-off between robustness of the ML model and its computational complexity. On the one hand, increasing the number of MRRs enhances the robustness of the model as shown in Table 2. On the other hand, the computational complexity grows linearly with the number of MRRs (each redundancy module incurs the same overhead as the victim model). Our proposed MRR defense mechanism outperforms existing state-of-the-art defenses both in terms of the detection success rate and the amount of perturbation required to fool

Table 2: Evaluation of PCL methodology against adaptive white-box attack. We compare our results with prior-art works including Magnet,[11] Efficient Defenses Against Adversarial Attacks,[15] and APE-GAN.[14] For each evaluation, the L_2 distortion is normalized to that of the attack without the presence of any defense mechanism. For fair comparison to prior work, we did not include our non-differentiable input defenders in this experiment. Note that highly disturbed images (with large L_2 distortions) can be easily detected using the input dictionaries/filters.

	MRR Methodology (White-box Attack)												**Prior-Art Defenses (Gray-box Attack)**		
Security Parameter	SP=1%						SP=5%						Magnet	Efficient Defenses	APE-GAN
Number of Defenders	N=0	N=1	N=2	N=4	N=8	N=16	N=0	N=1	N=2	N=4	N=8	N=16	N=16	-	-
Defense Success	-	43%	53%	64%	65%	66%	-	46%	63%	69%	81%	84%	1%	0%	0%
Normalized Distortion (L_2)	1.00	1.04	1.11	1.12	1.31	1.38	1.00	1.09	1.28	1.28	1.63	1.57	1.37	1.30	1.06
FP Rate	-	2.9%	4.4%	6.1%	7.8%	8.4%	-	6.9%	11.2%	16.2%	21.9%	27.6%	-	-	-

the defenders in a white-box setting.

We emphasize that training the defender module is carried out in an unsupervised setting, meaning that no adversarial sample is included in the training phase. We believe that leveraging an unsupervised learning approach is the key to having a generalizable defense scheme that is applicable to a wide class of adversarial machine learning attacks. To the best of our knowledge, our proposed MRR approach[1] is the first unsupervised countermeasure to withstand the existing adversarial attacks for (deep) ML models including Fast-Gradient-Sign, Jacobian Saliency Map Attack, Deepfool, and Carlini&WagnerL2 in both black-box and white-box settings. Details about the robustness of the MRR methodology against the aforementioned attack methods are available in our paper.[1]

Transferability

In the context of adversarial samples, transferability is defined as the ability of adversarial samples to deceive ML models that have not been used by the attack algorithm, i.e. their parameters and network structures were not revealed to the attacker. In other words, adversarial samples that are generated for a certain ML model can potentially deceive another model that has not been exposed to the attacker. Our proposed MRR methodology is robust against model transferability in a sense that the adversarial samples generated for the victim model using the best-known attack methodologies *are not transferred* to the defender modules.[1] This, in turn, guarantees the effective performance of our MRR method against both white-box and black-box[16] attacks.

Our key observation is that the majority of adversarial samples that can be easily transferred in between different models are crafted from legitimate samples that are inherently hard-to-classify due to the closeness to decision boundaries corresponding to such classes. For instance, in the MNIST digit recognition task, such adversarial samples mostly belong to class 5 that is misclassified to class 3, or class 4 misclassified as 9. These misclassifications are indeed the model approximation error which is well-understood due to the statistical nature of the models.

Figure 5 shows example adversarial samples generated by such hard-to-classify examples. As demonstrated, even a human observer might make a mistake in labeling such images. We believe that a more precise definition of adversarial samples is necessary to distinguish malicious samples form those that simply lie near the decision boundaries. Therefore, the notion of transferability should be redefined to differentiate between hard-to-classify samples and adversarial examples. ■

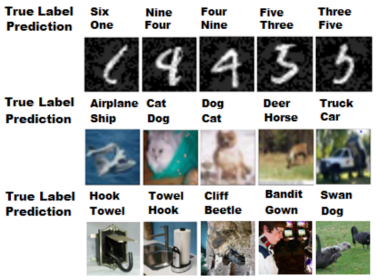


Figure 5: Example adversarial samples for which accurate detection is hard due to the closeness of decision boundaries for the corresponding data categories.

References

1. Bita Darvish Rouhani, Mohammad Samragh, Tara Javidi, and Farinaz Koushanfar. Curtail: Characterizing and thwarting adversarial deep learning. *arXiv preprint arXiv:1709.02538*, 2017.
2. Jonghoon Jin, Aysegul Dundar, and Eugenio Culurciello. Robust convolutional neural networks under adversarial noise. *arXiv preprint arXiv:1511.06306*, 2015.
3. Ruitong Huang, Bing Xu, Dale Schuurmans, and Csaba Szepesvári. Learning with a strong adversary. *arXiv preprint arXiv:1511.03034*, 2015.

4. Uri Shaham, Yutaro Yamada, and Sahand Negahban. Understanding adversarial training: Increasing local stability of neural nets through robust optimization. *arXiv preprint arXiv:1511.05432*, 2015.
5. Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.
6. Shixiang Gu and Luca Rigazio. Towards deep neural network architectures robust to adversarial examples. *arXiv preprint arXiv:1412.5068*, 2014.
7. Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, Ken Nakae, and Shin Ishii. Distributional smoothing with virtual adversarial training. *arXiv preprint arXiv:1507.00677*, 2015.
8. Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *Security and Privacy (SP), 2017 IEEE Symposium on*. IEEE, 2017.
9. Nicolas Papernot, Patrick McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami. Distillation as a defense to adversarial perturbations against deep neural networks. 2016.
10. Nicholas Carlini and David Wagner. Defensive distillation is not robust to adversarial examples. *arXiv preprint*, 2016.
11. Dongyu Meng and Hao Chen. Magnet: a two-pronged defense against adversarial examples. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*. ACM, 2017.
12. Nicholas Carlini and David Wagner. Magnet and" efficient defenses against adversarial attacks" are not robust to adversarial examples. *arXiv preprint arXiv:1711.08478*, 2017.
13. Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.
14. Shiwei Shen, Guoqing Jin, Ke Gao, and Yongdong Zhang. Ape-gan: Adversarial perturbation elimination with gan. *ICLR Submission, available on OpenReview*, 2017.
15. Valentina Zantedeschi, Maria-Irina Nicolae, and Ambrish Rawat. Efficient defenses against adversarial attacks. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*. ACM, 2017.
16. Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z Berkay Celik, and Ananthram Swami. Practical black-box attacks against deep learning systems using adversarial examples. *arXiv preprint arXiv:1602.02697*, 2016.

Bita Darvish Rouhani is a Ph.D. candidate in the Department of Electrical and Computer Engineering at the University of California San Diego. Contact her at bita@ucsd.edu.

Mohammad Samragh is a Ph.D. student in the Department of Electrical and Computer Engineering at the University of California San Diego. Contact him at msamragh@ucsd.edu.

Tara Javidi is a Professor in the Department of Electrical and Computer Engineering at the University of California San Diego. Contact her at tjavidi@ucsd.edu.

Farinaz Koushanfar is a Professor and Henry Booker Faculty Scholar in the Department of Electrical and Computer Engineering at the University of California San Diego. Contact her at farinaz@ucsd.edu.