

Analysing robustness of tiny deep neural networks

Hamid Mousavi¹[0000-0001-5710-1206], Ali Zoljodi¹[0000-0001-6889-5005], and
Masoud Daneshtalab^{1,2}[0000-0001-6289-1521]

¹ Mälardalen University, Universitetsplan 1, 722 20 Västerås, Sweden
{seyedhamidreza.mousavi, ali.zoljodi, masoud.daneshtalab}@mdu.se

² Tallinn University of Technology (Taltech), Tallinn, Akadeemia tee 15A, Estonia
masoud.daneshtalab@taltech.ee

Abstract. Real-world applications that are safety-critical and resource-constrained necessitate using compact and robust Deep Neural Networks (DNNs) against adversarial data perturbation. MobileNet-tiny has been introduced as a compact DNN to deploy on edge devices to reduce the size of networks. To make DNNs more robust against adversarial data, adversarial training methods have been proposed. However, recent research has investigated the robustness of large-scale DNNs (such as WideResNet), but the robustness of tiny DNNs has not been analysed. In this paper, we analyse how the width of the blocks in MobileNet-tiny affects the robustness of the network against adversarial data perturbation. Specifically, we evaluate natural accuracy, robust accuracy, and perturbation instability metrics on the MobileNet-tiny with various inverted bottleneck blocks with different configurations. We generate configurations for inverted bottleneck blocks using different width-multipliers and expand-ratio hyper-parameters. We discover that expanding the width of the blocks in MobileNet-tiny can improve the natural and robust accuracy but increases perturbation instability. In addition, after a certain threshold, increasing the width of the network does not have significant gains in robust accuracy and increases perturbation instability. We also analyse the relationship between the width-multipliers and expand-ratio hyper-parameters with the Lipchitz constant, both theoretically and empirically. It shows that wider inverted bottleneck blocks tend to have significant perturbation instability. These architectural insights can be useful in developing adversarially robust tiny DNNs for edge devices.

Keywords: · Robustness analysis · Adversarial training · Adversarial data perturbation · Lipchitz constant .

1 Introduction

Deep Neural Networks (DNNs) are increasingly employed in safety-critical applications [18]. Recent research indicates that DNNs are susceptible to adversarial data perturbations, which are small, imperceptible noises that add to the input data [19,6]. Adversarial data can be generated by different adversarial attack methods to fool the DNNs [7,14,4]. In addition, the resource-constraint edge

devices require to employ the tiny DNNs [1]. To this end, some tiny DNNs such as MobileNet-tiny [12] have been designed to deploy on edge devices with high accuracy on clean data. However, designing a robust and tiny DNN is a critical challenge in these applications [26]. To address the robustness of DNNs against adversarial data, adversarial training methods have been proposed as state-of-the-art defense methods [14,24,16]. Nevertheless, adversarial training methods have been analysed for large-scale DNNs such as WideResNet with high capacity (huge number of parameters) [25,22,10,23]. However, there are no comprehensive analyses of adversarial training for the tiny DNNs. In this paper, we present the first robustness analysis for tiny DNNs from the architectural perspective. Our analysis is based on MobileNet-tiny as the extensive architecture used for tiny applications. The baseline network architecture of MobileNet-tiny depicts in figure 1(a). It is composed of n inverted bottleneck block which is configured with two hyper-parameters: width-multiplier and expand-ratio. We analyse the impact of increasing the width of MobileNet-tiny on the performance of the network by expanding the width-multiplier and expand-ratio hyper-parameters of inverted bottleneck blocks. These expanded blocks are illustrated in figure 1(b). The width-multiplier increases the width of the network by extending the number of output channels of the blocks, as shown in figure 1(b)-left. The expand-ratio hyper-parameter extends the number of middle channels inside the inverted bottleneck blocks, as shown in figure 1(b)-right. To comprehensively evaluate the baseline and expanded networks, we leverage different metrics, including natural accuracy, robust accuracy, and perturbation instability (the mathematical definition of the metrics is presented in the section 3.1). Natural accuracy and robust accuracy measure the ratio of clean and adversarial data that can be correctly classified by trained DNNs. Perturbation instability shows the difference between the distribution of the predictions for natural and adversarial data without focusing on correct labels. To support our observations, we theoretically and empirically analyse the relationship between the width-multiplier and expand-ratio of inverted bottleneck blocks with the Lipschitz constant. The Lipschitz constant indicates the stability of the network output to data perturbations, and the larger Lipschitz constant value corresponds to the instability of the network. The following important insights have been discovered by our investigation:

