Fairness in Deep Learning: A Computational Perspective

Mengnan Du[1], Fan Yang[1], Na Zou[2], Xia Hu[1]
[1]Department of Computer Science and Engineering, Texas A&M University
[2]Department of Industrial and Systems Engineering, Texas A&M University
{dumengnan,nacoyang,nzou1,xiahu}@tamu.edu

ABSTRACT

Deep learning is increasingly being used in high-stake decision making applications that affect individual lives. However, deep learning models might exhibit algorithmic discrimination behaviors with respect to protected groups, potentially posing negative impacts on individuals and society. Therefore, fairness in deep learning has attracted tremendous attention recently. We provide a comprehensive review covering existing techniques to tackle algorithmic fairness problems from the computational perspective. Specifically, we show that interpretability can serve as a useful ingredient, which could be augmented into the biases detection and mitigation pipelines. We also discuss open research problems and future research directions, aiming to push forward the area of fairness in deep learning and build genuinely *fair, accountable, and transparent* deep learning systems.

1. INTRODUCTION

Machine learning algorithms have achieved dramatic progresses nowadays, and are increasingly being deployed in high-stake applications, including credit, employment, education, criminal justice, personalized medicine, *etc* [20]. 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 main prediction task [33]. Eventually they are capable of exhibiting discriminatory behaviors against certain subgroups. For example, a recruiting tool believes that men are more qualified and shows bias against women [26], facial recognition performs extremely poorly for darker skin females [5], recognition accuracy is very low for subgroup of people in pedestrian detection of self-driving cars [33]. The fairness problem might cause adverse impacts on individuals and society. It not only limits a person's opportunities which he/she is qualified, but also might further exacerbate social inequity.

Among different machine learning models, the fairness problem of *deep learning models* has especially attracted attention from academia and industry recently. First, deep learning models have outperformed conventional machine learning models and achieved 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 contains human biases, which reflect historical prejudices against certain social groups and existing demographic inequalities. 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 biases 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 [20]. We first introduce the fairness problem in deep learning, including the categorizations, measurements, as well as interpretation techniques which are closely relevant to fairness. We proceed by presenting bias detection methods, followed by approaches to mitigate bias and create fair models from the computational perspective. We will show that interpretability could significantly contribute to better understandings of the reasons that affect fairness, and enables designing of mitigation strategies to combat the adverse influence of unfairness. Finally, we propose open challenges and future research directions. Throughout this article, *we don't discriminate between deep learning and DNN* (Deep neural network) and use them interchangeably. Besides, we abstract from the exact DNN architectures, e.g., convolutional neural network (CNN), recurrent neural network (RNN), and multi-layer perceptron (MLP), and focus more on conceptual aspects which underlie the success of DNN bias detection and mitigation techniques.

2. DNN FAIRNESS

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

2.1 Fairness Problem Categorization

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

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	*Health Care*: 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.
Prediction Quality Disparity	*Computer Vision*: Facial recognition performs extremely poorly for darker skin females. *Natural Language Processing*: Language identification models perform significantly worse when processing text produced by people belonging to certain races. *Health Care*: ICU mortality and psychiatric 30-day readmission model prediction accuracy is significantly different across gender and insurance types.

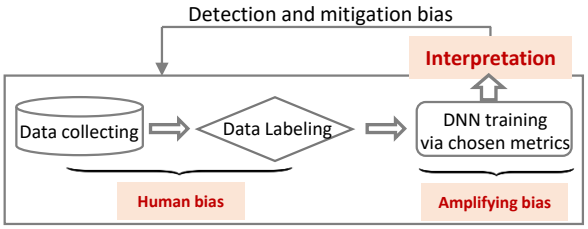

Figure 1: Bias exists in different stages of the DNN training pipeline, and interpretation could be utilized to detect and mitigate bias.

2.1.1 *Prediction Outcome Discrimination*

Discrimination refers to the phenomenon that DNN models produce unfavourable treatment of people due to the membership of certain demographic groups [20]. For instance, recruiting tool believes that men are more qualified and shows bias against women, and loan eligibility system negatively rates African Americans. Current DNN models generally follow the purely data-driven and end-to-end learning paradigm, which are trained with labeled data. The model training pipeline is illustrated in Fig. 1. Any training data may contain some biases, either intrinsic noise or additional signals inadvertently introduced by human annotators [21]. DNNs are designed to fit these skewed training data, and thus would naturally replicate the biases already existent in data. Even worse, DNNs not only rely on these biases to make decisions, but also make unwarranted implicit associations, and amplify societal stereotypes about people [4, 38]. This eventually results in trained models with algorithmic discrimination. Outcome discrimination can be further split into *Input* and *Representation* prospective. We present below detailed descriptions for these two subcategories, and give representative examples in Tab. 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 [24]. In the context of DNN systems, *protected attributes are often not observed in the input data*, mainly due to two reasons. Firstly, most DNN models rely on raw data, e.g., text, as input and thus protected attributes are not explicitly encoded in the input. Secondly, collecting protected attributes such as race and ethnicity information is often not allowed by the law in real world applications [24]. Despite the absence of explicit protected attributes, DNN models still could exhibit unintentional discrimination, mainly due to that there are some features highly correlated with class membership. For instance, ZIP code and surname could indicate race, many words within text input could be used to infer gender [24]. The model prediction might highly depends on the membership of a protected class. Eventually certain protected groups may be harmed by a model.

