Optimising Network Architectures for Provable Adversarial Robustness

Citation for published version:
Gouk, H & Hospedales, TM 2020, Optimising Network Architectures for Provable Adversarial Robustness. in *2020 Sensor Signal Processing for Defence Conference (SSPD).* Institute of Electrical and Electronics Engineers (IEEE), pp. 1-5, 9th International Conference of the Sensor Signal Processing for Defence , Virtual Conference, United Kingdom, 15/09/20. https://doi.org/10.1109/SSPD47486.2020.9272169

Digital Object Identifier (DOI):
10.1109/SSPD47486.2020.9272169

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:
2020 Sensor Signal Processing for Defence Conference (SSPD)

Optimising Network Architectures for Provable Adversarial Robustness

Henry Gouk
School of Informatics
University of Edinburgh
Edinburgh, United Kingdom
hgouk@inf.ed.ac.uk

Timothy M. Hospedales
School of Informatics
University of Edinburgh
Edinburgh, United Kingdom
thospedales@ed.ac.uk

Abstract—**Existing Lipschitz-based provable defences to adversarial examples only cover the ℓ_2 threat model. We introduce the first bound that makes use of Lipschitz continuity to provide a more general guarantee for threat models based on any ℓ_p norm. Additionally, a new strategy is proposed for designing network architectures that exhibit superior provable adversarial robustness over conventional convolutional neural networks. Experiments are conducted to validate our theoretical contributions, show that the assumptions made during the design of our novel architecture hold in practice, and quantify the empirical robustness of several Lipschitz-based adversarial defence methods.**

Index Terms—**Artificial Neural Network, Computer Vision**

I. Introduction & Related Work

The robustness of deep neural networks to adversarial attack [1] is an increasingly topical issue as deep models are becoming more widely deployed in practice. This paper focuses on the problem of ensuring that once a deep network trained for image classification has been deployed, one can be confident that an adversary has only a limited ability to maliciously impact model predictions when they tamper with the system inputs. Such malicious inputs, so-called adversarial examples, appear to humans as normal images, but in reality they have undergone imperceptible modifications that cause a model to make an incorrect prediction. The majority of research into adversarial examples is still based on empirical results that have been shown to be somewhat fragile [2], [3]. In contrast, we look to the more recent trends in provable adversarial robustness, where it one is able to compute a certificate for each prediction made by the model, ensuring that it is robust to some pre-specified family of attacks, known as the threat model [4], [5].

There are several papers in the literature on deep learning that address adversarial robustness through the use of Lipschitz continuity, but they focus solely on perturbations with bounded Euclidean norm. Tsuzuku et al. [4] present an efficient method for determining whether an example could have been tampered with at test time or, conversely, certify that a prediction has not been influenced by an adversarial attack. They compare the prediction margin normalised by the Lipschitz constant of the network to the magnitude of the largest perturbation allowed by the threat model, allowing them to determine whether the input could be an adversarial example. Farnia et al. [6] show how the adversarial risk can be bounded in terms of the training loss by adapting the bound of [7] to consider perturbations to the margin, using a similar technique to [4]. The analysis in this paper takes a similar high-level strategy—making use of margins and Lipschitz constants—but we extend this theory to threat models based on arbitrary p-norms, and provide a simpler proof than previous methods [4]. Huster et al. [8] demonstrate that current methods for regularising Lipschitz constants of networks have deficiencies when used for improving adversarial robustness. Specifically, it is shown that existing approaches for regularising the Lipschitz constant may be too restrictive because the bound on the Lipschitz constant is too loose, resulting in over-regularisation. We take an orthogonal approach: we provide theoretical and practical contributions that are compatible with arbitrary bounds on the Lipschitz constant.

