Fairness in Deep Learning: A Computational Perspective

Mengnan Du, Fan Yang, Na Zou, and Xia Hu, *Texas A&M University, College Station, TX, 77843, USA*

Fairness in deep learning has attracted tremendous attention recently, as deep learning is increasingly being used in high-stake decision making applications that affect individual lives. We provide a review covering recent progresses to tackle algorithmic fairness problems of deep learning from the computational perspective. Specifically, we show that interpretability can serve as a useful ingredient to diagnose the reasons that lead to algorithmic discrimination. We also discuss fairness mitigation approaches categorized according to three stages of deep learning life-cycle, aiming to push forward the area of fairness in deep learning and build genuinely fair and reliable deep learning systems.

Machine learning algorithms have achieved dramatic progress nowadays, and are increasingly being deployed in high-stake applications, including employment, criminal justice, personalized medicine, etc.[1] Nevertheless, *fairness in machine learning* remains a problem. Machine learning algorithms have the risk of amplifying societal stereotypes by over associating protected attributes, e.g., race and gender, with the prediction task.[2] Eventually, they are capable of exhibiting discriminatory behaviors against certain subgroups. For example, a recruiting tool for STEM jobs believes men are more qualified and shows bias against women,[3] facial recognition performs extremely poorly for female with darker skin.[4] The fairness problem might cause adverse impacts on individuals and society. It not only limits a person's opportunity that s/he is qualified, but also might further exacerbates social inequity.

Among different machine learning models, the fairness problem of *deep learning models* has attracted attention from academia and industry recently. First, deep learning models have achieved the state-of-the-art performance in many domains. Their success can partially be attributed to the data-driven learning paradigm, which enables the models to learn useful representations automatically from data. The data might contain human biases and the data-driven learning also inevitably causes deep learning models to replicate and even amplify biases present in data. Second, it remains a challenge to diagnose and address the deep learning fairness problem. Deep learning models are generally regarded as black-boxes, and their intermediate representations are opaque and hard to comprehend. This is problematic and makes it difficult to identify whether these models make decisions based on right and justified reasons, or due to biases. In addition, this makes it challenging to design bias detection and mitigation approaches.

In this article, we summarize *fairness in deep learning* work from the computational perspective, and do not discuss work from social science, law, and many other disciplines.[1] Particularly, we show that interpretability could significantly contribute to better understandings of the reasons that affect fairness. We also review fairness mitigation strategies categorized into three stages of deep learning life-cycle. Throughout this article, deep learning and Deep neural network (DNN) will be used interchangeably.

DNN FAIRNESS

In this section, we introduce the categorization of fairness problem, measurements of fairness, and interpretation methods closely relevant to understanding DNN fairness.

Fairness Problem Categorization

From the computational perspective, DNN unfairness can be generally categorized into two classes: *prediction outcome discrimination* and *prediction quality disparity*.

1541-1672 © 2020 IEEE
Digital Object Identifier 10.1109/MIS.2020.3000681
Date of publication 10 June 2020; date of current version 14 September 2021.

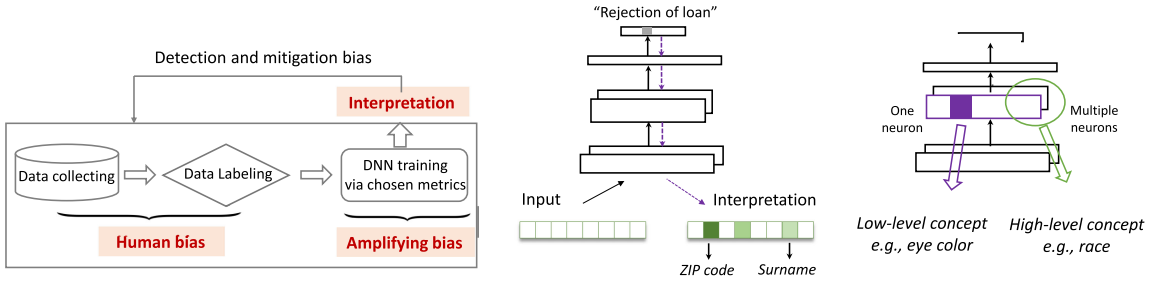

FIGURE 1. (a) Bias exists in different stages of the DNN training pipeline, and interpretation could be utilized to detect and mitigate bias. (b) DNN local interpretation. (c) DNN global interpretation.

