Fairness by Learning Orthogonal Disentangled Representations

Mhd Hasan Sarhan[1,2][0000−0003−0473−5461], Nassir Navab[1,3], Abouzar Eslami[1][0000−0001−8511−5541], and Shadi Albarqouni[1,4][0000−0003−2157−2211]

[1] Computer Aided Medical Procedures, Technical University of Munich, Munich, Germany
hasan.sarhan@tum.de
[2] Carl Zeiss Meditec AG, Munich, Germany
[3] Computer Aided Medical Procedures, Johns Hopkins University, Baltimore, USA
[4] Computer Vision Lab, ETH Zurich, Switzerland

Abstract. Learning discriminative powerful representations is a crucial step for machine learning systems. Introducing invariance against arbitrary nuisance or sensitive attributes while performing well on specific tasks is an important problem in representation learning. This is mostly approached by purging the sensitive information from learned representations. In this paper, we propose a novel disentanglement approach to invariant representation problem. We disentangle the meaningful and sensitive representations by enforcing orthogonality constraints as a proxy for independence. We explicitly enforce the meaningful representation to be agnostic to sensitive information by entropy maximization. The proposed approach is evaluated on five publicly available datasets and compared with state of the art methods for learning fairness and invariance achieving the state of the art performance on three datasets and comparable performance on the rest. Further, we perform an ablative study to evaluate the effect of each component.

Keywords: Representation learning, disentangled representation, fairness in machine learning

1 Introduction

Learning representations that are useful for downstream tasks yet robust against arbitrary nuisance factors is a challenging problem. Automated systems powered by machine learning techniques are corner stones for decision support systems such as granting loans, advertising, and medical diagnostics. Deep neural networks learn powerful representations that encapsulate the extracted variations in the data. Since these networks learn from historical data, they are prone to represent the past biases and the learnt representations might contain information that were not intended to be released. This has raised various concerns regarding fairness, bias and discrimination in statistical inference algorithms [17]. The European union has recently released their "Ethics guidelines for trustworthy

AI" report [5] where it is stated that unfairness and biases must be avoided.

Since a few years, the community has been investigating to learn a latent representation $\boldsymbol{z}$ that well describes a target observed variable $\boldsymbol{y}$ (e.g. Annual salary) while being robust against a sensitive attribute $\boldsymbol{s}$ (e.g. Gender or race). This nuisance could be independent from the target task which is termed as a domain adaptation problem. One example is the identification of faces $\boldsymbol{y}$ regardless of the illumination conditions $\boldsymbol{s}$. In the other case termed fair representation learning $\boldsymbol{s}$ and $\boldsymbol{y}$ are not independent. This could be the case with $\boldsymbol{y}$ being the credit risk of a person while $\boldsymbol{s}$ is age or gender. Such relation between these variables could be due to past biases that are inherently in the data. This independence is assumed to hold when building fair classification models. Although this assumption is over-optimistic as these factors are probably not independent, we wish to find a representation $\boldsymbol{z}$ that is independent from $\boldsymbol{s}$ which justifies the usage of such a prior belief [18]. This is mostly approached by approximations of mutual information scores between $\boldsymbol{z}$ and $\boldsymbol{s}$ and force the two variables to minimize this score either in an adversarial [25, 16] or non-adversarial [14, 18] manner. These methods while performing well on various datasets, are still limited by either convergence instability problems in case of adversarial solutions or hindered performance compared to the adversarial counterpart. Learning disentangled representations has been proven to be beneficial to learning fairer representations compared to general purpose representations [13]. We use this concept to disentangle the components of the learned representations. Moreover, we treat the $\boldsymbol{s}$ and $\boldsymbol{y}$ as separate independent generative factors and decompose the learned representation in such a way that each representation holds information related to the respective generative factor. This is achieved by enforcing orthogonality between the representations as a relaxation for the independence constraint. We hypothesize that decomposing the latent code into target code $\boldsymbol{z_T}$ and residual sensitive $\boldsymbol{z_S}$ code would be beneficial for limiting the leakage of sensitive information into $\boldsymbol{z_T}$ by redirecting it to $\boldsymbol{z_S}$ while keeping it informative about some target task that we are interested in.

We propose a framework for learning invariant fair representations by decomposing learned representations into target and residual/sensitive representations. We impose disentanglement on the components of each code and impose orthogonality constraint on the two learned representations as a proxy for independence. The learned target representation is explicitly enforced to be agnostic to sensitive information by maximizing the entropy of sensitive information in $\boldsymbol{z_T}$.

Our contributions are three-folds:

- Decomposition of target and sensitive data into two orthogonal representations to promote better mitigation of sensitive information leakage.
- Promote disentanglement property to split hidden generative factors of each learned code.

[5] Ethics guidelines for trustworthy AI, https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

– Enforce the target representation to be agnostic of sensitive information by maximizing the entropy.