Existing work that aims to provide theory-backed guarantees for adversarial robustness has resulted in several techniques able to certify whether a prediction for a particular example is immune to adversarial attack under a threat model based on ℓ_p-norm perturbation size. [9] propose a method that can only be applied to networks composed of fully connected layers with rectified linear units activation functions, and no batch normalisation. [10] present an approach based on solving an optimisation problem. While the robustness estimates they give are considerably tighter than many other certification methods, they scale very poorly to networks with large input images or feature maps. In contrast to these methods, our approach bounds the expected adversarial generalisation error, has virtually no test-time computational overhead, and can be applied to arbitrary feed-forward architectures. Bounding the expected generalisation error enables us to give guarantees about the level of robustness a model will have once it has been deployed. Existing approaches to provable robustness do not come with such guarantees, and can only provide certification for individual instances.

We begin by extending existing theory addressing the relationship between Lipschitz continuity and provable adversarial robustness. Using insights from the resulting bounds, it is shown how one can adjust network architectures in such a way that Lipschitz-based regularisation methods are more effective. Experimental results show that, while having little difference in clean performance compared to existing Lipschitz-based defences, our approach improves the level of *provable* robustness significantly.

II. Generalisation Under Attack

Methods for estimating the generalisation performance of learned models typically assume examples, $(\vec{x}, y)$, observed at both training and testing time are independently drawn from the same distribution, $\mathcal{D}$. Such methods estimate or bound the expected risk,

$$R^\ell(f) = \mathbb{E}_{(\vec{x},y)\sim\mathcal{D}}[\ell(f(\vec{x}), y)], \tag{1}$$

of a classifier, f, with respect to some loss function, ℓ. The standard technique for estimating the expected risk in deep learning is to use an empirical approximation measured on a set of held-out data. In the adversarial setting one must consider the expected risk when under the influence of an attacker that can add perturbations to feature vectors at test time,

$$\tilde{R}^\ell_{p,t}(f) = \mathbb{E}_{(\vec{x},y)\sim\mathcal{D}}\left[\max_{\vec{\epsilon}:\|\vec{\epsilon}\|_p\leq t} \ell(f(\vec{x}+\vec{\epsilon}), y)\right], \tag{2}$$

which is known as the adversarial risk [5]. In contrast to the expected risk, $\tilde{R}^\ell_{p,t}(f)$ cannot be reliably approximated from data when f is nonlinear, as one must find the globally optimal setting of $\vec{\epsilon}$ for each data point in the held-out set.

For a hypothesis, f, that produces a vector of real-valued scores, each associated with a possible class, we define the margin function as

$$m_f(\vec{x}, y) = f_y(\vec{x}) - \max_{j\neq y} f_j(\vec{x}), \tag{3}$$

where $f_i(\vec{x})$ is the ith component of the output of $f(\vec{x})$. Typical loss functions for measuring the performance of a model via composition with the margin function include the zero–one loss and the hinge. These compositions result in the classification error rate and the multi-class hinge loss variant proposed by [11], respectively.

Proposition 1. *If f is k-Lipschitz with respect to the p-norm and $\ell : \mathbb{R} \rightarrow \mathbb{R}^+$ is a monotonically decreasing loss function, then*

$$\max_{\vec{\epsilon}:\|\vec{\epsilon}\|_p\leq t} \ell(m_f(\vec{x}+\vec{\epsilon}, y)) \leq \ell(m_f(\vec{x}, y) - 2^{1/q}kt), \tag{4}$$

where q is defined such that $\|\cdot\|_q$ is the dual norm of $\|\cdot\|_p$.

Proof. The main idea behind the proof is to show that the Lipschitz constant of the network controls how much the margin can be influenced by an adversarial perturbation. Note that one can express the margin function given in Equation 3 as $m_f(\vec{x}, y) = m_{\mathbb{I}}(f(\vec{x}), y)$, where $\mathbb{I}$ is the identity function. The Lipschitz constant of $m_{\mathbb{I}}$ with respect to its first argument when using the p-norm is $\max_{\vec{x}} \|\nabla_{\vec{x}} m_{\mathbb{I}}(\vec{x}, y)\|_q$ [12, p. 133]. The gradient of $m_{\mathbb{I}}$ is a vector with all elements set to zero, except for those corresponding to the largest and second largest components of $\vec{x}$. These components of the gradient take the values of 1 and -1, respectively. Plugging these values into the definition of vector p-norms, one arrives at a Lipschitz constant of $2^{1/q}$. From the composition property of Lipschitz functions, we can say that m_f is $(2^{1/q}k)$-Lipschitz with respect to $\vec{x}$. The Lipschitz property of m_f can be used to bound the worst-case change in the output the margin function for a bounded change in the input, yielding

