TUNING NEURAL ODE NETWORKS TO INCREASE ADVERSARIAL ROBUSTNESS IN IMAGE FORENSICS

Roberto Caldelli[♣,♠], *Fabio Carrara*[◇], *Fabrizio Falchi*[◇]

[♣]National Inter-University Consortium for Telecommunications (CNIT), Florence, Italy,
[♠]Universitas Mercatorum, Rome, Italy, [◇]ISTI CNR, Pisa, Italy

ABSTRACT

Although deep-learning-based solutions are pervading different application sectors, many doubts have arisen about their reliability and, above all, their security against threats that can mislead their decision mechanisms. In this work, we considered a particular kind of deep neural network, the Neural Ordinary Differential Equations (N-ODE) networks, which have shown intrinsic robustness against adversarial samples by properly tuning their tolerance parameter at test time. Their behaviour has never been investigated in image forensics tasks such as distinguishing between an original and an altered image. Following this direction, we demonstrate how tuning the tolerance parameter during the prediction phase can control and increase N-ODE's robustness versus adversarial attacks. We performed experiments on basic image transformations used to generate tampered data, providing encouraging results in terms of adversarial rejection and preservation of the correct classification of pristine images.

***Index Terms*—** Image Forensics, Deep Learning, Neural ODE Networks, Adversarial Samples.

1. INTRODUCTION

Nowadays, it is becoming evident that deep-learning-based solutions will dominate many of the different sectors of our everyday life, mainly for their impressive performances compared to previous classical methodologies. On the contrary, diverse doubts have arisen about their reliability and, above all, their security against malevolent threats that can crucially mislead their decision mechanisms. In particular, input tampering with adversarial attacks represents one of the most studied threat, as it impacts virtually all deep neural models. For this reason, recent years experienced a florid literature of defenses against adversarial attacks. Proposed approaches to mitigate adversarial attacks are roughly divided in *adversarial detection* and *adversarial robustness* methodologies. The former aims at recognizing an attack and rejecting the induced malicious output, e.g., using auxiliary detection models [1, 2], statistical tests [3, 4], and neighborhood analysis [5]. The latter instead focuses on maintaining the model's performance even when under attack; examples in this category include adversarial training [6, 7], input or features smoothing/denoising [8, 9], and gradient masking [10].

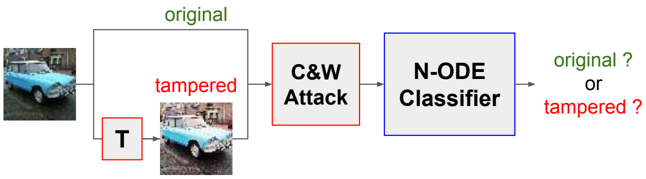

Fig. 1. **Operative framework.** The T block stands for tampering transformation. The model have to discern tampered from original data, while the attack tries to flip this decision.

This work follows the latter direction of increasing adversarial robustness in image forensic tasks. Specifically, we consider a particular kind of deep neural networks, the Neural Ordinary Differential Equations (N-ODE) networks, which has been demonstrated to be interesting both for their intrinsic efficiency and for a specific characteristic that allows improving their robustness against adversarial samples through a tunable tolerance parameter. It has been verified that decoupling the tolerance values used during training from the one adopted at test time can strongly decrease the attack success rate [11]. However, this kind of network has been tested for general image classification. In contrast, their behaviour has never been investigated in image forensics tasks to distinguish, in a binary way, between original and fraudulently altered contents. This is particularly interesting because, usually, the behaviour of trained models (e.g., based on CNNs) under adversarial attacks is not so similar to what happens in common image classification applications [12]. Following this direction, we demonstrate how using the tolerance parameter of N-ODEs during the prediction phase enables us to tune and increase the robustness of our model against state-of-the-art white-box adversarial attacks. Specifically, we analysed how a defensive strategy based on such a property can be devised to improve

This work was partially funded by AI4Media - A European Excellence Centre for Media, Society and Democracy (EC, H2020 n. 951911). The authors would like to thank Prof. Mauro Barni for helpful suggestions and discussions on this topic. Code is available at `https://github.com/fabiocarrara/node-forensic-adversarial`.

