Variational Fair Information Bottleneck

Sajad Norouzi
Department of Computer Science
University of Toronto
sajadn@cs.toronto.edu

Abstract

Learning a representation which is invariant to changes with respect to some specified factors are useful for a wide range of problems such as removing bias in classification. Based on variational information bottleneck [1] we propose a new framework to minimize the mutual information between a sensitive or nuisance factor and representation while keeping the required information for the classification task. We demonstrate the effectiveness of the proposed method both on a toy problem and Adult dataset. The implementation is available at: https://github.com/sajadn/Variational-Fair-Information-Bottleneck

1 Introduction

Data driven solutions are being used for wide range of tasks. They are making influencing consequential decisions such as bank loans, college admissions, and criminal sentences. This recent rise of applications brings some concerns about the fairness of these models.

In representation learning, one tries to find a lower dimensional manifold of data that makes performing tasks like classification easier. For instance, we can stack multiple convolutional layers with non-linearity to obtain a good representation for classification of images. However, due to the bias in training data the classifiers might be unfair to some groups in the data.

In fair representation learning, we try to find a representation of the data that are informative for a particular task while removing the factors that we have concerns about them. For example, we try to find a representation that help us for giving loan while there is no information about the race of individuals in the representation. Mathematically speaking, we want a latent representation z that is maximally informative about an observed random variable y (e.g., target label) while minimally informative about a sensitive or nuisance variable s.

In this work, we introduce a novel toy task for evaluation of fair representation methods. Furthermore, inspired by Variational Information Bottleneck, we introduce a new approach for learning a fair representation.

2 Background

Representation learning can be done both in unsupervised and supervised manner. Variational Autonecoder[2] (VAE) is a probabilistic framework tries to find the latent variable z which assign high probability to the observed data points. Using variational methods, they maximize a lower bound on log likelihood of the data. The objective of VAE is written below:

$$\max \; E_{q_\theta(z|x)}[\log p_\phi(x|z) - KL(q_\theta(z|x)||p(z))] \leq \log p(x) \tag{1}$$

The training procedure is unsupervised and no label is needed, but for evaluation of the learned representation we mostly use downstream tasks like classification which requires labels.

30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

Deep Variational Information Bottleneck[1] (DVIB) proposed an information-theoric supervised representation learning frameworks based on maximizing the mutual information between latent variable z and label y while minimizing the mutual information between z and observation x. Because of intractability of mutual information, they have used variational methods to lowerbound their objective. The final objective of DVIB is shown below:

$$\max \; E_{q_\theta(z|x)}[\log p_\phi(y|z) - \beta KL(q_\theta(z|x)||p(z))] \leq I(Y;Z) - \beta I(X;Z) \tag{2}$$

We can see that the objective of both techniques are pretty similar. DVIB replace the decoder network in VAE with a classifier to just keep the required information for the discriminative task.

These methods are general purpose representation learning approaches and they might suffer from unfairness with respect to some groups.

3 Related Works

Variational Fair Autoencoder (VFAE) [3] is a pivotal work trying to make VAE robust with respect to unfairness. They considered both supervised and semi-supervised circumstances. In unsupervised case, they lowerbounded $\log p(x|s)$ to store no information about the sensitive attribute s:

$$\max \; E_{q_\theta(z|x,s)}[\log p_\phi(x|s,z) - KL(q_\theta(z|x,s)||p(z))] \leq \log p(x|s) \tag{3}$$

This objective function also re-derived in [4] from an information theoric point of view. Modeling the problem in a unsupervised way can however be harmful to the performance of the model on the discriminative task, so they have suggested a semi-supervised version.

$$\begin{aligned} \max \; E_{q_\phi(z_{1_n}, z_{2_n}, y_n|x_n, s_n)}[&\log p(z_{2_n}) + \log p(y_n) + \log p_\theta(z_{1_n}|z_{2_n}, y_n) && (4)\\ &+ \log p_\theta(x_n|z_{1_n}, s_n) - \log q_\phi(z_{1_n}, z_{2_n}, y_n|x_n, s_n)] \leq \log p(x|s) && (5) \end{aligned}$$

Therefore using the above objective we can incorporate our knowledge about the label which hopefully leads to benefits on the performance of the model. Despite that, this loss function is too complex and we find it hard to use it in practice.