$$\ell(\min_{\vec{\epsilon}:\|\vec{\epsilon}\|<t} m_f(\vec{x}+\vec{\epsilon}, y)) \leq \ell(m_f(\vec{x}, y) - 2^{1/q}kt). \tag{5}$$

From the decreasing monotonicity of ℓ, we have that

$$\max_{\vec{\epsilon}:\|\vec{\epsilon}\|<t} \ell(m_f(\vec{x}+\vec{\epsilon}, y)) = \ell(\min_{\vec{\epsilon}:\|\vec{\epsilon}\|<t} m_f(\vec{x}+\vec{\epsilon}, y)), \tag{6}$$

which concludes the proof. □

This proposition bounds the worst-case change in loss for a single image in terms of prediction confidence, Lipschitz constant of the network, and the maximum allowable attack strength.

The relationship given in Proposition 1 is a more general form of the bound derived by [4], who consider only the Euclidean norm.

Proposition 1 can be extended to provide a non-trivial bound on the expected adversarial risk through the use of a held-out dataset and a simple application of McDiarmid's inequality.

Proposition 2. *If f is k-Lipschitz w.r.t. the p-norm, $\ell : \mathbb{R} \rightarrow [0, B)$ is a monotonically decreasing loss function, and $\{(\vec{x}_i, y_i) \sim \mathcal{D}\}_{i=1}^n$ is independent of f (i.e., held-out data), the following holds with probability at least $1-\delta$:*

$$\tilde{R}^\ell_{p,t}(f) \leq \frac{1}{n}\sum_{i=1}^n \ell(m_f(\vec{x}_i, y_i) - 2^{1/q}kt) + B\sqrt{\frac{\ln(2/\delta)}{2n}} \tag{7}$$

where q is defined such that $\|\cdot\|_q$ is the dual norm of $\|\cdot\|_p$.

Proof. Constructing a mean over loss terms,

$$L = \frac{1}{n}\sum_{i=1}^n \ell(m_f(\vec{x}_i, y_i) - 2^{1/q}kt), \tag{8}$$

results in a sequence where each term is bounded by $\frac{B}{n}$, allowing McDiarmid's inequality to probabilistically bound the deviation from its expected,

$$\mathbb{P}(|L - \mathbb{E}[L]| > \gamma) \leq 2\text{exp}\left(\frac{-2n\gamma^2}{B^2}\right). \tag{9}$$

Setting δ equal to the right-hand side of Inequality 9 and solving for γ yields

$$\gamma = B\sqrt{\frac{\ln(2/\delta)}{2n}}. \tag{10}$$

Thus, we can say with confidence $1-\delta$ that

$$\mathbb{E}[L] \leq L + B\sqrt{\frac{\ln(2/\delta)}{2n}}. \tag{11}$$

Applying Proposition 1 to each term of the summation, L, concludes the proof. □

Proposition 2 extends the result of Proposition 1 from the loss on a single instance to the expected risk.

In practice, this means that a practitioner can bound the worst-case adversarial performance of their model based on its (non-adversarial) validation-set performance and its Lipschitz constant, both of which can be measured efficiently. As we show later, this can lead to non-vacuous bounds on error rate, which in turn could allow a user to deploy a model with provable confidence about its performance under adversarial attack—without the hassle and computational expense of instance-wise certification at run-time [9], [10].

III. Architectures for Provable Robustness

The analysis in Section II motivates a high-level strategy for improving the adversarial robustness of neural networks: maximise the prediction margin while minimising the Lipschitz constant of the model. Several papers have proposed different methods for regularising the Lipschitz constant of a network, with various motivations, including improving robustness to adversarial exmaples [4], [13] and improving generalisation performance in the non-adversarial case [14].

