Learning Disentangled Representation for Fair Facial Attribute Classification via Fairness-aware Information Alignment

Sungho Park[1] **Sunhee Hwang**[1] **Dohyung Kim**[1] **Hyeran Byun**[1,2*]
[1] Department of Computer Science, Yonsei University
[2] Graduate School of AI, Yonsei University
{qkrtjdgh18, sunny16, dohkim02, hrbyun}@yonsei.ac.kr

Abstract

Although AI systems archive a great success in various societal fields, there still exists a challengeable issue of outputting discriminatory results with respect to protected attributes (*e.g.*, gender and age). The popular approach to solving the issue is to remove protected attribute information in the decision process. However, this approach has a limitation that beneficial information for target tasks may also be eliminated. To overcome the limitation, we propose Fairness-aware Disentangling Variational Auto-Encoder (FD-VAE) that disentangles data representation into three subspaces: 1) Target Attribute Latent (TAL), 2) Protected Attribute Latent (PAL), 3) Mutual Attribute Latent (MAL). On top of that, we propose a decorrelation loss that aligns the overall information into each subspace, instead of removing the protected attribute information. After learning the representation, we re-encode MAL to include only target information and combine it with TAL to perform downstream tasks. In our experiments on CelebA and UTK Face datasets, we show that the proposed method mitigates unfairness in facial attribute classification tasks with respect to gender and age. Ours outperforms previous methods by large margins on two standard fairness metrics, equal opportunity and equalized odds.

Introduction

Despite tremendous advances in AI systems, there still exists a problem of outputting discriminatory results regarding different demographic groups (Wang et al. 2019a; Alvi, Zisserman, and Nellåker 2018; Wang et al. 2019b). Particularly, unfair decisions in terms of protected attributes (*e.g.*, gender, age, and ethnicity) cause social-ethical problems (Dougherty 2015; Lomas 2018; J. Angwin and Kirchner 2016). For example, Google photos, an image recognition algorithm, recognizes an African-American as being gorillas (Dougherty 2015). Besides, FaceApp, a face editing application, deploys a racist *hot* filter that is intended to edit a user's face more attractive. When a user with a dark skin tone uses this filter, it makes the user look like Caucasian by brightening the user's skin tone (Lomas 2018). These ethical issues raise the necessity of AI models that make a fair decision in terms of the protected attributes.

To address this problem, many previous works (Wang et al. 2019b; Kim et al. 2019; Zhang, Lemoine, and Mitchell 2018) adversarially train models not to discriminate protected attribute labels. These works remove information related to protected attributes in data representation, thereby outputting invariant results in terms of the attributes. However, there is a limitation that they potentially remove some useful information for target tasks due to its correlation with the protected attributes (*cf.*, Figure 1(a)).

On the other hand, (Creager et al. 2019), the research most relevant to this work, tries to solve the unfairness problem by utilizing a disentanglement learning method (Kim and Mnih 2018). They disentangle data representation into subspaces relevant to protected attributes or not, then exclude the subspaces with protected attribute information for downstream tasks. However, it is difficult to disentangle protected attribute information and beneficial information for target tasks explicitly due to their correlation. For this reason, the information for target tasks is inevitably included in the subspaces for protected attributes, thus the information is removed in downstream tasks as the adversarial approaches. (*cf.*, Figure 1(b)).

In this paper, we aim to learn representation that is fair in terms of protected attributes and preserves beneficial information for target tasks. To this end, we propose a Fairness-aware Disentangling Variational Auto-Encoder (FD-VAE) that disentangles data representation into three subspaces. As shown in Figure 1(c), Target Attribute Latent (TAL) and Protected Attribute Latent (PAL) include target attribute information and protected attribute information, respectively. Mutual Attribute Latent (MAL) includes intersected information between target and protected attributes. This is derived from the intuition that there are the intersection and complementary sets of target and protected attribute information. The complementary set is not necessary to be concerned to achieve our goal, but the intersected information needs to be addressed carefully. If the intersected information is included in TAL, it causes unfair results to protected attributes. Conversely, being it included in PAL, it causes loss of useful information for target attributes. We solve this dilemma by introducing an additional subspace (MAL) that includes the intersected information.

Specifically, each subspace is learned to include its appropriate information by our proposed decorrelation loss. This loss encourages each subspace to maximize mutual infor-

[*] Corresponding author

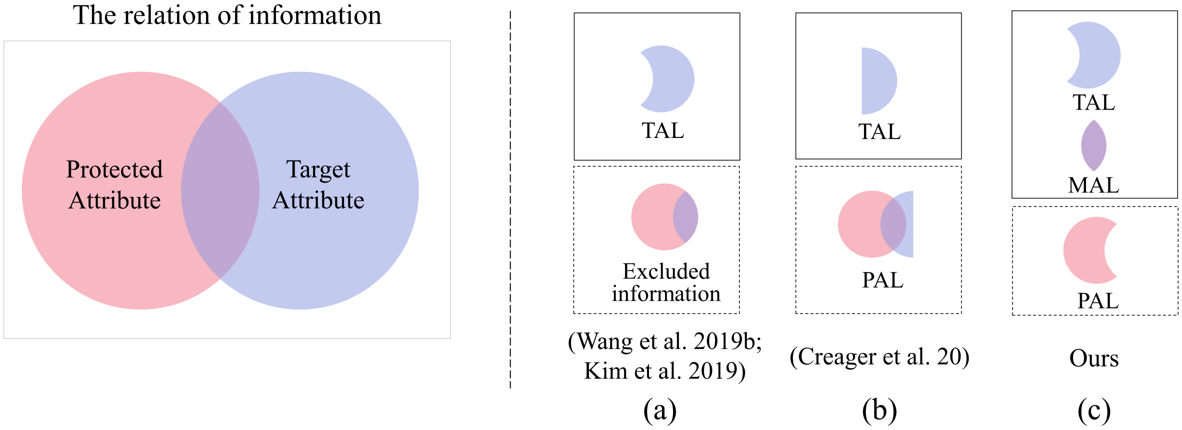


Figure 1: The motivation of our work. The red circle and blue circle denote protected and target attribute information, respectively. This figure conceptually shows how each latent space (or subspace) includes the information in the adversarial training method (Wang et al. 2019b; Kim et al. 2019) (a), disentanglement method (Creager et al. 2019) (b), and ours (c). The dotted boxes represent the information not used for target tasks.