To force the representation to have even less information about the sensitive attribute, a regularization term added to equation 5. The regularizer is basically a Maximum Mean Discrepancy (MMD) term which is defined below:

$$R_{MMD} = ||\frac{1}{N_0}\sum_{i=1}^{N_0}\psi(x_i) - \frac{1}{N_1}\sum_{i=1}^{N_1}\psi(x_i)||^2 \tag{6}$$

Where ψ is a feature extractor. We can apply a kernel trick to compute this function more efficiently. So the final loss of semi-supervised VFAE is the summation of equation 5 and 6.

The effectiveness of MMD is a bit arguable and [5] proposed to use $I(Z;S|X)$ rather than MMD. They show for a particular task this regularizer works better than MMD. Their final regularizer is:

$$\mathcal{L}_{MI} \quad = \quad E_{q_\phi(z_1|x,s)}[\log \frac{q_\phi(z_1|x,s)}{\sum_s q_\phi(s|x)q_\phi(z_1|x,s)}] \tag{7}$$

This introduce a new parametric form $q_\phi(z_1|x)$ to the trainable components.

Both of these regularizers can be added to our proposed method, but in this work we study the performance of VFAE vs our proposed method without any regularization term and further study of the regularization term left as a future work.

4 Method

Similar to [4] we define our objective through information theoric terms. We believe that the objective of learning a fair representation as we mentioned in the introduction can be learning a latent representation z which has high mutual information with target label y, but has low mutual information with s. The objective can be written as:

$$max\ I(Y;Z|S) - \beta I(Z;S) \tag{8}$$

Because of possibility of correlation between y and s, in order to make sure that there is no information about s in the learned representation we condition the first term on sensitive attribute.

However, computing the mutual information is intractable. Following [1] we try to lower bound equation 10. We can write the first term as:

$$I(Y;Z|S) = \int p(y,z|s) \log \frac{p(y,z|s)}{p(y|s)p(z|s)} dydzds \tag{9}$$

$$I(Y;Z|S) = \int p(y,z|s) \log \frac{p(y|z,s)}{p(y|s)} dydzds \tag{10}$$

$$I(Y;Z|S) = \int p(y,z|s) \log p(y|z,s) dydzds - \int p(y,z|s) \log p(y|s) dydzds \tag{11}$$

Using the fact that $0 \leq D_{KL}(p(y|z,s), q_\theta(y|z,s))$ we can write:

$$\int p(y|z,s) \log q_\theta(y|z,s) dy \leq \int p(y|z,s) \log p(y|z,s) dy \tag{12}$$

So we can write following lower bound on the first mutual information term:

$$\int P(y,z|s) \log \frac{q_\theta(y|z,s)}{p(y|s)} dydzds \leq I(Y;Z|S) \tag{13}$$

$$\int P(y,z|s) \log q_\theta(y|z,s) dydzds + H_p(Y|S) \leq I(Y;Z|S) \tag{14}$$

$$\int P(y,z|s,x) \log q_\theta(y|z,s) dydzdxds + H_p(Y|S) \leq I(Y;Z|S) \tag{15}$$

Notice that the entropy of our labels $H(Y \mid S)$ is independent of our optimization procedure and so can be ignored. Now let's look at the second term:

$$I(Z;S) < I(Z;S,X) \tag{16}$$

And this mutual information between Z and joint distribution of S, X can be upper-bounded by following derivation:

$$I(Z;S,X) = \int dxdsdz\ p(s,x)p(z|s,x) \log \frac{p(z,x,s)}{p(x,s)p(z)} \tag{17}$$

$$I(Z;S,X) = \int dxdsdz\ p(s,x)p(z|s,x) \log \frac{p(z|x,s)}{p(z)} \tag{18}$$

$$I(Z;S,X) = \int dxdsdz\ p(s,x)p(z|s,x) \log p(z|x,s) - \int dz p(z) \log p(z) \tag{19}$$