1. Extending the inverted bottleneck blocks in MobileNet-tiny with both width-multiplier and expand-ratio hyper-parameters improves the natural and robust accuracy and increases the perturbation instability.
2. There is a threshold for expanding the width of the network that improves the natural and robust accuracy. Beyond the threshold, the improvement is negligible, and perturbation instability significantly increases.
3. The theoretical and experimental values for the Lipschitz constant upper bound show that increasing the width and depth of the MobileNet-tiny increase the perturbation instability

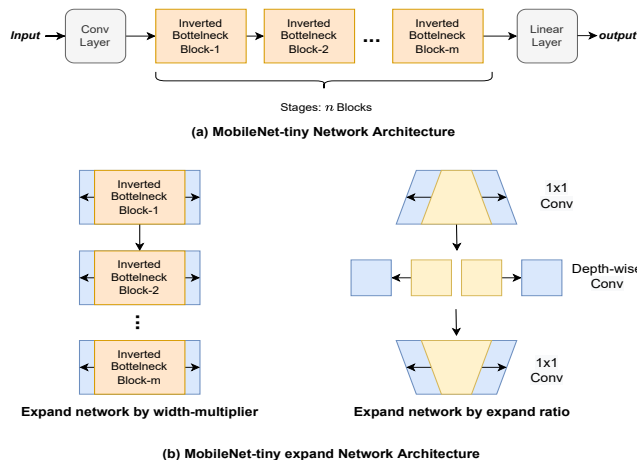


Fig. 1: (a) The baseline architecture of MobileNet-tiny with n inverted bottleneck blocks and (b) expanded networks by changing width-multiplier (left) and expand-ratio (right) hyper-parameters.

2 Related Works

2.1 Adversarial data and Defences

The pretrained deep neural networks are vulnerable to adversarial data, which can be generated by the *Fast Gradient Sign Method (FGSM)* [7], *Projected Gradient Descent (PGD)* [14], and *Carlini and Wagner (CW)* [3] approaches. In addition, *Auto Attack (AA)* [4] is the ensemble of four attacking methods that generate powerful adversarial perturbed data. Adversarial training is the current state-of-the-art defense method against adversarial data. The first adversarial training method employed clean and adversarial data to train a robust deep neural network [7]. The robustness of the network can be increased by encouraging similar logits for clean and adversarial data [11]. To enhance robustness, adversarial training is reformulated as a min-max optimization problem, and the network is trained exclusively on adversarial data [14]. Theoretically, TRADES [24] regularizes the loss function for clean data by incorporating a robust loss term and making a trade-off between them. Improved variants of TRADES have been proposed to consider regularization terms and reduce the distance between the distribution of natural data and their adversarial counterpart [20]. We use TRADES as the default adversarial training method because it uses both natural and robust loss terms to improve robustness.

2.2 Robustness from Architectural perspective

Researchers have investigated the relationship between robustness against adversarial data and the architecture of DNNs [22,10,23]. Xie et al. [23] have studied

the impact of the depth of large-scale DNN networks on adversarial training. They have found that the number of layers in the WideResNet network has a much more substantial effect on robust accuracy than natural accuracy. Their experiments shed light on the intricate relationship between DNN architecture and robustness against adversarial data. Furthermore, Xie et al. [5] have also explored the role of batch normalization layers in the performance of adversarial training, particularly in large-scale datasets such as ImageNet-1K [5]. Their experiments demonstrate that proper batch normalization techniques markedly impact robust accuracy. In addition to the number of layers, the impact of the width of robust accuracy of the large-scale WideResnet network has been studied in [22]. B. Wu et al. [22] showed that the robustness against adversarial data is related to natural accuracy and perturbation stability parameters. Their studies illustrated that increasing the width of WideResNet improves natural accuracy but disprove the perturbation stability. They also elaborate that increasing the DNN width in large-scale networks reduces the overall robust accuracy. H. Huang et al. [10] obtain a study on the impact of DNN architecture on robustness against adversarial data. According to their results, increasing the network capacity (number of parameters) does not necessarily increase its robustness against adversarial data. They also indicate that reducing the capacity in the last blocks of the network may increase the robustness. H. Huang et al. [10] prove that with a constant number of parameters, we can find a DNN architecture with the optimum robustness. Although the research as mentioned earlier studies analysed the robustness of DNNs against adversarial data, but they used large-scale networks such as WideResNet with 127 million parameters. Due to the increase in the number of edge devices, it is necessary to analyse the robustness of tiny networks such as MobileNet-tiny. In this paper, we analyse the robustness of these tiny networks to find meaningful insights into designing adversarially robust tiny networks.