mation with appropriate attributes and exclude information of inappropriate attributes based on adversarial training (*cf.* Figure 1(c)). Unlike previous works with the adversarial training (Wang et al. 2019b; Kim et al. 2019), the excluded information of each subspace is not removed in data representation but it is aligned to its proper subspace.

After learning this disentangled representation, we leverage TAL and MAL to perform downstream tasks. To this end, we propose a framework that re-encodes MAL to remain only beneficial information for the target tasks and combine it with TAL. This combined feature is informative to the target tasks and invariant to protected attributes, which enables fair classifications.

In the experiment section, we conduct facial attribute classifications on CelebA and UTK Face datasets (Liu et al. 2015; Zhang and Qi 2017). As indicated in previous studies, facial attributes are correlated with each other in the datasets (Kärkkäinen and Joo 2019; Celona, Bianco, and Schettini 2018; Torfason et al. 2016a), which causes unfair classification results in terms of protected attributes (Creager et al. 2019; Sattigeri et al. 2019; Wang et al. 2020). Our method is exploited to solve this problem and achieves the fairest results in comparison with previous methods. We measure the degree of fairness with two standard fairness metrics, equal opportunity and equalized odds (Hardt, Price, and Srebro 2016). Moreover, we propose a new metric, *equalized accuracy*, to fairly measure classification accuracy on a skewed test dataset. Through ablation study, we extensively validate the contribution of each component of our method on CelebA dataset.

The Main Contributions

We summarize our main contributions as follows: 1) We propose a novel FD-VAE that disentangles data representation into three independent subspaces: Target Attribute Latent (TAL), Protected Attribute Latent (PAL), and Mutual Attribute Latent (MAL). 2) The decorrelation loss encourages each subspace of our model to contain its appropriate information. 3) To utilize both MAL and TAL properly in downstream tasks, we introduce a downstream classification framework that re-encodes MAL and combines it with TAL. 4) To fairly measure classification accuracy in biased datasets, we propose a new metric, *equalized accuracy*.

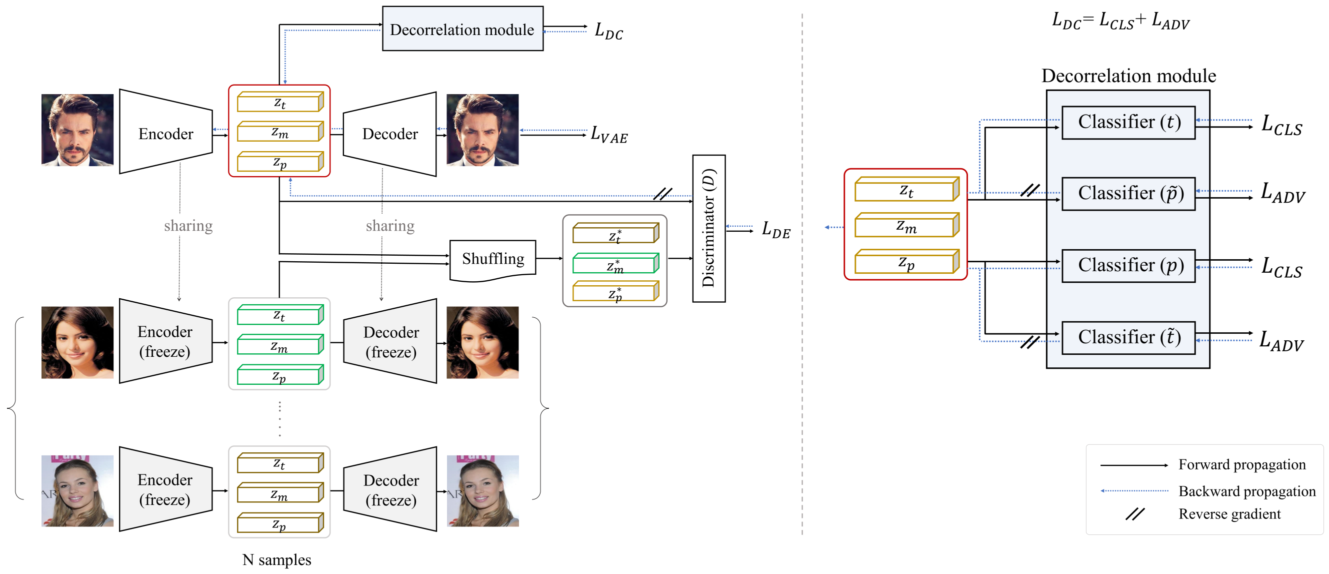
Related Works

Disentangled Representation Learning

Previous studies (Higgins et al. 2017; Kim and Mnih 2018; Chen et al. 2018) propose disentangled representation learning methods to learn latent variables to be independent of each other. (Higgins et al. 2017) proves KL-divergence term in the VAE objective function encourages latent variables to be disentangled and proposes β-VAE to weight this term with a larger hyper-parameter β (> 1). However, (Kim and Mnih 2018) indicates that β-VAE has a trade-off between a disentangling performance and reconstruction quality. To reduce the trade-off, they exploit Total Correlation (Watanabe 1960), a measure to estimate the dependency between latent variables, to learn disentangled representation. They approximate it with adversarial learning using a discriminator. (Chen et al. 2018) also optimizes the equivalent objective function to FactorVAE (Kim and Mnih 2018) but propose a new stochastic estimation method on Total Correlation, enabling more stable training than FactorVAE. In this paper, our method leverages the disentangling algorithm proposed in (Kim and Mnih 2018) to separate subspaces of representation.

