# Variational Fair Information Bottleneck

Sajad Norouzi

University of Toronto, Vector Institute

## Variational Fair Information Bottleneck

We propose an information theoretically motivated objective for learning a fair representations to solve downstream classification tasks.
To formalize the approach we first introduce some notation:

- X: Observation
- S: Sensitive Attribute
- Y: Label of the classification task (target label)
- Z: Representation of data that we want to learn

We believe the objective of fair representation can be written as:

$$\max \ I(Y;Z|S) - \beta I(Z;X,S)$$

Using variational methods we can define a lower bound for the above formulation, and the final objective for a given $x_i, y_i, s_i$ is:

$$\max E_{p_\theta(z|x_i,s_i)}[\log p_\theta(y_i|z,s)] - \beta KL(p_\theta(z|x_i,s_i)||p(z))$$

Due to similarity to Variational Information Bottleneck we named our method Variational Fair Information Bottleneck (VFIB). The general framework for learning a fair representation is shown in the figure below. Our proposed method doesn't have any decoder or adversary which makes it much simpler to train. One of the primary properties of our method is using sensitive attribute for training classifier (shown in pink line)

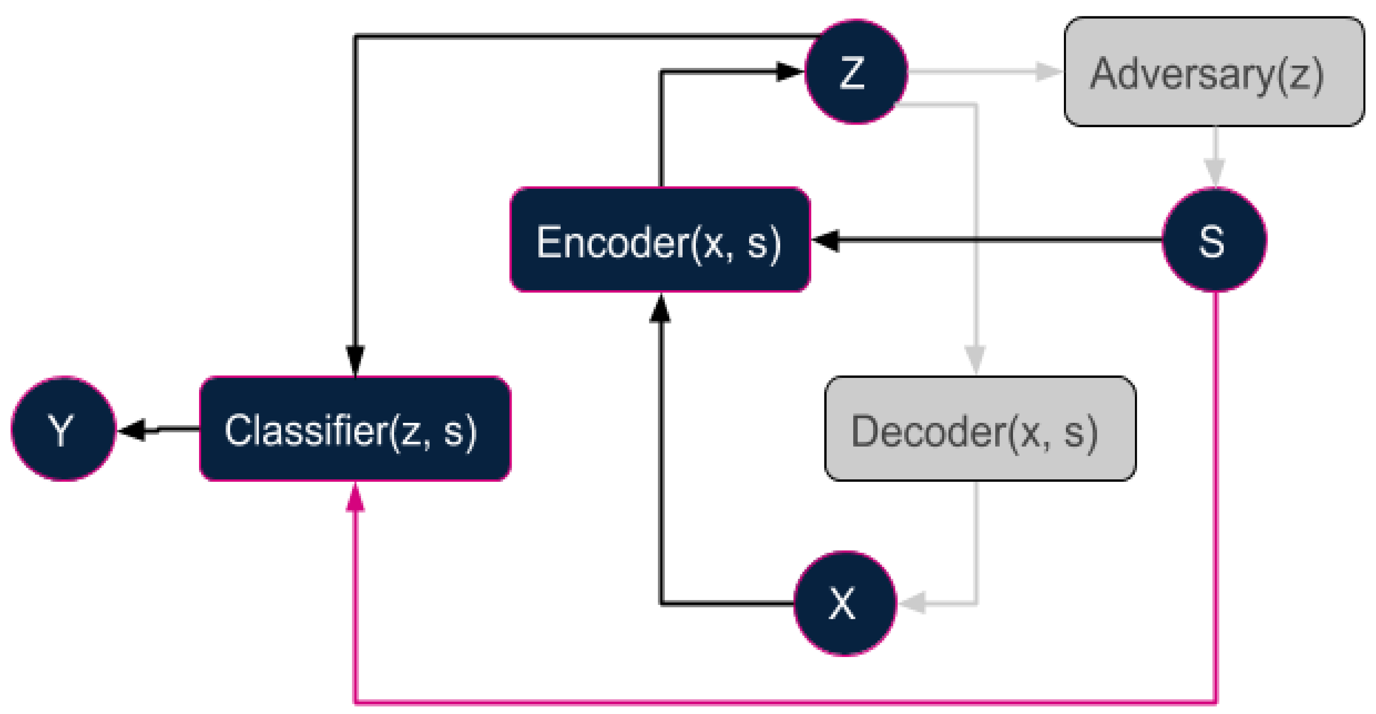



Figure: Diagram of fair representation learning framework, gray components are inactive in our method

## Related Works

The proposed objective has direction resemblances with Unsupervised variant of Variational Fair Autoencoder[1] (VFAE) and Deep Variational Information Bottleneck [2] (VIB).