We propose a strategy for modifying network architectures to make them more amenable to Lipschitz-based regularisers: splitting a single multi-class classification network into a collection of one-versus-all (OVA) classifiers that each produce a real-valued score. Unlike the conventional OVA method, where each component classifier is trained in isolation, the networks used in our approach are still trained jointly using a softmax composed with the cross entropy loss function. There are two requirements for this OVA scheme to have a benefit: each of the simpler binary classification subproblems must be solvable by a network with a smaller Lipschitz constant, and the Lipschitz constant of the multi-classifier system must grow slowly with the number of classes. [14] show that the Lipschitz constant is related to model capacity, so the subnetwork associated with each class should be able to achieve high accuracy with a smaller Lipschitz constant than a conventional multi-class classification network. For the second requirement, consider the vector-valued function,

$$f(\vec{x}) = [f_1(\vec{x}), f_2(\vec{x}), ..., f_C(\vec{x})], \tag{12}$$

where C is the number of classes, and f_i is k_i-Lipschitz. We have from the Lipschitz property of each f_i that

$$\|f(\vec{x}) - f(\vec{x}+\vec{v})\|_p \leq \|[k_1\|\vec{v}\|_p, k_2\|\vec{v}\|_p, ..., k_C\|\vec{v}\|_p]\|_p \tag{13}$$

$$= \|\vec{v}\|_p \|[k_1, k_2, ..., k_C]\|_p, \tag{14}$$

from which we can deduce that the Lipschitz constant of the one-versus-all classifier is the ℓ_p norm of the vector of Lipschitz constants corresponding to each binary classifier. In the case of the ∞-norm, the largest Lipschitz constant associated with a single binary classifier dictates the Lipschitz constant of the entire OVA classifier. From this, we can conclude that the second requirement is satisfied.

A. Lipschitz Regularization Training

We investigate two approaches to controlling the Lipschitz constant of neural networks. The first approach we use is to add the bound on the Lipschitz constant as a regularisation term to the objective function, resulting in

$$\frac{1}{n}\sum_{i=1}^{n} \ell(f(\vec{x}_i), y_i) + \prod_{l=1}^{d} \|W_l\|_p, \tag{15}$$

where d is the number of layers in the network and $\|\cdot\|_p$ is the operator norm induced by the vector p-norm. In the case where p is two, the operator norm is the largest singular value of the matrix (i.e., the spectral norm). For $p = \infty$, it is the maximum absolute row sum norm [14],

$$\|W\|_\infty = \max_i \sum_j |W_{i,j}|. \tag{16}$$

IV. Experiments

This section presents the results of numerical experiments that demonstrate the tightness of the bounds presented in Section II and provides evidence that the architecture proposed in Section III is inherently easier to optimise for provable robustness than conventional network architectures. The models used in these experiments were implemented using Keras [1], and the adversarial attacks were performed using the CleverHans toolkit [2].

A. Tightness of the Bound

The bound given in Proposition 2 provides a way to estimate the worst-case performance of a model when under the influence of an adversary. In order to validate this bound empirically, we train linear support vector machines with different levels of ℓ_2 regularisation on the MNIST dataset of hand-written digits. In the case of linear SVMs, the optimisation problem solved by iterative gradient-based attacks, such as the projected gradient descent method of [5], are convex and can therefore be solved globally. This means the the empirical adversarial risk can be computed exactly. Plots indicating the tightness of the bound for linear SVMs are given in Figure 1. These were generated by training models on the first 50,000 instances of the training set, using the other 10,000 training instances as the held-out data required for computing the bound, and using the PGD attack when evaluating the network on the test data. These plots confirm that the bound proposed in Proposition 2 is non-vacuous and has the potential to be useful in practice.