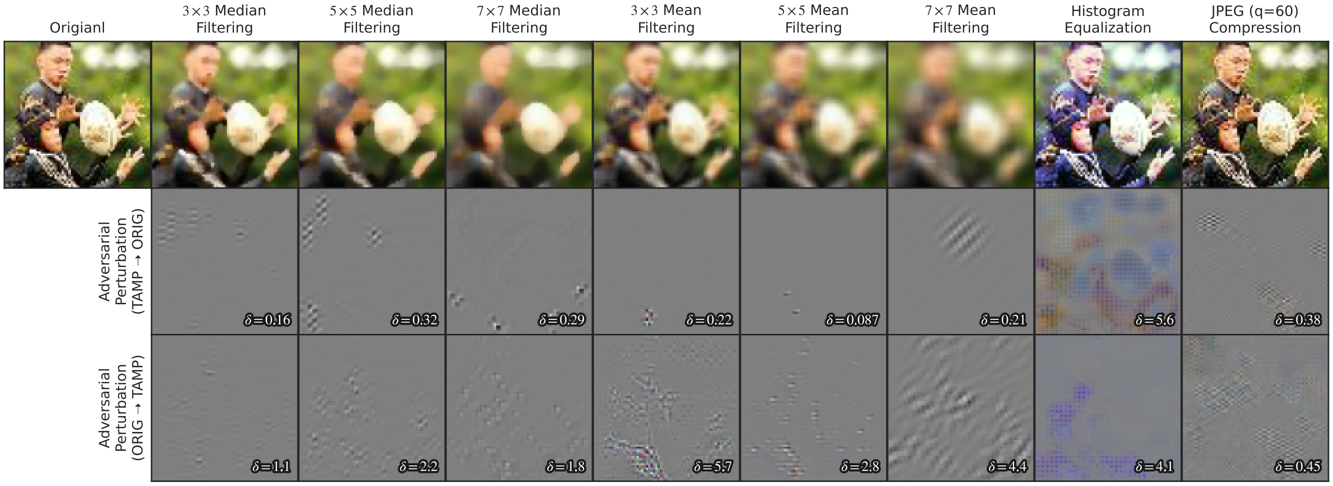

Fig. 2. **Tampering transformations and examples of adversarial perturbations found by the C&W attack.** The 2nd and 3rd rows show adversarial perturbations (δ = L_2-norm) changing the classification from tampered to original and viceversa.

robustness against *Carlini & Wagner (C&W)* [13] adversarial examples in an image forensic scenario. Experimental tests, carried out on basic image transformations used to generate tampered data, provide encouraging results both in terms of adversarial rejection and, at the same time, preservation of the correct classification of pristine images. The paper is organized as follows: after this introductory section, N-ODE networks are presented in Section 2, while Section 3 discusses the proposed defense strategy. Section 4 describes the experimental results and Section 5 draws conclusions.

2. N-ODE AND THE TOLERANCE PARAMETER

A Neural ODE (*Ordinary Differential Equations*) net is a parametric model which contains an *ODE block* whose computation is defined by a parametric ordinary differential equation [14]. The solution of the ODE comprises the output of the ODE block. Formally, let $\frac{d\mathbf{x}(t)}{dt} = f(\mathbf{x}(t), t, \theta)$ an ODE with state $\mathbf{x}(t) \in \mathbb{R}^n$ that continuously evolves through time following the dynamics defined by $f(\cdot)$ parametrized by θ. Let also $\mathbf{x}(t_0) = \mathbf{x}_0$ the input of the ODE block coinciding with the initial state at time t_0 of the ODE. The output of the ODE block is $\mathbf{x}(t_1)$ at time $t_1 > t_0$ computed by integration:

$$\mathbf{x}(t_1) - \mathbf{x}(t_0) = \int_{t_0}^{t_1} \mathrm{d}\mathbf{x}(t) = \int_{t_0}^{t_1} f(\mathbf{x}(t), t, \theta)\mathrm{d}t \qquad (1)$$