Fairness-aware Algorithms in Machine Learning

In this section, we describe fairness algorithms based on adversarial training and disentanglement learning, which are the most relevant to our work. Firstly, (Wang et al. 2019b; Kim et al. 2019; Zhang, Lemoine, and Mitchell 2018) adversarially train models not to discriminate protected attributes. It encourages the models to output fair results by not using



(a) Overall architecture

(b) Decorrelation module

Figure 2: The overview of our FD-VAE framework. (a) shows the overall architecture of FD-VAE and (b) shows details of the decorrelation module.

information related to the protected attributes. Inspired by (Ganin et al. 2016), (Wang et al. 2019b) apply a Gradient Reversal Layer (GRL) to various intermediate representation to remove unwanted bias in visual recognition tasks. Similarly, (Kim et al. 2019) remove unwanted bias by minimizing the mutual information between data representation and bias using the GRL. Besides, (Zhang, Lemoine, and Mitchell 2018) propose an adversarial debiasing method to maximize the ability of the predictor for target classes and to minimize the ability of the adversary network for protected attributes.

On the other hand, (Hwang et al. 2020; Gong, Liu, and Jain 2020; Creager et al. 2019; Sarhan et al. 2020) propose methods that disentangle data representation into subspaces including protected attributes information or not. (Hwang et al. 2020) learn disentangled representation to prevent unwanted translation of protected attributes in image-to-image translation tasks. In addition, (Gong, Liu, and Jain 2020) mitigate unfairness in face recognition tasks by separating representation of gender, ethnicity, age, and identity. FFVAE (Creager et al. 2019) exploits the disentangled representation learning method from (Kim and Mnih 2018) to separate representation into sensitive latents and non-sensitive latents which are relevant to protected attributes or not, respectively. Since the model is trained without target labels during the representation learning, it flexibly performs various downstream classification tasks only by excluding corresponding sensitive latents. Lastly, (Sarhan et al. 2020) propose an orthogonal constraint that disentangles representation into two orthogonal subspaces to learn invariant representation to protected attributes.

Proposed Method

When observed data $X = (x_1, ..., x_n)$, target attribute labels $Y_t = (y_{t_1}, ..., y_{t_n})$, and protected attribute labels $Y_p = (y_{p_1}, ..., y_{p_n})$ are given, our goal is to encode representation that is fair to Y_p and preserves beneficial information for Y_t. To this end, we design FD-VAE that disentangles representation z into three subspaces: Target Attribute Latent (z_t), Protected Attribute Latent (z_p), and Mutual Attribute Latent (z_m). As illustrated in Figure 2, our model is composed of a VAE network, discriminator, and decorrelation module. It is optimized by VAE objective function, disentanglement loss, discriminator loss, and decorrelation loss. We specify each loss function in this section.

VAE

Our model is based on Variational Auto-Encoder (VAE) (Kingma and Welling 2014), which is composed of an encoder and a decoder. We learn the VAE network by maximizing the Evidence Lower BOund (ELBO):

$$\begin{aligned}\mathcal{L}_{VAE} = \sum_{i=1}^{n} \mathbb{E}_{q_\Phi(z_t,z_p,z_m|x_i)}[\log p_\Theta(x_i|z_t,z_p,z_m)] \\ -KL[q_\Phi(z_t,z_p,z_m|x_i)||p(z_t,z_p,z_m)],\end{aligned} \quad (1)$$

where Φ and Θ are parameters of the encoder and decoder, respectively. The first term of Equation 1 denotes a reconstruction loss that encourages the encoder to map the observed data X into representation z and the decoder to reconstruct X from z. z is sampled from $q_\Phi(z|x) = N(\mu_{q_\Phi}(x), \sigma_{q_\phi}(x))$ using the reparameterization trick, where μ and σ are the outputs of the encoder. The second term indicates a regularization loss that makes the distribution $q_\Phi(z|x)$ similar to the Gaussian prior distribution $p(z)$ by KL divergence.

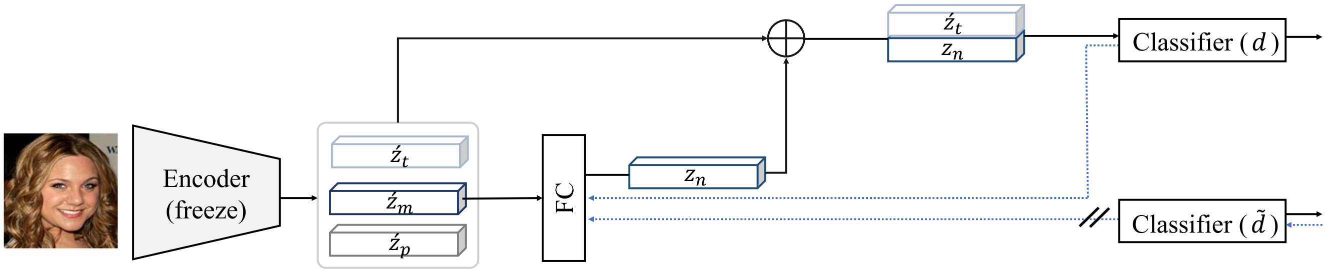


Figure 3: The architecture of the downstream classification framework. FC and $\oplus$ denote a single fully connected layer and element-wise summation.

