Group Fairness for Learning Representation on Coupled Variational Autoencoder

Fubao Wu
fubaowu@umass.edu

Abstract

Autoencoder is a good way to represent data and could be used for many application inference in classification and regression. Several variations of autoencoder are being proposed recently. The accuracy of the classification with autoencoder is heavily explored, however, the fairness and bias of classification with autoencoder representation is still an ongoing research. In this paper, we explore the recently proposed autoencoder "coupled variational autoencoder" which focuses on improving the accuracy and robustness of the probabilistic inference on the represented data. We explore the group fairness of the learned representation for classification. We consider several different sensitive attributes by the defined group fairness, and compare the difference with the general variational autoencoder. It shows this variational autoencoder has different degrees of biases with different sensitive attributes and most of them have good indications for fairness against gender on small credit card datasets, but not good on the fairness against race on the UTKFace dataset.

1 Introduction

Autoencoder [17, 23], as a specific type of neural network to reconstruct the input, has important applications in unsupervised learning to be the "informative" representation of data. Thus the representation is also effectively used for classification and regression. The accuracy on the learned representation is highly explored, but the fairness is still an ongoing research [6]. The bias/discrimination in the prediction tasks could be from the biased data or the algorithm. In this project, we explore the unexplored fairness property of a recently proposed machine learning model-coupled variational autoencoder. It was proposed by Cao et.al [8] in 2019, and claimed to improve the robust accuracy of the probabilistic inferences on represented data. The coupled variational autoencoder is modified with coupled cross-entropy and coupled KL divergence for loss function based on the general variational autoencoder. We investigate the group fairness property with regards to several sensitive attribute pairs.

The group fairness is defined as follows. Given label examples $(x, a, y) \sim p_{data}$ where $y \in Y$ are the labels, we want to predict. x and a are attributes where a is sensitive attributes, and x is non-sensitive features.

A classifier has a predictive $y = f(x, a)$ (or $y = f(x)$) that achieves a certain group fairness criteria with regard to sensitive attributes a. If a prediction $\hat{y}$ is completely independent of the sensitive attributes $\hat{y} \perp a$, we think the it achieved demographic parity fairness in general. If it is not complete independent, then we think it leads to certain unfairness. Therefore, we define a metric called demographic parity distance as the fairness distance to measure how the model is fair

$$D_{dp} = |\mathbb{E}[\bar{y} = 1|a = 1] - \mathbb{E}[\bar{y} = 1|a = 0]| \tag{1}$$

The smaller D_{dp}, the more fairness it is to the sensitive a.

Preprint. Under review.

We investigate how the learned representation for coupled variational autoencoder is fair to multiple attributes for classification based on the the definition of D_{dp}. The experiments are done on a small default of credit card, a German credit risk, and one UTKFace image dataset and we compare the fairness with the general variational autoencoder as the baseline model. The experimental results show the coupled variational model has very good fairness on first default of credit card dataset against gender, education level, but not good against the gender, job skill level, and race attributes on the German credit risk and UTKFace image dataset. It has also similar fairness on the gender and sensitive model around the two credit card dataset compared to the general variational autoencoder model, but performs worse than on the image dataset.

2 Novelty Statement

The recently proposed representation algorithm-coupled variational autoencoder [8] was modified from the general variational encoder. It is claimed to be robust and high accuracy, but the bias and fairness properties are not explored. We investigate the new property of group fairness on multiple sensitive attributes. We experimentally evaluate the group fairness on three datasets based on the definition of group fairness, and compare with the general variational autoencoder.

3 Related work

Machine learning algorithms have been widely applied in almost every circle of our life in healthcare, financial aid, insurance, online shopping, advertising etc. One of the problem in machine learning is the machine, like humans, are vulnerable to show bias or unfairness hidden from the results. The decision of a machine learning algorithm tends to be skewed or discriminated towards a certain individual or groups of people [13, 6, 19]. There are commonly two types of sources of bias in the machine learning area. The first one is the bias caused by the biased data [20]. The data especially big data are often generated by groups which tend to have their own characteristics or behaviors. The machine learning algorithms learned on these data might lead to unfair and biased prediction [7]. The second type is the algorithm bias itself. Some of of AI and machine learning algorithm that have been shown in lots of research that shows bias towards individual or certain groups in face recognition applications, and search engines [12].

For the fairness of machine learning algorithms [3], there are further three categories involved. (1) Pre-processing: some research focuses on try to process the data to remove the hidden discrimination [10, 11]. (2) In-processing: some research investigates to modify the machine learning algorithm to remove the discrimination during the training process [4, 15]. (3) Post-processing: this type of research treats the trained model a back box without modifying the training data or learning algorithms, and try to investigate the fairness property and further reassign the labels to improve the fairness [5].