The above integral can be computed with standard ODE solvers, such as Runge-Kutta or Multi-step methods. Thus, the computation performed by the ODE block can be formalized as a call to a generic ODE solver. Generally, in image classification applications, the function $f(\cdot)$ is implemented by means of a small trainable convolutional neural network. During the training phase, the gradients of the output $\mathbf{x}(t_1)$ with respect to the input $\mathbf{x}(t_0)$ and the parameter θ can be obtained using the adjoint sensitivity method. This consists of solving an additional ODE in the backward pass. Once the gradient is obtained, standard gradient-based optimization can be applied. In this work, we consider a N-ODE image classifier constituted by a single ODE block responsible for the whole feature extraction chain. This block is preceded by a limited pre-processing stage comprised of a single 256-filter 2-strided 3x3-kernels convolutional layer with no activation function that linearly maps the input image in the ODE state space. The $f(\cdot)$ function in the ODE block is implemented similarly as a standard residual block used in ResNets; it comprises the sequence of layers GN-GeLU-Conv-GN-GeLU-Conv-GN, where GN stands for *Group-Normalization* with group size of 32, *GeLU* is the Gaussian Error Linear Unit, and *Conv* is a 256-filter 3x3-kernels convolutional layer. After the ODE block, a classification head comprised of global average-pooling and a fully-connected layer provides the output logits. As ODE solver, we use an adaptive solver that chooses the best step size when integrating the ODE solution given a tolerance parameter τ. The tolerance controls the trade-off between the computational cost and precision of the solution. The use of an adaptive ODE solver induces some peculiar properties. It has been demonstrated [11] that such a parameter can also be used to increase N-ODE robustness to adversarial attacks without penalizing performances on pristine samples. In particular, this can be achieved by diversifying the value of τ_{attack} used by the attacker (that is generally set to the same adopted during the training of the model, but not necessarily) from that one used at test time τ_{test}. Stated in another way, when $\tau_{\text{test}} = \tau_{\text{attack}}$, we have the maximum attack success rate.

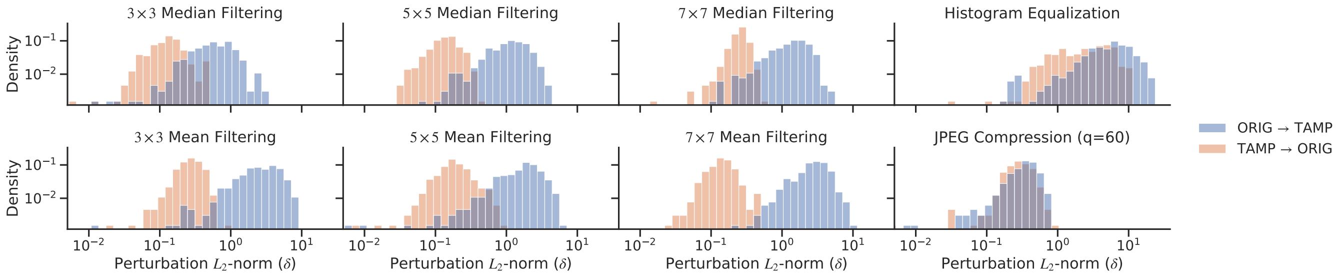

Fig. 3. **Distribution of perturbation norms** for different tampering transformations.