2 Related work

Learning fair and invariant representations has a long history. Earlier strategies involved changing the examples to ensure fair representation of the all groups. This relies on the assumption that equalized opportunities in the training set would generalize to the test set. Such techniques are referred to as data massaging techniques [9, 19]. These approaches may suffer of under-utilization of data or complications on the logistics of data collection. Later, Zemel *et al.* [26] proposed a semi-supervised fair clustering technique to learn a representation space where data points are clustered such that each cluster contains similar proportions of the protected groups. One drawback is that the clustering constraint limits the power of a distributed representation. To solve this, Louizos *et al.* [14] presented the Variational Fair Autoencoder (VFAE) where a model is trained to learn a representation that is informative enough yet invariant to some nuisance variables. This invariance is approached through Maximum Mean Discrepancy (MMD) penalty. The learned sensitive-information-free representation could be later used for any subsequent processing such as classification of a target task. After the success of Generative Adversarial Networks (GANs) [7], multiple approaches leveraged this learning paradigm to produce robust invariant representations [25, 27, 5, 16]. The problem setup in these approaches is a minimax game between an encoder that learns a representation for a target task and an adversary that extracts sensitive information from the learned representation. In this case, the encoder minimizes the negative log-likelihood of the adversary while the adversary is forced to extract sensitive information alternatively. While methods relying on adversarial zeros-sum game of negative log-liklihood minimization and maximization perform well in the literature, they sometimes suffer from convergence problems and require additional regularization terms to stabilize the training. To overcome these problems, Roy *et al.* [21] posed the problem as an adversarial non-zero sum game where the encoder and discriminator have competing objectives that optimize for different metrics. This is achieved by adding an entropy loss that forces the discriminator to be un-informed about sensitive information. It is worth noting that it is argued by [18] that adversarial training for fairness and invariance is unnecessary and sometimes leads to counter productive results. Hence, they approximated the mutual information between the latent representation and sensitive information using a variational upper bound. Lastly, Creager *et al.* [3] proposed a fair representation learning model by disentanglement, their model has the advantage of flexibly changing sensitive information at test time and combine multiple sensitive attributes to achieve subgroup fairness. In their work, independence is enforced adversarially by utilizing a discriminator to distinguish simulated independent representations (fake) from learned representations (real).

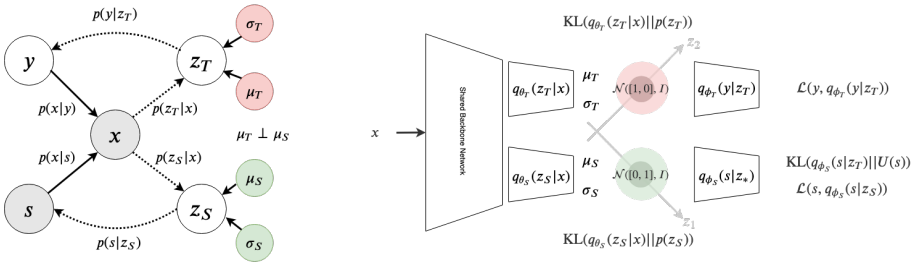


Fig. 1: Left: The graphical model of our proposed method. Right: Our framework encode the input data to intermediate target and residual (sensitive) representations, parameterized by μ and σ. Samples from the estimated posteriors are fed to the discriminators to predict the target and sensitive labels.

3 Methodology

let $\mathcal{X}$ be the dataset of individuals from all groups and $\boldsymbol{x} \in \mathbb{R}^D$ be an input sample. Each input is associated with a target attribute $\boldsymbol{y} = \{y_1, \ldots, y_n\} \in \mathbb{R}^n$ with n classes, and a sensitive attribute $\boldsymbol{s} = \{s_1, \ldots, s_m\} \in \mathbb{R}^m$ with m classes. Our goal is to learn an encoder that maps input $\boldsymbol{x}$ to two low-dimensional representations $\boldsymbol{z_T} \in \mathbb{R}^{d_T}$, $\boldsymbol{z_S} \in \mathbb{R}^{d_S}$. Ideally $\boldsymbol{z_T}$ must contain information regarding target attribute while mitigating leakage about the sensitive attribute and $\boldsymbol{z_S}$ contains residual information that is related to the sensitive attribute.

3.1 Fairness definition

One of the common definition of fairness that has been proposed in the literature [25, 21, 20, 1] is simply requiring the sensitive information to be statistically independent from the target. Mathematically, the prediction of a classifier $p(\boldsymbol{y}|\boldsymbol{x})$ must be independent from the sensitive information, which is expressed as follows

$$p(\boldsymbol{y}|\boldsymbol{x}) = p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{s}) \tag{1}$$

For example, in the German credit dataset, we need to predict the credit behaviour of the bank account holder regardless the sensitive information, such as gender, age ...etc. In other words, $p(\boldsymbol{y} = \text{good credit risk}|\boldsymbol{x}, \boldsymbol{s} = \text{male})$ should be equal to $p(\boldsymbol{y} = \text{good credit risk}|\boldsymbol{x}, \boldsymbol{s} = \text{female})$. The main objective is to learn fair data representations that are i) informative enough for the downstream task, and ii) independent from the sensitive information.

3.2 Problem Formulation

To promote the independence of the generative factors, *i.e.* target and sensitive information, we aim to maximize the log likelihood of the conditional distribution

$\log p(\boldsymbol{y}, \boldsymbol{s}|\boldsymbol{x})$, given the fairness assumption in Eq. 1

$$p(\boldsymbol{y}, \boldsymbol{s}|\boldsymbol{x}) = \frac{p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{s})p(\boldsymbol{x}|\boldsymbol{s})p(\boldsymbol{s})}{p(\boldsymbol{x})} = p(\boldsymbol{s}|\boldsymbol{x})p(\boldsymbol{y}|\boldsymbol{x}) \tag{2}$$

To enforce our aforementioned conditions, we let our model $f(\cdot)$ encode the observed input data $\boldsymbol{x}$ into target $\boldsymbol{z_T}$ and residual $\boldsymbol{z_S}$ intermediate representations on which the independence constraints are applied,

$$p(y, s|x) = \underbrace{p(y|z_T)}_{L_T}\underbrace{p(z_T|x)}_{L_{z_T}}\underbrace{p(s|z_S)}_{L_S}\underbrace{p(z_S|x)}_{L_{z_S}} \tag{3}$$

losses in Sec. 3.3 correspond to terms in Eq. 3 which are shown in the under brackets. The log likelihood is maximized given the following constraints; (i) $p(\boldsymbol{z_S}|\boldsymbol{x})$ is statistically independent from $p(\boldsymbol{z_T}|\boldsymbol{x})$, and (ii) $\boldsymbol{z_T}$ is agnostic to sensitive information $\boldsymbol{s}$. Our objective function J can be written as

$$J = -\log p(\boldsymbol{y}, \boldsymbol{s}|\boldsymbol{x}) \text{ s.t. } \mathrm{MI}(\boldsymbol{z_T}, \boldsymbol{z_S}) = 0 \text{ and } \mathrm{KL}(p(\boldsymbol{s}|\boldsymbol{z_T}),\mathcal{U}) = 0 \tag{4}$$

where $\mathcal{U}(\boldsymbol{s})$ is the uniform distribution.

3.3 Fairness by Learning Orthogonal and Disentangled Representations

As depicted in Fig. 1, our observed data $\boldsymbol{x}$ is fed to a shared encoder $f(\boldsymbol{x}; \theta)$, then projected into two subspaces producing our target, and residual (sensitive) representations using the encoders; $q_{\theta_T}(\boldsymbol{z_T}|\boldsymbol{x})$, and $q_{\theta_S}(\boldsymbol{z_S}|\boldsymbol{x})$, respectively, where θ is shared parameter, *i.e.* $\theta = \theta_S \cap \theta_T$. Each representation is fed to the corresponding discriminator; target discriminator, $q_{\phi_T}(\boldsymbol{y}|\boldsymbol{z_T})$, and sensitive discriminator $q_{\phi_S}(\boldsymbol{s}|\boldsymbol{z_S})$. Both discriminators and encoders are trained in supervised fashion to minimize the following loss,

$$\mathcal{L}_T(\theta_T, \phi_T) = \mathrm{KL}(p(\boldsymbol{y}|\boldsymbol{x}) \;||\; q_{\phi_T}(\boldsymbol{y}|\boldsymbol{z_T})), \tag{5}$$

$$\mathcal{L}_S(\theta_S^*, \phi_S) = \mathrm{KL}(p(\boldsymbol{s}|\boldsymbol{x}) \;||\; q_{\phi_S}(\boldsymbol{s}|\boldsymbol{z_S})), \tag{6}$$

where $\theta_S^* = \theta_S \backslash \theta$.

To ensure that our target representation does not encode any leakage of the sensitive information, we follow Roy *et al.* [21] in maximizing the entropy of the sensitive discriminator given the target representation $q_{\phi_S}(\boldsymbol{s}|\boldsymbol{z_T})$ as

$$\mathcal{L}_E(\phi_S, \theta_T) = \mathrm{KL}(q_{\phi_S}(\boldsymbol{s}|\boldsymbol{z_T}) \;||\; \mathcal{U}(\boldsymbol{s})). \tag{7}$$

We relax the independence assumption by enforcing i) disentanglement property, and ii) the orthogonality of the corresponding representations.

To promote the (i) disentanglement property on the target representation, we first need to estimate the distribution $p(\boldsymbol{z_T}|\boldsymbol{x})$ and enforce some sort of independence among the latent factors,

$$p(\boldsymbol{z_T}|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\boldsymbol{z_T})p(\boldsymbol{z_T})}{p(\boldsymbol{x})}, \text{ s.t. } p(\boldsymbol{z_T}) = \prod_{i=1}^{N_T} p(z_T^i). \tag{8}$$

Since $p(z_T|\boldsymbol{x})$ is intractable, we employ the Variational Inference, thanks to the re-paramterization trick [11], and let our model output the distribution parameters; $\boldsymbol{\mu_T}$, and $\boldsymbol{\sigma_T}$, and minimize the KL-divergence between posterior $q_{\theta_T}(\boldsymbol{z_T}|\boldsymbol{x})$ and prior $p(\boldsymbol{z_T})$ distributions as

$$\mathcal{L}_{\boldsymbol{z_T}}(\theta_T) = \mathrm{KL}(q_{\theta_T}(\boldsymbol{z_T}|\boldsymbol{x}) \;||\; p(\boldsymbol{z_T})), \tag{9}$$

where $p(\boldsymbol{z_T}) = \prod_{i=1}^{N_T} p(z_T^i) = \mathcal{N}(\mathbf{0}, I)$, and $q_{\theta_T}(\boldsymbol{z_T}|\boldsymbol{x}) = \mathcal{N}(\boldsymbol{z_T}; \boldsymbol{\mu_T}, diag(\boldsymbol{\sigma_T}^2))$. Similarly, we enforce the same constraints on the residual (sensitive) representation $\boldsymbol{z_S}$ and minimize the KL-divergence as $\mathcal{L}_{\boldsymbol{z_S}}(\theta_S) = \mathrm{KL}(q_{\theta_S}(\boldsymbol{z_S}|\boldsymbol{x}) \;||\; p(\boldsymbol{z_S}))$.

To enforce the (ii) orthogonality between the target and residual (sensitive) representations,*i.e.* $\boldsymbol{\mu_S} \perp \boldsymbol{\mu_T}$, we hard code the means of the prior distributions to orthogonal means. In this way, we implicitly enforce the weight parameters to project the representations into orthogonal subspaces. To illustrate this in 2-dimensional space, we set the prior distributions to $p(\boldsymbol{z_S}) = \mathcal{N}([0,1]^T, I)$, and $p(\boldsymbol{z_T}) = \mathcal{N}([1,0]^T, I)$ (*cf.* Fig. 1).

To summarize, an additional loss term is introduced to the objective function promoting both Orthogonality and Disentanglement properties, denoted *Orthogonal-Disentangled* loss,

$$\mathcal{L}_{OD}(\theta_T, \theta_S) = \mathcal{L}_{\boldsymbol{z_T}}(\theta_T) + \mathcal{L}_{\boldsymbol{z_S}}(\theta_S). \tag{10}$$

A variant of this loss without the property of orthogonality, denoted *Disentangled* loss, is also introduced for the purpose of ablative study (See Sec. 4.3).

3.4 Overall objective function

To summarize, our overall objective function is

$$arg \min_{\theta_T, \theta_S, \phi_T, \phi_S} \mathcal{L}_T(\theta_T, \phi_T) + \mathcal{L}_S(\theta_S^*, \phi_S) + \lambda_E \mathcal{L}_E(\phi_S, \theta_T) + \lambda_{OD} \mathcal{L}_{OD}(\theta_T, \theta_S) \tag{11}$$

where λ_E, and λ_{OD} are hyper-parameters to weigh the *Entropy* loss and the *Orthogonal-Disentangled* loss, respectively. A sensitivity analysis on the hyper-parameters is presented in Sec. 4.5.

4 Experiments

In this section, the performance of the learned representations using our method will be evaluated and compared against various state of the art methods in the

Algorithm 1 Learning Orthogonal Disentangled Fair Representations

Require: Maximum Epochs E_{max}, Step size t_s, $\lambda_{OD}, \lambda_E, \gamma_{OD}, \gamma_E, p(\boldsymbol{z_T}), p(\boldsymbol{z_S})$
Ensure: $\boldsymbol{z_S} \perp \boldsymbol{z_T}$
  Initialize: $\theta_T, \theta_S, \phi_T, \phi_S \leftarrow \theta_T^{(0)}, \theta_S^{(0)}, \phi_T^{(0)}, \phi_S^{(0)}$
  for $t = 1, 2, \ldots, E_{max}$ **do**
    $[\boldsymbol{\mu_T}, \boldsymbol{\sigma_T}] = q_{\theta_T}(\boldsymbol{z_T}|\boldsymbol{x})$
    $[\boldsymbol{\mu_S}, \boldsymbol{\sigma_S}] = q_{\theta_S}(\boldsymbol{z_S}|\boldsymbol{x})$
    sample $\boldsymbol{z_T} \sim \mathcal{N}(\boldsymbol{\mu_T}, diag(\boldsymbol{\sigma_T}^2))$
    sample $\boldsymbol{z_S} \sim \mathcal{N}(\boldsymbol{\mu_S}, diag(\boldsymbol{\sigma_S}^2))$
    compute $\mathcal{L}_{\boldsymbol{z_T}}(\theta_T) = \mathrm{KL}(q_{\theta_T}(\boldsymbol{z_T}|\boldsymbol{x}) \,||\, p(\boldsymbol{z_T}))$
    compute $\mathcal{L}_{\boldsymbol{z_S}}(\theta_S) = \mathrm{KL}(q_{\theta_S}(\boldsymbol{z_S}|\boldsymbol{x}) \,||\, p(\boldsymbol{z_S}))$
    compute $\mathcal{L}_T(\theta_T, \phi_T) = -\sum p(\boldsymbol{y}|\boldsymbol{x}) \log[q_{\phi_T}(\boldsymbol{y}|\boldsymbol{z_T})]$
    compute $\mathcal{L}_S(\theta_S^*, \phi_S) = -\sum p(\boldsymbol{s}|\boldsymbol{x}) \log[q_{\phi_S}(\boldsymbol{s}|\boldsymbol{z_S})]$
    compute $\mathcal{L}_E(\phi_S, \theta_T) = \sum q_{\phi_S}(\boldsymbol{s}|\boldsymbol{z_T}) \log[q_{\phi_S}(\boldsymbol{s}|\boldsymbol{z_T})]$
    update $\lambda_{OD} \leftarrow \lambda_{OD} \gamma_{OD}^{t/t_s}$
    update $\lambda_E \leftarrow \lambda_E \gamma_E^{t/t_s}$
    $\mathcal{L}_{OD}(\theta_T, \theta_S) = \mathcal{L}_{\boldsymbol{z_T}}(\theta_T) + \mathcal{L}_{\boldsymbol{z_S}}(\theta_S)$
    $J(\theta_T, \theta_S, \phi_T, \phi_S) = \mathcal{L}_T(\theta_T, \phi_T) + \mathcal{L}_S(\theta_S^*, \phi_S) + \lambda_E \mathcal{L}_E(\phi_S, \theta_T) + \lambda_{OD} \mathcal{L}_{OD}(\theta_T, \theta_S)$
    update $\theta_T, \theta_S, \phi_T, \phi_S \leftarrow arg\min J(\theta_T, \theta_S, \phi_T, \phi_S)$
  end for
  return $\theta_T, \theta_S, \phi_T, \phi_S$

domain. First, we present the experimental setup by describing the five datasets used for validation, the model implementation details for each dataset, and design of the experiments. We then compare the model performance with state of the art fair representation models on the datasets. We perform an ablative study to monitor the effect of each added component on the overall performance. We then evaluate the models qualitatively by showing t-SNE projections of the learned representations. Lastly, we perform a sensitivity analysis to study the effect of hyper-parameters on the training.

4.1 Experimental Setup

Tabular data: For evaluating fair classification, we use two datasets from the UCI repository [4], namely, the German and the Adult datasets. The German credit dataset consists of 1000 samples each with 20 attributes, and the target task is to classify a bank account holder having good or bad credit risk. The sensitive attribute is the gender of the bank account holder. The adult dataset contains 45,222 samples each with 14 attributes. The target task is a binary classification of annual income being more or less than $\$50,000$ and again gender is the sensitive attribute.

Visual data: To examine the model learned invariance on visual data, we have used the application of illumination invariant face classification. Ideally, we want the representation to contain information about the subject's identity without

holding information regarding illumination direction. For this purpose, the extended YaleB dataset is used [6]. The dataset contains the face images of 38 subjects under five different light source direction conditions (upper right, lower right, lower left, upper left, and front). The target task is the identification of the subject while the light source condition is considered the sensitive attribute.

CIFAR data: Following Roy *et al.* [21], we have created a binary target task from CIFAR-10 dataset [12]. The original dataset contains 10 classes we refer to as fine classes, we divide the 10 classes into two categories living and non-living classes and refer to this split as coarse classes. It is expected that living objects have common visual proprieties that differ from non-living ones. The target task is the classification of the coarse classes while not revealing information about the fine classes. With a similar concept, we divide the 100 fine classes of CIFAR-100 dataset into 20 coarse classes that cluster similar concepts into one category. For example, the coarse class 'aquatic mammals' contains the fine classes 'beaver', 'dolphin', 'otter', 'seal', and 'whale'. For the full details of the split, the reader is referred to [21] or the supplementary materials of this manuscript. The target task for CIFAR-100 is the classification of the coarse classes while mitigating information leakage regarding the sensitive fine classes.

Implementation details: For the Adult and German datasets, we follow the setup appeared in [21] by having a 1-hidden-layer neural network as encoder, the discriminator has two hidden layer and the target predictor is a logistic regression layer. Each hidden layer contains 64 units. The size of the representation is 2. The learning rate for all components is 10^{-3} and weight decay is 5×10^{-4}.
For the Extended YaleB dataset, we use an experimental setup similar to Xie *et al.* [25] and Louizos *et al.* [14] by using the same train/test split strategy. We used $38 \times 5 = 190$ samples for training and 1096 for testing. The model setup is similar to [25, 21], the encoder consisted of one layer, target predictor is one linear layer and the discriminator is neural network with two hidden layers each contains 100 units. The parameters are trained using Adam optimizer with a learning rate of 10^{-4} and weight decay of 5×10^{-2}.
Similar to [21], we employed ResNet-18 [8] architecture for training the encoder on the two CIFAR datasets. For the discriminator and target classifiers, we employed a neural network with two hidden layers (256 and 128 neurons). For the encoder, we set the learning rate to 10^{-4} and weight decay to 10^{-2}. For the target and discriminator networks, the learning rate and weight decay were set to 10^{-2} and 10^{-3},respectively. Adam optimizer [10] is used in all experiments.

Experiments design: We address two questions in the experiments. First, is *how much information about the sensitive attributes is retained in the learned representation* $\boldsymbol{z_T}$*?*. Ideally, $\boldsymbol{z_T}$ would not contain any sensitive attribute information. This is evaluated by training a classifier with the same architecture as the discriminator network on sensitive attributes classification task. The closer the accuracy to a naive majority label predictor, the better the model is. This classifier is trained with $\boldsymbol{z_T}$ as input after the encoder, target, and discriminator had

Table 1: Results on CIFAR-10 and CIFAR-100 datasets.

	CIFAR-10		CIFAR-100	
	Target Acc. ↑	Sensitive Acc. ↓	Target Acc. ↑	Sensitive Acc. ↓
Baseline	0.9775	0.2344	**0.7199**	0.3069
Xie et al. [25] (trade-off #1)	0.9752	0.2083	0.7132	0.1543
Roy et al. [21] (trade-off #1)	**0.9778**	0.2344	0.7117	0.1688
Xie et al. [25] (trade-off #2)	0.9735	0.2064	0.7040	0.1484
Roy et al. [21] (trade-off #2)	0.9679	0.2114	0.7050	0.1643
Ours	0.9725	**0.1907**	0.7074	**0.1447**

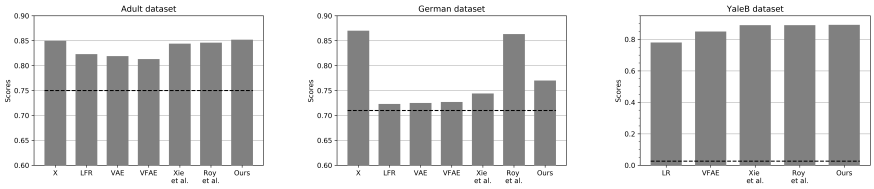
been trained and freezed. Second, is *how well the learned representation* $\boldsymbol{z_T}$ *performs in identifying target attributes?*. To this end, we train a classifier similar to the target on the learned representation $\boldsymbol{z_T}$ to detect the target attributes. We also visualize the representations $\boldsymbol{z_T}$ and $\boldsymbol{z_S}$ by using their t-SNE projections to show how the learned representations describe target attributes while being agnostic to the sensitive information.

4.2 Comparison with state of the art

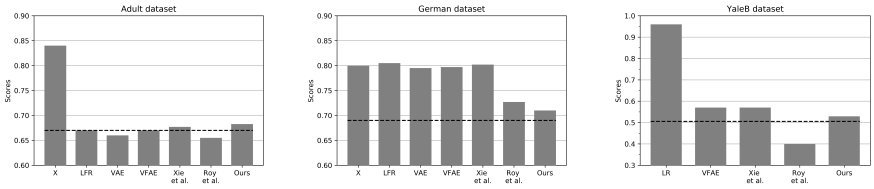
We compare the proposed approach against various state of the art methods on the five presented datasets. We first train the model with Algorithm 1 while changing hyper-parameters between runs.We choose the best performing model in terms of the trade-off between target and sensitive classification accuracy based on $\boldsymbol{z_T}$. We then compare it with various state of the art methods for sensitive information leakage and retaining target information.