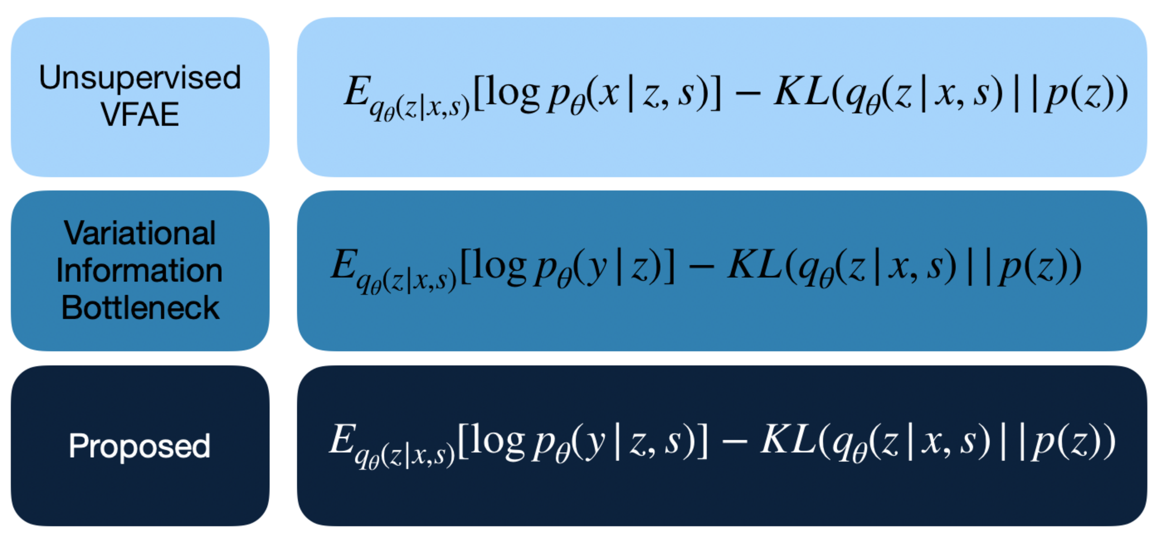



Figure: Objective of the proposed method vs VFAE and VIB

## Toy Problem: Black and White MNIST

We believe Toy problems can be extremely helpful for creation and validation of new ideas. Inspired by MNIST datset, we create a toy problem for fair classification. We considered digits two and three in the MNIST and contrasted the value of each pixel in 70% of images belong to digit two and did the same for 30% of images with label three. You can find samples from dataset in the figure below:

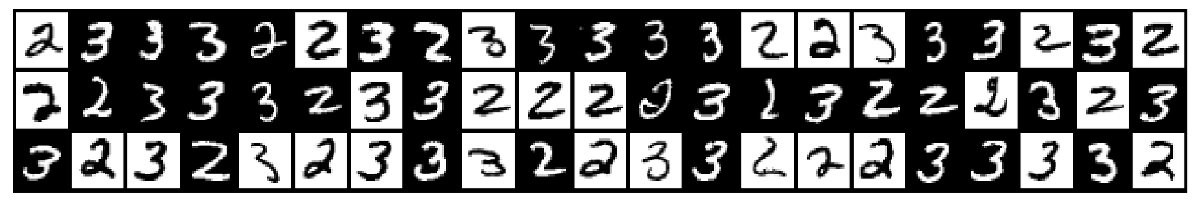


Figure: Samples from created dataset

A new label added to each image which is the color of digit. We considered digit color as a sensitive attribute and tried to solve classification task for image label.