Disentanglement Loss

To disentangle the representation z into subspaces as z_t, z_p, and z_m, we minimize Total Correlation (Watanabe 1960) by following objective function:

$$\begin{aligned}\mathcal{L}_{DE} &= KL[q_\Phi(z_t,z_p,z_m)||\prod_{j\subseteq S} q_\Phi(z_j)] \\ &= \mathbb{E}_{q_\Phi(z_t,z_p,z_m)}[\log \frac{q_\Phi(z_t,z_p,z_m)}{\prod_{j\subseteq S} q_\Phi(z_j)}],\end{aligned} \tag{2}$$

where $S = \{t,\ p,\ m\}$. Total Correlation is one of the most popular measures that estimate the dependency between latent variables and minimizing this term encourages latent variables to be disentangled (Kim and Mnih 2018; Chen et al. 2018; Creager et al. 2019). Following the methods in (Kim and Mnih 2018; Creager et al. 2019), we approximate log density ratio instead of optimizing KL divergence directly by leveraging a discriminator as the following equation:

$$\mathcal{L}_{DE} \approx \mathbb{E}_{q_\Phi(z_t,z_p,z_m)}[\log \frac{D(z_t,z_p,z_m)}{1-D(z_t,z_p,z_m)}], \tag{3}$$

where $D(z_t,z_p,z_m)$ is the output probability of the discriminator D that classifies the samples from $q_\Phi(z_t,z_p,z_m)$ as real and the samples from $\prod_{j\subseteq S} q_\Phi(z_j)$ as fake. The encoder is trained for the discriminator not to classify whether the samples are real or fake. The fake samples z^*=$[z_t^*;z_p^*;z_m^*]$ are generated by subspace-wise random shuffling within a mini-batch. The loss function for training the discriminator is as follows:

$$\begin{aligned}\mathcal{L}_D = -\mathbb{E}_{q_\Phi(z_t,z_p,z_m)}[&\log D(z_t,z_p,z_m) \\ &+\log(1-D(z_t^*,z_p^*,z_m^*))].\end{aligned} \tag{4}$$

Decorrelation Loss

On top of the disentanglement loss that encourages the subspaces to be independent of each other, we introduce a decorrelation loss that aligns appropriate information into each subspace. The decorrelation loss is composed of $\mathcal{L}_{CLS}$ and $\mathcal{L}_{ADV}$ as follows:

$$\mathcal{L}_{DC} = \mathcal{L}_{CLS} + \lambda\mathcal{L}_{ADV}, \tag{5}$$

$$\begin{aligned}\mathcal{L}_{CLS} = -\sum_{i=1}^{n} \mathbb{E}_{q_\Phi(z_t,z_p,z_m|x_i)}[&\log p_t(y_t|z_t)) \\ &+\log p_p(y_p|z_p))], \\ \mathcal{L}_{ADV} = \min_{\Phi}\max_{\tilde{t},\tilde{p}} \sum_{i=1}^{n} \mathbb{E}_{q_\Phi(z_t,z_p,z_m|x_i)}[&\log p_{\tilde{p}}(y_p|z_t)) \\ &+\log p_{\tilde{t}}(y_t|z_p))],\end{aligned} \tag{6}$$

where t and $\tilde{t}$ denote classifiers for target attributes, and p and $\tilde{p}$ denote classifiers for protected attributes. λ is a hyperparameter. $\sim$ means the classifiers are trained adversarially. $\mathcal{L}_{CLS}$ encourages z_t and z_p to contain target and protected attribute information, respectively. Meanwhile, $\mathcal{L}_{ADV}$ encourages z_t and z_p to exclude protected attribute and target attribute information, respectively, in an adversarial way. The intersected information of protected and target attributes is excluded in both TAL and PAL by $\mathcal{L}_{ADV}$. This information is included in MAL through the reconstruction loss.

Total Loss

In conclusion, the total loss function of FD-VAE is as follows:

$$\mathcal{L}_{TOTAL} = \mathcal{L}_{VAE} - (\alpha\mathcal{L}_{DE} + \beta\mathcal{L}_D + \gamma\mathcal{L}_{DC}), \tag{7}$$

where α, β, and γ are hyper-parameters.

Downstream Classification Framework

After learning the fair representation, we perform downstream classifications for the target attributes. The overall architecture of our downstream classification framework is shown in Figure 3. The loss function of the framework is defined by:

$$\begin{aligned}\min_{\tilde{d}}\max_{d,f} \sum_{i=1}^{n} \mathbb{E}_{q_\Phi(\acute{z}_t,\acute{z}_p,\acute{z}_m|x_i)}[&\log p_d(y_t|\acute{z}_t \oplus f(\acute{z}_m)) \\ &-\log p_{\tilde{d}}(y_p|f(\acute{z}_m))],\end{aligned} \tag{8}$$

where d, $\tilde{d}$, and $\oplus$ indicate a target attribute classifier, protected attribute classifier, and element-wise summation, respectively. The learned representation $\acute{z}_t$, $\acute{z}_p$, and $\acute{z}_m$ are fixed in downstream classification tasks.

