FARMUR: Fair Adversarial Retraining to Mitigate Unfairness in Robustness

Seyed Ali Mousavi[1][0009−0007−9278−9043], Hamid Mousavi[2][0000−0001−5710−1206], and Masoud Daneshtalab[2,3][0000−0001−6289−1521]

[1] Shahid Bahonar University, Kerman, Iran
mousavi.sayedali@math.uk.ac.ir

[2] Mälardalen University , Universitetsplan 1, 722 20 Västerås , Sweden
{seyedhamidreza.mousavi,masoud.daneshtalab}@mdu.se

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

Abstract. Deep Neural Networks (DNNs) have been deployed in safety-critical real-world applications, including automated decision-making systems. There are often concerns about two aspects of these systems: the fairness of the predictions and their robustness against adversarial attacks. In recent years, extensive studies have been devoted to addressing these issues independently through adversarial training and unfairness mitigation techniques. To consider fairness and robustness simultaneously, the robustness-bias concept is introduced, which means an attacker can more easily target a particular sub-partition in the dataset. However, there is no unified and mathematical definition for measuring fairness in the robustness of DNNs independent of the type of adversarial attacks. In this paper, we first provide a unified, precise, and mathematical theory and measurement for fairness in robustness independent of adversarial attacks for a DNN model. Finally, we proposed a fair adversarial retraining method (FARMUR) to mitigate unfairness in robustness that retrains the DNN models based on vulnerable and robust sub-partitions. In particular, FARMUR leverages different objective functions for vulnerable and robust sub-partitions to retrain the DNN. Experimental results demonstrate the effectiveness of FARMUR in mitigating the unfairness in robustness during adversarial training without significantly degrading robustness. FARMUR improves fairness in robustness by 19.18% with only 2.22% reduction in robustness in comparison with adversarial training on the UTKFace dataset, which is partitioned based on race attributes.

Keywords: Deep Neural Networks · Robustness · Robustness-bias · Fairness · Fairness-In-Robustness.

1 Introduction

Deep Neural Networks (DNNs) have been used dramatically in real-world applications over the past decade. These applications cover a wide range of automated decision-making systems, such as self-driving cars [15,31], social media [45], healthcare [27], targeted ads [35], and hiring [34]. Defense against adversarial attacks and the fairness of the predictions are two of the main concerns of these applications. Adversarial attacks can easily fool DNNs into predicting the wrong results by adding small, imperceptible noise into the input data (adversarial data). On the other side, unfair DNNs exacerbate societal stereotypes by using sensitive attributes, such as race and gender, in their predictions. These issues are addressed independently in the various research studies [38,6,10]. Adversarial training is a well-known and widely adopted defense method to increase the robustness of DNNs against adversarial attacks [25,43,38]. The goal of the AT methods is to train an adversarially robust DNN such that its predictions are locally invariant to a small neighborhood of the inputs. In terms of the fairness of DNNs, many studies have defined it based on the output accuracy for different sub-partitions of the test data. The accuracy of the model on some sub-partitions may be significantly lower than the average accuracy of the model on all test data. This is interpreted as the existence of a kind of unfairness in the DNN. This can happen for a variety of reasons, including the existence of some spurious correlations [24] and the data-driven training algorithm with unfair data used in DNNs [9]. A lot of works have been proposed to quantify, measure, and mitigate unfairness in DNNs [1,7,11,14,33]. To consider fairness and robustness jointly, the robustness-bias concept is defined such that DNNs may have similar accuracy across the sub-partitions in the dataset, but be more vulnerable to adversarial attacks on certain sub-partitions [30]. Based on this definition, different sub-partitions of the dataset can have different levels of robustness, which leads to unfairness in the robustness. However, there is no unified and clear mathematical definition for fairness in robustness, and previous metrics measured it based on adversarial attacks or computed it for only one sub-partition of data. Therefore, previous measures are highly dependent on the type of adversarial attacks and do not consider the whole DNN to measure fairness in robustness. In addition, the robustness-bias problem is investigated in the case of natural training of DNNs [30], but this issue also needs to be investigated for the adversarial training methods.