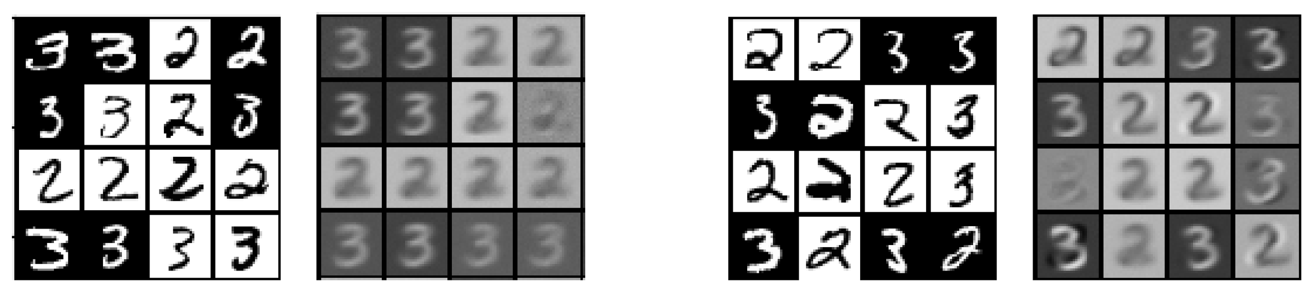
## Results on Toy Problem

To evaluate the proposed method we applied VFIB on the toy problem and compared it with VFAE. After training the encoder, we trained two classifiers for sensitive attribute and the image label. Moreover, we trained a decoder network to reconstruct the original data

| Method \ Label | Y | S |
|---|---|---|
| VFAE | **99.20** $\pm$ 0.1 | 70.96 $\pm$ 0.2 |
| VFIB | 98.31 $\pm$ 0.2 | **70.16** $\pm$ 0.5 |

Table: Black White MNIST classification accuracies

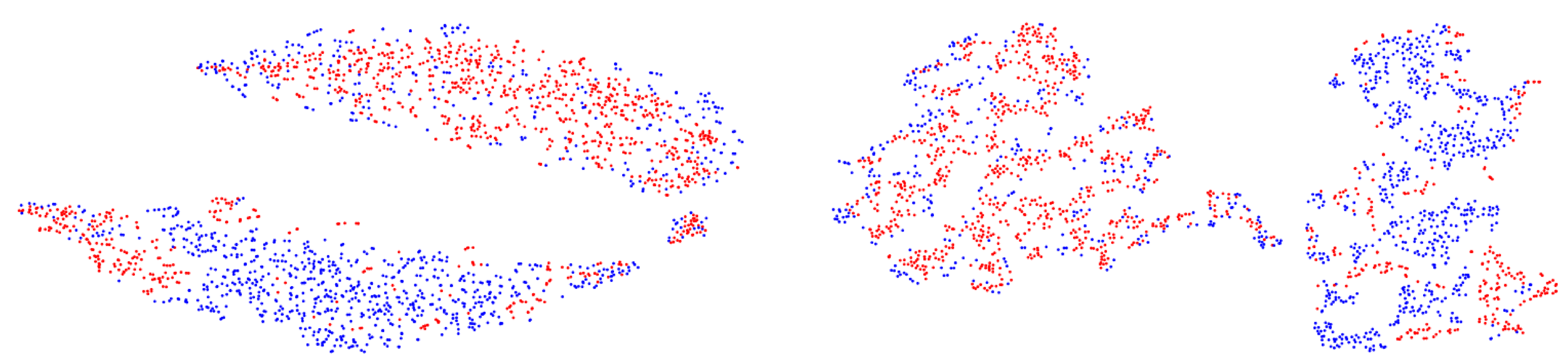
While the accuracy of predicting label for VFAE is higher, its accuracy for sensitive attribute is also higher which is a non-desirable property in fair classification.



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

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

The reconstruction network shows that both methods were able to omit most of the information related to digit color. However, the VFAE kept more information about the styles of the digits which is not related to the current task



(a) VFIB (b) VFAE

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

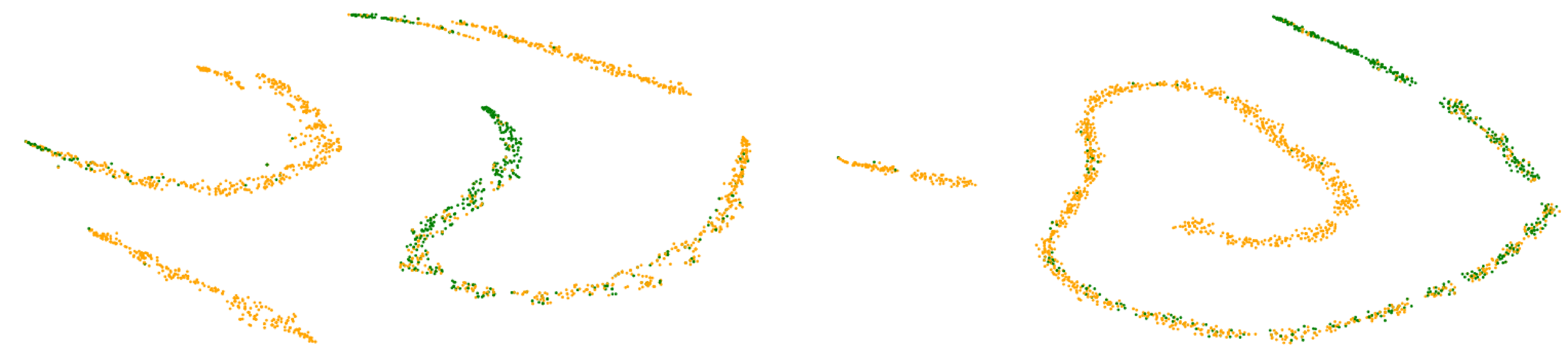
## Results on Adult Dataset

We binarized the adult dataset based on the mean and considered the age as a sensitive attribute. The target label is predicting whether the person has a income above 50k or not.

| Method \ Label | Y | S | Discrimination |
|---|---|---|---|
| VFAE | 84.05 $\pm$ 0.07 | 66.27 $\pm$ 0.2 | 0.2024 |
| VFIB | **84.06** $\pm$ 0.08 | **65.15** $\pm$ 0.5 | **0.1956** |

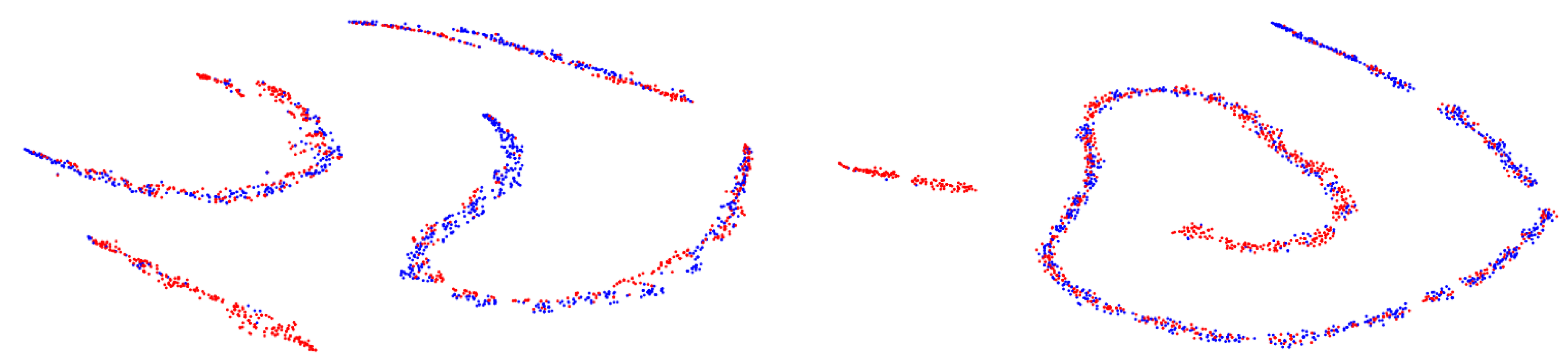
Table: Adult dataset evaluation for prediction of income

The fully supervised non-fair classification accuracy is 84.35. The proposed VFIB can keep the accuracy as high as VFAE, but remove more information about age from the latent which is our goal.



(a) VFIB (b) VFAE

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



(a) VFIB (b) VFAE

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

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.

## Conclusion

We proposed a new framework for learning a fair representation. We showed the effectiveness of the method on both Toy dataset and real datasets.

## Future Works

- Apply this method on more complex datasets
- Use Vamppriorfor modeling the prior
- Extend the framework to semi-supervised problems

## References

(1) Louizos, Swersky, Li, Welling, & Zemel, The Varitional Fair Autoencoder
(2) Alemi, Fischer, Dillon & Kevin Murphy, Deep Variational Information Bottleneck

Code: https://github.com/sajadn/Variational-Fair-Information-Bottleneck