A Generative Approach to Mitigate Bias in Face Matching using Learned Latent Structure

Jamal Alasadi
Rutgers University, USA
and University of Thiqar, Iraq
jamal.alasadi@rutgers.edu

Ahmed AlHilli
Al-Fural Al-Awsat
Technical University, Iraq
ahmed.alhilli@atu.edu.iq

Pradeep K. Atrey
State University of New York
Albany, USA
patrey@albany.edu

Vivek K. Singh
Rutgers University
State University of New Jersey, USA
v.singh@rutgers.edu

2022 IEEE Eighth International Conference on Multimedia Big Data (BigMM) | 978-1-6654-5963-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/BIGMM55396.2022.00033

***Abstract*—This work tackles the problem of bias in face matching algorithms. Face matching refers to the task of matching a low-resolution face image of a person with a high-resolution face image of the same person and has applications in security and personalization. Algorithmic bias is the difference in performance of an algorithm based on demographic descriptors of various users. Such bias can significantly reify and amplify societal biases and make certain advancements in technology benefit one section of the society while hurting the other. This work proposes a generative AI framework that can counter multiple kinds bias (e.g., gender bias and age bias) at the same time. The framework consists of two major components: a variational auto-encoder (VAE) that converts the images into their more generic underlying representation, and second, a neural network architecture that uses the above representations to undertake multi-label classification. A generative approach is useful in ensuring that the system learns to deal with the underlying (latent) structure of the data for better generalizability and bias reduction. The approach is tested over a public image dataset and found to be effective at reducing bias while maintaining high accuracy.**

***Keywords*-Fairness, Bias in Machine Learning, Face Matching, Face Detection, Neural Network**

I. Introduction

Multimedia computing has grown exponentially in the recent past and is now impacting human lives in fields ranging from healthcare to security. Hence, it is important to ensure that the algorithms adopted are not only accurate but also fair, i.e., free of bias. Bias in algorithmic settings is often defined as the difference in performance of an algorithm based on demographic descriptors of different users. Such bias can significantly reify and amplify societal biases and make certain advancements in technology benefit one section of the society while hurting the other [1], [2].

Face matching is the problem of matching a low-resolution face image of a person with a high-resolution face image of the same person [3]. While long-term face image datasets (e.g., an organization's database of employees) are captured at high resolution, they often need to be matched with low-resolution face images (e.g., those captured by a security camera) in different practical tasks. Such face matching is now increasingly being used to decide who gets access to buildings (and prosecute trespassers), to search for criminals, and for customizing digital applications. Hence, fairness in this process is very important to ensure that certain sections of society are not unfairly denied access to resources or criminalized simply because of their demographic characteristics.

There have been very few attempts to date to quantify and ameliorate the issue of bias in face-matching algorithms and especially, there is a lack of generative approaches that can handle intersectional bias, i.e., bias due to multiple demographic attributes at the same time (e.g., age and gender). A generative approach is useful in ensuring that the system learns to deal with the underlying (latent) structure of the data and any under/over representation in the training data is countered elegantly. Dealing with multiple biases at the same time is important because in real world multiple biases (e.g., gender bias, racial bias, age bias) often co-occur and hence countering them in the same framework is important.

Specifically, we present a new framework for the fair matching of low-resolution and high-resolution facial images. The framework consists of two parts: a variational auto-encoder (VAE) that converts the images into their more generic underlying representation, and second, a neural network architecture that uses the above representations to undertake a multi-label classification. While the primary classification task is about face matching (same/different) of the person in the LR image, the secondary labels are about demographic properties of the person. The loss function has been designed so as to support high accuracy at face matching while ensuring *low accuracy* at demography detection. The approach punishes the network if a demographic property (e.g., gender, age) is predictable from the matching prediction because that would indicate that the prediction is not independent of the demography of the person being considered. To give an analogy, if just knowing the result of a college admission algorithm (accept/reject) is enough to infer the gender of the applicant, then the algorithm is likely biased towards a gender identity.