In this paper, we first introduce a unified, precise, and mathematical theory and measurement for fairness in robustness of a DNN model. Based on this metric, we empirically show that adversarial training increases the robustness-bias problem in the DNNs compared to natural training. To solve the robustness-bias problem in adversarial training, we propose the fair adversarial retraining (FARMUR) method to mitigate unfairness in robustness. It finds the vulnerable and robust sub-partitions of a partitioned dataset by analyzing their robustness and retrains the DNN on this data based on a new objective function. The new objective function consists of the robust and natural loss functions for vulnerable and robust sub-partitions. We evaluate FARMUR by comparing the fairness in robustness metric for the DNNs before and after fair adversarial retraining on various datasets.
Key contributions: The main contributions are as follows:

1. We introduce a unified, precise, and mathematical theory and measurement for fairness in robustness of the DNN models (Definition 4) (section 3 and section 4).
2. We analyze the fairness in robustness for natural and adversarial training methods and demonstrate that adversarial training increases unfairness in robustness.
3. We propose the fair adversarial retraining method (FARMUR) to mitigate unfairness in the robustness of adversarial training methods (section 5).
4. We evaluate the FARMUR method on MNIST, Cifar-10, and UTKFace datasets. Experimentally we show that our method improves the fairness of the DNNs in terms of robustness for adversarial training methods.

Throughout this paper, we will denote the dataset, partition, and sub-partition by D, $\mathcal{P}$, and P, respectively. Also, we demonstrate the DNN by f.

2 Related Works

2.1 Fairness in DNNs

DNNs are generally regarded as high-performance black-box models. They used data-driven learning algorithms to learn useful representations from the raw data. Because the data might have biases, this data-driven algorithm in DNNs causes to amplify the biases in the data and make unfairness in the DNN model [9]. When we talk about fairness in DNNs, we need to be clear about what we are talking about in terms of fairness. In recent research works, the concept of fairness is considered as *fair predictions* [26]. It means that the DNNs predictions should be the same for different sub-partitions of data. There are different metrics to measure the fairness of a DNN. They can be categorized into individual and group fairness measurements. Individual fairness metrics consider the rule that similar inputs should have similar predictions [10]. Group fairness partitions the data based on a sensitive attribute and computes the statistic measures for each sub-partition and compares it across all sub-partitions.[3]. Moreover, Some mitigation techniques have been proposed for removing bias from DNNs [21,6,33,42]. These methods can be divided into pre-processing, in-processing, and post-processing methods. Pre-processing methods try to remove the biases from training data [4]. In-processing approaches regularize the loss function by adding fairness metric to the overall objective function [8]. Post-processing techniques try to calibrate the predictions of trained DNN models [16]. The comprehensive overview of fairness in DNNs in terms of definition, measurement, and mitigation methods can be found in [26,9]. In this paper, we look at the fairness of DNNs from the robustness perspective. Therefore, we introduce a theory, measurement, and mitigation method for fairness in robustness.

2.2 Robustness and Adversarial Training

Despite the high accuracy of DNNs, the presence of adversarial examples [13] has raised concerns about the use of these models for sensitive tasks. This sensitivity can be related to various aspects such as safety (e.g. autonomous driving vehicles [29]) and fairness (e.g. hiring [34]). A growing body of research shows that neural networks are vulnerable to adversarial examples generated with adversarial attacks [13,18,41,32,39,12]. Therefore, it requires first to define the robustness of DNNs and find the method to defend against adversarial attacks. The robustness of DNNs to has been defined in different but related ways. A commonly used definition is *robust accuracy*, which monitors the accuracy of the model on adversarial examples [44,37,5]. This definition is highly dependent on the type of adversarial attacks. Another definition, introduced by [28], is the average of the normalized distance of each point to the decision boundary. In terms of defense against adversarial attacks, adversarial training is a well-known and widely adopted defense method to increase the robustness of classifiers [25,43,38] Adversarial training is formulated as a min-max optimization problem, and the model is trained exclusively on adversarial images [25]. To find a trade-off between accuracy and robustness, TRADES [44] regularizes the loss function for clean data by incorporating a robust loss term and making a trade-off between them. Authors in [40] have shown that adversarial training can cause a serious disparity in both standard accuracy and adversarial robustness between different classes of data. The main goal of this paper is to find a trade-off between robustness and the fairness in robustness of the DNN model.

2.3 Fairness in Robustness

Recently, the concept of fairness has been integrated with robustness by using robust accuracy measurement [2,40,36]. These researches show that robustness may be at odds with fairness [2] and adversarial training algorithms. In addition, they demonstrate that adversarial training tends to introduce severe disparity of accuracy and robustness between different sub-partitions of data [40]. In terms of robustness measurement based on the distance of data points to the decision boundary, [30] show unfairness in the robustness for the DNNs that are naturally trained on clean data. It discusses fairness through robustness and has introduced the concept of robustness-bias based on the geometry of the decision boundary. However, it considers the natural training method and computes the robustness-bias for one sub-partition, not for the entire DNN. In this paper, we introduce a new unified, prices and mathematical theory and measurement for fairness in robustness based on the geometry of decision boundary for evaluating the fairness of the DNN models. Our new metric also shows that adversarial training decreases fairness in robustness. In addition, we improve the fairness in robustness by fair adversarial retraining (FARMUR) without compromising robustness significantly.