First, we exclude $\acute{z}_p$ in order not to exploit the protected attribute information in downstream tasks. Then, we re-encode $\acute{z}_m$ to latent variables z_n that includes only information related to the target attributes through a single fully connected

Method	TA	PA		Opp. ↓	Odds ↓	Acc. ↑	TA	PA		Opp. ↓	Odds ↓	Acc. ↑
		M=1	M=0			EAcc. ↑		Y=1	Y=0			EAcc. ↑
VAE (Kingma and Welling 2014)	A=1	54.3	81.5	27.2	28.7	70.6	A=1	77.4	67.6	9.7	11.3	70.9
	A=0	77.8	47.5			65.3	A=0	60.8	72.7			69.6
$\beta-$VAE (Higgins et al. 2017)	A=1	57.7	78.8	21.1	22.1	68.7	A=1	72.9	65.6	7.2	5.8	67.5
	A=0	73.1	48.8			64.6	A=0	61.1	65.4			66.4
FactorVAE (Kim and Mnih 2018)	A=1	59.7	80.4	20.7	23.4	69.0	A=1	76.0	66.8	9.2	9.1	68.9
	A=0	72.8	46.6			64.9	A=0	58.9	68.1			67.5
FFVAE (Creager et al. 2019)	A=1	53.9	68.9	15.0	16.6	62.2	A=1	68.3	62.4	5.8	3.4	63.7
	A=0	66.1	47.9			59.2	A=0	60.2	59.1			62.5
FFVAE (Creager et al. 2019)+TAC	A=1	60.4	76.1	15.6	17.2	66.3	A=1	71.7	65.0	6.7	4.6	65.4
	A=0	67.6	48.7			63.2	A=0	58.8	61.3			64.2
Adversarial training (Ganin et al. 2016)	A=1	63.9	71.6	7.6	10.7	64.5	A=1	69.9	64.1	5.8	4.9	65.5
	A=0	64.9	51.0			62.8	A=0	59.9	64.0			64.4
Ours	A=1	66.2	67.6	**1.3**	**4.9**	64.1	A=1	76.5	72.7	**3.7**	**1.9**	65.5
	A=0	64.4	55.8			63.5	A=0	55	55.2			64.8

Table 1: Classification results on CelebA dataset. TA and PA are the abbreviations of the target and protected attributes, respectively. We utilize four metrics: equal opportunity (Opp.), equalized odds (Odds), accuracy (ACC.), and *equalized accuracy* (EAcc.). A, M, and Y denote *attractiveness*, *male*, and *young*, respectively. The third and fourth columns show TPR and TNR of each demographic group.

layer f. f and $\tilde{d}$ are adversarially trained so that z_n can not classify Y_p. Finally, z_n and $\acute{z_t}$ are element-wise summed and feed into d for the target attribute classification.

Experiments

Dataset

CelebA dataset: It consists of about 200k face images with 40 binary attribute annotations. Among the attributes, we first set *male* and *young* to the protected attributes. Next, we set *attractiveness* to the target attribute since it has high Pearson correlation with *male* (p=-0.40) and *young* (p=-0.39) (Torfason et al. 2016b).

UTK Face dataset: It has a total of 23,705 images without the specification of train, validation, and test dataset. Among its annotations of *age*, *ethnicity*, and *gender*, we set *gender* to the protected attribute and the others to the target attributes. We conduct binary attribute classification tasks with the following two settings: Caucasians and the others, age under 35 and the others. To establish a high correlation between target and protected attributes, we compose train dataset (10k) as follows. For *ethnicity*, the ratio of Caucasian females and males is 1:4, and the ratio of the other ethnicity group is opposite. Likewise, the ratio of males and females with age under 35 is 1:4, and the ratio of the other group is opposite. Meanwhile, we compose the validation (2.4k) and test (2.4k) datasets to be balanced sets for a fair comparison. (*cf.*, Appendix A).

Comparable Models

We note that all networks of comparable models have the same structures as ours.

Beta-VAE (Higgins et al. 2017), Factor-VAE (Kim and Mnih 2018): Both models disentangle representation without protected attribute labels, thus it is not explicitly known which subspaces include protected attribute information. Therefore, we manually remove a few subspaces most correlated to protected attributes to perform downstream tasks fairly as in (Creager et al. 2019).

Adversarial training (Ganin et al. 2016): As previous works (Wang et al. 2019b; Kim et al. 2019), this model is composed of one encoder followed by two classifiers for target and protected attribute predictions. By adding a Gradient Reversal Layer (GRL) (Ganin et al. 2016) to the protected attribute classification branch, we remove protected attribute information in data representation.

FFVAE (Creager et al. 2019), FFVAE+TAC: Both models disentangle representation into two subspaces with the same dimensionality. Since FFVAE is trained without target attribute labels, for fair comparison, we construct the model (FFVAE+TAC) by adding a target attribute classifier (TAC) to classify the labels using non-sensitive latents.

Evaluation Metrics

In our experiments, we measure the degree of fairness with two metrics, equal opportunity and equalized odds (Hardt, Price, and Srebro 2016). Equal opportunity represents the parity of True Positive Rate (TPR) between the groups with different protected attribute labels (p_0 and p_1). Furthermore, equalized odds represents the parity of TPR and True Negative Rate (TNR) between the groups. These metrics are formulated as $|\text{TPR}_{p_0} - \text{TPR}_{p_1}|$ and $\frac{1}{2}[|\text{TPR}_{p_0} - \text{TPR}_{p_1}| + |\text{TNR}_{p_0} - \text{TNR}_{p_1}|]$, respectively.

On the other hand, demographic parity (Dwork et al. 2012; Kusner et al. 2017), which represents parity in the proportion of positive decisions between p_0 and p_1, is also one of the popular metrics for measuring fairness. However, since the proportion of positive target labels (ground truth) is different between p_0 and p_1 in our tasks, this metric does not ensure fairness of models. Therefore, we show experimental results measured by demographic parity only in Appendix B.