Contributions. The contributions of this paper can be summarized as follows:

- A novel convolutional neural network based approach

978-1-6654-5963-1/22/$31.00 ©2022 IEEE
DOI 10.1109/BigMM55396.2022.00033

for face matching that utilizes a generative AI approach. The generative VAE approach allows for learned latent variables to modify the respective probabilities distribution of individual data points while training for better generalizability and bias reduction.
- Handling multiple kinds of bias (e.g., gender bias, age bias) at the same time via the use of multi-label classification framework. The loss function in the framework incentivizes learning identity while disentivizing learning demographic features.

Paper Organization. The rest of the paper is organized as follows. Section II describes the related work and motivates the problem. Section III, explains the proposed framework and describes the case scenarios that we consider. Our evaluation results are presented in Section IV and discussed in Section V. Subsequently, Section VI concludes the paper.

II. Related Work

Despite impressive advancements, multiple machine learning algorithms have recently been reported to be biased. These include algorithms for predicting recidivism, search results, policing, and facial analysis [4], [1], [5]. For instance, Buolamwini et al., reported that the performance of various facial analysis systems has been affected by various biases [2].

Consequently, various types of interventions try to introduce fairness into the machine learning pipelines. These include pre-processing (i.e., processing the data before going into the algorithm), in-processing, and post-processing approaches. A frequently described reason for the existence of bias is the imbalance in the training data for different demographic groups, resulting in limited training opportunities for certain groups. Consequently, mutliple researchers have proposed methods to counter this imbalance. In some cases, the researchers normalize the incoming data across different groups to reduce bias [6] or resample [7] for fairness. Unfortunately, these approaches focus on class imbalances instead of the underlying mechanisms that explain the variability within a class.

A long history of machine learning has shown that learning the structure of data is a common component of learning, including expectation maximization [8], topic modeling [9], latent-SVM [10], and more recently, Variational autoencoders [11], [12]. The proposed algorithm takes advantage of the latent structure of the data and automatically debiases it whenever training is performed. It does not require any pre-processing or annotations before being tested or trained.

Recent developments in data transformation [6] and generative models [13] have allowed for fairer training data generation [14]. In a paper, Sattigeri and colleagues [13] show how a generative adversarial network can produce a reconstructed dataset with more accurate and fair attributes. Despite the existence of methods that can be used to minimize discrimination in data [6], these methods are not learned in a way that is adaptive enough during training.

Supervised learning techniques have been used to analyze the biases in data sets. These include clustering techniques to identify clusters in the data and resampling the training data to produce smaller sets of representative examples. [15]. This method cannot be used to analyze large datasets such as images due to the lack of a cluster. In order to overcome these limitations, we use a variational approach to learn the underlying structure of the data.

To handle multiple types of bias at the same time, this work builds upon the recent advances in multi-label classification and adversarial modeling [16]. The framework tries to undertake multiple classifications at the same time but the loss function in the framework incentivizes learning the identity of the person while disincentivizing the learning of demographic atrributes. In recent efforts, Gong et al., proposed a debiasing network that adversarially learns to generate disentangled representations for unbiased face and demographics recognition [17] This disincentivization of learning demographic attributes has also previously been undertaken using generative adversarial networks by Alasadi et al., [3]. However, none of these works employ a generative process to learn the latent structure of the facial data to support bias reduction in a generalizable manner.

III. Proposed Framework

A. Quantifying Bias

This work focuses on group fairness, which is a type of fairness that divides the world into groups defined by one or multiple high-level sensitive attributes. It needs a particular relevant statistic (e.g., accuracy, true positive rate) about the classifier to be the same across those combinations. We describe the popular definitions of these kinds used in recent research [18]. In one such definition, a classifier is considered to make a fair decision if the prediction $\hat{Y}$ from features X is independent of the protected attributes S (e.g., gender) [19] i.e.

$$P(\hat{Y}=1|S=0)=P(\hat{Y}=1|S=1) \tag{1}$$

Such an absolute notion of fairness is rarely achieved in practical systems. Following relevant literature, here we focus on the notion of equality of accuracy and "equalized odds" [4]. A model is considered fair when across both demographic groups ($S=0$ and $S=1$), the predictor $\hat{y}$ has equal TPR and equal FPR [20]. This enforces that the accuracy is equally high for particular sub-populations within the overall population because the rate of positive and negative classification is the same across such groups.

$$P(\hat{Y}=1|Y=y,S=0)=P(\hat{Y}=1|Y=y,S=1) \tag{2}$$

where $y \in (0,1)$.