B. Provable Robustness

We first experiment on MNIST to determining whether our proposed OVA networks achieve better provable robustness than conventional convolutional neural networks. To control for the potentially confounding factor of model capacity, a series of networks with different widths are trained. We define the width of a conventional convolutional network as the number of

[1] https://github.com/keras-team/keras

[2] https://github.com/tensorflow/cleverhans

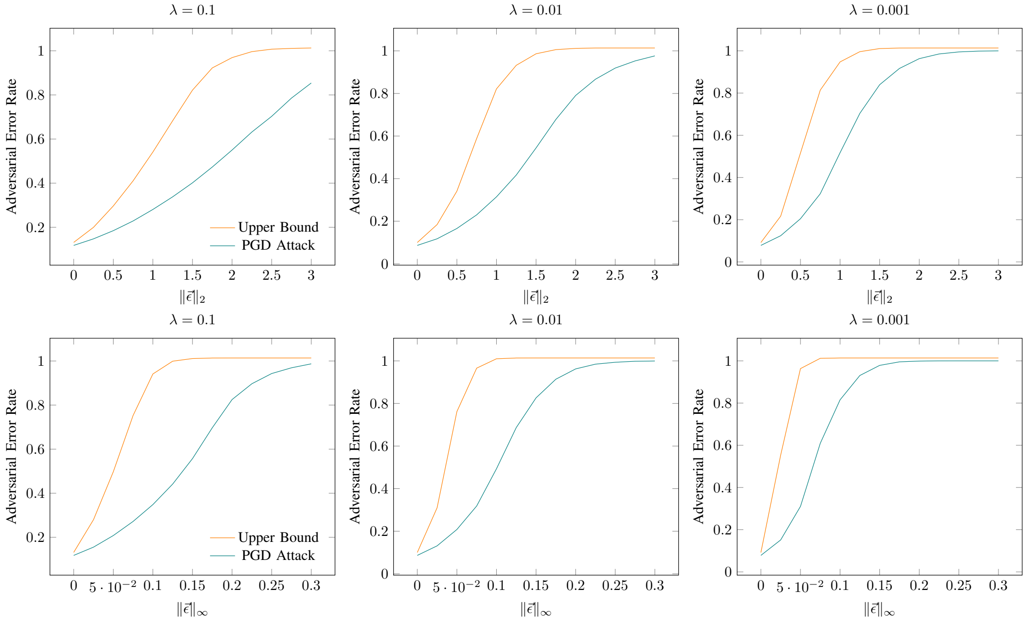

Fig. 1. Plots demonstrating the relationship between the provable upper bound on adversarial risk, and the actual misclassification rate on the test set under adversarial attack. Linear SVM recognition on MNIST with ℓ_2 (top row) and ℓ_∞ (bottom row) threat models and regularization strength λ.

feature maps produced by the first convolutional layer. For OVA networks, the width is the number of feature maps produced by the first layer in a single binary classifier, multiplied by the number of binary classifiers. For both network types, the chosen architectures contain two convolutional layers, the second of which has twice the number of feature maps as the first. Each convolutional layer contains 5×5 kernels, rectified linear unit activation functions, and is followed by a 2×2 max pooling layer. After the convolutional layers are two fully connected layers: one with 128 hidden units, and another with either 10 units (for conventional networks), or one unit (for OVA networks).

Figure 2 shows how the number of model parameters impacts the provable adversarial robustness for threat models based on the ℓ_2 and ℓ_∞ norms. The models in these plots are regularised using the Lipschitz penalty method proposed in Section III. These figures show that: (1) Regularised OVA networks exhibit superior provable robustness compared to regularized conventional CNNs at comparable model sizes, (2) The magnitude of this margin becomes more pronounced as model size increases, (3) All methods have low error rate for unperturbed examples (left plots).