Method	TA	PA		Opp. ↓	Odds ↓	Acc. ↑	TA	PA		Opp. ↓	Odds ↓	Acc. ↑
		G=1	G=0					G=1	G=0			
VAE (Kingma and Welling 2014)	E=1	70.5	54.3	16.1	17.4	65.2	A=1	50.8	75.5	24.6	29.8	60.7
	E=0	58.6	77.4				A=0	75.8	40.7			
$\beta-$VAE (Higgins et al. 2017)	E=1	72.1	59.4	12.7	12.1	60.1	A=1	45.3	62.5	17.1	20.8	54.3
	E=0	48.7	60.2				A=0	67.1	42.5			
FactorVAE (Kim and Mnih 2018)	E=1	72.5	60.3	12.1	12.9	59.4	A=1	43.5	66.7	23.2	27.1	54.6
	E=0	45.8	59.5				A=0	69.6	38.5			
FFVAE (Creager et al. 2019)	E=1	70.0	63.4	6.6	8.3	59.4	A=1	45.2	58.9	13.6	17.2	54.5
	E=0	47.3	57.3				A=0	67.4	46.5			
FFVAE (Creager et al. 2019)+TAC	E=1	69.0	60.6	8.3	9.7	59.9	A=1	43.6	61.3	17.6	22.2	54.1
	E=0	49.4	61.1				A=0	68.8	42.1			
Adversarial training (Ganin et al. 2016)	E=1	59.3	52.7	6.5	4.6	60.8	A=1	26.1	37.6	11.5	13.2	54.7
	E=0	64.0	66.8				A=0	85.6	70.5			
Ours	E=1	65.8	63.5	**2.3**	**1.1**	60.3	A=1	47.9	45.8	**2.0**	**2.9**	54.1
	E=0	56.0	56.0				A=0	63.3	59.4			

Table 2: Classification results on UTK Face dataset. G, E, and A indicate *gender* (1: male, 0: female), *ethnicity* (1: Caucasian, 0: others), and *age* (1: under 35, 0: others), respectively.

FFVAE (Creager et al. 2019)		
Subspace	TA (Acc. ↑)	PA (Acc. ↑)
Non-sensitive Latents	62.2	69.1
Sensitive Latents	67.4	76.7
FD-VAE (Ours)		
Subspace	TA (Acc. ↑)	PA (Acc. ↑)
TAL	60.3	63.0
PAL	56.6	67.3
MAL	67.2	68.9

Table 3: Amount of attribute information in each subspace. We perform target attribute (TA) and protected attribute (PA) classifications using each the learned subspace. TA and PA are set to *attractiveness* and *male*, respectively.

In addition, we propose a new metric *equalized accuracy* to fairly measure the classification accuracy in a biased test dataset. If train and test datasets have similar data distributions, the over-fitted model to the distribution of train dataset is favorable in standard classification accuracy. In UTK Face dataset, we solve this problem by comprising a balanced test dataset. However, since test dataset in CelebA has a similar distribution as train dataset, we utilize *equalized accuracy* which has the same effect as testing on the balanced dataset. The proposed metric is defined as $\frac{1}{4}[\text{TPR}_{p_0} + \text{TNR}_{p_0} + \text{TPR}_{p_1} + \text{TNR}_{p_1}]$. More details are described in Appendix C.

Implementation Details

The detailed structures of our networks are specified in Appendix D. For a fair comparison, the dimensionality of data representation is set to 60 identically in all models. In our model, the dimensionality of TAL, PAL, and MAL is set to 20. The hyper-parameters are found by the grid search method: α=50, β=1, γ=5, and λ= 2. We note that all results presented in our table are the best performances in terms of fairness and converged enough.

Comparison with Previous Methods

To validate the effectiveness of the proposed method, we compare the performance of ours with previous methods (Higgins et al. 2017; Kim and Mnih 2018; Creager et al. 2019). Table 1 shows the classification accuracy and fairness scores on CelebA dataset. VAE, β-VAE (Higgins et al. 2017), and FactorVAE (Kim and Mnih 2018) show highly discriminatory results in terms of *male* and *young*, since they do not consider the protected attributes in representation learning. FFVAE (Creager et al. 2019) improves both equal opportunity (Opp.) and equalized odds (Odds.). However, it shows the lowest classification accuracy (Acc.) and *equalized accuracy* (EAcc.) due to a large loss of target attribute information in the disentangling process. FFVAE+TAC (Creager et al. 2019) and Adversarial training (Ganin et al. 2016) leverage both protected attribute and target attribute labels as ours. FFVAE+TAC increases Acc. and EAcc. by increasing mutual information between non-sensitive latents and target attribute labels. However, it degrades fairness scores since protected attribute information correlated to the target attributes is partially included in the latents. Although Adversarial training (Ganin et al. 2016) shows comparable Acc. and EAcc. to ours, it shows a large trade-off between fairness and accuracy. Our method significantly outperforms all the previous methods above with comparable Acc. and EAcc. The better trade-off of our results indicates that our method can remove more protected attribute information while preserving a similar amount of target attribute information.

The classification results on UTK Face dataset are shown in Table 2. Similarly to CelebA dataset, VAE, β-VAE, and FactorVAE output unfair results in terms of equal opportunity and equalized odds. Although previous fairness-aware methods (Creager et al. 2019; Ganin et al. 2016) improve

Representation learning		Downstream Classification			TA	PA		Opp. ↓	Odds ↓	Acc. ↑
L_{DC}	z_m	z_t'	z_m'	z_n		M=1	M=0			
		✓			A=1	53.9	68.9	15.0	16.6	62.2
					A=0	66.1	47.9			
✓		✓			A=1	75.7	72.3	3.4	5.7	64.1
					A=0	58.7	50.8			
✓	✓	✓			A=1	68.1	66.3	1.8	**2.1**	60.3
					A=0	55.1	52.6			
✓	✓	✓	✓		A=1	57.9	77.7	19.8	20.8	67.1
					A=0	69.8	47.8			
✓	✓	✓		✓	A=1	66.2	67.6	**1.3**	4.9	64.1
					A=0	64.4	55.8			