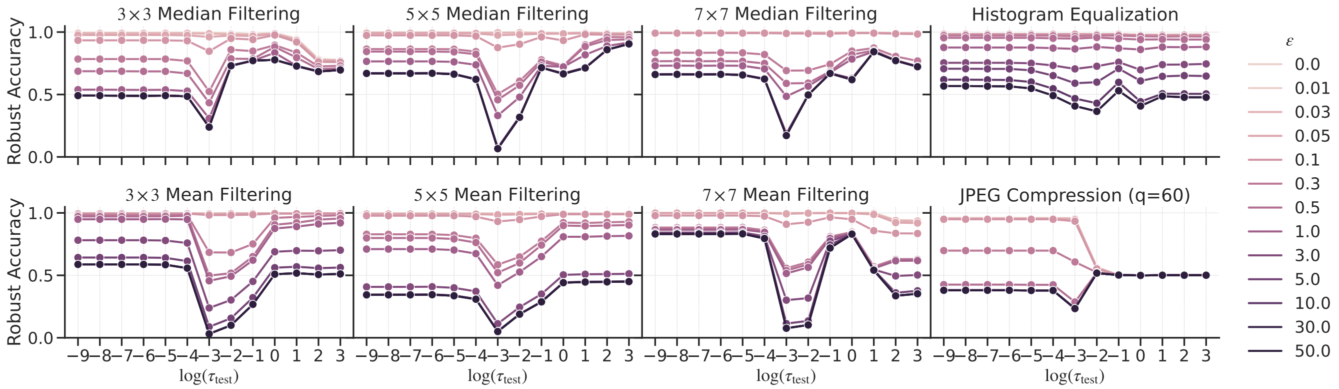

Fig. 4. **Robust Accuracy vs Test-time Tolerance τ_{test}** for different tampering transformations and maximum attack power ε.

3. C&W ATTACK AND DEFENSE STRATEGY

Despite their unique properties and good performances in image classification, also N-ODE nets suffer from white-box gradient-based adversarial attacks [15]. However, to the best of our knowledge, N-ODEs have never been tested in a classical image forensic task, that is, distinguishing between a pristine image and an altered one, neither from the point of view of accuracy nor concerning the robustness to adversarial attacks. To evaluate this issue, we considered the framework depicted in Fig.1 where an N-ODE net, trained to classify original and tampered images, is called to make a decision on adversarial samples generated by means of one of the most efficient state-of-the-art white-box attack — the *Carlini&Wagner ($C\&W$)* method [13, 16]. The rationale behind the $C\&W$ attack is to minimize, at each iteration, the highest confidence among non-target classes while keeping the smallest possible distortion. Such a minimization is performed in the $tanh$ space to help regularizing the gradient in the extreme regions of the perturbation space. We consider certain computational and perceptual budgets the attacker can spend, respectively given by $n_{\text{it}} = 1000$, the maximum number of iterations of the $C\&W$ algorithm, and ε, the maximum allowed L_2-norm of the perturbation, that we vary in our experiments. The $C\&W$ attack is usually very powerful and achieves a high success rate with a very limited distortion; to apply $C\&W$, the attacker needs access to the neural network and, in the case of N-ODE nets, to the value of the tolerance parameter τ_{train} used during the training phase of the model. Without defense strategies, the victim model would operate using $\tau_{\text{test}} = \tau_{\text{train}}$, and thus the best configuration for the $C\&W$ attack is $\tau_{\text{attack}} = \tau_{\text{train}}$.

4. TEST SET-UP AND EXPERIMENTAL RESULTS

In this section, we introduce the experimental setup and discuss achieved results to verify the capacity of N-ODE nets to increase their robustness to adversarial examples by tuning the tolerance parameter τ.

4.1. The test set-up

The experimental setup has been structured to have three sets of images: original images, tampered images, and a set of adversarial samples generated from the first two sets. We considered the *TinyImageNet*[17] dataset as source for original samples. Tampered images have been generated from original ones by applying simple distortions commonly used for malevolent image modification in a forensic scenario. Such transformations are filtering (mean or median with different window sizes), histogram equalization, and JPEG compression, as shown in the top row of Figure 2. For each tampering transformation, we prepared a training dataset composed of 2000 original images (10 random images per class in the TinyImageNet train set) and their tampered version, for a total of 4000 training images. In the same manner, we built a validation set and a test set picking 2000 and 200 images from the validation and test set of Tiny ImageNet respectively, for a total of 4000 validation images and 400 test images. We trained N-ODE binary image classifiers (one for each tampering transformation) to discern original images from tampered ones using the respective train and validation subsets. During training, we used a solver tolerance $\tau_{\text{train}} = 10^{-3}$ and dropout with probability 0.2 after the global average-pooling layer of the classification head. The binary cross-entropy loss is minimized with the Adam optimizer and a cosine annealing learning rate scheduling with an initial value of 10^{-3} until the validation loss plateaus. After the network converged, we applied the $C\&W$ attack to the test subset to create the corresponding adversarial examples for both kinds of images (original and tampered), fooling the classifier and exchanging the predicted class respectively. During the attack, we set the N-ODE solver tolerance $\tau_{\text{attack}} = 10^{-3}$ and varied the maximum perturbation L_2-norm $\varepsilon \in [0, 50]$. We refer to the L_2-norm of the perturbation obtained by the attack as δ, and consider an attack succesful only if $\delta \leq \varepsilon$. Figure 2 shows an example of the obtained adversarial samples. It is worth noting from Figure 3 how the attack generally needs smaller adversarial perturbations (see red bars with lower values of δ) when acting on tampered images to make them look original.