To investigate how well OVANets scale to larger networks and more challenging datasets, additional experiments are run on the CIFAR-10 dataset, using VGG-style networks [15] as the base architecture. The baseline CNN uses the VGG11 architecture, and each subnetwork of the OVANet architecture is a VGG11 network with half the number of feature maps in each layer. Table I provides probabilistic (95% confidence) bounds on the worst-case adversarial error rate using Proposition 2. Table II shows the corresponding provable robustness results for SVHN benchmark. From the results we can see that: (1) Lipschitz penalty training improves the adversarial error rate for both vanilla VGG11 and VGG11-OVANet (performance improves with λ); (2) VGG11-OVANet generally has superior provable robustness compared to vanilla VGG11 for corresponding regularisation strength, especially for strong attacks. (3) Meanwhile, regularized OVANet achieves comparable results to a regularized CNN in terms of clean data performance.

TABLE I
BOUNDS ON THE ERROR RATE FOR VGG MODELS TRAINED ON CIFAR-10. THE BOUNDS WERE COMPUTED WITH PROPOSITION 2 AT THE 95% CONFIDENCE LEVEL AND THE ℓ_2 THREAT MODEL.

			Perturbation Size (ℓ_2)			
Model	λ	Clean	1/255	2/255	3/255	4/255
VGG11-CNN	0	14.50	100.00	100.00	100.00	100.00
	0.0001	14.22	47.61	79.22	95.87	100.00
	0.0005	16.00	29.00	42.74	56.49	69.84
	0.001	17.64	26.80	35.60	44.75	53.66
VGG11-OVA	0	17.18	100.00	100.00	100.00	100.00
	0.0001	15.58	44.54	73.49	93.11	99.99
	0.0005	15.86	27.68	39.01	51.85	63.68
	0.001	17.09	25.00	32.54	40.35	48.53

V. Conclusions

This paper presents a p-norm-agnostic theoretical analysis of provable adversarial robustness via Lipschitz regularisation. A new architecture, the OVA network, is proposed, motivated by insights of how Lipschitz constants can be bounded for different architecture design choices. It is shown that OVA networks achieve similar empirical performance to conventional neural networks but, as network size increases, OVA networks are able to achieve significantly better certifiable robustness. This

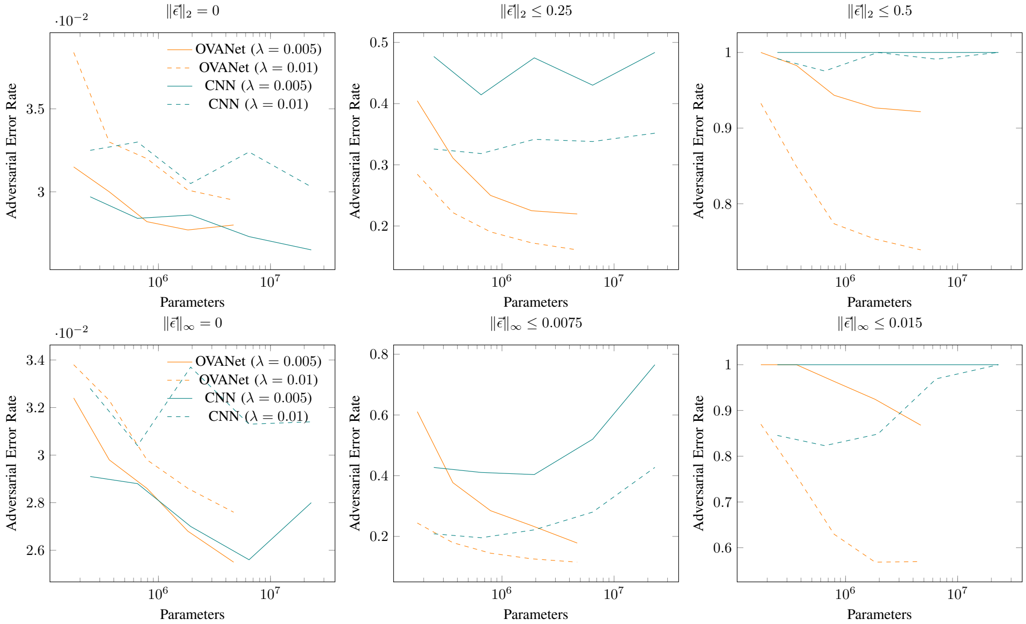

Fig. 2. Comparison of provable adversarial risk for conventional CNN versus OVANet trained with Lipschitz penalty regularization over a range of model sizes. The ℓ_2 (top row) and ℓ_∞ (bottom row) Lipschitz constants are used for regularisation and computing the bound. OVANet shows superior provable robustness, especially at larger model sizes.