2.3 Tiny Deep Learning

By increasing the usage of tiny edge devices, the demand for DNNs with lower resource consumption and inference time is growing [17]. Recently, tiny deep learning networks [2,15,13] have been proposed to reduce DNNs computation cost and latency. Network pruning [8] and weight quantization [13] are two common approaches to compress existing networks without manipulating the number of layers and hyper-parameters. On the other hand, designing tiny DNN networks from scratch [17,2,15] is another widely used technique in tiny deep learning. Tiny DNN networks can be either designed manually [17] or using AutoML approaches [2]. We use MobileNet-tiny, which is designed based on neural architecture search method [12]. It has only 0.4M parameters and is significantly smaller than WideResNet networks. From an architectural perspective, there has not been any robustness exploration for these tiny networks. This paper analyses their robustness and finds some insights into designing tiny networks.

3 Exploring Robustness

To analyse robustness for configurations of MobileNet-tiny, we need to define the baseline architecture and metrics used for evaluation (Section 3.1). In sections 3.3 and 3.4, we demonstrate the results for expanding the network based on the width-multiplier and expand-ratio. In section 3.5, we theoretically and empirically show the relation between these hyper-parameters with the Lipschitz constant.

3.1 Baseline Network and Evaluation Metrics

We take the MobileNet-tiny [12] network as the baseline architecture. Figure 1(a) show the overall architecture of this network. This architecture consists of 6 inverted bottleneck blocks that we expand them to generate a wider network. We denote the width-multiplier and expand-ratio for all inverted bottleneck blocks as W and E , respectively. For the baseline network, the width-multiplier and expand-ratio are set to 0.35 and 6, respectively. We explore the impact of W and E while other hyper-parameters are fixed. In terms of metrics, we consider different aspects of the performance of the tiny network as follows.

Natural Accuracy: the ratio of examples that are correctly classified as:

$$Acc_{Nat} = \frac{\#\{x : \forall x \in D, f(\theta; x) == y\}}{\#examples} \quad (1)$$

D , $f(\theta; \cdot)$, and y indicate the test dataset, network, and correct labels.

Robust Accuracy: ratio of adversarial data that are correctly classified as:

$$Acc_{Rob} = \frac{\#\{x : \forall \hat{x} \in \mathcal{B}(x, \epsilon), f(\theta; \hat{x}) == y\}}{\#examples} \quad (2)$$

where \mathcal{B} indicates the l_p norm ball around the natural example x .

Perturbation instability: the difference between the prediction of the network for natural and adversarial examples. We use the KL -divergence statistical measure to compute the perturbation instability as:

$$Pert_{Inst} = \mathbb{E}_{x \sim D} KL(f(\theta; x), f(\theta; \hat{x})) \quad (3)$$

Where \mathbb{E} and \hat{x} indicate the expectation function and adversarial example.

3.2 Experimental setting

We train the explored networks using TRADES [24] on the CIFAR-10 training data. For adversarial training settings, we use l_∞ norm by setting the maximum perturbation size to $\epsilon = 8/255$, and use 10-steps PGD with step size $\alpha = 2/255$. For robustness evaluation, we use a 20-PGD attack to generate adversarial data with the same perturbation size ($\epsilon = 8/255$) on CIFAR-10 test data.

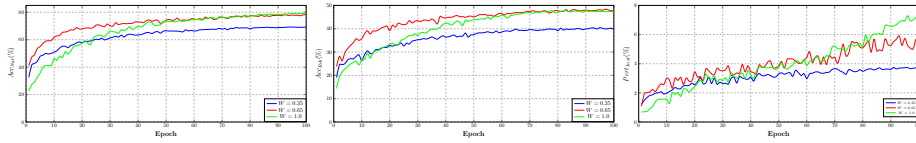


Fig. 2: The dynamics of natural accuracy, robust accuracy and perturbation instability with regard to training epochs and fixed expand-ratio ($E = 6$) on CIFAR-10 dataset.