Prediction Outcome Discrimination

Discrimination refers to the phenomenon that DNN models produce unfavorable treatment of people due to the membership of certain demographic groups.[1] For instance, a recruiting tool believes that men are more qualified and shows bias against women. Current DNNs generally follow the purely data-driven learning paradigm, and the model training pipeline is illustrated in Figure 1(a). Any training data may contain some biases, either intrinsic noise or additional signals inadvertently introduced by human annotators. DNNs are designed to fit these skewed training data, and thus would naturally replicate the biases existed in data. Even worse, DNNs might make unwanted implicit associations and amplify societal stereotypes about people.[6] This eventually results in trained models with algorithmic discrimination. Outcome discrimination could be further split into *Input* and *Representation* subcategories, with representative examples in Table 1.

Discrimination via Input: Prediction outcome discrimination could be traced back to the input. Even though a DNN model does not explicitly take *protected attributes* as input, e.g., race, gender, and age, it may still induce prediction discrimination. In the context of DNN systems, *protected attributes are often not observed in the input data*, mainly due to two reasons. First, most DNN models rely on raw data, e.g., text, as input and thus protected attributes are not explicitly encoded in the input. Second, collecting protected attributes is often not allowed by the law in real-world applications. Despite the absence of explicit protected attributes, DNNs still could exhibit unintentional discrimination, since there are some features highly correlated with class membership. For instance, ZIP

TABLE 1. DNN fairness problem categorization and representative examples.

Class	Representative examples
Discrimination via Input	*Employment:* Recruiting tool believes that men are more qualified and shows bias against women. *Loan Approval:* Loan eligibility system negatively rates people belonging to certain ZIP code, causing discrimination for certain races. *Criminal Justice:* Recidivism prediction system predicts black inmates are three times more likely to be classified as "high risk" than white inmates.
Discrimination via Representation	*Medical Image Diagnosis:* CNN model could identify patients' self-reported sex from a retina image, and shows discrimination based on gender. *Credit Scoring:* Using raw texts as input, demographic information of authors is encoded in the intermediate representations DNN-based credit scoring classifiers.[5]
Prediction Quality Disparity	*Facial Recognition:* Facial recognition performs very poorly for females with darker skin. *Language processing:* Language identification models perform significantly worse when processing text produced by people belonging to certain races. *Readmission:* ICU mortality and psychiatric 30-day readmission model prediction accuracy is significantly different across gender and insurance types.

code and surname could indicate race. The model prediction might highly depend on the class memberships, and eventually shows discrimination to certain demographic group.

Discrimination via Representation: Sometimes prediction outcome discrimination needs to be diagnosed and mitigated from the representation prospective. In some cases, attributing the bias to input is nearly impossible, e.g., for image input. For instance, CNN model could identify patients' self-reported sex from a retina image, while humans even ophthalmologists cannot identify cues from the input image. In those settings, different demographic groups would have distinct DNN intermediate representations. The class memberships of different protected attributes could be encoded in deep representations. DNN model will make decisions based on the implicitly learned membership information and produce discriminate classification outcomes. Thus, prediction outcome discrimination could be detected and removed from the deep representation perspective.

Prediction Quality Disparity

Prediction quality difference of models for different protected groups is another important category of unfairness. DNN systems have shown lower quality for some groups of people as opposed to other groups. Different from prediction outcome discrimination which is about *resources and opportunities allocations harm* in high-stake applications such as hiring, loan, and credit, this category is about *quality of services harm* that usually happen in general applications, e.g., facial recognition and language processing (see Table 1). This is usually due to the underrepresentation problem, where data may be less informative or less reliably collected for certain parts of the population. Take the Imagenet dataset as an example: females comprise only 41.62% of images, people over 60 are almost nonexistent.[7] The typical objective of DNN training is to minimize the overall error. If the model cannot simultaneously fit all populations optimally, it will fit the majority group. Although this may maximize overall prediction accuracy, it might come at the expense of the underrepresented populations and leads to poor performance for those groups.

Measurements of Fairness

Many metrics have been proposed to measure model fairness, including group fairness and individual fairness. In this article, we focus on *group fairness*, where examples are grouped according to a particular protected attribute, and statistics about model predictions is calculated for each group and compared across groups. We introduce below three mostly used group fairness measurements.