Our project is related to the post-processing research which focuses on a state-of-the-art algorithm–coupled variational autoencoder, investigating the unexplored bias and fairness property. Out of the machine learning domain, autoencoder fairness has been explored in the following researches. Louizos et.al [16] proposes a fair variational encoder by removing the dependence of sensitive and latent factors of variation by corporating an additional penalty term-"Maximum Mean Discrepancy". The authors in [1] fused the original learning task with a variational autoencoder to learn the latent structure within the dataset and then adaptively uses the learned latent distributions to re-weight the importance of certain data points while training. Creager et.al [9] proposed an algorithm for learning compact representation and claim the flexibly fair representation with respect to multiple sensitive attributes. Our work is most close to this one and we investigate the recently proposed robust coupled variational autoencoder and explore the fairness of representation learning with multiple sensitive attributes.

4 Methodology

In this project, we investigate the novel property–bias and fairness of a recently proposed machine learning algorithm-coupled variational autoencoder. We study the group fairness demographic parity with respect to multiple sensitive attributes.

4.1 Learning Representation with Coupled Variation Autoencoder

In this section, we simply introduce the recently proposed coupled variational autoencoder and the learning representation for the classification tasks used for evaluating the fairness.

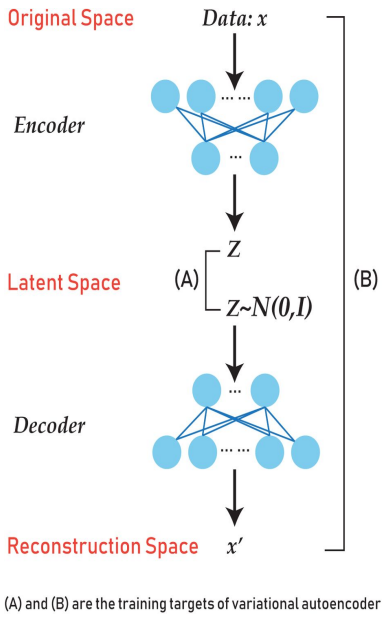


Figure 1: Varitional autoencoder model

The coupled VAE model incorporates the positive coupling for the cross-entropy and divergence costs of the variational autoencoder which improves the learning of a robust inference model. It is based on the variational encoder [14]. The basic structure of variational encoder is shown in Figure 1. The whole process in (B) in this figure consists of an encoder, a decoder and a loss function. The input data x in the original space is input into an encoder, which is usually a neural network that converts x into a low dimensional latent space z, then the latent space is sampled usually with a Gaussian distribution [z], and input again into the decoder. The decoder is usually a neural-network model which decode the z into the reconstructed space x'.

The objective loss function is to minimize the loss in the encoding and decoding process. It is measured by the log likelihood $logP_\phi(x|z)$, which is input data given the model and decoder parameters. It is defined as follows.

$$L(x^i) = -D_{KL}(q(z|x^{(i)})||p(z)) + \mathbb{E}_{q(z|x^{(i)})}[logp(x^{(i)}|z)] \tag{2}$$

The loss is established with the negative Kullback-Leibler divergence loss between the variational approximation q and the intractable posterior p plus the expected reconstruction error. After simplifying with the the prior and posterior distribution of z with a Gaussian distribution, we get:

$$L(x^i) = -D_{KL}(q(z|x^{(i)})||p(z)) + \frac{1}{L}\sum_{l=1}^{L}(logp(x^{(i)}|z^{(i,l)})) \tag{3}$$

In the proposed coupled variational autoencoder, the loss function is modified by coupled generalizations of the KL-divergence and cross-entropy to improve the robustness of the VAE model. The coupled entropy derives from the Tsallis entropy which utilizes a modified transformation. The nonlinear statistical coupling (or simply the coupling) has been shown to quantify the relative variance of a superstatistics model in which the variance of exponential distribution fluctuates according to a gamma distribution, and is equal to the inverse of the degree of freedom of the Student's t distribution. The coupling is related to the risk bias by the expression $r = \frac{-2k}{1+k}$. It uses a generalized mean $(\sum p_i^{1-\frac{2k}{1+k}})^{-\frac{1+k}{2k}}$ to model the long range correlations between the states. The mathematical form of

coupled entropy function [18] with power $\alpha = 2$ and coupling k is defined as