Table 4: Ablation study on CelebA dataset. We set FFVAE (Creager et al. 2019) to our baseline (the first row). L_{DC} and z_m denote the decorrelation loss and MAL in representation learning, and z_t',z_m', and z_n denote learned TAL, MAL, and re-encoded MAL in the downstream classification, respectively. We set the protected attribute to *male* (M) and the target attribute to *attractiveness* (A)

Method	Opp .↓	Odds ↓	Acc. ↑
FFVAE (Creager et al. 2019)	15.0	16.6	62.2
FFVAE † (Creager et al. 2019)	15.3	17.6	63.5
Ours	**1.3**	**4.9**	**64.1**

Table 5: The role of our downstream classification framework. FFVAE † is FFAVE with the downstream framework. We set the target and protected attributes to *attractiveness* and *male* on CelebA dataset, respectively.

the fairness scores, they still have a large trade-off between fairness and accuracy. Our method surpasses the previous methods in both *ethnicity* and *age* classifications, achieving the lowest scores of 1.3% and 4.9% at equal opportunity and 3.7% and 1.9% at equalized odds with comparable accuracy.

Qualitative Analysis

In Table 3, we intuitively show our method decorrelates the target and protected attribute information better than FFVAE. In FFVAE, sensitive latents classifies the target attribute as well as the protected attribute better than non-sensitive latents. This indicates that beneficial information for the target attributes is included in the sensitive latents as our assumption. Meanwhile, TAL achieves a better result in the target attribute classification than PAL, and PAL classifies the protected attribute better than TAL. It demonstrates that the target attribute information is less included in PAL than non-sensitive latents of FFVAE. MAL classifies both the attributes well since it includes the intersected information.

Ablation Study

We conduct ablation study on CelebA dataset to validate the contribution of each component in our model as shown in Table 4. We set FFVAE (Creager et al. 2019) to a baseline and add each component step-by-step. First, we apply the decorrlation loss $\mathcal{L}_{DC}$ to the baseline. It improves both the classification accuracy and fairness scores by aligning the target and protected attributes information to appropriate subspaces. Then, we add MAL (z_m) to this model. This subspace encourages TAL to be more independent of protected attributes, showing the fairest score in terms of equalized odds (Odds.). The lowest accuracy indicates that it is necessary to utilize the useful information in MAL for the downstream task. Therefore, we leverage MAL (z_m') with TAL for the downstream task. It shows that the intersected information in z_m' improves the classification accuracy but significantly degrades fairness scores. Ours shows that re-encoded MAL (z_n) mitigates the degradation of the fairness scores. It achieves the fairest equalized opportunity (Opp.) while maintaining higher accuracy than the baseline.

Besides, we apply our downstream classification framework to FFVAE as shown in Table 5. We re-encode sensitive latents and combine it with non-sensitive latents by element-wise summation. In this experiment, the classification accuracy is improved, but equal opportunity and equalized odds are slightly degraded. It indicates that our FD-VAE framework and decorrelation loss are major factors in the improvement of fairness and the role of the downstream framework is only to increase the classification accuracy.

Conclusion

The objective of our work is to encode fair representation in terms of protected attributes while preserving beneficial information for target tasks. To this end, we proposed Fairness-aware Disentangling Variational Auto-Encoder (FD-VAE) that disentangles representation into three subspaces. TAL and PAL include target and protected attribute information, respectively, and MAL encourages TAL and PAL to be more independent by including the intersected information. On top of that, our decorrelation loss aligns appropriate information to each subspace. We leveraged the disentangled representation to perform facial attribute classifications on CelebA and UTK datasets. By re-encoding MAL and combining it with TAL, we performed fair and accurate classifications. In all the experiments in both datasets, our method shows the fairest results in terms of equal opportunity and equalized odds with comparable accuracy and *equalized accuracy*.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No.2019R1A2C2003760) and Institute for Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (Development of framework for analyzing, detecting, mitigating of bias in AI model and training data) under Grant 2019-0-01396 and (Artificial Intelligence Graduate School Program (YONSEI UNIVERSITY)) under Grant 2020-0-01361.

References

Alvi, M. S.; Zisserman, A.; and Nellåker, C. 2018. Turning a Blind Eye: Explicit Removal of Biases and Variation from Deep Neural Network Embeddings. In Leal-Taixé, L.; and Roth, S., eds., *Computer Vision - ECCV 2018 Workshops - Munich, Germany, September 8-14, 2018, Proceedings, Part I*, volume 11129 of *Lecture Notes in Computer Science*, 556–572. Springer. doi:10.1007/978-3-030-11009-3_34. URL https://doi.org/10.1007/978-3-030-11009-3_34.

Celona, L.; Bianco, S.; and Schettini, R. 2018. Fine-Grained Face Annotation Using Deep Multi-Task CNN. *Sensors (Basel, Switzerland)* 18.

Chen, R. T. Q.; Li, X.; Grosse, R. B.; and Duvenaud, D. K. 2018. Isolating Sources of Disentanglement in Variational Autoencoders. In Bengio, S.; Wallach, H.; Larochelle, H.; Grauman, K.; Cesa-Bianchi, N.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 31*, 2610–2620. Curran Associates, Inc. URL http://papers.nips.cc/paper/7527-isolating-sources-of-disentanglement-in-variational-autoencoders.pdf.

Creager, E.; Madras, D.; Jacobsen, J.-H.; Weis, M.; Swersky, K.; Pitassi, T.; and Zemel, R. 2019. Flexibly Fair Representation Learning by Disentanglement. In Chaudhuri, K.; and Salakhutdinov, R., eds., *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, 1436–1445. Long Beach, California, USA: PMLR. URL http://proceedings.mlr.press/v97/creager19a.html.

Dougherty, C. 2015. Google photos mistakenly labels black people gorillas. Twitter. URL https://bits.blogs.nytimes.com/2015/07/01/google-photos-mistakenly-labels-black-people-gorillas.