4.2. Experimental results

Figure 4 shows the accuracy of the classifier when varying the test-time tolerance τ_{test} for different modifications and for different $C\&W$ attack strength (maximum perturbation L_2-norm permitted ε). The accuracy on original images is reported as $\varepsilon = 0$. Note that the maximum attack success rate occurs when $\tau_{\text{test}} = \tau_{\text{attack}} = 10^{-3}$, witnessed by the minimum value of the accuracy for each tampering transformation. For instance, for the case of 5×5 median filtering, the accuracy is drastically decreased to almost zero for higher values of attack power ($\varepsilon \geq 5.0$). Correspondingly, it can be favorably appreciated how the accuracy can be restored by decreasing (left on the x-axis) or increasing (right on the x-axis) τ_{test} in order to create a mismatch with respect to τ_{attack}. Acceptable values of accuracy around 70% or higher can be reached in most cases. In most cases, increasing τ_{test} (and thus injecting an approximation in the ODE solution) do not degrade the performance on original images (see $\varepsilon = 0$ lines in Figure 4), except for extreme tolerance values ($\tau_{\text{test}} \geq 10$ in 3×3 median and 7×7 mean filtering). The only exceptions seem to be the cases of histogram equalization and JPEG tampering. The former, though it slightly shows a behaviour similar to mean and median filtering, appears to be intrinsically robust to the $C\&W$ attack with higher overall accuracy values on average. The latter behaves as expected for $\tau_{\text{test}} < \tau_{\text{train}}$; on the contrary, when we introduce a coarser approximation than the one used to train the ODE-Net ($\tau_{\text{test}} > \tau_{\text{train}}$), the classifier is not capable anymore of grasping the differences between original and JPEG-tampered samples, and the performance drastically collapses independently from the attacker actions. We plan to get a deeper insight into this case by exploring more quality factors values in future work. The same phenomena observed in Figure 4 also occurs when measuring AuROC values that we report in Table 1 for some transformations to avoid redundancy. Orange and red cells, in correspondence of $\tau_{\text{test}} = \tau_{\text{attack}}$, highlight a high effectiveness of the adversarial attack that can be recovered by increasing (moving right) or decreasing (moving left) τ_{test} as well.

		$\log_{10}(\tau_{\text{test}})$									
T	ε	-6	-5	-4	-3*	-2	-1	0	1	2	3
7×7 Mean	-	100	100	100	100	100	100	100	100	100	100
	1.0	98.4	98.4	98.2	94.0	94.1	98.6	99.4	96.7	96.2	96.5
	3.0	88.9	88.8	86.6	49.4	45.7	93.1	96.7	76.6	60.2	61.6
	5.0	81.9	81.9	77.8	12.8	10.9	85.8	94.5	62.2	32.7	34.4
	10.0	80.6	80.5	76.0	5.5	4.6	82.3	94.2	59.4	28.0	29.7
5×5 Median	-	100	100	100	100	100	100	100	100	100	100
	0.5	94.0	93.9	93.6	87.5	91.5	95.8	98.3	96.9	98.6	99.0
	1.0	83.5	83.4	81.6	59.6	70.7	90.4	96.0	90.5	97.0	98.1
	3.0	69.3	68.9	64.6	5.5	35.4	87.5	83.6	80.0	95.7	97.3
	5.0	68.8	68.4	64.1	4.0	34.2	87.5	82.7	79.6	95.7	97.3
7×7 Median	-	99.9	99.9	99.9	99.9	99.9	99.9	99.8	100	99.8	99.8
	0.5	95.1	95.0	94.8	91.4	95.0	95.6	97.3	97.5	98.0	97.3
	1.0	86.8	86.7	85.9	70.8	86.5	92.7	91.2	95.8	96.5	95.2
	3.0	70.2	69.9	66.8	10.9	65.2	91.1	65.0	94.8	95.5	93.0
	5.0	69.7	69.5	66.3	8.1	64.8	90.5	63.3	94.8	95.5	92.9

$^* = \tau_{\text{attack}}$

Table 1. AuROC (%) vs ODE Solver Tolerance τ_{test}. Adversarial attacks are performed with a solver tolerance $\tau_{\text{attack}} = 10^{-3}$ and a maximum L_2 perturbation norm of ε.