3 Fairness in Robustness: Theory

Unfairness in terms of robustness means that there are data sub-partitions that are most vulnerable to adversarial attacks. Therefore, the data on those sub-partitions is closer to the decision boundary of the DNN and can easily convert to an adversarial example. A partition ($\mathcal{P}$) consists of the non-empty sub-partitions of the dataset in which each data point is contained in exactly one sub-partition. In a classification problem, data can be partitioned in different ways. The simple way is to leverage the original labels as the sensitive attributes to partition data. However, data can be partitioned based on other sensitive attributes apart from the original labels. For the UTKFace dataset, data can partition based on age (original label) or other sensitive attributes such as race, gender, or ethnicity. Let D, $\mathcal{P}$, and f demonstrate the dataset, the partition of the dataset, and the DNN classifier, respectively. To define a new unified and precise metric for fairness in robustness, we first define the robustness of each sub-partition of $\mathcal{P}$ as follows.

Definition 1. *For every sub-partition P of a partition $\mathcal{P}$, we consider $Corr_f(P)$ as all data in P that are classified correctly by DNN classifier f and define $I_P : [0, \infty] \to [0, 1]$ as:*

$$I_P(\tau) = \frac{|\{(x, y) \in Corr_f(P) : d_\theta(x) > \tau\}|}{|P|} \tag{1}$$

where $|.|$ indicates the number of data of a sub-partition and $d_\theta(x)$ is the distance of x to decision boundary of DNN classifier. τ is a threshold value for the distance.

To make the comparison between the robustness of the sub-partitions in a partition, we define the AUC (Area Under Curve) metric that is independent of variable τ as follows.

Definition 2. *(robustness of a sub-partition) For every I_P in definition 1 we define:*

$$AUC(I_P) = \int_0^\infty I_P(\tau) d\tau \tag{2}$$

$$Rob(P) = AUC(I_P) \tag{3}$$

The AUC (Area Under Curve) metric is a well-defined function and converges to a precise value due to the following properties of I_P. For every $P \in \mathcal{P}$: 1) the I_P is a non-increasing function and 2) As $\tau \to \infty$ then $I_P(\tau) \to 0$. Based on this definition, we can sort the sub-partitions according to their robustness $Rob(P)$. To evaluate the robustness of each DNN classifier, we define $Rob(f)$ as follows:

Definition 3. *(robustness of the DNN classifier) Suppose $\mathcal{P} = \{P_1, P_2, \dots, P_n\}$ is a partition of dataset and f be our DNN classifier. We consider $I_f : [0, \infty] \longrightarrow [0, 1]$ as*

$$I_f(\tau) = mean\{I_{P_i}(\tau) : i = 1, 2, \dots, n\} \tag{4}$$

and define the robustness of f as

$$Rob(f) = AUC(I_f) \tag{5}$$

Where $mean$ indicates the average function on $I_{P_i}(\tau)$.

This definition differs from the other papers that used robust accuracy metrics, which evaluate the proportion of correctly classified attacked data relative to the total data. The robust accuracy metric depends on the type and features of the adversarial attacks. However, our definition is based on the distance of data to the decision boundary and therefore depends only on the DNN classifier. To evaluate the fairness in robustness of the DNN classifier f, we define a metric for **fairness in robustness** as follows.

Definition 4. *(Metric for fairness in robustness) Let $\mathcal{P}$ be a partition of data set D and f be a DNN classifier on D. The metric defines as:*

$$V_f(\mathcal{P}) = Var\{Rob(P) : P \in \mathcal{P}\}. \tag{6}$$

Where $Var\{.\}$ shows the variance of the robustness values. The smaller value of the $V_f(\mathcal{P})$, shows better "fairness in robustness" of the DNN classifier relative to the partition $\mathcal{P}$. Because the smaller values indicate that the distance between each sub-partition and the decision boundary is close to the average distance between all sub-partitions.