Demographic Parity: It asserts that average of algorithmic decisions should be similar across different groups: $\frac{p(\hat{y}=1|z=0)}{p(\hat{y}=1|z=1)} \geq \tau$, where τ is a given threshold, usually set as 0.8,[8] $\hat{y}$ is a model prediction, 1 denotes favorable outcome, z denotes *protected attribute*, e.g., race, gender. Demographic parity is independent of the ground truth labels. This is useful especially when reliable ground truth information is not available, e.g., employment and criminal justice.[1]

Equality of Opportunity: This metric has taken into consideration that different groups could have different distribution in terms of label y. It is defined as: $p(\hat{y} = 1|z = 0, y = 1) - p(\hat{y} = 1|z = 1, y = 1)$, where y is the ground truth label.[9] Essentially this is comparing the *true positive* rate across different groups. A symmetric measurement can be calculated for *false positive* rate: $p(\hat{y} = 1|z = 0, y = 0) - p(\hat{y} = 1|z = 1, y = 0)$. Putting them together will result the *Equality of Odds* metric.[9]

Predictive Quality Parity: This metric measures prediction quality difference between different subgroups. The quality denotes quantitative model performance in terms of model predictions and ground truth, and in this article we focus on *accuracy* measurement for multiclass classification.[4]

Interpretability for Addressing Fairness Problem

DNNs are often regarded as black-boxes and criticized by the lack of interpretability. Interpretability could be utilized as an effective debugging tool to analyze the models, and enhance the transparency and fairness of models[10] [see Figure 1(a)]. It can be grouped into following two categories.

Local Interpretation: Local interpretation could illustrate how the model arrives at a certain prediction for a specific input [see Figure 1(b)]. It is achieved by attributing model's prediction in terms of its input features. The final interpretation is illustrated in the format of *feature importance* visualization. Take loan prediction for example. The model input is a vector containing categorical features, and the interpretation result is a heat map, where the features with higher scores represent higher relevance for the prediction.

Global Interpretation: The goal is to provide a global understanding about what knowledge has been captured by a pretrained DNN, and illuminate the learned representations in an intuitive manner to

humans [see Figure 1(c)]. The simplest way is to comprehend the concept captured by a single neuron, which is the representation derived from a specific channel at a specific layer. The combination of multiple neurons of different channels or even different layers could represent more abstract concepts.[11] Those protected concepts are usually based on multiple elementary low-level concepts. For instance, race concept can be indicated via multiple local clues such as eye color and hair color.

DETECTION OF MODELING BIAS

In this section, we present methods for detecting and understanding algorithmic discrimination, by making use of DNN interpretability as an effective computational tool.

Discrimination via Input

The source of prediction outcome discrimination could be traced back to the input features. As discussed in the section "Prediction Outcome Discrimination," protected attributes are often not observed in the input data. Due to the redundant encodings, other seemingly innocuous features may be highly correlated with protected attribute and cause model bias.[9] The goal here is to locate these features via local DNN interpretation.

The first solution is performed in a top-down manner, where local interpretation is employed to generate feature importance vector. After getting feature importance for all input features, we can take out those with relatively high importance scores and further analyze them. Among this subset of features, the focus is to identify those fairness sensitive features (in contrast to task relevant features). Take the loan application for example. If the features contributing most to DNN prediction include surname and ZIP code of applicants, we can assert that this model has discrimination toward certain race, and surname and ZIP code here are fairness sensitive features [see Figure 1(b)]. The second solution is implemented in the bottom-up manner. Humans first prechoose features which they are skeptical to be associated with protected attributes, and then analyze feature importance of the identified features.[3] These subset of features are perturbed to generate new data samples, i.e., counterfactual(s). We then feed the counterfactual to the DNN and observe the model prediction difference. If the perturbation of those suspected fairness sensitive features causes significant model prediction change, it can be asserted that the DNN has made biased decisions based on protected attributes.

Discrimination via Representation

Sometimes it is hard to identify bias from the input perspective, and detecting model bias from the deep representations is more convenient. DNN global interpretation could be exploited as a debugging tool to analyze the deep representations. The goal is to identify whether a protected attribute has been captured by the intermediate representation, and the degree to which this protected attribute contributes to the model prediction. Thus, a two-stage scheme could be applied to detect discrimination.