This objective can be measured by a metric that determines the disparate mistreatment [21]. It calculates the total differences between TPRs and FPRs for both demographics, given by D_{TPR} and D_{FPR}, respectively, which are computed as follows:

$$D_{TPR} : |P(\hat{Y} = 1|Y = 1, S = 1) - P(\hat{Y} = 1|Y = 1, S = 0)| \tag{3}$$

$$D_{FPR} : |P(\hat{Y} = 1|Y = 0, S = 1) - P(\hat{Y} = 1|Y = 0, S = 0)| \tag{4}$$

B. Learning Model: Learning Latent Structure with Variational Autoencoders

Here, we introduce a debiasing-VAE (DB-VAE) network architecture that allows us to perform unsupervised learning of latent variables while training. The VAE learns the true distribution of the variables given a data point by performing an estimate, i.e., $q(z|x)$. Instead of using classical VAE architecture, we introduce an output variable d where $\hat{x} \in \mathbf{R}^d$ and k is a latent variable. This allows us to perform unsupervised learning of variable distributions.

In order to maintain the original learning task, we explicitly supervise the output variables. This changes the traditional VAE model from an unsupervised to a semi-supervised model, where some of the variables are learned by reconstructing the input and the others are supervised for a specific task (e.g., classification). For instance, if we wanted to train a binary classifier our DB-VAE model would learn k latent variables and a variable (i.e., $\hat{y} \in \mathbf{0}, \mathbf{1}$) for classification. This method allows us to perform unsupervised learning of variable distributions while training. It also allows us to debias the data collected during the training.

The network training using backpropagation works with a three-component loss function comprised of the supervised latent loss function, a reconstruction loss, and a latent loss for the unsupervised variational autoencoder. For instance, the supervised loss l_y is given by the cross entropy loss, the reconstructed is given by the l_x norm between the input and the reconstructed output, and the latent loss l_{KL} is given by the Kullback-Liebler (KL) divergence. The total loss is a weighted combination of these three losses:

$$\begin{aligned} L_{tot} = [\frac{1}{2}\sum_{j=0}^{k-1}(\sigma_j + \mu_j^2 - 1 - \log(\sigma_j))] + [||x - \hat{x}||_p] \\ - \left[\sum_{i \in \{0,1\}} y_i \log\left(\frac{1}{\hat{y}_i}\right)\right] \\ = l_y(y, \hat{y}) + l_x(x, \hat{x}) + l_{kl}(\mu, \sigma) \end{aligned} \tag{5}$$

The supervised loss function (l_y) consists of a loss for the primary prediction task label (e.g., same/different) and the losses in predicting any sensitive variables (e.g., age, gender). Given that we want the model to do well at the primary task but ***not*** well at learning the sensitive variables, they are assigned a negative weight. Finally, c includes the weight coefficients of the loss functions used to determine their relative importance. In the current implementation where the primary task is to learn whether the facial images belong to the same/different person, and the sensitive variables are age and gender, this results in:

$$l_y(y, \hat{y}) = c(l_{y\_same}) - (l_{y\_gender}) - (l_{y\_age}) \tag{6}$$

The baseline model for a given task is similar to the DB-VAE model, except that it does not have a multi-task network. Rather, it focuses on a single task i.e., predicting same/different .

$$l_{y\_baseline}(y, \hat{y}) = l_{y\_same} \tag{7}$$

For classes that involve the optimization of the supervised loss and unsupervised loss, the gradients from the latent space and decoder should not be stopped. This ensures that the training algorithm only focuses on optimizing the supervised loss for classifiers.