3.3 Exploring Different width-multipliers

We first explore the impact of different width-multipliers on the baseline MobileNet-tiny architecture. For each inverted bottleneck block with the same expand ratios ($E = 6$), we explore different width multipliers ($W = \{0.35, 0.65, 1.0\}$). Since we need to have a tiny network that is suitable for edge devices, we do not use a larger value than 1.0 for the width-multiplier. The dynamics of natural accuracy, robust accuracy, and perturbation instability measures with regard to the training epochs for these three adversarially trained networks are plotted in figure 2. By following the dynamics of the metrics in different epochs, we find that increasing the width-multiplier leads to improve natural and robust accuracy, but it also increases the perturbation instability. The other important finding is that the improvement of natural and robust accuracy is negligible (or sometimes reduced) after a threshold for width-multiplier. Table 1 shows the results for training MobilNet-tiny with different width-multiplier and fixed expand-ratio ($E = 6$). As shown in the table, increasing the width multiplier from 0.35 in the baseline network to 0.65 improves the natural and robust accuracy by 9.01% and 8.09% but moving from 0.65 to 1.0 have 0.91% improvement in natural and hurt robust accuracy by 0.89%. In addition by increasing the width-multiplier from 0.35 to 0.65 and 1.0 the perturbation instability increase by 20.13% and 26.46%. It means that adversarial training increases the difference between the prediction of the network for natural and adversarial examples.

3.4 Exploring Different expand-ratios

We also explore the impact of the expand-ratio hyper-parameter on the robustness of MobileNet-tiny. The baseline MobileNet-tiny has a fixed width-multiplier of $W = 0.35$ for all inverted bottleneck blocks. We investigate different values for expand-ratio as $E = \{6, 10, 20, 29\}$. Like the width-multiplier, we do not significantly alter the expand-ratio to remain in the tiny regime. The dynamics of natural accuracy, robust accuracy, and perturbation instability metrics in the training epochs are plotted in figure 3. Furthermore, table 2 indicates the best results for different expand-ratios. We find that the expand-ratio has a similar effect as the width-multiplier. However, increasing it until a threshold improves the natural and robust accuracy, but it compromises the perturbation stability. To compare the impact of the width-multiplier and expand-ratio, we set these hyper-parameters to have a similar number of parameters. To this end, we

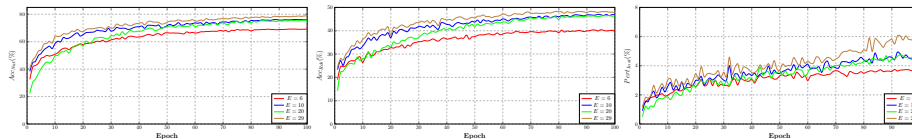


Fig. 3: The dynamics of natural accuracy, robust accuracy and perturbation instability with regard to training epochs and fixed width-multiplier ($W = 0.35$) on CIFAR-10 dataset.

made a network by increasing the width-multiplier to 0.65 with the same expand ratio as the baseline and created another network by using 29 for expand-ratio and the same width multiplier as the baseline. Both networks have almost 1.08 million parameters. The expanded network with width-multiplier shows 0.17% better robust accuracy than the expanded network with expand-ratio. In terms of perturbation instability, expanding the network with width-multiplier, increases the instability by 2.59% compared to the expand-ratio.

3.5 Theoretical and Empirical Lipschitz constant

Recent works [9,21] formally prove the relation between Lipschitzness and perturbation instability. They show that smaller Lipschitzness (small Lipschitz constant) leads to decreased perturbation instability and improved robustness. In this section, we first theoretically show the relation between width-multiplier and expand-ratio hyper-parameters in MobileNet-tiny with perturbation instability. Then we empirically analyse this relation to support theoretical findings. The Lipschitz constant L of MobileNet-tiny architecture measures the rate of change in the output of the network by changing the input as:

$$\|f(\theta; x) - f(\theta; \hat{x})\| \leq L \cdot \|x - \hat{x}\| \quad (4)$$

The expected Lipschitz constant for MobileNet-tiny with n inverted bottleneck blocks with width h and m middle channels is upper bounded by:

$$L(f(\theta; x) \leq (\sqrt{W \cdot h} + \sqrt{E \cdot m})^n \quad (5)$$

where W and E show the width-multiplier and expand-ratio hyper-parameters in MobileNet-tiny. This formulation is the conclusion of the theorem in [10]

Table 1: The results of the expanded network by altering width-multiplier and fixed expand-ratio ($E = 6$) (Last-checkpoint)

Expand-Ratio	Width-Multiplier	#MACs	#Params	Acc _{Nat} (%)	Acc _{Rob} (%)	Pert _{Inst}	Lipchitz L
$E = 6$	$W = 0.35$	15.98	0.404	69.23	39.91	3.75	67.24
	$W = 0.65$	45.08	1.088	78.24	48.00	5.63	80.77
	$W = 0.1$	90.68	2.278	79.15	47.11	7.27	85.03

Table 2: The results of the expanded network by altering expand-ratio and fixed width-multiplier ($W = 0.35$) (Last-checkpoint).