TABLE II
BOUNDS ON THE ERROR RATE FOR VGG MODELS TRAINED ON SVHN. THE BOUNDS WERE COMPUTED WITH PROPOSITION 2 AT THE 95% CONFIDENCE LEVEL AND THE ℓ_2 THREAT MODEL.

			Perturbation Size (ℓ_2)			
Model	λ	Clean	1/255	2/255	3/255	4/255
VGG11-CNN	0	7.29	100.00	100.00	100.00	100.00
	0.0001	7.15	13.16	21.86	33.59	47.10
	0.0005	8.46	11.38	14.84	19.33	24.49
	0.001	9.41	11.76	14.95	17.76	21.48
VGG11-OVA	0	7.69	100.00	100.00	100.00	100.00
	0.0001	7.45	12.10	19.19	28.38	39.17
	0.0005	8.25	10.64	13.02	16.11	19.82
	0.001	9.02	10.92	12.86	15.05	17.81

is a useful result for practitioners, who can use a Lipschitz regulariser and our bound in order to train models with a certifiable level of robustness against adversarial attack.

REFERENCES

[1] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," in *arXiv:1312.6199 [Cs]*, 2014.

[2] 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*, ser. AISec '17. New York, NY, USA: ACM, 2017, pp. 3–14.

[3] A. Athalye, N. Carlini, and D. Wagner, "Obfuscated Gradients Give a False Sense of Security: Circumventing Defenses to Adversarial Examples," in *International Conference on Machine Learning*, Jul. 2018, pp. 274–283.

[4] Y. Tsuzuku, I. Sato, and M. Sugiyama, "Lipschitz-Margin Training: Scalable Certification of Perturbation Invariance for Deep Neural Networks," in *Advances in Neural Information Processing Systems*, 2018.

[5] 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*, Feb. 2018.

[6] F. Farnia, J. M. Zhang, and D. Tse, "Generalizable Adversarial Training via Spectral Normalization," *arXiv:1811.07457 [cs, stat]*, Nov. 2018.

[7] P. L. Bartlett, D. J. Foster, and M. J. Telgarsky, "Spectrally-normalized margin bounds for neural networks," in *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., 2017, pp. 6240–6249.

[8] T. Huster, C.-Y. J. Chiang, and R. Chadha, "Limitations of the Lipschitz constant as a defense against adversarial examples," *arXiv:1807.09705*, Jul. 2018.

[9] T.-W. Weng, H. Zhang, H. Chen, Z. Song, C.-J. Hsieh, D. Boning, I. S. Dhillon, and L. Daniel, "Towards fast computation of certified robustness for relu networks," in *International Conference on Machine Learning*, 2018.

[10] E. Wong and J. Z. Kolter, "Provable defenses against adversarial examples via the convex outer adversarial polytope," in *International Conference on Machine Learning*, 2018.

[11] K. Crammer and Y. Singer, "On the Algorithmic Implementation of Multiclass Kernel-based Vector Machines," *Journal of Machine Learning Research*, vol. 2, no. Dec, pp. 265–292, 2001.

[12] S. Shalev-Shwartz, "Online Learning and Online Convex Optimization," *Foundations and Trends® in Machine Learning*, vol. 4, no. 2, pp. 107–194, Mar. 2012.

[13] M. Cisse, P. Bojanowski, E. Grave, Y. Dauphin, and N. Usunier, "Parseval Networks: Improving Robustness to Adversarial Examples," in *International Conference on Machine Learning*, Jul. 2017, pp. 854–863.

[14] H. Gouk, E. Frank, B. Pfahringer, and M. Cree, "Regularisation of Neural Networks by Enforcing Lipschitz Continuity," *arXiv preprint arXiv:1804.04368*, 2018.

[15] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.