C. Problem Formulation

Following the relevant literature [22], we consider a classifier to be biased if its performance changes based on any particular sensitive characteristic (e.g., age, gender) of the data. This means that the algorithm is considered fair with respect to a particular variable Z if the classifier's output is the same whether we condition on that variable or not so for example if we have a single binary variable Z. The likelihood of the prediction being correct should be the same whether or not z = 0 or z = 1.

For example, The values of various latent variables, such as the gender and the age of the individual should not affect the ultimate decision of the classifier.

$$|P(\hat{Y} = 1|z = 0, y = 1) = P(\hat{Y} = 1|z = 1, y = 1) \tag{8}$$

A classifier, $f_\theta(x)$ is biased if its decision changes after being exposed to additional sensitive feature inputs, it is fair with respect to variables z if:

$$f_\theta(x) = f_\theta(x, z) \tag{9}$$

Hence, the problem we encounter is that of finding the right parameters θ such that the overall loss function is minimized

$$\begin{aligned} \theta^* = \arg\min_\theta l_{kl}(\mu, \sigma) + l_x(x, \hat{x}) + c(l_{y\_same}) - \\ ((l_{y\_gender}) + (l_{y\_age})) \end{aligned} \tag{10}$$

D. System Implementation

Our model is implemented on Colab using the PyTorch open source framework [23], [24]. We validate our method using the Celeb A dataset made available by [25]. The CelebA dataset contains 202,599 face images and 5 landmark locations. It features 40 binary attribute annotations per

Figure 1: Sample images from the dataset.

image. Celeb A dataset includes demographic information in terms of gender (male/female) and age (young, not young), among other terms. Some of the sample images from the dataset are shown in Figure 1.

An overview of the architecture for Low-High resolutions variational autoencoder to mitigate the bias is shown in Figures 2 (baseline) and 3 (proposed). Both architectures are similar in terms of having two VAE pathways, which result in a short-hand representation (latent vectors) for the high-resolution and the low resolution image. This short-hand representation is combined, and is followed by a fully connected network that tries to predict the variables needed. While the baseline model focuses only on predicting whether the facial images belong to the same/different person, the proposed model has an added fairness component. In the fairness component, the problem becomes that of multi-label classification and using a weighted average for the loss function in back propagation. Details of the architecture(s) are as follows.

The feature maps obtained from the layers are processed by the CNN network that consists of five convolutional layers with ReLU activation function after each layer except the last convolutional layer for the second path. The first path consists of five convolutional layers for encoding and four de-convolution layers for decoding. The second path for the low-resolution image consists of three convolution layers for encoding and four de-convolution layers for decoding. The latent vector for both networks combines in a concatenation layer which means vector mean 1 and mean 2, variance 1 and variance 2 are fed in one classifier with the baseline model and three classifiers in the proposed model. Each classifier has four convolution layers.

We implement our models using the CUDA package version 11.2. The batch size of the training is 50 and for validation set is 100. Adam optimization with a learning rate of 0.00001 with a weight decay of 0.0005 is used in the training phase. We train the network for 5 epochs, and the output label is set to -1 and 1 nodes for each classifier (same/different, male/female, and young/old). The training took approximately 8 hours on Google Colab platform.

After pre-processing the training are ready to train the network. The training parameters are set based on the stochastic gradient descent with a patch size of 50. The latent variable size for the high-resolution network is 500 and 60 for the low-resolution network.

In each training trial, an image is taken from the training set, and with 50% probability a low-resolution version of the same person's face image is constructed. In the other 50% cases, different face images are used to construct the low-resolution version. Both high and low-resolution images are fed to the network, and the output label is set to 1 if the low-resolution image is the same (otherwise 0). We train the network for 5 epochs.

In the testing phase, the pre-processed images are provided along with their labels to the network, and the accuracy of classification is calculated after processing all the images in the testing set.

IV. Results

We evaluate the performance of our bias-reduction models relative to the accuracy metric using the confusion matrix as a starting point. The primary task for which are measuring accuracy is face matching and the two sensitive attributes considered are age and gender. Please note that we recognize gender to include multiple non-binary variants but focus on the two genders as a proof of concept based on the currently available annotations in the datasets. Similarly, we consider age as a binary variable due to the dataset limitations.