Dwork, C.; Hardt, M.; Pitassi, T.; Reingold, O.; and Zemel, R. 2012. Fairness through Awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, ITCS '12, 214–226. New York, NY, USA: Association for Computing Machinery. ISBN 9781450311151. doi:10.1145/2090236.2090255. URL https://doi.org/10.1145/2090236.2090255.

Ganin, Y.; Ustinova, E.; Ajakan, H.; Germain, P.; Larochelle, H.; Laviolette, F.; Marchand, M.; and Lempitsky, V. 2016. Domain-Adversarial Training of Neural Networks. *J. Mach. Learn. Res.* 17(1): 2096–2030. ISSN 1532-4435.

Gong, S.; Liu, X.; and Jain, A. 2020. Jointly De-biasing Face Recognition and Demographic Attribute Estimation. In *In Proceeding of European Conference on Computer Vision*. Virtual.

Hardt, M.; Price, E.; and Srebro, N. 2016. Equality of Opportunity in Supervised Learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, 3323–3331. Red Hook, NY, USA: Curran Associates Inc. ISBN 9781510838819.

Higgins, I.; Matthey, L.; Pal, A.; Burgess, C.; Glorot, X.; Botvinick, M. M.; Mohamed, S.; and Lerchner, A. 2017. beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. In *ICLR*.

Hwang, S.; Park, S.; Kim, D.; Do, M.; and Byun, H. 2020. FairFaceGAN: Fairness-aware Facial Image-to-Image Translation. In *BMVC*, volume 2020.

J. Angwin, J. Larson, S. M.; and Kirchner, L. 2016. Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks. ProPublica. URL https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Kärkkäinen, K.; and Joo, J. 2019. FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age. *CoRR* abs/1908.04913. URL http://arxiv.org/abs/1908.04913.

Kim, B.; Kim, H.; Kim, K.; Kim, S.; and Kim, J. 2019. Learning Not to Learn: Training Deep Neural Networks With Biased Data. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Kim, H.; and Mnih, A. 2018. Disentangling by Factorising. In Dy, J.; and Krause, A., eds., *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, 2649–2658. Stockholmsmässan, Stockholm Sweden: PMLR.

Kingma, D. P.; and Welling, M. 2014. Auto-Encoding Variational Bayes. In Bengio, Y.; and LeCun, Y., eds., *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*. URL http://arxiv.org/abs/1312.6114.

Kusner, M. J.; Loftus, J.; Russell, C.; and Silva, R. 2017. Counterfactual Fairness. In Guyon, I.; Luxburg, U. V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 30*, 4066–4076. Curran Associates, Inc. URL http://papers.nips.cc/paper/6995-counterfactual-fairness.pdf.

Liu, Z.; Luo, P.; Wang, X.; and Tang, X. 2015. Deep Learning Face Attributes in the Wild. In *Proceedings of International Conference on Computer Vision (ICCV)*.

Lomas, N. 2018. FaceApp apologizes for building a racist AI. TechCrunch. URL https://techcrunch.com/2017/04/25/faceappapologises-for-building-a-racist-ai.

Sarhan, M. H.; Navab, N.; Eslami, A.; and Albarqouni, S. 2020. Fairness by Learning Orthogonal Disentangled Representations. In Vedaldi, A.; Bischof, H.; Brox, T.; and Frahm, J., eds., *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXIX*, volume 12374 of *Lecture Notes in Computer Science*, 746–761. Springer. doi:10.1007/978-3-030-58526-6_44. URL https://doi.org/10.1007/978-3-030-58526-6_44.

Sattigeri, P.; Hoffman, S. C.; Chenthamarakshan, V.; and Varshney, K. R. 2019. Fairness GAN: Generating datasets with fairness properties using a generative adversarial network. *IBM Journal of Research and Development* 63(4/5): 3:1–3:9.

Torfason, R.; Agustsson, E.; Rothe, R.; and Timofte, R. 2016a. From face images and attributes to attributes. In *Asian Conference on Computer Vision*, 313–329. Springer.

Torfason, R.; Agustsson, E.; Rothe, R.; and Timofte, R. 2016b. From Face Images and Attributes to Attributes. In *ACCV*.

Wang, M.; Deng, W.; Hu, J.; Tao, X.; and Huang, Y. 2019a. Racial Faces in the Wild: Reducing Racial Bias by Information Maximization Adaptation Network. In *The IEEE International Conference on Computer Vision (ICCV)*.

Wang, T.; Zhao, J.; Yatskar, M.; Chang, K.-W.; and Ordonez, V. 2019b. Balanced Datasets Are Not Enough: Estimating and Mitigating Gender Bias in Deep Image Representations. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)* 5309–5318.

Wang, Z.; Qinami, K.; Karakozis, I. C.; Genova, K.; Nair, P.; Hata, K.; and Russakovsky, O. 2020. Towards Fairness in Visual Recognition: Effective Strategies for Bias Mitigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Watanabe, S. 1960. Information Theoretical Analysis of Multivariate Correlation. *IBM J. Res. Dev.* 4(1): 66–82. ISSN 0018-8646. doi:10.1147/rd.41.0066. URL https://doi.org/10.1147/rd.41.0066.

Zhang, B. H.; Lemoine, B.; and Mitchell, M. 2018. Mitigating Unwanted Biases with Adversarial Learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, 335–340. New York, NY, USA: Association for Computing Machinery. ISBN 9781450360128. doi:10.1145/3278721.3278779. URL https://doi.org/10.1145/3278721.3278779.

Zhang, Zhifei, S. Y.; and Qi, H. 2017. Age Progression/Regression by Conditional Adversarial Autoencoder. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE.