Transfer Robustness to Downstream Tasks Through Sampling Adversarial Perturbations

Ivan Reyes-Amezcua
CINVESTAV
Guadalajara, Mexico.
ivan.reyes@cinvestav.mx

Gilberto Ochoa-Ruiz
Tecnologico de Monterrey
Guadalajara, Mexico
gilberto.ochoa@tec.mx

Andres Mendez-Vazquez
CINVESTAV
Guadalajara, Mexico
andres.mendez@cinvestav.mx

Abstract

Due to the vulnerability of deep neural networks to adversarial attacks, adversarial robustness has grown to be a crucial problem in deep learning. Recent research has demonstrated that even small perturbations to the input data can have a large impact on the model's output, exposing them susceptible to malicious attacks. In this work, we propose Delta Data Augmentation (DDA), a data augmentation method for enhancing transfer robustness by sampling extracted perturbations from trained models against adversarial attacks. The main idea of our work is to generate adversarial perturbations and to apply them to downstream datasets in a data augmentation fashion. Here we demonstrate, through extensive experimentation the advantages of our data augmentation method over the current State-of-the-Art in Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD) attacks for CIFAR10 dataset.

1. Introduction

The research in the field of adversarial robustness for deep learning aims to increase the robustness of models to adversarial attacks [3, 4, 11]. These attacks are deliberate attempts to trick a model by purposefully introducing undetectable perturbations to the input data, leading the algorithm to misclassify or make inaccurate predictions [1, 9]. Applications like autonomous vehicles [38], medical diagnosis [2, 36], and fraud detection [8] are all susceptible to adversarial attacks, which can have major repercussions. Thus, it has become crucial to conduct research on increasing the adversarial robustness of deep learning models in order to make such systems safe and useful in real settings.

The recent literature has seen a surging interest in this field, resulting in a large number of techniques to improve the robustness of deep learning models to adversarial attacks, such as adversarial training [3, 15, 35], and data augmentation [12, 13]. For example, to enhance the model's capacity to recognize and fend off hostile attacks, adversarial training entails supplementing the training data with adversarial samples. Conversely, data augmentation creates new data from existing data in order to expand the size of a dataset [28, 33]. This leads to an overall improvement in the robustness of the model by exposing it to a wider range of variations.

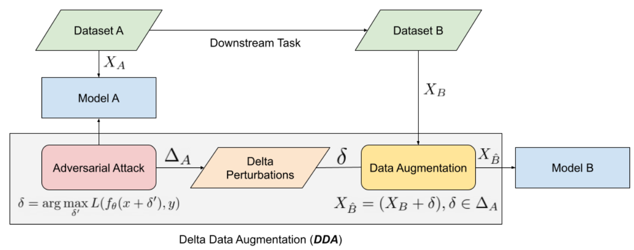


Figure 1. Overview of Delta Data Augmentation (DDA). A method for data augmentation through sampling adversarial perturbations δ from upstream trained models to downstream tasks.

Our proposal, Delta Data Augmentation (DDA) (Fig. 1) solves the crucial problem of transfer robustness in deep learning. When there is a lack of labeled data, our transfer robustness approach only requires training a model on one dataset and then applying the obtained knowledge to another dataset. By incorporating perturbations sampled from trained models that are resistant to adversarial attacks, DDA is designed to improve transfer robustness. Given its design, our approach is able to incorporate samples that have been generated by the addition of perturbations of previous datasets. Leading to more diverse training examples that can better reflect the heterogeneity of the target dataset. Compared to other approaches in the literature, DDA does not require additional labeled data or knowledge of the target dataset. Instead, it makes use of the robust model's acquired knowledge to produce perturbations that are pertinent to the target domain.

The rest of the paper is organized as follows: In Section 2, we describe the related work on the adversarial robustness problem. In Section 3, we describe the proposed method. We first explain the adversarial training procedure and then explain the generation of adversarial perturbations for transfer robustness. Next, we discuss the design choices we made for DDA for sampling adversarial perturbations. In Section 4, we detail the experimental setup used for implementing the models. In Section 5, we discuss the performance obtained by using DDA. Finally, in Section 6, we present our conclusions and discuss future work.

2. Related Work and Motivation

Due to deep neural network's susceptibility to adversarial attacks, adversarial robustness has grown to be a crucial research area in deep learning (DL) [1, 6, 10, 15]. It has been demonstrated that DL models are susceptible to adversarial instances, which are intentionally constructed inputs that can lead the model to produce wrong predictions [21].

Although several proposals for mitigating adversarial risks for DL models have been investigated [18, 20], there is still a need for enhanced robustness in many settings. Recently, research on adversarial robustness has focused on creating adversarial attacks to fool a model [15]. These attacks can be classified as white-box or black-box attacks. In the former, the attacker has full knowledge of the model weights and architecture and can generate adversarial examples by optimizing a certain loss function. In the latter, the attacker has limited knowledge about the model and can only create adversarial examples by feeding the network with inputs and analyzing the outputs.

Similarly, adversarial defense methods have been proposed in recent years [7, 9]. Defense methods can be mainly classified into two categories: *pre-processing* defense, which aims to modify input data before feeding the model. Popular techniques under this category include data augmentation [12, 13] and input denoising [25]. The category, known as *post-processing* defenses, involves treating the output of a model. Examples of these techniques include adversarial training [15, 21], defensive distillation [26], and model ensembles [37]. However, none of these can guarantee full coverage in terms of security and robustness against all adversarial inputs. Thus, robust adversarial defenses are still a high-demanding and high-priority requirement for designing trustable ML solutions.

Additionally, the accuracy under adversarial attacks is the most commonly used metric to evaluate the robustness of a model [21]. This metric determines the percentage of correctly classified examples under a particular attack. Similarly, the robustness radius is a metric that measures the maximum magnitude of adversarial perturbation that a model can withstand [11]. Moreover, minimum distortion is a metric that assesses the minimum magnitude of adversarial perturbation needed to fool a model [5].

The security and dependability of deep learning systems are seriously threatened by adversarial robustness in critical applications. Many defense mechanisms have improved the robustness of models, but considerable work still needs to be done before these models can tolerate various adversarial attacks and still keep their natural accuracy. Therefore, reducing this gap between accuracy and robustness is still a research problem [17, 29, 34].

The main contribution of this work is the data augmentation method based on adversarial attacks and transfer learning to enhance model robustness. DDA is designed by using adversarial perturbations that are effective on a larger and more complex task to transfer them to downstream tasks.

3. Methodology

In deep learning, the term *adversarial robustness* describes a model's capacity to continue operating effectively even in the minimum engineered changes to the input data intended to trick the model [6]. Let X be the set of possible input data, and let Y be the set of possible output labels. A supervised learning model can be represented as a function $f : X \to Y$ that maps input data to output labels. Adversarial examples can be generated by adding a small perturbation δ to the input data x, such that $x' = x + \delta$. The perturbation is typically constrained to have a small ℓ_p-norm, where p is a positive integer (e.g., $p = 2$ corresponds to the Euclidean distance).

The utilization of adversarial examples as a means of data augmentation during the training phase constitutes a technique referred to as *Adversarial Training*. This technique aims to enhance the robustness of deep learning models against adversarial examples.

3.1. Adversarial Training

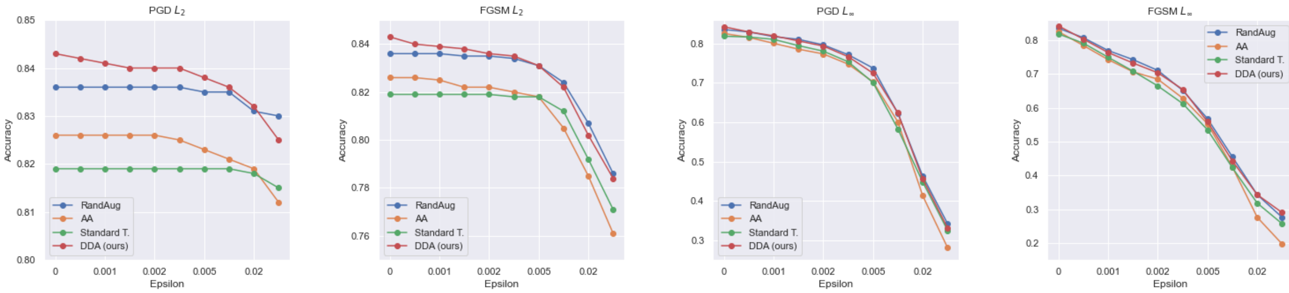
Let $x \in X$ be an input data vector, and $y \in Y$ be its corresponding label. The loss function is typically defined as the cross-entropy loss between the predicted output of the model and the true label (Eq. 1).

$$L(f_\theta(x), y) = -\sum_{i=1}^{|Y|} y_i \log f_\theta(x)_i, \tag{1}$$

where $f\theta(x)_i$ is the i-th output of the model for input x.

To generate an adversarial perturbation δ for input x, the first step is to compute the δ that maximizes the loss function (Eq. 2), subject to a constraint on the ℓ_p-norm of the perturbation, such that $|\delta|_p \leq \epsilon$. In adversarial attacks, ϵ is a parameter used to define a constraint on the magnitude of the perturbation that can be applied to the input data x.

$$\delta = \arg\max_{\delta'} L(f_\theta(x + \delta'), y), \text{s.t.} |\delta|_p \leq \epsilon \tag{2}$$



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

Figure 2. Accuracy comparison for PGD with L_2 (2a) and L_∞ (2c), and FGSM with L_2 (2b) and L_∞ (2d), for CIFAR-10 dataset with different methods of data augmentation: Delta Data Augmentation (ours), RandAugment [13], AutoAugment [12], Standard Training with no data augmentation. An epsilon of 0 means natural accuracy.

One approach to generating adversarial examples is to use an iterative optimization algorithm such as Fast Gradient Descent Method (FGSM) [15] or Projected Gradient Descent (PGD) [21] to compute a perturbation. The resulting adversarial example $x + \delta$ is then added to the original training data along with its corresponding label y, creating an augmented training dataset. Then, the empirical risk over the augmented training data is used as the final objective function for adversarial training, as shown in Equation 3.

$$\min_{\theta} \frac{1}{n+m} \sum_{i=1}^{n} L(f_\theta(x_i), y_i) + \frac{1}{n+m} \sum_{i=1}^{m} L(f_\theta(x_i + \delta_i), y_i), \tag{3}$$

where n is the size of the original training data, m is the size of the augmented training data, and (x_i, y_i) and $(x_i + \delta_i, y_i)$ are pairs of original and adversarial training examples, respectively.

3.2. Delta Data Augmentation (DDA)

Transfer Learning (TL) is a technique used in deep learning to transfer knowledge learned from one model to another [34]. In TL, a pre-trained model is used as a starting point for a new model rather than beginning from scratch [17]. For this, $g_\phi : X' \rightarrow Y'$ is a pre-trained model parameterized by ϕ, where X' and Y' may or may not be the same as X and Y. The goal of transfer learning is to initialize the parameters of f_θ using the pre-trained parameters ϕ and then fine-tune f_θ using a small amount of data from the new task [32]. Then, *transfer robustness* of g_ϕ is defined as the ability of f_θ to maintain its performance on a new task under adversarial attacks when initialized with the pre-trained parameters ϕ [24,32]. Now, instead of using pre-trained parameters, we look for transfer adversarial perturbations that are effective on a larger and more complex model.

In accordance with this notion, a model may enhance its performance by incorporating a greater variety of data into the training phase [28]. The augmentation of training data via adversarial examples can result in an improvement in model generalization. Nevertheless, adversarial training is a computationally demanding and time-consuming undertaking. One approach to address the challenges of adversarial training is the use of universal adversarial perturbations [10, 22]. These perturbations can be generated once and applied to any image, which makes the process more efficient compared to generating adversarial examples for each image individually. Incorporating such perturbations into the training data can enhance model robustness and improve its generalization performance [28]. However, generating these perturbations can also be a computationally demanding task.

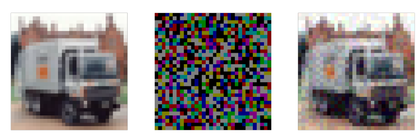


Figure 3. Example of DDA for CIFAR10. First, is the original image, then, in the second image is the perturbation extracted by DDA. Finally, the third image is the augmented image.