First, global interpretation is utilized to analyze whether a DNN has learned a protected concept. This is usually achieved by pointing to a direction in the activation space of DNN's intermediate layers. A typical example is the *concept activation vector (CAV)* method.[11] Here, CAV defines a high-level concept using a set of example inputs. For example, to define concept *African American race*, a set of darker skin Congoid images could be used. The CAV vector is the direction of activation values for the set of examples corresponding to that concept. This vector is obtained by training a linear classifier between the concept examples and a set of random counterexamples, where the vector is the direction orthogonal to the decision boundary. Second, after confirming that a DNN has learned a protected concept, we proceed to test the contribution of this concept toward the model's final prediction. Different strategies can be used to quantify the conceptual sensitivity, including the top-down manner which calculates derivative of DNN's prediction to the concept vector,[11] or the bottom-up manner which adds this concept vector to different inputs' intermediate activation and then observe the change of model predictions. Ultimately, the representation bias level for a protected attribute is described using a numerical score. The higher the numerical sensitivity score, the more significantly this concept contributes to DNN's prediction.

Prediction Quality Disparity

There usually happens that some groups appear more frequently than others in training data. The DNN model will optimize for those groups in order to boost the overall model performance, leading to low prediction accuracy for the minority group.

The detection of prediction quality disparity is typically performed in a two-step manner: splitting data into subgroups according to sensitive attributes and calculating the accuracy for each demographic groups. For instance, facial recognition systems are analyzed in terms of their prediction quality.[4] Human

TABLE 2. Representative algorithms for mitigating unfairness in DNN models.

Class	Preprocessing	In-processing	Postprocessing
Discrimination via Input	Sensitive features removal	Attribution regularization[14,15]	Calibrated distribution[6]
	Sensitive features replacement	Reduction game[16]	Calibrated equalized odds[9]
	Reweighing[17]	Prejudice remover[18]	
	Optimized pre-processing[19]		
Discrimination via Representation	Balanced dataset collection	Adversarial training[2,5]	Troubling neurons turn OFF
		Adversarial fairness desideratum[20]	
		Semantic constraints[21]	
		Distance metrics[22,23]	
Prediction Quality	Diverse dataset collection[24]	Transfer learning[25]	
Disparity	Synthetic data generation[26]	Multitask learning[27]	

Preprocessing, in-processing, and postprocessing correspond to three stages of deep learning pipeline: dataset construction, model training, and model inference.

face images are classified into four categories: darker skin males, darker skin females, lighter skin males, and lighter skin females. Three gender classification systems are evaluated for the four groups, and substantial accuracy disparities are observed. For all three systems, the darker skin females group yields the highest misclassification rate, with error rate up to 34.7%. In contrast, the maximum error rate for lighter skin males is 0.8%. These results confirm that the model has violated *predictive quality parity* metric and raise an urgent need for building fair facial analysis systems.

Beyond the verification accuracy, model interpretability could be used to analyze the reasons of discrimination. A decomposition-based local DNN interpretation method, i.e., *class activation maps (CAM)*,[12] is used to investigate the regions of interest attended by the DNN models when making decisions. CAM is utilized to analyze two groups: lighter skin and darker skin group.[13] The visualization shows that the model needs to focus on eye region for lighter skin group, while focus on the nose region and chin region for darker skin group. It suggests different strategies are needed to make decisions for different demographic groups. If the training dataset has inadequate samples for darker skin group, the trained model may capture representation preference for the majority group and fail to learn effective classification strategy for minority darker skin group, thus leading to poor performance for minority group.

MITIGATION OF MODELING BIAS

After presenting bias detection approaches, we introduce below methods which could mitigate against adverse biases. A typical and simplified deep learning pipeline could be split into three stages: dataset construction, model training, and inference. Mitigation methods could be correspondingly divided into three broad groups: preprocessing, inprocessing, and postprocessing[28] (see Table 2). Preprocessing tries to debias and increase the quality of training set. Inprocessing adds auxiliary regularization term to the overall objective function during training, explicitly, or implicitly enforcing constraints for certain fairness metric. Postprocessing is performed after model training to calibrate the predictions of trained models.

Discrimination via Input

In this section, we introduce some representative mitigation methods as well as their empirical evaluation.

Preprocessing

A straightforward solution is to remove those fairness sensitive features from training data. For instance, surname and ZIP code can be deleted to reduce the discrimination of DNNs toward a certain race. A drawback of directly removing features is that this might lead to poor model performance and thus reduces model utility. We can replace these fairness sensitive features with alternative values. Take the sentence *"The conversation with Malik was heartbreaking"* for example, we can replace *"Malik"* with *"IDENTITY"* to reduce the possibility that DNN model shows discrimination based on races. In addition, there are some data-agnostic preprocessing transformation techniques. For instance, the weights of each training sample are given differently to ensure fairness

before model training.[17] Another transformation is formulated in a probabilistic framework, where features and labels are edited to ensure group fairness.[19]

In-Processing

An alternative approach to mitigate discrimination is via model regularization. The regularization implicitly or explicitly optimizes a fairness metric.