CIFAR datasets: We compare the proposed approach with two other state of the art methods on the CIFAR-10 and CIFAR-100 datasets, namely Xie *et al.* [25] and Roy *et al.* [21]. We examine two different trade-off points of both approaches. The first trade-off point is the one with best target accuracy reported by the model while the second trade-off point is the one with the target accuracy closest to ours for a more fair comparison. The lower the target accuracy in the trade-off the better (lower) the sensitive accuracy is. We can see when the target accuracies are comparable, our model performs better in preventing sensitive information leakage to the representation $\boldsymbol{z_T}$. Hence, the proposed method has a better trade-off on the target and sensitive accuracy for both CIFAR-10 and CIFAR-100 datasets. However, the peak target performance is comparable but lower than the peak target performance of the studied methods.

Extended YaleB dataset: For the illumination invariant classification task on the extended YaleB dataset, the proposed method is compared with the logistic regression baseline (LR), variational fair autoencoder VFAE [14], Xie *et al.* [25] and Roy *et al.* [21]. The results are shown in Fig. 2 on the right hand side. The proposed model performs best on the target attribute classification while having



(a) Target attribute classification accuracy.



(b) Sensitive attribute classification accuracy.

Fig. 2: Results on Adult, German, and extended YaleB datasets. The dashed black line represent a naive majority classifier that predicts the majority label.

the closest performance to the majority classification line (dashed line in Fig. 2). The majority line is the trivial baseline of predicting the majority label. The closer the sensitive accuracy to the majority line the better the model is in hiding sensitive information from $\boldsymbol{z_T}$. This means the learned representation is powerful at identifying subject in the images regardless of illumination conditions. To assess this visually, refer to Sec. 4.4 for qualitative analysis.

Tabular datasets: On the Adult and German datasets, we compare with LFR [26], vanilla VAE [11], variational fair autoencoder [14], Xie *et al.* [25] and Roy *et al.* [21]. The results of these comparisons are shown in Fig. 2. On the German dataset, we observe a very good performance in hiding sensitive information with 71% accuracy compared to 72.7% in [21]. On the target task, the model performs well compared to other models except for [21] which does marginally better than the rest. On the Adult dataset, our proposed model performs better than the aforementioned models on the target task while leaking slightly more information compared to other methods and the majority line at 67%. Our method has 68.26% sensitive accuracy while LFR, VAE, vFAE, Xie *et al.* , and Roy *et al.* have 67%, 66%, 67%, 67.7%, and 65.5% sensitive accuracy, respectively.

Generally, we observe that the proposed model performs well on all datasets with state of the art performance on visual datasets (CIFAR-10, CIFAR-100, YaleB). This suggests that such a model could lead to more fair/invariant representation without large sacrifices on downstream tasks.

4.3 Ablative study

In this section, we evaluate the contributions provided in the paper by eliminating parts of the loss function and study how each part affects the training in terms of target and sensitive accuracy. To this end, we used the best performing models after hyper-parameter search when training for all contributions for each dataset. The models are trained with the same settings and architectures described in Sec. 4.1. We compare five different variations for each model alongside the baseline classifier:

1. **Baseline**: Training a deterministic classifier for the target task and evaluate the information leakage about the sensitive attribute.
2. **Entropy w/o KL**: Entropy loss $\mathcal{L}_E$ is incorporated (Equation 7) in the loss while $\mathcal{L}_{OD}$ is not included (Equation 10).
3. **KL Orth. w/o Entropy**: Entropy loss $\mathcal{L}_E$ is not used (Equation 7) while $\mathcal{L}_{OD}$ is used for target and sensitive representations with orthogonal means (Equation 10).
4. **w/o Entropy w/o KL**: Neither entropy loss nor KL divergence are used in the loss. This case is similar ti multi-task learning with the tasks being the classification of target and sensitive attributes.
5. **Entropy + KL w/o Orth.**: Entropy loss $\mathcal{L}_E$ is used and *disentangled loss* is used with similar means. Hence, there might be some disentanglement of generative factors in the components of each latent code but no constraints are applied to force disentanglement of the two representations.
6. **Entropy + KL Orth.**: All contributions are included.

The results of the ablative study are shown in Figure 3.

- For the *sensitive class accuracy*, it is desirable to have a lower accuracy in distinguishing sensitive attributes. Compared to the baseline, we observe that adding entropy loss and orthogonality constraints on the representations lowers the discriminative power of the learned representation regarding sensitive information. This is valid on all studied datasets except for CIFAR-10 where orthogonality constraint without entropy produced better representations for hiding sensitive information with a small drop (0.26%) on the target task performance. In the rest of the cases, having either entropy loss or KL loss only does not bring noticeable performance gains compared to a multi-task learning paradigm. This could be attributed to the fact that orthogonality on its own does not enforce independence of random variables and another constraint is needed to encourage independent latent variables (*i.e.* entropy loss).
- Comparing baseline with **w/o Entropy w/o KL** case answers the important question *"Does multi-task learning with no constraints on representations bring any added value in mitigating sensitive information leakage?"*. In three out of the five studied datasets, it is the case. We can see lower accuracy in identifying sensitive information by using the learned target representation as input to a classifier while having no constraints on the relationship

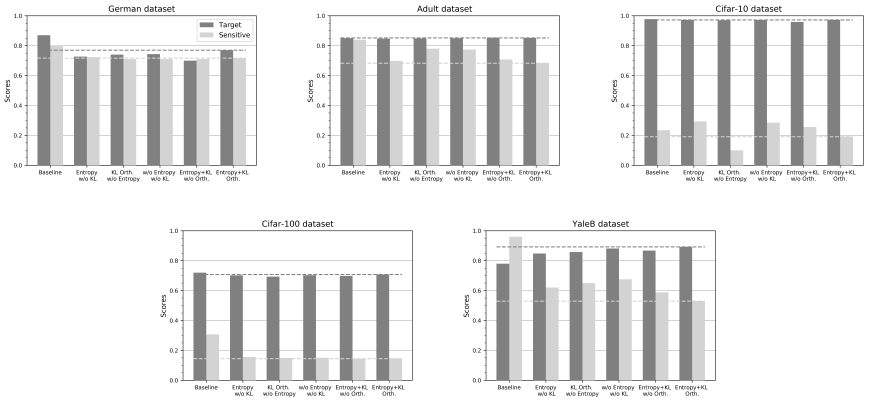


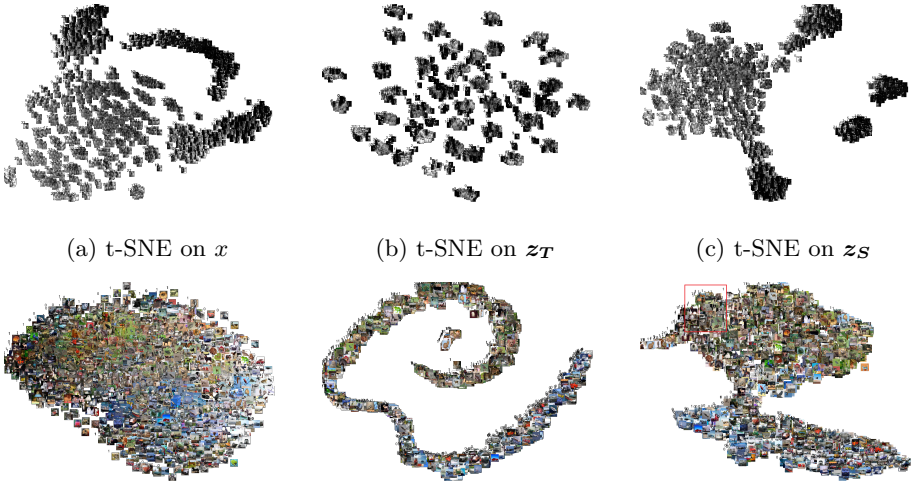
Fig. 3: Ablative study. Dark gray and light gray dashed lines represent the accuracy results on the target and sensitive task respectively for the "Entropy + KL Orth." model.

between the sensitive and target representations during the training process of the encoder. Simply, adding an auxiliary classifier to the target classifier and force it to learn information about sensitive attributes hides some sensitive data from the target classifier.

– Regarding **target accuracy**, the proposed model does not suffer from large drops in target performance when disentangling target from sensitive information. This could be seen by comparing target accuracy between the baseline and **Entropy+KL Orth.** columns. The largest drop in target performance compared to no privacy baseline is seen on the German dataset. This could be because of the very high dependence between gender and granting good or bad credit to a subject in the dataset and the small amount of subjects in the dataset.

4.4 Qualitative analysis

We visualize the learned embeddings using t-SNE [15] projections for the extended YaleB and CIFAR-10 datasets (*cf.* Fig. 4. We use the image space, $\boldsymbol{z_T}$, $\boldsymbol{z_S}$ as inputs to the projection to visualize what type of information is held within each representation. We also show the label of each image with regards to the target task to make it easier to investigate the clusters. For the extended YaleB, we see that, using the image space $\boldsymbol{x}$, the images are clustered mostly depending on their illumination conditions. However, when using $\boldsymbol{z_T}$, the images are not clustered according lighting conditions but rather, mostly based on the subject identity. Moreover, the visualization of representation $\boldsymbol{z_S}$ shows that the representation contains information about the sensitive class. For the CIFAR-10 dataset, using the image space basically clusters the images on the dominant



(a) t-SNE on x (b) t-SNE on $\boldsymbol{z_T}$ (c) t-SNE on $\boldsymbol{z_S}$

(d) t-SNE on x (e) t-SNE on $\boldsymbol{z_T}$ (f) t-SNE on $\boldsymbol{z_S}$

Fig. 4: t-SNE visualization of the extended YaleB faces (top) and CIFAR-10 (bottom) images. Figure is better seen in color and high resolution.

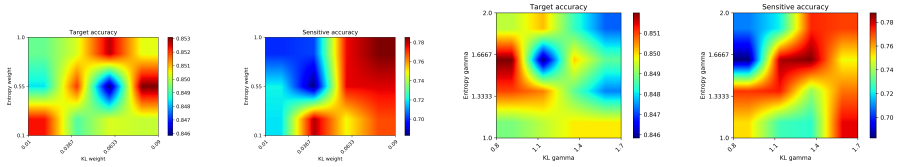


Fig. 5: Sensitivity analysis on the Adult dataset

color. When using $\boldsymbol{z_T}$, it is clear that the target information is separated where the right side represent the non-living objects, and the left to inside part represents the living objects. What should be observed in $\boldsymbol{z_T}$, is that within each target class, the fine classes are mixed and indistinguishable as we see cars, boats and trucks mixed in the right hand side of the figure, for example. The representation $\boldsymbol{z_S}$ has some information about the target class and also has the residual information about the fine classes as we see in the annotated red rectangle. A group of horses images are clustered together, then few dogs' images are clustered under it, then followed by birds. This shows that $\boldsymbol{z_S}$ has captured some sensitive information while $\boldsymbol{z_T}$ is more agnostic to the sensitive fine classes.

4.5 Sensitivity analysis

To analyze the effect of hyper-parameters choices on the sensitive and target accuracy, we show heatmaps of how the performance changes when the studied hyper-parameters are changed. The investigated hyper-parameters are KL

weight (λ_{OD}), Entropy Weight (λ_E), KL gamma (γ_{OD}), and Entropy gamma (γ_E). We show the results on the Adult dataset. We can see that the sensitive accuracy is sensitive to λ_{OD} more than λ_E as changes in λ_E do not induce much change on the sensitive accuracy. A similar trend is not visible on the target accuracy. Regarding the choice of γ_{OD} and γ_E, we can see that the sensitive leakage is highly affected by these hyper-parameters and the results vary when changed. However, a more robust performance is observed on the target classification task.

5 Conclusions and future work

In this work, we have proposed a novel model for learning invariant representations by decomposing the learned codes into sensitive and target representation. We imposed orthogonality and disentanglement constraints on the representations and forced the target representation to be uninformative of the sensitive information by maximizing sensitive entropy. The proposed approach is evaluated on five datasets and compared with the state of the art models. The results show that our proposed model performs better than state of the art models on three datasets and performed comparably on the other two. We observe better hiding of sensitive information while affecting the target accuracy minimally. This goes in line with our hypothesis that decomposing the two representations and enforcing orthogonality could solve the information leakage problem by redirecting the information into the sensitive representation. One current limitation of this work is that it requires a target task to learn the disentanglement which could be avoided by learning the reconstruction as an auxiliary task similar to other privacy-preserving applications [24]. A direction worth investigating is replacing the pre-definition of the orthogonal sub-spaces priory by learning orthogonality intrinsically with low-rank constraints on the learned representations [22]. Another direction for future work could be focusing on the disentanglement part of the framework. The current disentanglement of factors of generation in the learned representations could be improved by using other disentanglement frameworks [2, 23] that are capable of better disentanglement.

Acknowledgments

S.A. is supported by the PRIME programme of the German Academic Exchange Service (DAAD) with funds from the German Federal Ministry of Education and Research (BMBF).

References

1. Barocas, S., Hardt, M., Narayanan, A.: Fairness and Machine Learning. fairmlbook.org (2019), http://www.fairmlbook.org
2. Chen, R.T., Li, X., Grosse, R.B., Duvenaud, D.K.: Isolating sources of disentanglement in variational autoencoders. In: Advances in Neural Information Processing Systems. pp. 2610–2620 (2018)

3. Creager, E., Madras, D., Jacobsen, J.H., Weis, M.A., Swersky, K., Pitassi, T., Zemel, R.: Flexibly fair representation learning by disentanglement. arXiv preprint arXiv:1906.02589 (2019)
4. Dua, D., Graff, C.: UCI machine learning repository (2017)
5. Edwards, H., Storkey, A.: Censoring representations with an adversary. arXiv preprint arXiv:1511.05897 (2015)
6. Georghiades, A.S., Belhumeur, P.N., Kriegman, D.J.: From few to many: Illumination cone models for face recognition under variable lighting and pose. IEEE Transactions on Pattern Analysis and Machine Intelligence **23**(6), 643–660 (2001)
7. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Advances in neural information processing systems. pp. 2672–2680 (2014)
8. He, K., Zhang, X., Ren, S., Sun, J.: Identity mappings in deep residual networks. In: European Conference on Computer Vision. pp. 630–645. Springer (2016)
9. Kamiran, F., Calders, T.: Classifying without discriminating. In: 2009 2nd International Conference on Computer, Control and Communication. pp. 1–6. IEEE (2009)
10. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
11. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
12. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images (2009)
13. Locatello, F., Abbati, G., Rainforth, T., Bauer, S., Schölkopf, B., Bachem, O.: On the fairness of disentangled representations. In: Advances in Neural Information Processing Systems. pp. 14584–14597 (2019)
14. Louizos, C., Swersky, K., Li, Y., Welling, M., Zemel, R.: The variational fair autoencoder. arXiv preprint arXiv:1511.00830 (2015)
15. Maaten, L.v.d., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research **9**(Nov), 2579–2605 (2008)
16. Madras, D., Creager, E., Pitassi, T., Zemel, R.: Learning adversarially fair and transferable representations. arXiv preprint arXiv:1802.06309 (2018)
17. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: A survey on bias and fairness in machine learning. arXiv preprint arXiv:1908.09635 (2019)
18. Moyer, D., Gao, S., Brekelmans, R., Galstyan, A., Ver Steeg, G.: Invariant representations without adversarial training. In: Advances in Neural Information Processing Systems. pp. 9084–9093 (2018)
19. Pedreshi, D., Ruggieri, S., Turini, F.: Discrimination-aware data mining. In: Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 560–568 (2008)
20. Quadrianto, N., Sharmanska, V., Thomas, O.: Discovering fair representations in the data domain. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8227–8236 (2019)
21. Roy, P.C., Boddeti, V.N.: Mitigating information leakage in image representations: A maximum entropy approach. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2586–2594 (2019)
22. Sanyal, A., Kanade, V., Torr, P.H., Dokania, P.K.: Robustness via deep low-rank representations. arXiv preprint arXiv:1804.07090 (2018)
23. Sarhan, M.H., Eslami, A., Navab, N., Albarqouni, S.: Learning interpretable disentangled representations using adversarial vaes. In: Domain Adaptation and Rep-

resentation Transfer and Medical Image Learning with Less Labels and Imperfect Data, pp. 37–44. Springer (2019)
24. Xiao, T., Tsai, Y.H., Sohn, K., Chandraker, M., Yang, M.H.: Adversarial learning of privacy-preserving and task-oriented representations. arXiv preprint arXiv:1911.10143 (2019)
25. Xie, Q., Dai, Z., Du, Y., Hovy, E., Neubig, G.: Controllable invariance through adversarial feature learning. In: Advances in Neural Information Processing Systems. pp. 585–596 (2017)
26. Zemel, R., Wu, Y., Swersky, K., Pitassi, T., Dwork, C.: Learning fair representations. In: International Conference on Machine Learning. pp. 325–333 (2013)
27. Zhang, B.H., Lemoine, B., Mitchell, M.: Mitigating unwanted biases with adversarial learning. In: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society. pp. 335–340 (2018)