4 Fairness in Robustness: Calculation

The main challenge to evaluate the robustness of sub-partitions and the robustness of a DNN classifier is the calculation of $d_\theta(x)$; that is the distance of data point x to the decision boundary of the DNN classifier. For a linear classifier $f(x) = W^T x + b$ we can exactly compute this distance as:

$$d_\theta(x) = \frac{|f_{\hat{l}(x)}(x) - f_{\hat{k}(x)}(x)|}{||w_{\hat{l}(x)} - w_{\hat{k}(x)}||_2^2} \tag{7}$$

where $\hat{k}(x)$ is the ground label of x and

$$\hat{l}(x) = \underset{k \neq \hat{k}(x)}{\arg\min} \frac{|f_k(x) - f_{\hat{k}(x)}(x)|}{||w_k - w_{\hat{k}(x)}||_2^2}$$

The non-linear DNN classifiers have non-convex decision boundaries. Therefore, we do not have an exact formula for calculating $d_\theta(x)$. In this case, we approximate this distance by using iterative linearization of the decision boundary [28]. This algorithm modifies data values until the classification label is changed. In each iteration of the approximation algorithm, we compute the distance between the generated data point and the decision boundary. To compute the accurate distance, we consider r_{max} and r_{tot} as the distance to the decision boundary for the two last iterations (before changing the classification label and after that). We approximate $d_\theta(x)$ by averaging these two distances as:

$$d_\theta(x) = \frac{\|r_{max}\|_2 + \|r_{tot}\|_2}{2} \tag{8}$$

5 Fair Adversarial Retraining to Mitigate Unfairness in Robustness (FARMUR)

We analyze the quantity of $V_f(\mathcal{P})$ to investigate the fairness in robustness for natural and adversarial training. Table 1 demonstrates that the $V_f(\mathcal{P})$ in natural training is smaller than in adversarial training. It shows that if we want to increase the robustness of the network through adversarial training, the fairness metric for robustness is reduced. More precisely, for a DNN classifier f and a partition $\mathcal{P}$ on the dataset, if we apply adversarial training on the model and obtain a new DNN classifier f', then we have $V_{f'}(\mathcal{P}) > V_f(\mathcal{P})$ (lower $V_f(\mathcal{P})$ shows better fairness in robustness) (see Table 1).

To improve fairness in robustness in the case of adversarial training, we propose the "Fair Adversarial Retraining" method. To this end, we first find the vulnerable and robust sub-partitions in the partition $\mathcal{P}$ of the dataset as follows.

Definition 5. *Let $\mathcal{P}$ and f be the partition of the dataset and the DNN classifier respectively. A sub-partition $P \in \mathcal{P}$ is vulnerable if $Rob(P) < Rob(f)$. The other sub-partitions are robust.*

After identifying the vulnerable and robust sub-partitions, we split the dataset D into two subsets, $D = D^{vul} \cup D^{rob}$. Then we apply “fair adversarial retraining” on these subsets with the following loss functions. For vulnerable subset (D^{vul}) we use TRADES [44] loss function as:

$$L_{TRADES} = \mathrm{E}_{(X,Y)\sim D^{vul}}[\mathcal{L}(f(X),Y) + \beta \max_{X' \in \mathcal{B}(X,\epsilon)} \mathcal{L}(f(X), f(X')] \tag{9}$$

Where $\mathcal{B}(X,\epsilon)$ and β denote the neighborhood of the decision boundary of f and regularization hyperparameter. For robust subset (D^{rob}) we leverage natural loss function as:

$$L_{Natural} = \mathrm{E}_{(X,Y)\sim D^{rob}}[\mathcal{L}(f(X),Y)] \tag{10}$$

In our fair adversarial retraining method, we use the sum of these two loss functions to retrain the model.

$$L_{FARMUR} = L_{TRADES} + L_{Natural} \tag{11}$$

Figure 1 and Algorithm 1 demonstrate the main steps of our fair adversarial retraining method. The main aim of our method is to generate a DNN with high robustness and fairness in the robustness.

6 Experiments

6.1 Experimental Setup

Datasets and Partition To evaluate our method, we use MNIST [22], CIFAR-10 [20], and UTKFace [46] datasets. The UTKFace dataset is a large-scale face dataset with a long age span (ranging from 0 to 116 years old). It consists of

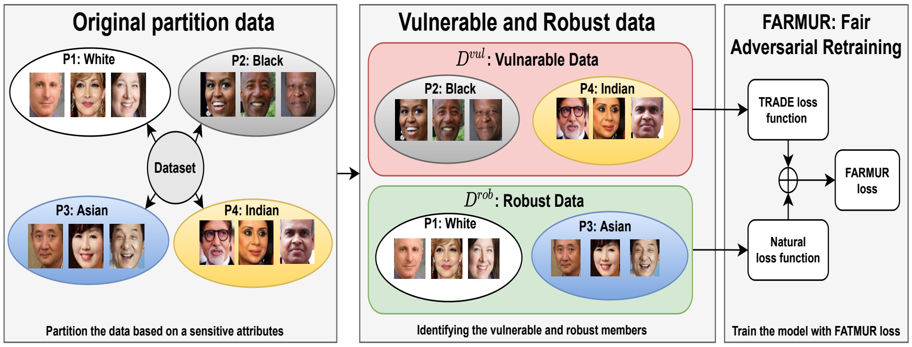
Original partition data
P1: White
P2: Black
Dataset
P3: Asian
P4: Indian
Partition the data based on a sensitive attributes
Vulnerable and Robust data
D^{vul}: Vulnerable Data
P2: Black
P4: Indian
D^{rob}: Robust Data
P1: White
P3: Asian
Identifying the vulnerable and robust members
FARMUR: Fair Adversarial Retraining
TRADE loss function
FARMUR loss
Natural loss function
Train the model with FATMUR loss

Fig. 1: The overview of FARMUR. It first finds the vulnerable and robust sub-partitions D^{vul}, D^{rob} and retrains the DNN based on the new loss function.

20,000 face images with annotations of age, gender, and race/ethnicity. Partitions in MNIST and CIFAR-10 are exactly the same as the labels in the datasets. However, in UTKFace we select the partition differently from the labels of the dataset. In this case, although the classifier has trained on the age labels, we select the race attribute for partitioning the data.

Models We use LeNet-5 [23] and AlexNet [19] to evaluate our method on the MNIST dataset. For CIFAR-10 and UTK-Face, we use ResNet18 [17] and UTKFace classifier [30], respectively. For natural training, we use Stochastic Gradient Decent (SGD) for 120 epochs with a batch size of 64. For adversarial training, we use TRADES [44] algorithm. We set perturbation size $\epsilon = 0.031$, perturbation steps 0.007, learning rate 0.1, batch size 64, and run 120 epochs on the training dataset. For fair adversarial retraining, we first adversarially train the DNN model for 100 epochs. Then, we use the proposed FARMUR loss function on vulnerable and robust sub-partitions and we continue the training for 20 epochs with the same hyperparameters of adversarial training.

6.2 Experimental Results

Table 1 compares the performance of FARMUR against natural and adversarial training. In terms of fairness in robustness, we evaluate the natural, adversarial, and fair retraining (FARMUR) methods with the proposed metric which is defined in 4. Although adversarial training increases the robustness of the DNN in comparison with natural training ($Rob(f)$), it decreases the fairness in robustness (increase $V_f(\mathcal{P})$. It decreases the fairness in robustness by 1.46×, 6.67×, and 10.43× For MNIST (AlexNet), Cifar-10, and UTKFace datasets. FARMUR improves the fairness in robustness by 29.17% and 15.39%, and 19.18% In comparison with adversarial training for MNIST (AlexNet) and CIFAR-10 and UTKFace datasets. The robustness of the DNN classifiers remains almost the same

Algorithm 1 Fair Adversarial Retraining

```
Input: Adversarially trained DNN model $f$, Partition dataset $\mathcal{P}$, number of epochs $T$
  for retraining, learning rate $\gamma$, training dataset $D$.
  Find vulnerable and robust data
  Step 1: Find vulnerable subset: $D^{vul} = \{P \in \mathcal{P} : Rob(P) < Rob(f)\}$
  Step 2: Find robust subset: $D^{rob} = \{P \in \mathcal{P} : Rob(P) > Rob(f)\}$
  Step 3: Split the dataset $D$ into two subsets, $D = D^{vul} \cup D^{rob}$.
  Retrain DNN with new loss function
  for 1:T do
      for mini-batches $(b^{vul}, y^{batch}) \subseteq D^{vul}, (b^{rob}, y^{batch}) \subseteq D^{rob}$ do
          Step 1: Calculate TRADES_loss for $(b^{vul}, y^{batch})$ by equation (9)
          Step 2: Calculate Natural loss for $(b^{rob}, y^{batch})$ by equation (10)
          Step 3: Calculate FARMUR loss based on equation  (11)
          Step 4: Retrain model with loss in Step 3.
      end for
  end for
Output: Retrained model
```

as the adversarial training in FARMUR retraining. It denotes that FARMUR not only improves the fairness in the robustness but also does not change significantly the robustness level of the DNN. To illustrate the effectiveness of FARMUR, we conduct some case studies on three different models and datasets. Figure 2 show the variation of $I_P(\tau)$ with respect to τ for the DNNs trained with natural, adversarial, and FARMUR retraining methods. All the figures show that our method has a smaller variance in the robustness (lower $V_f(\mathcal{P})$) in comparison to the adversarial training method. For the UTKFace dataset, the last row of the figure determines that our proposed method can increase fairness in terms of robustness for different races. It means the attacker cannot attack some specific race more simply than others. FARMUR consists of two separate phases in terms of time complexity: finding vulnerable and robust data and retraining the DNN. On a single NVIDIA ®RTX A4000, identifying vulnerable and robust data for the UTKFace dataset requires approximately 6 GPU hours, and retraining the DNN requires approximately 4 GPU hours.

Table 1: The accuracy, robustness ($Rob(f)$), and Fairness in Robustness ($V_f(\mathcal{P})$) of different DNN classifiers on different datasets. The sensitive attributes for MNIST and Cifar-10 are same as the classification labels. For UTKFace dataset, the race attributes use for partition data.

		Accuracy (%) ↑			**Robustness ($Rob(f)$) ↑**			**Fairness in Robustness ($V_f(\mathcal{P})$) ↓**		
Model (f)	Dataset	Natural	Adversarial	FARMUR	Natural	Adversarial	FARMUR	Natural	Adversarial	FARMUR
Lenet-5	MNIST	99.08	98.74	98.91	0.79	1.26	1.24	0.014	0.019	0.009
AlexNet	MNIST	99.30	99.41	99.38	1.47	1.96	1.95	0.048	0.070	0.034
ResNet-18	Cifra-10	93.77	93.58	90.98	0.18	0.48	0.45	0.003	0.02	0.011
UTKClassifier	UTKFace	66.70	65.60	65.66	0.33	1.34	1.31	0.007	0.073	0.059

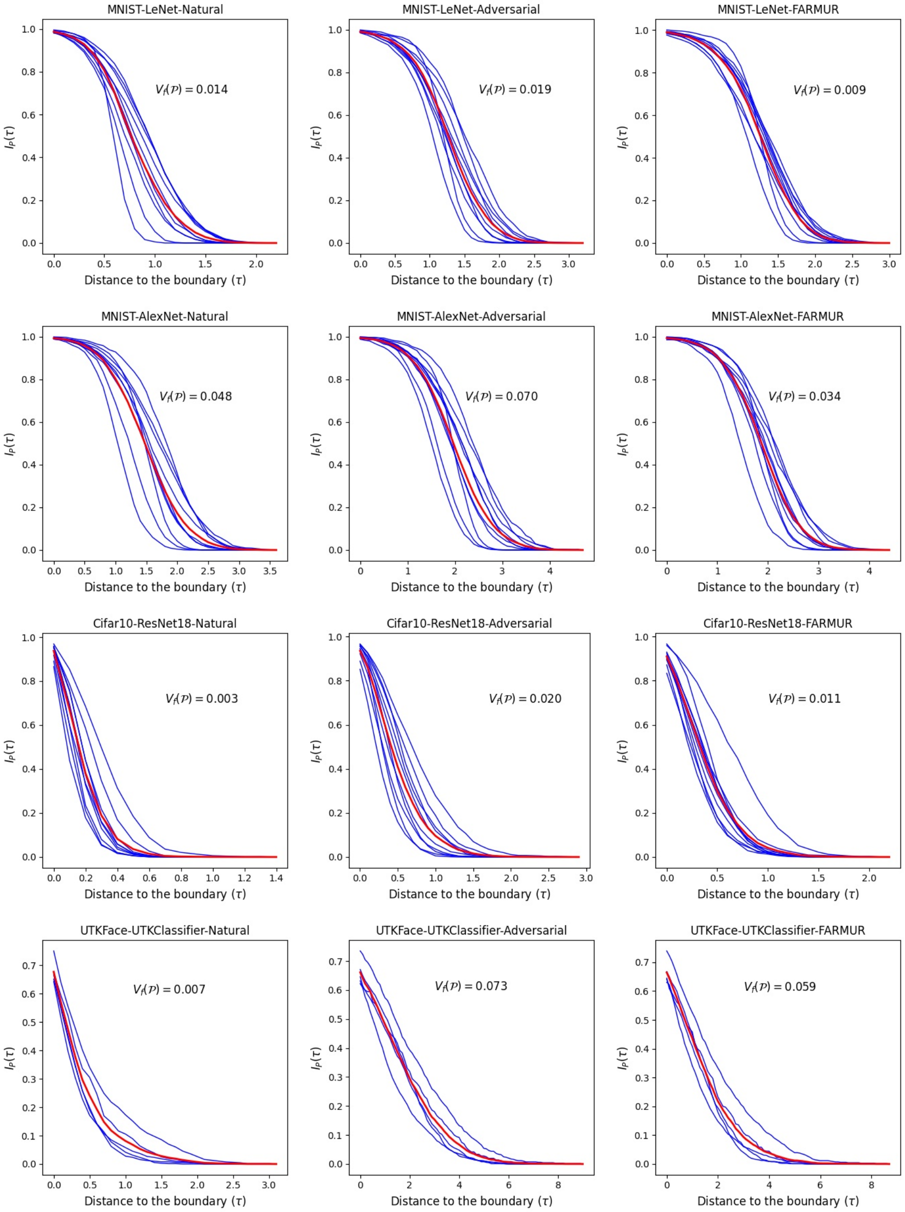

Fig. 2: We plot $I_P(\tau)$ for each sub-partition P in each dataset. Each blue line represents one sub-partition. The red line represents the mean of the blue lines.

7 Conclusion and Future Work

In this paper, we introduce a new theory and metric to measure fairness in robustness based on the distance of the data to the decision boundary. Unlike other metrics, our metric is independent of the type of adversarial attacks and evaluates the fairness for the entire DNN model. Based on this metric, we demonstrate that the adversarial training methods reduce the fairness in robustness of the DNNs. Then we proposed FARMUR as a fair adversarial retraining method to mitigate unfairness in robustness of the adversarial training method. FARMUR tackles the unfairness issue in robustness by splitting the partitions into vulnerable and robust sub-partitions and retraining the model. Experimental results confirm the efficiency of our method to improve the fairness in the robustness of DNNs. For future work, we try to develop a new data evaluation mechanism to find appropriate data for adversarial training. In the other direction, we try to leverage neural architecture search to find DNNs with higher fairness in robustness.

8 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. Adel, T., Valera, I., Ghahramani, Z., Weller, A.: One-network adversarial fairness. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 2412–2420 (2019)
2. Benz, P., Zhang, C., Karjauv, A., Kweon, I.S.: Robustness may be at odds with fairness: An empirical study on class-wise accuracy. In: NeurIPS 2020 Workshop on Pre-registration in Machine Learning. pp. 325–342. PMLR (2021)
3. Beutel, A., Chen, J., Doshi, T., Qian, H., Woodruff, A., Luu, C., Kreitmann, P., Bischof, J., Chi, E.H.: Putting fairness principles into practice: Challenges, metrics, and improvements. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society. pp. 453–459 (2019)
4. Calmon, F., Wei, D., Vinzamuri, B., Natesan Ramamurthy, K., Varshney, K.R.: Optimized pre-processing for discrimination prevention. Advances in neural information processing systems **30** (2017)
5. 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)
6. Chouldechova, A.: Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Big data **5**(2), 153–163 (2017)
7. Donini, M., Oneto, L., Ben-David, S., Shawe-Taylor, J., Pontil, M.: Empirical risk minimization under fairness constraints. arXiv preprint arXiv:1802.08626 (2018)

8. Du, M., Liu, N., Yang, F., Hu, X.: Learning credible deep neural networks with rationale regularization. In: 2019 IEEE International Conference on Data Mining (ICDM). pp. 150–159. IEEE (2019)
9. Du, M., Yang, F., Zou, N., Hu, X.: Fairness in deep learning: A computational perspective. IEEE Intelligent Systems **36**(4), 25–34 (2020)
10. Dwork, C., Hardt, M., Pitassi, T., Reingold, O., Zemel, R.: Fairness through awareness. In: Proceedings of the 3rd innovations in theoretical computer science conference. pp. 214–226 (2012)
11. Dwork, C., Ilvento, C.: Fairness under composition. arXiv preprint arXiv:1806.06122 (2018)
12. Geraeinejad, V., Sinaei, S., Modarressi, M., Daneshtalab, M.: Roco-nas: Robust and compact neural architecture search. In: 2021 International Joint Conference on Neural Networks (IJCNN). pp. 1–8. IEEE (2021)
13. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572 (2014)
14. Grgić-Hlača, N., Zafar, M.B., Gummadi, K.P., Weller, A.: Beyond distributive fairness in algorithmic decision making: Feature selection for procedurally fair learning. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 32 (2018)
15. Grigorescu, S., Trasnea, B., Cocias, T., Macesanu, G.: A survey of deep learning techniques for autonomous driving. Journal of Field Robotics **37**(3), 362–386 (2020)
16. Hardt, M., Price, E., Srebro, N.: Equality of opportunity in supervised learning. Advances in neural information processing systems **29** (2016)
17. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
18. Ilyas, A., Santurkar, S., Tsipras, D., Engstrom, L., Tran, B., Madry, A.: Adversarial examples are not bugs, they are features. arXiv preprint arXiv:1905.02175 (2019)
19. Krizhevsky, A.: One weird trick for parallelizing convolutional neural networks. arXiv preprint arXiv:1404.5997 (2014)
20. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images (2009)
21. Leben, D.: Normative principles for evaluating fairness in machine learning. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. pp. 86–92 (2020)
22. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proceedings of the IEEE **86**(11), 2278–2324 (1998)
23. LeCun, Y., Haffner, P., Bottou, L., Bengio, Y.: Object recognition with gradient-based learning. In: Shape, contour and grouping in computer vision, pp. 319–345. Springer (1999)
24. Liu, E.Z., Haghgoo, B., Chen, A.S., Raghunathan, A., Koh, P.W., Sagawa, S., Liang, P., Finn, C.: Just train twice: Improving group robustness without training group information. In: International Conference on Machine Learning. pp. 6781–6792. PMLR (2021)
25. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083 (2017)
26. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: A survey on bias and fairness in machine learning. ACM Computing Surveys (CSUR) **54**(6), 1–35 (2021)

27. Miotto, R., Wang, F., Wang, S., Jiang, X., Dudley, J.T.: Deep learning for healthcare: review, opportunities and challenges. Briefings in bioinformatics **19**(6), 1236–1246 (2018)
28. Moosavi-Dezfooli, S.M., Fawzi, A., Frossard, P.: Deepfool: a simple and accurate method to fool deep neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2574–2582 (2016)
29. Morgulis, N., Kreines, A., Mendelowitz, S., Weisglass, Y.: Fooling a real car with adversarial traffic signs. arXiv preprint arXiv:1907.00374 (2019)
30. Nanda, V., Dooley, S., Singla, S., Feizi, S., Dickerson, J.P.: Fairness through robustness: Investigating robustness disparity in deep learning. In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. pp. 466–477 (2021)
31. Paden, B., Čáp, M., Yong, S.Z., Yershov, D., Frazzoli, E.: A survey of motion planning and control techniques for self-driving urban vehicles. IEEE Transactions on intelligent vehicles **1**(1), 33–55 (2016)
32. Papernot, N., McDaniel, P., Wu, X., Jha, S., Swami, A.: Distillation as a defense to adversarial perturbations against deep neural networks. In: 2016 IEEE symposium on security and privacy (SP). pp. 582–597. IEEE (2016)
33. Saha, D., Schumann, C., Mcelfresh, D., Dickerson, J., Mazurek, M., Tschantz, M.: Measuring non-expert comprehension of machine learning fairness metrics. In: International Conference on Machine Learning. pp. 8377–8387. PMLR (2020)
34. Schumann, C., Foster, J., Mattei, N., Dickerson, J.: We need fairness and explainability in algorithmic hiring. In: International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS) (2020)
35. Speicher, T., Ali, M., Venkatadri, G., Ribeiro, F.N., Arvanitakis, G., Benevenuto, F., Gummadi, K.P., Loiseau, P., Mislove, A.: Potential for discrimination in online targeted advertising. In: Conference on Fairness, Accountability and Transparency. pp. 5–19. PMLR (2018)
36. Tian, Q., Kuang, K., Jiang, K., Wu, F., Wang, Y.: Analysis and applications of class-wise robustness in adversarial training. arXiv preprint arXiv:2105.14240 (2021)
37. Tsipras, D., Santurkar, S., Engstrom, L., Turner, A., Madry, A.: Robustness may be at odds with accuracy. arXiv preprint arXiv:1805.12152 (2018)
38. Wang, J., Zhang, H.: Bilateral adversarial training: Towards fast training of more robust models against adversarial attacks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6629–6638 (2019)
39. Xie, C., Wang, J., Zhang, Z., Ren, Z., Yuille, A.: Mitigating adversarial effects through randomization. arXiv preprint arXiv:1711.01991 (2017)
40. Xu, H., Liu, X., Li, Y., Jain, A., Tang, J.: To be robust or to be fair: Towards fairness in adversarial training. In: International Conference on Machine Learning. pp. 11492–11501. PMLR (2021)
41. Yuan, X., He, P., Zhu, Q., Li, X.: Adversarial examples: Attacks and defenses for deep learning. IEEE transactions on neural networks and learning systems **30**(9), 2805–2824 (2019)
42. Zemel, R., Wu, Y., Swersky, K., Pitassi, T., Dwork, C.: Learning fair representations. In: International conference on machine learning. pp. 325–333. PMLR (2013)
43. Zhang, D., Zhang, T., Lu, Y., Zhu, Z., Dong, B.: You only propagate once: Accelerating adversarial training via maximal principle. arXiv preprint arXiv:1905.00877 (2019)
44. 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)

Preprint accepted in 27th European Conference on Advances in Databases and Information Systems (ADBIS 2023)

45. Zhang, Z., He, Q., Gao, J., Ni, M.: A deep learning approach for detecting traffic accidents from social media data. Transportation research part C: emerging technologies **86**, 580–596 (2018)
46. Zhang, Z., Song, Y., Qi, H.: Age progression/regression by conditional adversarial autoencoder. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5810–5818 (2017)