5. CONCLUSIONS

In this work, we considered N-ODE networks and investigated their intrinsic robustness against adversarial samples in image forensic tasks such as distinguishing between an original and an altered image. We demonstrated that by properly tuning the N-ODE tolerance parameter at test time with respect to that used by the attacker, it is possible to increase robustness versus $C\&W$ attack. Experiments on basic image transformations used to generate tampered data provided satisfactory results in adversarial rejection and maintained classification of pristine images. Future works will comprise testing on more challenging image tampering operations.

6. REFERENCES

[1] Fei Zuo and Qiang Zeng, "Exploiting the sensitivity of l2 adversarial examples to erase-and-restore," in *Proceedings of the 2021 ACM Asia Conference on Computer and Communications Security*, 2021, pp. 40–51.

[2] Fabio Carrara, Fabrizio Falchi, Roberto Caldelli, Giuseppe Amato, and Rudy Becarelli, "Adversarial image detection in deep neural networks," *Multimedia Tools and Applications*, vol. 78, no. 3, pp. 2815–2835, 2019.

[3] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin, "A simple unified framework for detecting out-of-distribution samples and adversarial attacks," *Advances in neural information processing systems*, vol. 31, 2018.

[4] Kathrin Grosse, Praveen Manoharan, Nicolas Papernot, Michael Backes, and Patrick D. McDaniel, "On the (statistical) detection of adversarial examples," *CoRR*, vol. abs/1702.06280, 2017.

[5] Gilad Cohen, Guillermo Sapiro, and Raja Giryes, "Detecting adversarial samples using influence functions and nearest neighbors," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 14453–14462.

[6] Cihang Xie, Mingxing Tan, Boqing Gong, Alan Yuille, and Quoc V Le, "Smooth adversarial training," *arXiv preprint arXiv:2006.14536*, 2020.

[7] Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel, "Ensemble adversarial training: Attacks and defenses," *arXiv preprint arXiv:1705.07204*, 2017.

[8] Aamir Mustafa, Salman H Khan, Munawar Hayat, Jianbing Shen, and Ling Shao, "Image super-resolution as a defense against adversarial attacks," *IEEE Transactions on Image Processing*, vol. 29, pp. 1711–1724, 2019.

[9] Yassine Bakhti, Sid Ahmed Fezza, Wassim Hamidouche, and Olivier Déforges, "Ddsa: A defense against adversarial attacks using deep denoising sparse autoencoder," *IEEE Access*, vol. 7, pp. 160397–160407, 2019.

[10] Nicolas Papernot, Patrick McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami, "Distillation as a defense to adversarial perturbations against deep neural networks," in *2016 IEEE symposium on security and privacy (SP)*. IEEE, 2016, pp. 582–597.

[11] Fabio Carrara, Roberto Caldelli, Fabrizio Falchi, and Giuseppe Amato, "Defending neural ODE image classifiers from adversarial attacks with tolerance randomization," in *Pattern Recognition. ICPR International Workshops and Challenges*, Cham, 2021, pp. 425–438, Springer International Publishing.

[12] M. Barni, K. Kallas, E. Nowroozi, and B. Tondi, "On the transferability of adversarial examples against cnn-based image forensics," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 8286–8290.

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

[14] Tian Qi Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud, "Neural ordinary differential equations," in *Advances in Neural Information Processing Systems*, 2018, pp. 6572–6583.

[15] Hanshu Yan, Jiawei Du, Vincent Y. F. Tan, and Jiashi Feng, "On robustness of neural ordinary differential equations," in *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. 2020, OpenReview.net.

[16] Nicholas Carlini and David Wagner, "Towards evaluating the robustness of neural networks," in *2017 IEEE SP*. IEEE, 2017, pp. 39–57.

[17] Patryk Chrabaszcz, Ilya Loshchilov, and Frank Hutter, "A downsampled variant of imagenet as an alternative to the cifar datasets," *arXiv preprint arXiv:1707.08819*, 2017.