<i>Width-Multiplier</i>	<i>Expand-Ratio</i>	<i>#MACs</i>	<i>#Params</i>	<i>Acc_{Nat} (%)</i>	<i>Acc_{Rob}(%)</i>	<i>Pert_{Inst}</i>	<i>Lipchitz L</i>
$W = 0.35$	$E = 6$	15.979	0.404	69.23	39.91	3.75	67.24
	$E = 10$	23.73	0.523	75.98	46.81	4.76	73.4
	$E = 20$	43.206	0.8218	75.54	46.3	4.51	71.26
	$E = 29$	58.63	1.086	78.72	47.83	6.28	-82.88

for WideResNet large-scale network. This establishes the connection between hyper-parameters in inverted bottleneck blocks and the Lipschitz constant and perturbation instability. This theoretical analysis shows that increasing the width-multiplier and expand-ratio increases the perturbation instability. Additionally, this formulation shows that adding more inverted blocks to the baseline network (more depth) exponentially increases the perturbation instability. Our empirical Lipschitz constant evaluation supports our theoretical findings:

$$L = \mathbb{E}_{x \sim D} \max_{\hat{x} \in \mathcal{X}} \frac{\|f(\theta; x) - f(\theta; \hat{x})\|}{\|x - \hat{x}\|} \quad (6)$$

Where \mathcal{X} is the ϵ -ball around the x and \hat{x} is adversarial data generated by PGD. We compute this metric for different hyper-parameter configurations. The results are indicated in figure 4 and tables 2,1. We can observe that when the width of the network increase by using a larger width-multiplier and expand-ratio, the empirical Lipschitz constant also increases. Theoretical and empirical analysis of network perturbation instability agrees.

4 Conclusion

This paper analyses the robustness of tiny deep neural networks (MobileNet-tiny) from an architectural perspective. Specifically, we explore how the width of the inverted bottleneck blocks affects the robustness. To generate different architectures, we change the width-multiplier and expand-ratio hyper-parameters

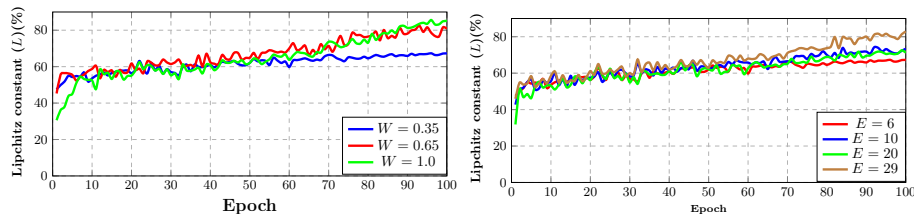


Fig. 4: The dynamics of Lipschitz constant with regard to training epochs and altering width-multipliers (left) and expand-ratios(right)

that increase the number of channels in inverted bottleneck blocks. Our findings are: 1) Although increasing the width of the blocks in MobileNet-tiny can improve the natural and robust accuracy, it also increases the perturbation instability. 2) After a threshold, expanding the width of the network cannot only improve the natural and robust accuracy but also increase the perturbation instability. We also find theoretically and empirically the relationship between width-multiplier and expand-ratio with the Lipchitz constant, which directly relates to perturbation instability. It shows that increasing the number of blocks and expanding the width of the network increase the Lipchitz constant. Our work provides valuable insights into designing robust tiny networks against adversarial data.

5 Acknowledgement

This work was supported in part by the European Union through European Social Fund in the frames of the “Information and Communication Technologies (ICT) program” and by the Swedish Innovation Agency VINNOVA project “AutoDeep” and “SafeDeep”. The computations were enabled by the supercomputing resource Berzelius provided by National Supercomputer Centre at Linköping University and the Knut and Alice Wallenberg foundation.