Computing $p(z) = \int p(z|s,x)p(x,s)dxds$ can be difficult, so we again use the non-negtivity of $0 \leq D_{KL}[p(Z), r_\theta(Z)]$, so we can write:

$$-\int dzp(z)\log p(z) \leq -\int dzp(z)\log r_\theta(z) \tag{20}$$

So the final loss is:

$$\mathcal{L}(x_i, s_i, y_i) = E_{p(z|x_i,s_i)}[\log q_\theta(y_i|z,s) - \beta \log \frac{p_\phi(z|x_i,s_i)}{r_\theta(z)}] \tag{21}$$

Here we assume $r_\theta(z)$ to be equal to a standard Gaussian, but one can try to extend this work by using a parametric $r_\theta(z)$ to have a better approximation of $p(z)$. Because of similarity of this objective to Variational Information Bottleneck we call the proposed method Variational Fair Information Bottleneck (VFIB). The proposed algorithm contains an encoder and a classifier during the training. The classifier receives both x and sensitive attribute s, but after the training is done we throw away the classifier and just utilize the trained encoder to obtain the fair representation. Therefore to solve a downstream classification task we need to train a classifier as a post processing. We also train a classifier to predict the sensitive attribute in order to understand how much information about sensitive attribute is kept in the learned representation.

We want to emphasize that our proposed method doesn't have any adversary or decoder. The main difference of our method is feeding the sensitive attribute to the classifier which enforce having less information in learned latent. This connection is shown in figure 2 with a pink color.

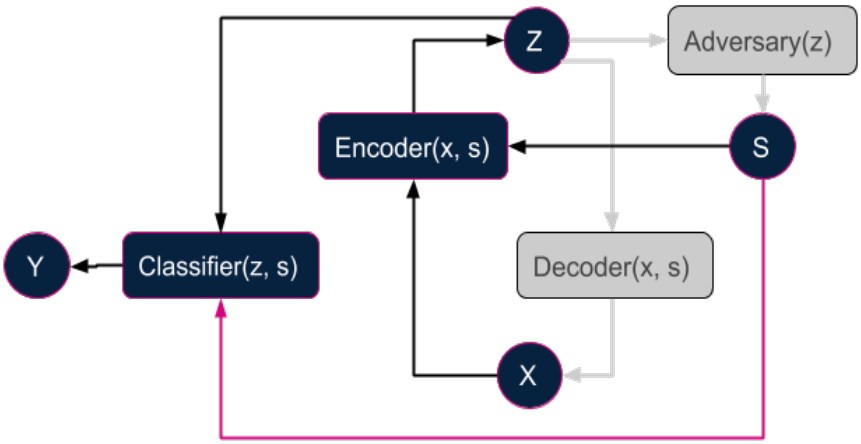


Figure 1: Diagram of fair representation learning framework, gray components are inactive in our method

Finally you can see a comparison between objective of VFAE, VIB, and VFIB in figure below:

5 Experiment

We believe that Toy problems are important for evaluating new techniques, but we're not aware of any toy problem for fairness. We first propose a toy problem based on MNIST and show the effectiveness of the proposed method on that. We further extend our experimental section by evaluating the method on Adult dataset.

5.1 Toy Problem

We first selected images belong to digits two and three from MNIST dataset. We define an operation called inversion which is changing the pixel values from black to white and white to black. We

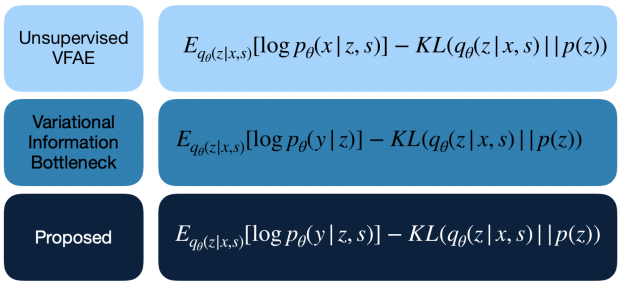


Figure 2: Objective of the proposed method vs related works

inverted 70% of digits two and 30% of digits three. You can find the samples from the new dataset below:

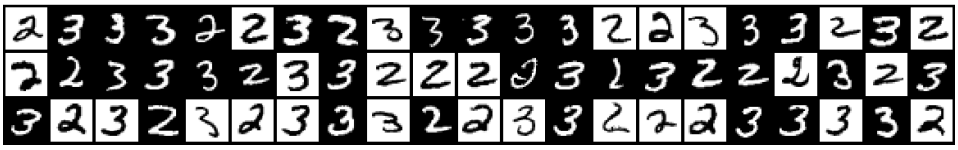


Figure 3: Samples from toy problem data

In this toy problem the sensitive attribute is the color and the target label is the label of the digit in the image. We apply our proposed method and also VFAE on this dataset and you can find the results in table 2.

For the architectures of the networks we followed VFAE paper. The latent size is 50 and both encoder and decoder (just for VFAE) is a two layer fully-connected network with 100 hidden neurons. The classifier is a logistic regression on top of latent space. We trained both models with Adam for 200 epochs with learning rate equals 0.0001. We trained a classifier supervisedly without any fairness constraints. As mentioned in section 4, we need to train our classifier in a post processing manner, so to be fair we did the same for VFAE.

We trained the VFIB model on this dataset and changed the β see if our method can effectively remove the information of the sensitive attribute from latent, in figure below you can find the results for different values of β.

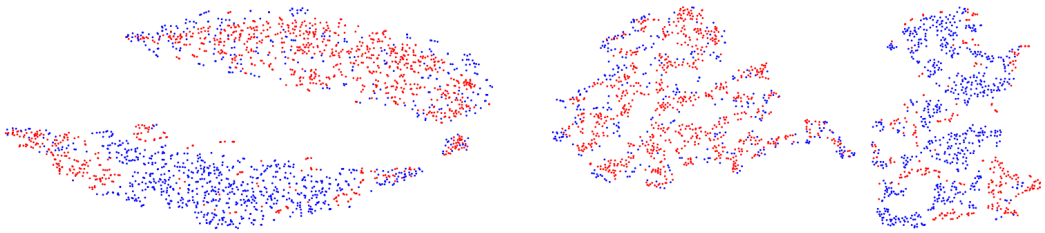
Label / Method	Y	S	Discrimination
VFIB-$\beta = 1$	98.31 ± 0.1	70.16 ± 0.61	0.39 ± 0.006
VFIB-$\beta = 2$	97.17 ± 0.41	69.42 ± 1.1	0.38 ± 0.01
VFIB-$\beta = 5$	94.37 ± 0.26	67.18 ± 0.29	0.31 ± 0.01
VFIB-$\beta = 10$	88.25 ± 0.12	62.33 ± 0.4	0.21 ± 0.007
VFIB-$\beta = 50$	50.03 ± 1.01	50.56 ± 0.5	0.00 ± 0.0002

Table 1: Numerical results of VFIB for different values of β

Label / Method	Y	S	Discrimination
VFAE	99.20 ± 0.1	70.96 ± 0.2	0.39 ± 0.007
VFIB-$\beta = 1$	98.31 ± 0.2	70.16 ± 0.5	0.39 ± 0.006
Supervised	99.44 ± 0.07	99.92 ± 0.08	0.40 ± 0.006

Table 2: Black White MNIST classification accuracies

Moreover we visualized the latent space of both models in figure 4. The points are colored based on the sensitive attribute and as you can see they are blended and separating them based on the sensitive attribute it not easy to achieve.

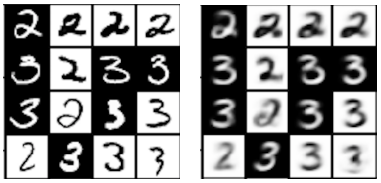


(a) VFIB (b) VFAE

Figure 4: t-SNE visualization of learned representation for 2000 randomly selected test data points (colored based on sensitive attribute)

Furthermore, we trained a decoder to reconstruct the data as a post-processing. You can find the results in figure 5.

(a) Random Data (b) VFIB Reconst (c) Random Data (d) VFAE Reconst



(e) Random Data (f) supervised Reconst

Figure 5: Results of a reconstruction network trained on top of learned representations

5.2 Adult Dataset

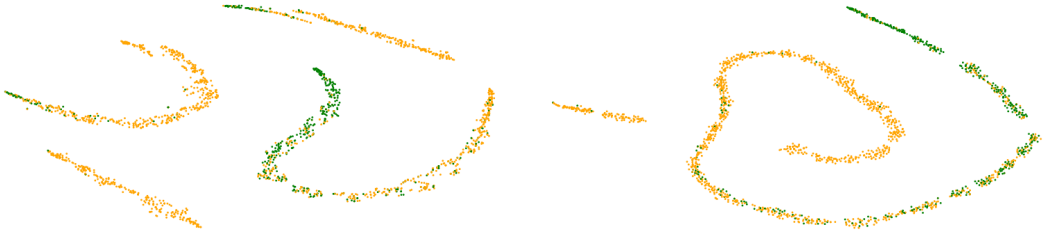
One of the famous datasets for fair classification is an Adult dataset. The Adult income dataset contains 45, 222 entries and describes whether an account holder has over \$50, 000 dollars in their account. The sensitive variable is age. This dataset obtained from the UCI machine learning repository. We binarized the continuous attributes of the dataset based on the mean and selected 30k randomly selected data points for training. We used the same set up of architecture and optimization as section 5.1 which is based on Variational Fair autoencoder paper.

Method \ Label	Y	S	Discrimination
VFAE	84.05 ± 0.07	66.27 ± 0.2	0.2024
VFIB-$\beta = 1$	84.06 ± 0.08	65.15 ± 0.5	0.1956
supervised	84.35 ± 0.08	93.16 ± 0.5	0.2109

Table 3: Adult dataset evaluation for prediction of income

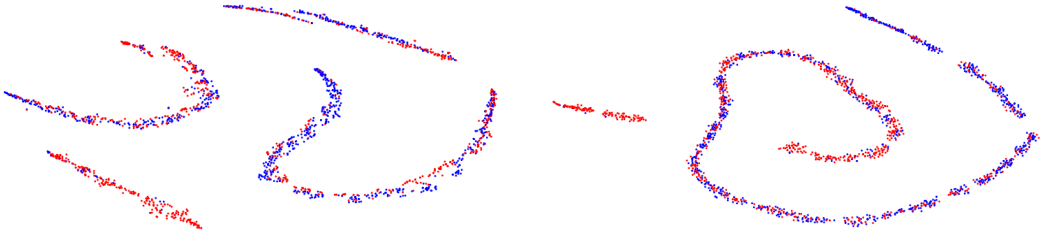
From table 3 we can see that the proposed VFIB can keep the accuracy as high as VFAE, but remove more information about age from the latent which is our goal.

t-SNE visualization of the latent shows that both models could effectively separate the data points based on the label, but the sensitive attribute doesn't cluster the datapoints.



(a) VFIB (b) VFAE

Figure 6: t-SNE visualization of learned representation colored based on the target label



(a) VFIB (b) VFAE

Figure 7: t-SNE visualization of learned representation colored based on sensitive attribute

6 conclusion

We proposed a new fair representation learning framework based on information bottleneck. We created a new toy problem for fair classification and demonstrated competitive performance on both real and synthetic data. The simplicity of our objective can be a primary benefit over similar methods.

7 Future Works

The proposed method is supervised but someone can extend it to semi-supervised cases. Besides, we also used a simple gaussian for modeling the prior but more complex parametric forms can improve the performance of the model.

References

[1] Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. Deep variational information bottleneck. *arXiv preprint arXiv:1612.00410*, 2016.

[2] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[3] Christos Louizos, Kevin Swersky, Yujia Li, Max Welling, and Richard Zemel. The variational fair autoencoder. *arXiv preprint arXiv:1511.00830*, 2015.

[4] Jiaming Song, Pratyusha Kalluri, Aditya Grover, Shengjia Zhao, and Stefano Ermon. Learning controllable fair representations. *arXiv preprint arXiv:1812.04218*, 2018.

[5] Philip Botros and Jakub M Tomczak. Hierarchical vampprior variational fair auto-encoder. *arXiv preprint arXiv:1806.09918*, 2018.