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. Besides, in some scenarios, finding the sensitive input attributes are challenging if the input dimension is too large [2]. In those settings, class memberships of certain protected attributes could be encoded in the DNN deep representations. DNN model will make decisions based on this information and produce discrimination. Thus prediction outcome discrimination could be detected and removed from the deep representation perspective.

2.1.2 *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 discriminations* which are mostly related to high-stake applications, this category also contains general applications, e.g., image recognition and natural language processing (See Tab. 1). Examples include the language identification systems perform significantly worse when processing text produced by people belonging to certain races [3, 23], health care applications including ICU mortality and psychiatric 30-day readmission model prediction accuracy is significantly different across gender and insurance types [7]. 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 widely used Imagenet dataset (ILSVRC 2012 subset with 1000 categories) as an example. Females comprise only 41.62% of images, people over 60 are almost non-existent, and the number of certain races is much fewer compared to others [14]. The typical objective of DNN model 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 model prediction accuracy, it might come at the expense of the under-represented populations and leads to poor performance for those groups.

2.2 Measurements of Fairness

Many different metrics have been proposed to measure the fairness of machine learning models. One line of work measures *individual fairness*, which follows the philosophy that similar inputs should yield similar predictions. Nevertheless, this leaves the open question of how to define input similarity [1, 20]. Another line of work focus on *group fairness*, where examples are grouped according to a particular sensitive attribute, and statistics about model predictions is calculated for each group and compared across groups [1]. Comparing to individual fairness, group fairness is more widely adopted in fairness research, and thus is the focus of this article. Different kinds of group fairness measurements have been proposed, and we will introduce below three mostly used ones.

Demographic Parity It asserts that average of algorithmic decisions should be the same across different groups: $\frac{p(\hat{y}=1|z=0)}{p(\hat{y}=1|z=1)} = 1$, where $\hat{y}$ is a model prediction, and z denotes *protected attribute*, e.g., race, gender, religion [6]. One issue of this metric is that it fails to consider that different groups could have very different labels y. Thus it may produce discrimination against qualified candidates from non-protected groups. A relaxed version of this metric is measured using: $\frac{p(\hat{y}=1|z=0)}{p(\hat{y}=1|z=1)} \geq \tau$, where τ is a given threshold, usually set as 0.8 [17]. Demographic parity is independent of the ground truth labels. This is useful especially when reliable ground truth information is not available, e.g., in applications like employment, credit, as well as criminal justice [20].

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: $\frac{p(\hat{y}=1|z=0,y=1)}{p(\hat{y}=1|z=1,y=1)} = 1$, where y is the ground truth label [22]. Essentially this is comparing the *true positive* rate across different groups. A symmetric measurement can be calculated for *false positive* rate: $\frac{p(\hat{y}=1|z=0,y=0)}{p(\hat{y}=1|z=1,y=0)} = 1$. Putting them together will result the *Equality of Odds* metric [22].

Predictive Quality Parity This metric measures the prediction quality difference between different protected subgroups. The quality here denotes quantitative model performance in terms of model predictions and ground truth, such as accuracy for multi-class classification, or precision, recall, F1 for binary classification application [5]. Thus it means that this metric also requires ground truth labels. It is desirable that a model has equal predictive quality across different demographic subgroups.

For a more comprehensive discussion of measurements, we refer interested readers to the work [20]. It is worth noting that different applications require different measurements which satisfy their specific ethical and legal requirements.

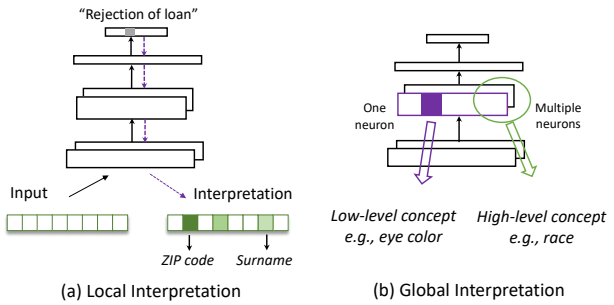

Figure 2: Illustration of DNN local interpretation as well as global interpretation.

2.3 Interpretability for Addressing DNN Fairness Problem

DNN models are often regarded as black-boxes and criticized by the lack of interpretability, since these models cannot provide meaningful interpretation on how a certain prediction is made. Interpretability could be utilized as an effective debugging tool to analyze the models (Fig. 1), ultimately enhancing the transparency and guaranteeing the fairness of models. DNN interpretability can generally be grouped into two categories: local interpretation and global interpretation, depending on whether the goal is comprehending a specific prediction locally or understanding how concepts are captured by deep representations globally [10].

Local Interpretation Local interpretation could illustrate how the model arrives at a certain prediction for a specific input (Fig. 2(a)). 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 as an example. The input to the model is a vector containing categorical features, and the interpretation result will be the heat map (or attribution map), where the features with higher scores represent higher relevance for the classification task.

Global Interpretation The goal is to provide a global understanding about what knowledge has been captured by a pre-trained DNN, and illuminate the learned representations in an intuitive manner to humans (Fig. 2(b)). The interpretation can be regarded as a function $f_{global} : E_h \rightarrow E_m$, mapping from intermediate representations E_h to human understandable concepts E_m [25]. The simplest way is to comprehend the concept captured by a single neuron. In this case, E_h is the representation derived from a specific channel at a specific layer. For instance, the (13×13) activations for the 151^{st} channel of *conv5* layer for a CNN responds to animal and human faces [35]. The combination of multiple neurons could represent more abstract concepts [18]. Here E_h corresponds to the combinations of different channels or even different layers. Especially for those protected concepts, they are usually based on multiple elementary low-level concepts. For instance, gender and race concept can be indicated via multiple local clues for a face image recognition application. Thus comparing to concepts learned by a single neuron, concepts yielded by a combination of neurons are more relevant to DNN fairness.

We will introduce below how DNN interpretability could act as a useful ingredient, which could be augmented into the biases detection and mitigation pipelines (Tab. 2).

Table 2: It is worth noting that DNN bias detection and mitigation solutions we review are not fully based on interpretability. Interpretability, however, could be used as an effective ingredient, which could be augmented into DNN model bias detection and mitigation pipelines.

	DNN fairness problem categorization		
Debias	Discrimination via Input	Discrimination via Representation	Prediction Quality Disparity
Bias detection	Local interpretation	Global interpretation	Local & global interpretation
Bias mitigation	Local interpretation	Global interpretation	Local & global interpretation

3. DETECTION OF MODELING BIAS

In this section, we present detection methods for aforementioned different kinds of model bias, by making use of DNN interpretability as an effective debugging tool.

3.1 Discrimination via Input

The source of prediction outcome discrimination could be traced back to the input features. As mentioned earlier, protected attributes, e.g., gender and race, 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 thus could be utilized to predict protected attribute and cause model bias [22]. 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, which is then further analyzed. Specifically, local interpretation methods can be roughly classified into four categories: *local approximation based* [28], *perturbation based* [11], *back-propagation based*, and *decomposition based methods* [13], all of which could be exploited to generate feature importance vector for an input. Local approximation based methods are based on the assumption that the behaviors of complex and opaque DNN model at the neighborhood of an input can be approximated by a simple and transparent white-box model [10]. For instance, a sparse linear model is exploited as the local model and the weight vector of the linear model is taken as feature importance interpretation for the input [28]. Perturbation based methods follow the philosophy that the contribution of a feature can be calculated by measuring how prediction probability changes when that feature is altered [11]. The motivation of back-propagation based methods is that the contribution of features could be derived by calculating the gradient of its variants of the output with respect to the model input, using back-propagation. The line of decomposition based methods usually derive the contribution of features via prediction decomposition. For instance, through modeling the information flowing process of the hidden representation vectors in RNN models, the RNN prediction is decomposed into additive contribution of each word in the input text [13]. All of aforementioned four kinds of methods are widely applied to generate feature importance interpretation for DNN models, and utilizing which approach highly depends on specific architectures (e.g., CNN, LSTM, MLP) and the application domains. 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 algorithmic discrimination towards certain race, and surname and ZIP code here are fairness sensitive features (Fig. 2(a)).

The second solution is implemented in the bottom-up manner. Humans first pre-choose features which they are skeptical to be associated with protected attributes, and then analyze feature importance of the identified features [26]. These fairness sensitive features are perturbed, either through omission where the feature is deleted directly or occlusion where feature is replaced with alternative features. We then feed the perturbed input to the DNN and observe the model prediction difference. If eventually the perturbation of those suspected fairness sensitive features causes significant model prediction change, it can be asserted that the DNN model has captured bias and makes decisions based on protected attributes. Note that statistical differences are calculated over a set of similar instances, so as to validate whether the model has violated *group fairness*.

A representative example is using local interpretation to detect the input bias in sentiment analysis systems [26]. To examine race bias, common African American first names (e.g., Ebony, Malik, Latisha, Jerome) and common European American first names (e.g., Ellen, Frank, Katie, Ryan) are chosen as race sensitive features. The comparison is between average prediction scores of sentences with African American first name (e.g., '*The conversation with Malik was heartbreaking*') and average prediction scores of sentences with European American first name (e.g., '*The conversation with Frank was heartbreaking*'). Except for the race sensitive first names, all other words within the compared sentences are the same. The results demonstrate that the DNN systems show statistically significant race bias. The systems consistently yield slightly higher sentiment prediction with African American name on the tasks of anger, fear and sadness intensity prediction. In contrast, on the task of joy and valence prediction, systems assign higher score to sentences with European American names. The results indicate the models have violated *demographic parity* metrics. They also to some extent reflect the stereotypes that African Americans are relevant to negative emotions.

3.2 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. Even when each input sample does not contain fairness sensitive features, DNN models still can discriminate based on classes like gender and race. For instance, CNN model could identify patients' self-reported sex from a retina image even though ophthalmologist cannot do this. DNN global interpretation could be exploited as a debugging tool to analyze the deep representations. The goal

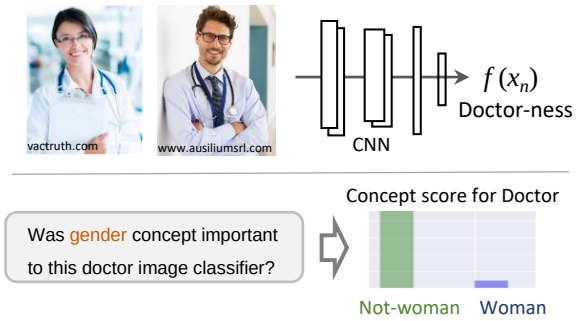

Figure 3: For a CNN based doctorness classifier, gender bias can be detected using global interpretation method CAV. Results show that this CNN has captured gender concept, and the *not-woman* concept would significantly increase doctor prediction confidence of the CNN classifier. Thus it indicates the CNN's discrimination towards woman.

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 representation discrimination.

Firstly, global interpretation is utilized to analyze the degree to which a DNN model has learned the concept relevant to a protected attribute. This is usually achieved by pointing to a direction in the activation space of DNN's intermediate layers [25, 40, 18]. Typical example for verifying whether DNN has captured a given concept is the *concept activation vector (CAV)* method [25]. 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. Note that this procedure could potentially lead to a meaningless CAV, since a random set of images are used to generate CAV. To guard against spurious results and to further confirm that the model indeed has captured the concept, this method proceeds to perform the statistical significant test. Instead of using a single set of random images and learning a single vector, multiple sets of random images are used to learn multiple CAVs, e.g., 500. The consistency among these CAVs would eventually indicate that the model indeed learns the concept.

Secondly, after confirming that a DNN has learned a protected concept, we would further proceed to test the contribution of this concept towards model's final prediction. Different strategies can be used to quantify the conceptual sensitivity, including the top-down manner which calculates the directional derivative of DNN's prediction to the concept vector [25], or the bottom-up manner which adds this concept vector to different inputs' intermediate activation and then observe the change of model prediction [31]. Ultimately the representation bias level for a protected attribute is described using a numerical score. For both manners, the higher of the numerical sensitivity score, the more significantly that this concept contributes to DNN's prediction.

A representative example is detecting the gender bias in deep representations of a CNN doctorness classifier, which predicts whether the person within an input image is a doctor or not (Fig. 3). This model has violated *demographic parity* metric and shows discrimination towards women. Global interpretation method CAV indicates that the model indeed has captured the gender concept, even though the main task is irrelevant to gender at all. Through calculating directional derivative of CNN's prediction to the CAV vector, the result further shows that the gender concept has a significant contribution for model prediction, and the *not-woman* concept would dramatically increase the model's prediction confidence of doctorness. The findings have conformed the CNN's discrimination towards women. They also to some extent reflect the commonly held gender stereotype that doctors are men.

3.3 Prediction Quality Disparity

Due to inadequate data collection for minority group, it usually happens that some groups appear more frequently than others. The DNN model will optimize for those groups in order to boost the overall model performance. This might lead to low prediction accuracy for the minority group, and thus causes unfairness and inequities.

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 [5]. Human 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 mis-classification rate, with error rate up to 34.7%. In contrast, the maximum error rate for lighter skin males is 0.8%. These results conform 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 [27]. A decomposition-based local DNN interpretation method, i.e., *class activation maps (CAM)* [39], is used to shed insight into the face recognition process and investigate the regions of interest attended by the DNN models when making decisions. CAM is utilized to analyze two groups: lighter skin group (Caucasoid), and darker skin group (Congoid) [27]. The visualization shows that the model needs to focus on eye region for lighter skin group, while focus on the nose region (below the eyes) and chin region for darker skin group. It suggests different strategies are needed to make decisions for different demographic groups. If the training dataset have 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.

4. MITIGATION OF MODELING BIAS

After presenting bias detection approaches, we introduce below methods which could mitigate against adverse biases and ensure fairness of DNNs. Mitigation strategies can gen-

erally be grouped into dataset refinement and model training two categories, agnostic of bias types. The former one tries to increase the quality of training set, while the latter one is achieved by adding auxiliary regularization term to the overall objective function, which explicitly or implicitly enforces constraints for certain measurements of fairness.

4.1 Discrimination via Input

For mitigation of input bias, the key idea is to enforce DNN models to pay more attention to correct features which are more relevant to the task at hand, rather than capture spurious correlations between the prediction task and fairness sensitive features, e.g., gender [32]. Attempts can be categorized as data-based and model-based approaches.

Data propocessing could be applied to reduce DNN model prediction discrimination. A straightforward solution is to remove those fairness sensitive features from training data. For instance, surname and geolocation (e.g., ZIP code) can be deleted to reduce the possibility that the DNN system estimates the race membership and thus reduce discrimination towards certain race. A drawback of directly removing features is that this might lead to poor model performance and thus reduces model utility. Instead of removing features, we can replace these fairness sensitive features with alternative values. Take the input sentence '*The conversation with Malik was heartbreaking*' for example, we can replace the feature *'Malik'* with *'IDENTITY'* to reduce the possibility that DNN model shows discrimination based on races.

An alternative approach to mitigate the input bias is via model regularization, taking into consideration of the fairness goal in the overall model objective function. Specifically, the model training is regularized with local DNN interpretation [12, 29]. Beyond ground truth y for the whole input x, this regularization also needs feature-wise annotations r, specifying whether each feature within the input correlates with protected attributes or not. For instance, the annotation r for input sentence '*The conversation with Malik was heartbreaking*' is $[0, 0, 0, 1, 0, 0]$, indicating that word 'Malik' is correlated with protected attribute (i.e., race), while the rest words are considered as task relevant. The annotation could be further incorporated into the training process, aiming to train a fair model. The overall loss function can be denoted as follows:

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

where d_1 is normal classification loss function, e.g., cross entropy loss, and $\mathcal{R}(\theta)$ is a regularization term. Function $f_{loc}(x)$ 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 the local interpretation method $f_{loc}(x)$ needs to be end-to-end differentiable, amenable for training with back-propagation and updating DNN parameters. A representative example is using input gradient interpretation method $\frac{\partial \hat{y}}{\partial x}$ for $f_{loc}(x)$ and $L2$ norm as distance metric for d_2 [29]. The intuition is to shrink the model's attention towards those features highly relevant to sensitive attributes, so as to reduce the model's dependence on these fairness sensitive features. The resulting fair model

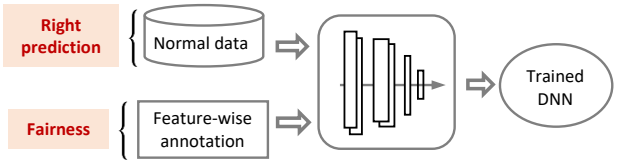

Figure 4: Using regularization for mitigation of discrimination via input. Besides normal training data and ground truth, feature-wise annotations are also needed, specifying which subset of features is fairness sensitive, and which subset is task relevant. The key idea is to enforce DNNs to depend less on sensitive features.

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 comparing to models without fairness term.

4.2 Discrimination via Representation

The goal is to reduce representation bias while at the same time preserve useful prediction properties of DNNs. Corresponding efforts can be grouped into dataset and model training two categories.

Collecting balanced dataset is a possible way is alleviate representation bias. This is effective mainly due to the reason that model prediction discrimination is partially caused by 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 towards another gender. The union of the original and gender-swapping dataset would be gender balanced, which can be used to retrain DNN models. This process typically is achieved in a crowdsourcing manner, e.g., by crowd workers from Amazon Mechanical Turk. However, it is still not guaranteed that balanced dataset could eliminate the representation bias. Previous studies show that even training data is balanced in terms of protected attributes, DNNs still could capture information like gender, race in intermediate representation [15, 33]. Thus more fundamental changes in the DNN models are needed to further reduce discrimination.

From model training perspective, adversarial training is a representative solution to remove information about sensitive attributes from the intermediate representation of DNNs and yield a fair classifier [15, 33, 36]. The goal is to learn a high-level input representation which are maximally informative for the major prediction task, while at the same time minimally predictive of the protected attributes (Fig. 5). A predictor and an adversarial classifier are learned simultaneously, where the role of the adversarial classifier is to minimize the predictor's ability to predict the protected attribute. The DNN model can be 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 model $f(x)$ can be arbitrary DNN which is learned through back-propagation, e.g., CNN and LSTM. The protected attribute we want to examine is denoted using z. Note that the main task $f(x) = c(h(x))$ is not inherently collated with protected attribute z. An adversarial classifier $g(h(x))$ is also constructed to predict

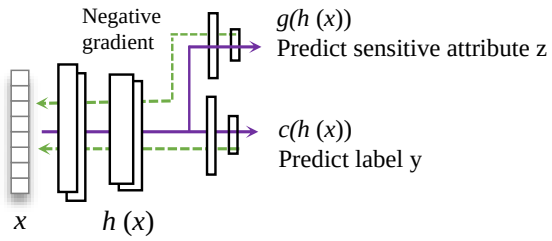

Figure 5: Using adversarial training for mitigation of discrimination via representation. The intuition is to enforce deep representation to maximally predict main task labels, while at the same time minimally predict sensitive attributes.

protected attribute z from representation $h(x)$. The adversarial training process can be denoted as follows:

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

where the adversarial classifier is to penalize the representation of $h(x)$ if protected attribute z is predictable, parameter λ is used to balance the prediction maximization and protected attribute suppression. The training is iteratively performed between the main classifier $f(x)$ and the adversarial classifier $g(h(x))$. After a certain number of iterations, we could obtain a debiased DNN model. Note that it is not sufficient to verify whether the model is debiased by using only the adversarial classifier $g(h(x))$. External verification, such as aforementioned global interpretation method CAV [25], should be used for sanity check so as to make sure the model indeed has removed representation bias.

Adversarial training is widely applicable for different DNN architectures and different kinds of input formats, including CNN with image data [33], RNN with text data [15], and MLP with categorical data [36, 2]. There are several interesting findings. Firstly, the distribution of examples over the sensitive attributes is crucial to the final model performance [2]. Using dataset which is balanced in terms of label distribution conditioned on sensitive attributes could significantly improve the fairness in the DNN model. Secondly, adversarial training could ensure fairness in terms of *demographic parity* and *equality of opportunity* metrics. On the other hand, this could possibly harm the model *accuracy* performance. Thirdly, carefully tuning the hyper-parameter λ may achieve amenable trade-off between increasing of demographic parity (or equality of opportunity) and decrease of model prediction accuracy.

4.3 Prediction Quality Disparity

The prediction quality for underrepresented minorities can be increased from two perspectives: increasing training set quality and regularizing model training.

Since limited data availability leads to the underrepresentation problem, thus 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 [16]. In some domains, e.g., health care, collecting data might be expensive or impractical. For those scenarios, Generative Adversarial Networks (GANs) could be used to generate synthetic data for those minorities [19]. This could increase the prediction quality of the minorities, while at the same time not affect prediction performance of the non-protected groups, thus avoiding discrimination for those groups. For instance, GAN is utilized to generate face images across all age ranges [37]. 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.

Regularizing model training is another perspective to increase accuracy of the minority groups. The regularization 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 [30]. A CNN model is firstly 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. A main benefit of transfer learning is that it could promote the overall accuracy as well as accuracy for gender and race subgroups. Model training regularization could also be achieved under the multi-task learning setting. For instance, a multi-task learning framework is designed for joint classification of gender, race, and age of faces images [8]. This can yield significant accuracy improvement for different demographic subgroups, thus promoting model fairness in terms of *predictive quality parity* measurement.

As a final step, sanity check is required to evaluate whether the trained DNN has indeed solved the underrepresentation problem for minority group. Both local and global interpretation could serve this purpose. On one hand, local interpretation such as *class activation maps (CAM)* could be used to analyze the attention of the DNN model [39]. A debiased model is supposed to focus on the correct locations for the minority group. On the other hand, global model interpretation could also be used to analyze the learned deep representation. If it turns out that DNN model has captured some distinct characteristics relevant to the minority group, it can be claimed that the underrepresentation problem is indeed solved.

5. RESEARCH CHALLENGES

Despite significant research progresses for fairness in deep learning, there are still some urgent challenges which deserve further research from the community.

5.1 Benchmark Datasets

Benchmark datasets are lacking to systematically examine the inappropriate biases in trained DNN systems [26]. Benchmark dataset here means the dataset which has been teased out biases towards certain protected groups. Current practice of testing the performance of a DNN model is using the hold-out test sets, which are likely to contain the same biases as the training set. Test sets might fail to unveil the unfairness problem of trained models. Thus it is recommended for each trained DNN to evaluate its performance on the benchmark dataset, serving as a supplementary test set beyond the normal test set. To facilitate the construc-

tion of benchmark datasets, it is also encouraged that statistics information including geography, gender, ethnicity and other demographic information should be provided, for those datasets containing information about people.

5.2 Intersectional Fairness

The investigation of intersectional fairness, i.e., combination of multiple sensitive attributes, is relatively lacking in current research. Take bias mitigation for example, current work generally focus on one kind of bias. Although this may increase model fairness in terms of a specific bias, it is highly possible that the model is still biased from the intersectional perspective. For instance, a DNN classifier is fair to women, while exhibiting discrimination towards a subdivision group, e.g., African American women or women over the age of 60. Similarly for DNN based job recruiting tool, even if the debiased model is free of gender bias, it is hard to guarantee that the model is not biased towards other protected attributes, e.g., race, age. More work is needed to figure out methods which are effective for identification and mitigation of intersectional biases.

5.3 Fairness and Utility Trade-off

The removal of bias could possibly hurt the model's ability for main prediction task. For instance, adversarial training could increase fairness with respect to demographic parity measurement. A possible deficiency of this mitigation solution is that it could compromise overall prediction accuracy, especially the accuracy for non-protected groups. Thus this might undermine the principle of beneficence. It remains a challenge to simultaneously reduce unintented bias and maintain satisfactory model prediction performance.

5.4 Formalization of Fairness

As the field of fairness machine learning is evolving quickly, there is still no consensus about the measurements of fairness. In certain cases, some measurements could be conflicting with others. A model may be fair in terms of one metric, but may lead to other sorts of unfairness. For instance, a loan approval tool may satisfy demographic parity measurement, while violating equality of opportunity measurement. There is no silver bullet, and each application domain calls for a fairness measurement or combination of measurements which meet its specific requirements [20].

5.5 Fairness in Large-scale Training

Large-scale training is employed in some domains to boost model performance. Take NLP domain for instance, current paradigm is to pre-train language models (e.g., BERT [9] and XLNet [34]) on large-scale text corpus, which will be further fine-tuned on downstream tasks such as machine translation. These powerful language models could capture biases and propagate them to other tasks. Since these models need to be trained on corpus with billion-scale words and are typically trained for days, bias mitigation either through data preprocessing or training regularization remains a challenge and more research is needed in this direction.

6. CONCLUSIONS

With increasing adoption of DNNs in high-stake real world applications, e.g., job hunting, criminal justice and loan approval, the undesirable algorithmic unfairness problem has attracted much attention recently. We present a clear categorization of unfairness and introduce the most widely used measurements of fairness. By introducing interpretability as an essential ingredient, we also give a comprehensive overview of existing bias detection and mitigation techniques from the computational perspective. The model bias to some extent exposes biases present in our society. To really benefit our society, DNN models are supposed to reduce these biases instead of amplifying biases. 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 all groups.

7. REFERENCES

[1] A. Beutel, J. Chen, T. Doshi, H. Qian, A. Woodruff, C. Luu, P. Kreitmann, J. Bischof, and E. H. Chi. Putting fairness principles into practice: Challenges, metrics, and improvements. *AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society (AIES)*, 2019.

[2] A. Beutel, J. Chen, Z. Zhao, and E. H. Chi. Data decisions and theoretical implications when adversarially learning fair representations. *Fairness, Accountability, and Transparency in Machine Learning (FAT/ML)*, 2017.

[3] S. L. Blodgett, L. Green, and B. O'Connor. Demographic dialectal variation in social media: A case study of african-american english. *2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2016.

[4] T. Bolukbasi, K.-W. Chang, J. Y. Zou, V. Saligrama, and A. T. Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Thirtieth Conference on Neural Information Processing Systems (NIPS)*, 2016.

[5] J. Buolamwini and T. Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency (FAT*)*, pages 77–91, 2018.

[6] T. Calders and S. Verwer. Three naive bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery*, pages 277–292, 2010.

[7] I. Y. Chen, P. Szolovits, and M. Ghassemi. Can ai help reduce disparities in general medical and mental health care? *AMA journal of ethics*, 2019.

[8] A. Das, A. Dantcheva, and F. Bremond. Mitigating bias in gender, age and ethnicity classification: a multi-task convolution neural network approach. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[9] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2019.

[10] M. Du, N. Liu, and X. Hu. Techniques for interpretable machine learning. *arXiv preprint arXiv:1808.00033*, 2018.

[11] M. Du, N. Liu, Q. Song, and X. Hu. Towards explanation of dnn-based prediction with guided feature inversion. *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2018.

[12] M. Du, N. Liu, F. Yang, and X. Hu. Learning credible deep neural networks with rationale regularization. In *IEEE International Conference on Data Mining (ICDM)*, 2019.

[13] M. Du, N. Liu, F. Yang, S. Ji, and X. Hu. On attribution of recurrent neural network predictions via additive decomposition. In *The World Wide Web Conference (WWW)*, 2019.

[14] C. Dulhanty and A. Wong. Auditing imagenet: Towards a model-driven framework for annotating demographic

attributes of large-scale image datasets. *arXiv preprint arXiv:1905.01347*, 2019.

[15] Y. Elazar and Y. Goldberg. Adversarial removal of demographic attributes from text data. *2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2018.

[16] S. Escalera, M. Torres Torres, B. Martinez, X. Baró, H. Jair Escalante, I. Guyon, G. Tzimiropoulos, C. Corneou, M. Oliu, M. Ali Bagheri, et al. Chalearn looking at people and faces of the world: Face analysis workshop and challenge 2016. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2016.

[17] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*. ACM, 2015.

[18] R. Fong and A. Vedaldi. Net2vec: Quantifying and explaining how concepts are encoded by filters in deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[19] M. Frid-Adar, E. Klang, M. Amitai, J. Goldberger, and H. Greenspan. Synthetic data augmentation using gan for improved liver lesion classification. In *IEEE international symposium on biomedical imaging (ISBI)*, 2018.

[20] P. Gajane and M. Pechenizkiy. On formalizing fairness in prediction with machine learning. *Fairness, Accountability, and Transparency in Machine Learning (FAT/ML)*, 2018.

[21] S. Gururangan, S. Swayamdipta, O. Levy, R. Schwartz, S. R. Bowman, and N. A. Smith. Annotation artifacts in natural language inference data. *North American Chapter of the Association for Computational Linguistics (NAACL)*, 2018.

[22] M. Hardt, E. Price, N. Srebro, et al. Equality of opportunity in supervised learning. In *Advances in neural information processing systems (NIPS)*, 2016.

[23] D. Jurgens, Y. Tsvetkov, and D. Jurafsky. Incorporating dialectal variability for socially equitable language identification. In *The 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2017.

[24] N. Kallus, X. Mao, and A. Zhou. Assessing algorithmic fairness with unobserved protected class using data combination. *arXiv preprint arXiv:1906.00285*, 2019.

[25] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viegas, and R. Sayres. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). *International Conference on Machine Learning (ICML)*, 2018.

[26] S. Kiritchenko and S. M. Mohammad. Examining gender and race bias in two hundred sentiment analysis systems. *Proceedings of the 7th Joint Conference on Lexical and Computational Semantics*, 2018.

[27] S. Nagpal, M. Singh, R. Singh, M. Vatsa, and N. Ratha. Deep learning for face recognition: Pride or prejudiced? *arXiv preprint arXiv:1904.01219*, 2019.

[28] M. T. Ribeiro, S. Singh, and C. Guestrin. Why should i trust you?: Explaining the predictions of any classifier. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*. ACM, 2016.

[29] A. S. Ross, M. C. Hughes, and F. Doshi-Velez. Right for the right reasons: Training differentiable models by constraining their explanations. *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI)*, 2017.

[30] H. J. Ryu, H. Adam, and M. Mitchell. Inclusivefacenet: Improving face attribute detection with race and gender diversity. *Fairness, Accountability, and Transparency in Machine Learning (FAT/ML)*, 2018.

[31] P. Upchurch, J. Gardner, G. Pleiss, R. Pless, N. Snavely, K. Bala, and K. Weinberger. Deep feature interpolation for image content changes. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2017.

[32] J. Wang, J. Oh, H. Wang, and J. Wiens. Learning credible models. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD)*, 2018.

[33] 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. *International Conference on Computer Vision(ICCV)*, 2019.

[34] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q. V. Le. Xlnet: Generalized autoregressive pretraining for language understanding. *arXiv preprint arXiv:1906.08237*, 2019.

[35] J. Yosinski, J. Clune, A. Nguyen, T. Fuchs, and H. Lipson. Understanding neural networks through deep visualization. *ICLR workshop*, 2015.

[36] B. H. Zhang, B. Lemoine, and M. Mitchell. Mitigating unwanted biases with adversarial learning. In *AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society (AIES)*, 2018.

[37] Z. Zhang, Y. Song, and H. Qi. Age progression/regression by conditional adversarial autoencoder. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[38] J. Zhao, T. Wang, M. Yatskar, V. Ordonez, and K.-W. Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2017.

[39] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Learning deep features for discriminative localization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[40] B. Zhou, Y. Sun, D. Bau, and A. Torralba. Interpretable basis decomposition for visual explanation. In *European Conference on Computer Vision (ECCV)*, 2018.