The first approach in this framework is to train a learning system that utilizes multi-label classification in an adversarial setting. However, it now works with a generative process (VAE) and can handle multiple sensitive attributes at the same time.

Specifically, the prediction network has multiple heads, one of which is the prediction of the target attribute y and the others are for the prediction of the sensitive attributes. The goal of this system is to try and remove the effects of the sensitive attribute on the final decision based on the negative weights given to their loss.

Our training set is around 70 percent images, validation 15 percent and our test set is 15 percent images.

A. *Performance of the Face Matching Algorithm*

Tables I and II show the results for accuracy as well as TPR (True Positive Rate) and FPR (False Positive Rate) for the task of matching low-resolution and high-resolution face images.

We notice that the proposed approach yields higher accuracy and TPR and lower FPR scores than the baseline condition. Hence, besides the fairness motivation, the proposed multi-label architecture might also be better at ensuring accuracy, potentially due to the rejection of gender-based signal.

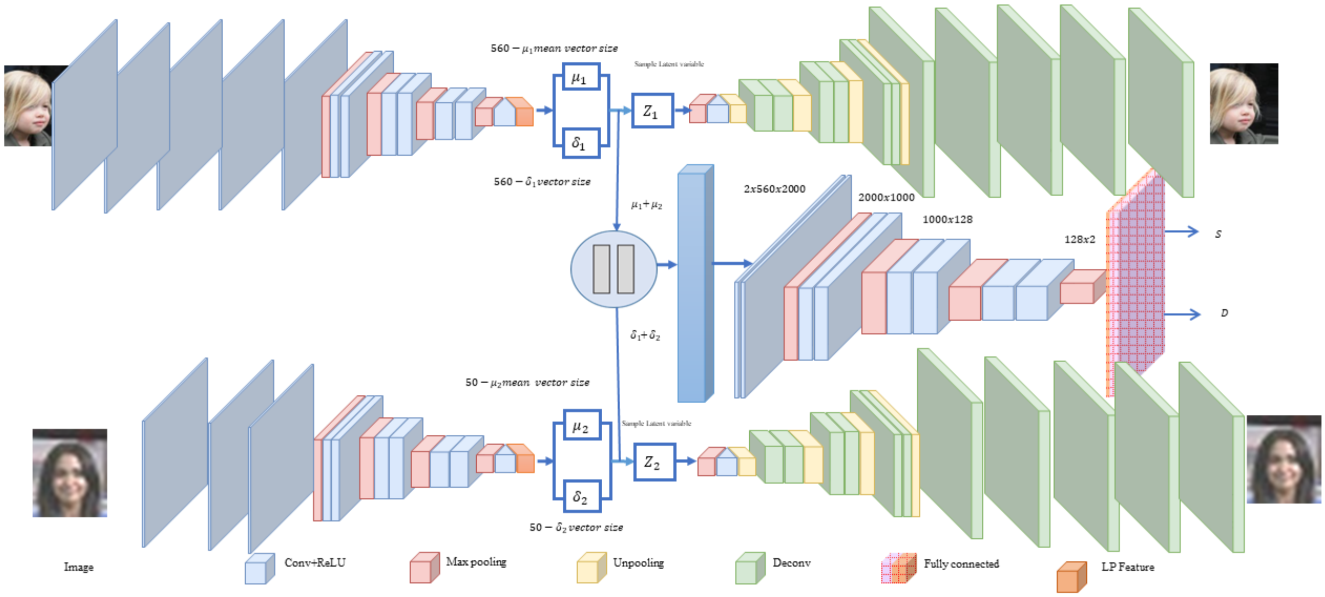

Figure 2: **Baseline Architecture** for matching low resolution and high resolution images using VAE.

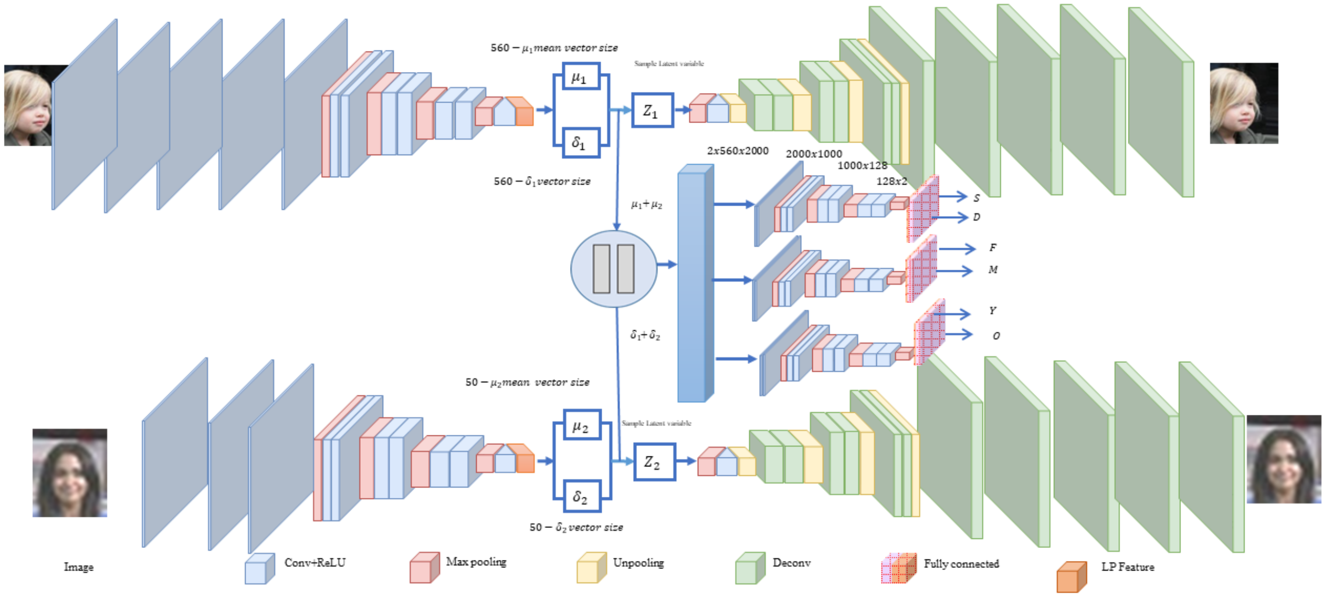

Figure 3: **Proposed Architecture** for matching low resolution and high resolution images using VAE. Note the multi-label classification component in the proposed architecture that is not present in the baseline.

Table I: Performance of the baseline architecture on face matching

Attributes	ACC	TPR	FPR
Same/Different	87.84	85.63	9.59

Table II: Performance of the proposed architecture on face matching

Attributes	ACC	TPR	FPR
Same/Different	95.79	94.93	3.41

B. Auditing for Bias in terms of Age and Gender

To understand the fairness aspect of the problem we consider the results as obtained for the test images belonging to specific demographic groups (e.g., male, female, young, old). Table III shows the results for the baseline and the proposed architecture. We notice significant differences in the performance of the baseline architecture based on demographic descriptors. For instance, the accuracy at face matching was 96.42% for male subjects, it was 91.78% for the female subjects, thereby indicating a difference (ΔACC) of 4.64%. In terms of age, face matching algorithm worked accurately for 92.76% for young subjects but only 78.48% for old participants. This marks a noticeable difference of 14.28%. These differences are summarized in Table IV.

C. Reduction of Bias in terms of Age and Gender

The proposed architecture aims to reduce bias levels. The results for the proposed architecture are also summarized in Tables III and IV.

The results indicate that the proposed approach resulted in a lower discrepancy between demographic groups in terms of ΔACC, ΔTPR, and ΔFPR for both age and gender. Hence, across different definitions of demographic groups and in terms of multiple fairness metrics, the proposed architecture yields fairer results. This combined with the observation that the accuracy of the proposed architecture is also higher than the baseline model, suggests that the proposed approach is indeed a useful way to design accurate and fair face matching algorithms.

Table III: Performance Male/Female,Young/Old samples on face matching algorithm (%)

	Baseline Model			Proposed Model		
Attributes	ACC	TPR	FPR	ACC	TPR	FPR
Male	96.42	95.70	2.8	99.96	99.96	0.04
Female	91.78	89.53	5.7	99.89	99.87	0.08
Young	92.76	90.49	5.1	93.93	95.23	8.33
Old	78.48	75.52	17.80	90.91	88.23	6.25

Table IV: Comparison of fairness metrics in the baseline and proposed models

	Baseline Model			Proposed Model		
Attributes	ΔACC	ΔTPR	ΔFPR	ΔACC	ΔTPR	ΔFPR
Male/ Female	4.64	6.17	2.9	0.07	0.09	0.04
Young/ Old	14.28	14.97	12.70	3.02	7.00	2.08

V. Discussion

A. The Value of Variational Approach

Generative models are specifically designed to learn and uncover the underlying variables in a dataset. For instance, in facial detection, if we have a large number of faces, we may not know which of the various latent variables in the dataset is going to be distributed fairly evenly. This could lead to various biases in one's model. With the help of these latent variables, we can then automatically identify areas of the latent landscape that are not represented by the data.

If we're given a data set with many different faces we may not know what the exact distribution of particular latent variables in this data set is going to be and there could be imbalances with respect to these different variables for example face pose, skin tone, gender and age attributes that could end up resulting in unwanted biases in our downstream model. Using generative models, the system can actually learn these latent variables and use this information to automatically uncover underrepresented and over-represented feature and regions of the latent landscape and use this information to mitigate some of these biases.

From the learned latent structure, we can then estimate the distribution of each of these learned latent variables which means the distribution of values that these latent variables can take, and certain instances are going to be over-represented so for example if our dataset has many images of faces of a certain female those are going to be overrepresented and thus the likelihood of selecting a particular image that has this particular female during training will be unfairly high which could result in unwanted biases in favor of these types of faces. Conversely faces with rare features like, male, old, shadows, darker skin, glasses, and hats may be under-represented in the data and thus the likelihood of selecting instances with these features to actually train the model will be low resulting in unwanted bias.

The VAE model could actually adaptively adjust the sampling probabilities of individual data instances to re-weight them during the training process itself such that these latent distributions and this resampling approach could be used to adaptively generate a more fair and more representative dataset for training. So the idea is that we want to find truth distribution for our observations and is a generative model with continuous latent variable, is simple and fast and the application is that can be used for generative models for instance generating images or classification image, reducing noise or adding noise. This approach has been found to be effective in the current setting.

B. Limitations

This work has some limitations. First, gender and age and have been operationalized as binary variables due to the labels available in the dataset. These descriptors exist as a continuum in the real world. We hope that future studies consider more diverse set of options for age and gender labels. We also note that the demographic labels were not available for all images in the dataset. The results presented for different demographic groups are based on the subset of images for whom we have that demographic label (age or gender) available.

VI. Conclusion

In this work, we present a novel method that allows us to debias the various probability distribution of data by fusing multi-network feature maps. This method is ideal for large datasets and for learning new features without having to explicitly label them. We present a method that aims to reduce hidden biases in training data by implementing a debiasing face-matching algorithm. We show that these models can improve classification accuracy and reduce categorical bias. We also provide a concrete implementation of this model. The development and deployment of unbiased and fair AI systems are very important to prevent the emergence of discrimination. This work aims to introduce a framework that will help AI systems to be more ethical and stable.

Acknowledgment

The work of Jamal Alasadi was supported by the MOHESR, Iraq.

References

[1] V. K. Singh, E. André, S. Boll, M. Hildebrandt, D. A. Shamma, and T.-S. Chua, "Legal and ethical challenges in multimedia research," in *Proceedings of the 27th ACM International Conference on Multimedia*, 2019, pp. 2514–2515.

[2] J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," in *Conference on fairness, accountability and transparency*. PMLR, 2018, pp. 77–91.

[3] J. Alasadi, A. Al Hilli, and V. K. Singh, "Toward fairness in face matching algorithms," in *Proceedings of the 1st International Workshop on Fairness, Accountability, and Transparency in MultiMedia*, 2019, pp. 19–25.

[4] D. Pessach and E. Shmueli, "Algorithmic fairness," *arXiv preprint arXiv:2001.09784*, 2020.

[5] V. K. Singh, M. Chayko, R. Inamdar, and D. Floegel, "Female librarians and male computer programmers? gender bias in occupational images on digital media platforms," *Journal of the Association for Information Science and Technology*, vol. 71, no. 11, pp. 1281–1294, 2020.

[6] F. Calmon, D. Wei, B. Vinzamuri, K. Natesan Ramamurthy, and K. R. Varshney, "Optimized pre-processing for discrimination prevention," *Advances in neural information processing systems*, vol. 30, 2017.

[7] A. More, "Survey of resampling techniques for improving classification performance in unbalanced datasets," *arXiv preprint arXiv:1608.06048*, 2016.

[8] T. L. Bailey, C. Elkan *et al.*, "Fitting a mixture model by expectation maximization to discover motifs in bipolymers," 1994.

[9] D. M. Blei, "Probabilistic topic models," *Communications of the ACM*, vol. 55, no. 4, pp. 77–84, 2012.

[10] P. Felzenszwalb, D. McAllester, and D. Ramanan, "A discriminatively trained, multiscale, deformable part model," in *2008 IEEE conference on computer vision and pattern recognition*. Ieee, 2008, pp. 1–8.

[11] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[12] D. J. Rezende, S. Mohamed, and D. Wierstra, "Stochastic backpropagation and approximate inference in deep generative models," in *International conference on machine learning*. PMLR, 2014, pp. 1278–1286.

[13] P. Sattigeri, S. C. Hoffman, V. Chenthamarakshan, and K. R. Varshney, "Fairness gan," *arXiv preprint arXiv:1805.09910*, 2018.

[14] R. Singh, P. Majumdar, S. Mittal, and M. Vatsa, "Anatomizing bias in facial analysis," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 11, 2022, pp. 12 351–12 358.

[15] G. H. Nguyen, A. Bouzerdoum, and S. L. Phung, "A supervised learning approach for imbalanced data sets," in *2008 19th international conference on pattern recognition*. IEEE, 2008, pp. 1–4.

[16] C.-K. Yeh, W.-C. Wu, W.-J. Ko, and Y.-C. F. Wang, "Learning deep latent space for multi-label classification," in *Thirty-first AAAI conference on artificial intelligence*, 2017.

[17] S. Gong, X. Liu, and A. K. Jain, "Jointly de-biasing face recognition and demographic attribute estimation," in *European conference on computer vision*. Springer, 2020, pp. 330–347.

[18] V. Grari, B. Ruf, S. Lamprier, and M. Detyniecki, "Fair adversarial gradient tree boosting," in *2019 IEEE International Conference on Data Mining (ICDM)*. IEEE, 2019, pp. 1060–1065.

[19] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, "Fairness through awareness," in *Proceedings of the 3rd innovations in theoretical computer science conference*, 2012, pp. 214–226.

[20] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," *Advances in neural information processing systems*, vol. 29, 2016.

[21] M. B. Zafar, I. Valera, M. G. Rogriguez, and K. P. Gummadi, "Fairness constraints: Mechanisms for fair classification," in *Artificial intelligence and statistics*. PMLR, 2017, pp. 962–970.

[22] E. Raisi and B. Huang, "Reduced-bias co-trained ensembles for weakly supervised cyberbullying detection," in *International Conference on Computational Data and Social Networks*. Springer, 2019, pp. 293–306.

[23] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," 2017.

[24] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in python," *the Journal of machine Learning research*, vol. 12, pp. 2825–2830, 2011.

[25] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.