Instead of attacking a model to create a set of adversarial examples, we propose to gather adversarial perturbations by attacking upstream model tasks (e.g. ImageNet [14]). This approach will yield sample adversarial noise that is effective across other models. We aim to collect adversarial noise δ and apply it to downstream tasks in a data augmentation fashion. We call this method *Delta Data Augmentation* (DDA) (Fig. 1). In DDA, a pre-trained model that is trained on an upstream task, such as ImageNet Classification, is used to sample adversarial perturbations δ given an adversarial attack (Fig. 3). The objective of this process is to obtain a representative sample of perturbations that re-

Norm	Attack	Attack Intensity ϵ													
		0	0.0005	0.001	0.0015	0.002	0.003	0.005	0.01	0.02	0.03	0.1	0.3	0.5	1
	FGSM		0.804	0.764	0.733	0.704	0.655	0.559	0.443	0.343	0.29	0.223	0.179	0.146	0.088
	PGD		0.83	0.82	0.807	0.794	0.767	0.725	0.624	0.456	0.331	0.024	0.002	0	0
ℓ_∞	*BIA*		0.801	0.762	0.725	0.69	0.615	0.488	0.233	0.037	0.015	0.002	0	0	0.001
	AUNA		**0.843**	**0.841**	**0.842**	**0.84**	**0.84**	**0.841**	**0.838**	**0.842**	**0.838**	**0.81**	**0.653**	**0.499**	**0.243**
	DFA	**0.843**	0.813	0.782	0.761	0.735	0.685	0.579	0.331	0.186	0.139	0.093	0.068	0.054	0.06
	FGSM		0.84	0.839	0.838	0.836	0.835	0.831	0.822	0.802	0.784	0.671	0.485	0.404	0.32
	PGD		0.842	0.841	0.84	0.84	0.84	0.838	0.836	0.832	0.825	0.791	0.711	0.613	0.45
ℓ_2	*BIA*		0.84	0.839	0.838	0.838	0.835	0.831	0.822	0.801	0.781	0.648	0.316	0.12	0.015
	AUNA		**0.843**	**0.843**	**0.843**	**0.843**	**0.843**	**0.843**	**0.843**	**0.843**	**0.842**	**0.841**	**0.842**	**0.84**	**0.838**
	DFA		0.84	0.84	0.84	0.84	0.839	0.837	0.835	0.819	0.806	0.706	0.422	0.245	0.154
	Average	0.843	0.8296	0.8171	0.8067	0.796	0.7754	0.7372	0.6627	0.5961	0.5651	0.4809	0.3678	0.2921	0.2169

Table 1. Results of Delta Data Augmentation (DDA) on CIFAR10 with ResNet18 on each adversarial attacks: Fast Gradient Sign Method (FGSM) [15], Projected Gradient Descent (PGD) [21], Basic Iterative Attack (BAI), Additive Uniform Noise Attack (AUNA) [30], and Deep Fool Attack (DFA) [23] for different epsilon ϵ perturbation intensities, $\epsilon = 0$ means natural accuracy (no attack). In bold are the highest robust scores for each ϵ, and ℓ_∞ ℓ_2 norms respectively.

flects the same underlying structure, which can be used to make downstream training datasets more adversarially diverse and thus more robust.

Following Eq. 2, in DDA, we sample perturbations δ_A from an upstream pre-trained model A trained on dataset X_A. Then, we apply these perturbations to downstream training dataset X_B for model B.

$$X'_B = X_B + \delta_A \tag{4}$$

Thus, DDA can collect and create more complex transformations on data rather than traditional techniques, (rotation, scaling, flipping, etc.). Furthermore, the training time used for data augmentation of model B is reduced due to the absence of adversarial training, and the gap between natural accuracy and robust accuracy is minimized as we forgo learning explicit adversarial perturbations, preventing overfitting to specific types of adversarial attacks.

4. Experimental Setup

We adhere to the fundamental framework and employ the PyTorch [27] implementation. For the main architecture, we use a ResNet18 [16] pre-trained on ImageNet [14]. We compare our method against two common and popular data augmentation techniques: RandAugment [13] and AutoAugment [12]. Also, we develop a baseline on ResNet18 by standard training with no data augmentation at all, cross-entropy loss, and Adam optimizer with 0.001 as a starting learning rate and 30 epochs (Fig. 2).

We test our method (Table 1) on the CIFAR10 [19] dataset with natural accuracy and robust accuracy using popular state-of-the-art adversarial attacks: Fast Gradient Sign Method (FGSM) [15], Projected Gradient Descent (PGD) [21], Basic Iterative Attack (BAI), Additive Uniform Noise Attack (AUNA) [30], and Deep Fool Attack (DFA) [23]. All these attacks are performed on ℓ_2-norm and ℓ_∞-norm using the implementations of *FoolBox* [31].

5. Discussion

The comparison of accuracy across various data augmentation techniques reveals that DDA performs better than the others in terms of robust accuracy against PGD and FGSM attacks. Particularly, DDA outperforms other approaches in terms of robust accuracy, achieving values of 76.7% and 84% for PGD attack with $\epsilon = .003$ for ℓ_∞ and ℓ_2 respectively. Also, for FGSM 65.5% and 83.5% for ℓ_∞ and ℓ_2 respectively with same ϵ.

The comparison of PGD and FGSM attacks with various ϵ values further demonstrates that the resilience of the model is significantly impacted by the choice of attack strength (Table 1). As expected given that greater attacks bring larger perturbations that are more challenging to recover from, our results indicate that stronger attacks result in lower robust accuracies (Fig. 2).

Overall, the findings imply that DDA is a successful technique for boosting the robustness of deep neural networks against adversarial attacks. The suggested method can be used to increase the robustness of models in a variety of applications and is simple to integrate into current training pipelines.

6. Conclusions and Future Work

In this study, we compared a variety of data augmentation strategies, such as DDA, RandAugment, AutoAugment, and Standard Training with no data augmentation, to examine the performance of various adversarial attack methods on the CIFAR10 dataset. Our findings demonstrated that, in terms of adversarial robustness, our approach performed better than or equal to State-of-the-Art approaches. DDA improves the transferability of robustness against adversarial attacks by reducing the gap between natural and robust accuracy. Future studies can examine the use of DDA with additional datasets and see how well it defends against increasingly sophisticated adversarial attacks.

References

[1] Naveed Akhtar and Ajmal Mian. Threat of adversarial attacks on deep learning in computer vision: A survey. *Ieee Access*, 6:14410–14430, 2018. 1, 2

[2] Kyriakos D Apostolidis and George A Papakostas. A survey on adversarial deep learning robustness in medical image analysis. *Electronics*, 10(17):2132, 2021. 1

[3] Tao Bai, Jinqi Luo, Jun Zhao, Bihan Wen, and Qian Wang. Recent advances in adversarial training for adversarial robustness. *arXiv preprint arXiv:2102.01356*, 2021. 1

[4] Nicholas Carlini, Anish Athalye, Nicolas Papernot, Wieland Brendel, Jonas Rauber, Dimitris Tsipras, Ian Goodfellow, Aleksander Madry, and Alexey Kurakin. On evaluating adversarial robustness. *arXiv preprint arXiv:1902.06705*, 2019. 1

[5] Nicholas Carlini, Guy Katz, Clark Barrett, and David L Dill. Provably minimally-distorted adversarial examples. *arXiv preprint arXiv:1709.10207*, 2017. 2

[6] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pages 39–57. Ieee, 2017. 2

[7] Yair Carmon, Aditi Raghunathan, Ludwig Schmidt, John C Duchi, and Percy S Liang. Unlabeled data improves adversarial robustness. *Advances in neural information processing systems*, 32, 2019. 2

[8] Francesco Cartella, Orlando Anunciacao, Yuki Funabiki, Daisuke Yamaguchi, Toru Akishita, and Olivier Elshocht. Adversarial attacks for tabular data: Application to fraud detection and imbalanced data. *arXiv preprint arXiv:2101.08030*, 2021. 1

[9] Anirban Chakraborty, Manaar Alam, Vishal Dey, Anupam Chattopadhyay, and Debdeep Mukhopadhyay. Adversarial attacks and defences: A survey. *arXiv preprint arXiv:1810.00069*, 2018. 1, 2

[10] Ashutosh Chaubey, Nikhil Agrawal, Kavya Barnwal, Keerat K Guliani, and Pramod Mehta. Universal adversarial perturbations: A survey. *arXiv preprint arXiv:2005.08087*, 2020. 2, 3

[11] Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In *international conference on machine learning*, pages 1310–1320. PMLR, 2019. 1, 2

[12] Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation policies from data. *arXiv preprint arXiv:1805.09501*, 2018. 1, 2, 3, 4

[13] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 702–703, 2020. 1, 2, 3, 4

[14] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 3, 4

[15] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014. 1, 2, 3, 4

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4

[17] Minyoung Huh, Pulkit Agrawal, and Alexei A Efros. What makes imagenet good for transfer learning? *arXiv preprint arXiv:1608.08614*, 2016. 2, 3

[18] Yunseok Jang, Tianchen Zhao, Seunghoon Hong, and Honglak Lee. Adversarial defense via learning to generate diverse attacks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2740–2749, 2019. 2

[19] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 4

[20] Hongshuo Liang, Erlu He, Yangyang Zhao, Zhe Jia, and Hao Li. Adversarial attack and defense: A survey. *Electronics*, 11(8):1283, 2022. 2

[21] 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. 2, 3, 4

[22] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Universal adversarial perturbations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1765–1773, 2017. 3

[23] 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*, pages 2574–2582, 2016. 4

[24] Awais Muhammad and Sung-Ho Bae. A survey on efficient methods for adversarial robustness. *IEEE Access*, 10:118815–118830, 2022. 3

[25] Weili Nie, Brandon Guo, Yujia Huang, Chaowei Xiao, Arash Vahdat, and Anima Anandkumar. Diffusion models for adversarial purification. *arXiv preprint arXiv:2205.07460*, 2022. 2

[26] 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)*, pages 582–597. IEEE, 2016. 2

[27] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. 4

[28] Luis Perez and Jason Wang. The effectiveness of data augmentation in image classification using deep learning. *arXiv preprint arXiv:1712.04621*, 2017. 1, 3

[29] Aditi Raghunathan, Sang Michael Xie, Fanny Yang, John Duchi, and Percy Liang. Understanding and mitigating the tradeoff between robustness and accuracy. *arXiv preprint arXiv:2002.10716*, 2020. 2

[30] Jonas Rauber and Matthias Bethge. Fast differentiable clipping-aware normalization and rescaling. *arXiv preprint arXiv:2007.07677*, 2020. 4

[31] Jonas Rauber, Wieland Brendel, and Matthias Bethge. Foolbox: A python toolbox to benchmark the robustness of machine learning models. In *Reliable Machine Learning in the Wild Workshop, 34th International Conference on Machine Learning*, 2017. 4

[32] Hadi Salman, Andrew Ilyas, Logan Engstrom, Ashish Kapoor, and Aleksander Madry. Do adversarially robust imagenet models transfer better? *Advances in Neural Information Processing Systems*, 33:3533–3545, 2020. 3

[33] Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of big data*, 6(1):1–48, 2019. 1

[34] Chuanqi Tan, Fuchun Sun, Tao Kong, Wenchang Zhang, Chao Yang, and Chunfang Liu. A survey on deep transfer learning. In *Artificial Neural Networks and Machine Learning–ICANN 2018: 27th International Conference on Artificial Neural Networks, Rhodes, Greece, October 4-7, 2018, Proceedings, Part III 27*, pages 270–279. Springer, 2018. 2, 3

[35] Eric Wong, Leslie Rice, and J Zico Kolter. Fast is better than free: Revisiting adversarial training. *arXiv preprint arXiv:2001.03994*, 2020. 1

[36] Mengting Xu, Tao Zhang, Zhongnian Li, Mingxia Liu, and Daoqiang Zhang. Towards evaluating the robustness of deep diagnostic models by adversarial attack. *Medical Image Analysis*, 69:101977, 2021. 1

[37] Zhuolin Yang, Linyi Li, Xiaojun Xu, Bhavya Kailkhura, Tao Xie, and Bo Li. On the certified robustness for ensemble models and beyond. *arXiv preprint arXiv:2107.10873*, 2021. 2

[38] Qingzhao Zhang, Shengtuo Hu, Jiachen Sun, Qi Alfred Chen, and Z Morley Mao. On adversarial robustness of trajectory prediction for autonomous vehicles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15159–15168, 2022. 1