Implicit Regularization: The first category adds implicit constraints, which disentangle the association between model prediction and fairness sensitive attributes. It enforces DNN models to pay more attention to correct features relevant to prediction task, rather than capture spurious correlations between prediction task and protected attributes. Specifically, the model training is regularized with local DNN interpretation.[14,15] Beyond ground truth y for the input x, the regularization also needs featurewise annotations r, specifying whether each feature within the input correlates with protected attributes or not. The annotation r either could be labeled by domain experts or identified through the detection methods discussed in the section "Discrimination via Input." For instance, the annotation r for input "*The conversation with Malik was heartbreaking*" is $[0, 0, 0, 1, 0, 0]$, indicating that "Malik" is correlated with race. The overall loss function is denoted as

$$L(\theta, x, y, r) = \underbrace{d_1(y, \hat{y})}_{\text{Prediction}} + \underbrace{\lambda_1 d_2(f_{\text{loc}}(x), r)}_{\text{Fairness}} + \underbrace{\lambda_2 \mathcal{R}(\theta)}_{\text{Regularizer}} \quad (1)$$

where d_1 is the normal classification loss function, e.g., cross entropy loss and $\mathcal{R}(\theta)$ is a regularization term. Function $f_{\text{loc}}(x)$ is local interpretation method and d_2 is a distance metric function. The three terms are used to guide the DNN model to make right prediction, make decision based on right and unbiased evidences, and not overfit to training set, respectively. Hyperparameters λ_1 and λ_2 are used to balance three terms. Note that $f_{\text{loc}}(x)$ needs to be end-to-end differentiable, amenable for training with backpropagation and updating DNN parameters. The resulting fair model depends more on holistic information which is task relevant, while at the same time conditions less on sensitive attributes. Besides, the trained models also satisfy better *demographic parity* criteria.

Explicit Regularization: This category adds explicit constraints through updating model's loss function to minimize the performance difference between different groups.[16,18] They optimize the tradeoff between accuracy and a specific kind of fairness metric given training-time access to protected attributes. A representative example combines demographic parity and equality of odds into overall objective function.[16] Specifically, it define a "reduction" that treats the accuracy-fairness tradeoff as a sequential game between two players. At each step in the gaming sequence, one player maximizes accuracy and the other player imposes a particular amount of fairness. This method is advantageous in that it is model agnostic and could be applied to different DNN architectures.

Postprocessing

Postprocessing calibration takes the model's prediction and protected attribute to calibrate model's prediction during the inference time.[6,9] The goal is to enforce prediction distribution to approach either the training distribution, or a specific fairness metric. First, corpus-level constraints are utilized to enforce the model prediction distribution to follow the training data distribution.[6] Second, the calibration could also be performed toward a fairness metric. For instance, one technique takes as input an existing classifier and the sensitive feature, and derives a monotone transformation of the classifier's prediction to enforce the specified equalized of odds constraint.[9] These methods allow for diverse fairness metrics and prove to be effective to reduce discrimination. On the other hand, these methods could be problematic since they require inference-time access to protected attributes, which however usually are not available during inference time in real-world applications.

Evaluation of Mitigation Algorithms

We conduct experiments to evaluate the performance of different mitigation algorithms. We use *Adult Census Income*[a] and *COMPAS*[b] two datasets, containing 48 842 and 6167 instances, respectively. We use gender and race as protected attribute for the two datasets, respectively. Each dataset is split into 50% for training, 20% for validation, and 30% for testing. The base DNN model is a multilayer perceptron (MLP) with three layers.[c] We evaluate the metrics with the best performing model on validation set. The results are displayed in Table 3. Note that we average each number over three runs to eliminate influence of random initialization to DNN performance. We have the following key observations. First, without using the debiasing algorithms, DNN models would amplify bias

[a]https://archive.ics.uci.edu/ml/datasets/Adult
[b]https://github.com/propublica/compas-analysis
[c]https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPClassifier.html

TABLE 3. Mitigation comparison between five methods for discrimination via input.

	Adult Income				COMPAS			
Model/Data	Acc	Parity	Opty	Odds	Acc	Parity	Opty	Odds
Dataset_bias	n/a	0.386	n/a	n/a	n/a	0.747	n/a	n/a
DNN_original	0.836	0.347	-0.094	-0.089	0.658	0.741	−0.160	−0.136
Reweighting[17]	0.832	0.654	−0.106	−0.090	0.652	0.788	−0.186	−0.149
Optimized_pre[19]	0.778	0.573	−0.107	-0.088	0.665	0.959	−0.018	−0.024
Prejudice_rem[18]	0.817	0.961	0.005	0.039	0.635	0.937	0.008	0.062
Calibrated_odds[9]	0.804	0.546	0.148	0.052	0.639	0.819	0.036	0.150

For accuracy and demographic parity, the close to 1 the better. For equality of opportunity and equality of odds, the close to 0 the better.

existing in training data, as shown by the comparison of Parity value between DNN and training set. Second, there is fairness utility tradeoff, where most mitigation algorithms could compromise overall model accuracy. Third, fairness measurements could be conflicting with others. Some mitigation methods may be fair in terms of demographic parity, but may result in unfairness with regard to equality of opportunity/odds. Fourth, mitigation could possibly lead to discrimination toward majority groups, where equality of opportunity/odds metrics switch from negative values to positive values.

Discrimination via Representation

The goal is to reduce representation bias while at the same time preserve useful prediction properties of DNNs.

Preprocessing

Collecting balanced dataset is a possible way to alleviate representation bias, since prediction discrimination is partially caused by the difference of label distribution conditioning on protected features in the training data. Take text dataset, for example, gender swapping can be used to create a dataset which is identical to the original one but biased toward another gender. The union of the original and gender-swapping dataset would be gender balanced, which can be used to retrain DNN models. However, it is still not guaranteed that balanced dataset could eliminate the representation bias. Previous studies show that even training data are balanced, DNNs still could capture information like gender, race in intermediate representation.[2,5] Thus, more fundamental changes in DNN models are needed to further reduce discrimination.

In-Processing

Adversarial Learning. From model training perspective, adversarial training is a representative solution to remove information about sensitive attributes from intermediate representation of DNNs.[2,5] A predictor and an adversarial classifier are learned simultaneously. The goal of the predictor is to learn a high-level representation, which is maximally informative for the major prediction task, while the role of adversarial classifier is to minimize the predictor's ability to predict the protected attribute. The DNN is denoted as $f(x) = c(h(x))$, where $h(x)$ is the intermediate representation for input x, and $c(\cdot)$ is responsible to map intermediate representation to final model prediction. The protected attribute is denoted using z. An adversarial classifier $g(h(x))$ is also constructed to predict protected attribute z from representation $h(x)$. The adversarial training process is denoted as follows:

$$\begin{aligned} &\arg\min_{g} \quad L(g(h(x)), z) \arg\min_{h,c} \\ &L(c(h(x)), y) - \lambda L(g(h(x)), z) \end{aligned} \tag{2}$$

where the adversarial classifier is to penalize the representation of $h(x)$ if protected attribute z is predictable, parameter λ is used to negotiate the tradeoff between maximizing utility and fairness. The training is iteratively performed between the main classifier $f(x)$ and the adversarial classifier $g(h(x))$. Some methods implement adversarial using general cross entropy loss,[2,5] while some others use advanced adversarial objectives according to fairness desideratum.[20] The adversarial frameworks show improved performance on metrics like *demographic parity* and *equality of opportunity*. In the meantime, adversarial training has some pitfalls. First, it could not fully retain the semantic meaning of the data,[21] thus could harm model *accuracy*, especially when adding a strong regularization, i.e., a large λ. Second, it is also hard to stabilize the training, similar like adversarial training in other applications.

Beyond Adversarial Learning: Besides adversarial framework, some other advanced fair representation

learning methods are proposed recently. For instance, residual decomposition is used for fair representation learning.[21] Beyond enforcing inner representations to suppress protected attribute and predict the main task label, this method also adds regularization term to ensure that the debiased representation lies in the same space with original input. With the semantic meaning constraints, such representation learning methods thus could achieve better tradeoff between fairness metrics and accuracy. In addition, there exist nonadversarial methods using distance metrics, such as maximum mean discrepancy[22] and Wasserstein distance,[23] aiming to learn fair representations and eliminate disparities between different sensitive groups.

Postprocessing

The mitigation of discrimination through representation could also be implemented at the inference stage. The key idea is to suppress the neurons that have captured protected attributes. The process is split into two stages. First, global interpretation methods are used to locate the neurons that are highly related with protected attributes.[11] Second, the activation values flowing out from those neurons are set to zero, so as to turn OFF the correlation between protected attribute and DNN model prediction.

Prediction Quality Disparity

In this section, we introduce methods which could increase the prediction quality for underrepresented minorities.

Preprocessing

From data perspective, one straightforward idea to increase the prediction quality of underrepresented group is to enforce the training dataset to be diverse. This can be achieved by collecting data from more comprehensive data sources. For instance, the *Faces of the World* dataset is developed, aiming to achieve a uniform distribution of face images across two genders and four ethnic groups.[24] In some domains collecting data might be expensive or impractical, and generative adversarial networks (GANs) could be used to generate synthetic data. For instance, GANs are utilized to generate face images across all age ranges.[26] DNN models trained on this dataset are able to achieve equal predictive quality even for those previously minority age groups, e.g., age over 60.

In-Processing

Regularizing model training is another perspective to increase the accuracy of the minority groups. This could be implemented using the transfer learning framework. For instance, transfer learning is proposed to solve the problem of unequal face attribute detection performance across different race and gender subgroups.[25] A CNN model is first trained using source domain dataset, which is rich with data for the minority group, i.e., the aforementioned *Faces of the World* dataset. Then, the trained CNN model is transferred to the target domain, i.e., face attribute detection, to improve accuracy of the minority group. Transfer learning could promote both the overall accuracy and gender/race subgroup accuracy. Model training regularization could also be achieved under the multitask learning setting. For instance, a multitask learning framework is designed for joint classification of gender, race, and age of faces images.[27] This can yield significant accuracy improvement for different demographic subgroups, thus promoting model fairness in terms of *predictive quality parity* measurement.

CONCLUSION

With increasing adoption of DNNs in high-stake applications, their unfairness problem has attracted much attention recently. We give an overview of recent DNN bias detection and mitigation techniques from the computational perspective, with a particular focus on interpretability. Despite significant progresses for fairness in deep learning, several open problems remain around bias detection and mitigation. This includes 1) designing benchmark datasets for each DNN application, 2) the investigation of intersectional fairness, i.e., combination of multiple sensitive attributes, 3) the balance between fairness and utility tradeoff, 4) fairness measurements which could meet the specific requirements of each application domain. In future, endeavor from different disciplines, including computer science, statistics, cognitive science, should be joined together to eliminate disparity and promote fairness. In this way, DNN systems could be readily applied for fairness sensitive applications and really improve benefits of our society.

ACKNOWLEDGMENTS

This work was supported by National Science Foundation under Grant IIS-1657196, Grant IIS-1718840, Grant IIS-1939716, and DARPA Grant N66001-17-2-4031.

REFERENCES

1. P. Gajane and M. Pechenizkiy, "On formalizing fairness in prediction with machine learning," *Fairness, Accountability, Transparency Mach. Learn.*, Workshop, FAT/ML, 2018.

2. T. Wang, J. Zhao, M. Yatskar, K.-W. Chang, and V. Ordonez, "Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations," in *Proc. IEEE/CVF Int. Conf. Comput. Vision*, 2019, pp. 5309–5318.
3. S. Kiritchenko and S. M. Mohammad, "Examining gender and race bias in two hundred sentiment analysis systems," in *Proc. 7th Joint Conf. Lexical Comput. Semantics*, 2018, pp. 43–53.
4. J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," in *Proc. Conf. Fairness, Accountability Transparency*, 2018, pp. 77–91.
5. Y. Elazar and Y. Goldberg, "Adversarial removal of demographic attributes from text data," *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2018, pp. 11–21.
6. J. Zhao, T. Wang, M. Yatskar, V. Ordonez, and K.-W. Chang, "Men also like shopping: Reducing gender bias amplification using corpus-level constraints," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2017, pp. 2979–2989.
7. C. Dulhanty and A. Wong, "Auditing imagenet: Towards a model-driven framework for annotating demographic attributes of large-scale image datasets," *Workshop Fairness Accountability Transparency Ethics Comput. Vis. (FATE CV)*, 2019.
8. M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian, "Certifying and removing disparate impact," in *Proc. 21st ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2015, pp. 259–268.
9. M. Hardt *et al.*, "Equality of opportunity in supervised learning," in *Proc. 30th Int. Conf. Adv. Neural Inf. Process. Syst.*, 2016, pp. 3323–3331.
10. M. Du, N. Liu, and X. Hu, "Techniques for interpretable machine learning," *Commun. ACM*, vol. 63, no. 1, pp. 68–77, 2020.
11. B. Kim *et al.*, "Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV)," in *Proc. Int. Conf. Mach. Learn.*, 2018.
12. B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2016, pp. 2921–2929.
13. S. Nagpal, M. Singh, R. Singh, M. Vatsa, and N. Ratha, "Deep learning for face recognition: Pride or prejudiced?" 2019, *arXiv:1904.01219*.
14. A. S. Ross, M. C. Hughes, and F. Doshi-Velez, "Right for the right reasons: Training differentiable models by constraining their explanations," in *Proc. 26th Int. Joint Conf. Artif. Intell.*, 2017, pp. 2662–2670.
15. F. Liu and B. Avci, "Incorporating priors with feature attribution on text classification," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 6274–6283.
16. A. Agarwal, A. Beygelzimer, M. Dudík, J. Langford, and H. Wallach, "A reductions approach to fair classification," in *Proc. 35th Int. Conf. Mach. Learn.*, 2018, pp. 60–69.
17. F. Kamiran and T. Calders, "Data preprocessing techniques for classification without discrimination," *Knowl. Inf. Syst.*, vol. 33, pp. 1–33, 2012.
18. T. Kamishima, S. Akaho, H. Asoh, and J. Sakuma, "Fairness-aware classifier with prejudice remover regularizer," in *Proc. Joint Eur. Conf. Mach. Learn. Knowl. Discovery Databases*, 2012, pp. 35–50.
19. F. Calmon, D. Wei, B. Vinzamuri, K. N. Ramamurthy, and K. R. Varshney, "Optimized pre-processing for discrimination prevention," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 3992–4001.
20. D. Madras, E. Creager, T. Pitassi, and R. Zemel, "Learning adversarially fair and transferable representations," in *Proc. 35th Int. Conf. Mach. Learn.*, 2018, pp. 3384–3393.
21. N. Quadrianto, V. Sharmanska, and O. Thomas, "Discovering fair representations in the data domain," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2019, pp. 8219–8228.
22. C. Louizos, K. Swersky, Y. Li, M. Welling, and R. Zemel, "The variational fair autoencoder," in *Proc. 4th Int. Conf. Learn. Representations*, 2016, *arXiv:1511.00830*.
23. R. Jiang, A. Pacchiano, T. Stepleton, H. Jiang, and S. Chiappa, "Wasserstein fair classification," 2019.
24. S. Escalera *et al.*, "Chalearn looking at people and faces of the world: Face analysis workshop and challenge 2016," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit. Workshops*, 2016, pp. 706–713.
25. H. J. Ryu, H. Adam, and M. Mitchell, "Inclusivefacenet: Improving face attribute detection with race and gender diversity," *Fairness, Accountability, Transparency Mach. Learn.*, 2018.
26. Z. Zhang, Y. Song, and H. Qi, "Age progression/regression by conditional adversarial autoencoder," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2017, pp. 4352–4360.
27. A. Das, A. Dantcheva, and F. Bremond, "Mitigating bias in gender, age and ethnicity classification: A multi-task convolution neural network approach," in *Proc. Eur. Conf. Comput. Vision*, 2018, pp. 573–585.

28. R. K. Bellamy *et al.*, "AI fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias," *IBM J. Res. Develop.*, vol. 63, no. 4/5, pp. 4:1–4:15, Jul.–Sep. 2019.

MENGNAN DU is currently working toward the PhD degree with the Department of Computer Science and Engineering, Texas A&M University. His research interests include interpretable machine learning and fairness in deep learning. Contact him at dumengnan@tamu.edu.

FAN YANG is currently working toward the PhD degree with the Department of Computer Science and Engineering, Texas A&M University. His research interests include interpretable machine learning and fairness in deep learning. Contact him at nacoyang@tamu.edu.

NA ZOU is currently an assistant professor with the Department of Engineering Technology and Industrial Distribution, Texas A&M University, College Station, TX, USA. Her research interests include on data-driven modeling and knowledge discovery for tackling data challenges raised by large-scale, dynamic and networked data from various real-world information systems. Contact her at nzou1@tamu.edu.

XIA HU is currently an assistant professor and Lynn '84 and Bill Crane '83 Faculty Fellow with the Department of Computer Science and Engineering, Texas A&M University, College Station, TX, USA. His research interests include explainable artificial intelligence, automated machine learning, network analytics, and anomaly detection. Contact him at xiahu@tamu.edu.