References

1. Banbury, C., Zhou, C., Fedorov, I., Matas, R., Thakker, U., Gope, D., Janapa Reddi, V., Mattina, M., Whatmough, P.: Micronets: Neural network architectures for deploying tinymml applications on commodity microcontrollers. *Proceedings of machine learning and systems* **3**, 517–532 (2021)
2. Cai, H., Zhu, L., Han, S.: Proxylessnas: Direct neural architecture search on target task and hardware. *arXiv preprint arXiv:1812.00332* (2018)
3. Carlini, N., Wagner, D.: Towards evaluating the robustness of neural networks. In: *2017 IEEE Symposium on Security and Privacy (SP)*. pp. 39–57. *Ieee* (2017)
4. Croce, F., Hein, M.: Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In: *International conference on machine learning*. pp. 2206–2216. *PMLR* (2020)
5. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*. pp. 248–255 (2009). <https://doi.org/10.1109/CVPR.2009.5206848>
6. Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Xiao, C., Prakash, A., Kohno, T., Song, D.: Robust physical-world attacks on deep learning visual classification. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 1625–1634 (2018)
7. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572* (2014)
8. Han, S., Mao, H., Dally, W.J.: Deep compression: Compressing deep neural networks with pruning, trained quantization and Huffman coding. *arXiv preprint arXiv:1510.00149* (2015)

9. Hein, M., Andriushchenko, M.: Formal guarantees on the robustness of a classifier against adversarial manipulation. *Advances in neural information processing systems* **30** (2017)
10. Huang, H., Wang, Y., Erfani, S., Gu, Q., Bailey, J., Ma, X.: Exploring architectural ingredients of adversarially robust deep neural networks. *Advances in Neural Information Processing Systems* **34**, 5545–5559 (2021)
11. Kannan, H., Kurakin, A., Goodfellow, I.: Adversarial logit pairing. *arXiv preprint arXiv:1803.06373* (2018)
12. Lin, J., Chen, W.M., Lin, Y., Gan, C., Han, S., et al.: Mccnet: Tiny deep learning on iot devices. *Advances in Neural Information Processing Systems* **33**, 11711–11722 (2020)
13. Loni, M., Mousavi, H., Riazati, M., Daneshtalab, M., Sjödin, M.: Tas: ternarized neural architecture search for resource-constrained edge devices. In: *2022 Design, Automation & Test in Europe Conference & Exhibition (DATE)*. pp. 1115–1118. *IEEE* (2022)
14. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083* (2017)
15. Mousavi, H., Loni, M., Alibeigi, M., Daneshtalab, M.: Pr-darts: Pruning-based differentiable architecture search. *arXiv preprint arXiv:2207.06968* (2022)
16. Rade, R., Moosavi-Dezfooli, S.M.: Reducing excessive margin to achieve a better accuracy vs. robustness trade-off. In: *International Conference on Learning Representations* (2022), <https://openreview.net/forum?id=Azh9QBQ4tR7>
17. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C.: Mobilenetv2: Inverted residuals and linear bottlenecks. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 4510–4520 (2018)
18. Shafique, M., Naseer, M., Theodoridis, T., Kyrkou, C., Mutlu, O., Orosa, L., Choi, J.: Robust machine learning systems: Challenges, current trends, perspectives, and the road ahead. *IEEE Design & Test* **37**(2), 30–57 (2020)
19. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R.: Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199* (2013)
20. Wang, Y., Zou, D., Yi, J., Bailey, J., Ma, X., Gu, Q.: Improving adversarial robustness requires revisiting misclassified examples. In: *International Conference on Learning Representations* (2020)
21. Weng, T.W., Zhang, H., Chen, P.Y., Yi, J., Su, D., Gao, Y., Hsieh, C.J., Daniel, L.: Evaluating the robustness of neural networks: An extreme value theory approach. *arXiv preprint arXiv:1801.10578* (2018)
22. Wu, B., Chen, J., Cai, D., He, X., Gu, Q.: Do wider neural networks really help adversarial robustness? *Advances in Neural Information Processing Systems* **34**, 7054–7067 (2021)
23. Xie, C., Yuille, A.: Intriguing properties of adversarial training at scale. *arXiv preprint arXiv:1906.03787* (2019)
24. Zhang, H., Yu, Y., Jiao, J., Xing, E., El Ghaoui, L., Jordan, M.: Theoretically principled trade-off between robustness and accuracy. In: *International conference on machine learning*. pp. 7472–7482. *PMLR* (2019)
25. Zhu, Z., Liu, F., Chrysos, G., Cevher, V.: Robustness in deep learning: The good (width), the bad (depth), and the ugly (initialization). *Advances in Neural Information Processing Systems* **35**, 36094–36107 (2022)
26. Zi, B., Zhao, S., Ma, X., Jiang, Y.G.: Revisiting adversarial robustness distillation: Robust soft labels make student better. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 16443–16452 (2021)