$$S_k(p) = \frac{1}{2} log_k((\sum p_i^{1+\frac{2k}{1+k}})^{\frac{-1}{k}}) = \frac{1}{k}((\sum p_i^{\frac{1+3k}{1+k}})^{-1} - 1) \tag{4}$$

where $ln_k(x) = \frac{1}{k}(x^k - 1)$ is the generalization of the logarithm function, known as the coupled logarithm function. Therefore, the modified loss function contains two terms: negative coupled divergence and coupled cross-entropy. Coupled divergence is the generalization of KL divergence in equation

$$D_k p(z)||q(z) \equiv \prod_{i=1}^{D1} = \int_{-\infty}^{\infty} \frac{p(z_i)^{1+\frac{2k}{1+k}}}{p(z_i)^{1+\frac{2k}{1+k}} dz_i} \frac{1}{2}(log_k(q(z_i) - log_k(p(z_i)^{-\frac{2}{1+k}}))dz_i \tag{5}$$

Where D1 is the dimensionality of z. Coupled cross-entropy is the generalization of crossentropy term, which is defined as

$$H_k(x) = \sum_{i=1}^{D2} x_i \frac{1}{2} log_k(y_i)^{\frac{2}{1+k}} - (1 - x_i)\frac{1}{2} log_k((1 - y_i)^{\frac{2}{1+k}}) \tag{6}$$

Where D2 is the dimensionality of z. Then the new loss function is the coupled loss function as

$$L(x^i) = -D_k(p(z)||q(z)) + \frac{1}{L}\sum_{l=1}^{L} H_k^l \tag{7}$$

4.2 Fairness Evaluation of Learned Representation with Coupled Variational Autoencoder

The group fairness is already defined in the Section 1. We explore the demographic parity distance D_{dp} for classification tasks with the learned representation data of coupled variational autoencoder.

(1) We investigate the group fairness of the learned representation of coupled variational autoencoder with multiple sensitive attributes. Specifically, we examine (a) whether the representation learned from the model allows users to easily adapt the representation to a variety of fair classification settings, where a task may have a different task label y and sensitive attributes. (b) whether the representation learning from the model can be fair with respect to pairs of multiple sensitive attributes (e.g. if a classifier is fair to women but not fair for men).

(2)We randomly split a dataset into audit dataset and test dataset. The audit dataset will be randomly split into training set and validation set and make them balanced. We train the coupled variational autoencoder on the training dataset and find the best learning representation model on the rest validation dataset with grid search of 6 different latent dimensions.

(3) We evaluate the models with the test dataset. For the audit dataset, we remove and does not remove the sensitive dimensions of attributes to train the model, respectively. We show the prediction performance of a multilayer perceptron (MLP) classifier trained on the encoded audit dataset for fairness evaluation, and test the demographic parity distance D_{dp} with different sensitive attributes on test dataset.

(4) We compare the group fairness of coupled variational autoencoder model with the general variational autoencoder model and show the results in the next Section 5.

5 Experimental Evaluation

We show the datasets used for autoencoder learning, the group fairness evaluation of the coupled variational autoencoder (CVAE), the prediction performance, and compare the fairness with the general variational autoencoder (VAE).

5.1 Data Sets

We use three public datasets effectively for the fairness evaluation with multiple sensitive attributes. The dataset statistics inforamtion are shown in Table 1.

Table 1: Dataset statistics information

Dataset	Instance number	Total attributes	Sensitive attributes
Default of Credit Card	30,000	24	2
German Credit Risk	1,000	10	2
UTK Faces	10,000	3 + image pixels (200x200)	2

Default of Credit Card Dataset: The default of credit card dataset [21] contains information on default payments, demographic factors, credit data, history of payment, and bill statements of credit card clients in Taiwan from April 2005 to September 2005. The attributes include the given credit limit, gender, education marital status, age, statements and default payment next month. We predict an individual has default payment next month or not and consider sensitive attributes gender and education level here.

German Credit Risk Dataset: This German credit dataset [2] contains 1000 credit records of indindividuals. The attributes includes personal status, sex, credit score, credit amount, housing status, etc. We use the learned representation to classify the individual record of credit good or bad. The sensitive features we consider are the gender or job skill level.

UTK Faces Dataset: This is a larger dataset of faces [22]. The diversity in face dataset is designed especially for fairness research in faces. It contains 21k annoted images. There are age, gender, race and other attributes of each face image. For fairness evaluation, we consider the sensitive attribute of race or gender given the other attributes for the classification. Due to the computing resources and time limit, here we use 10k images instances for the training and test in our experiments.

Some of the sensitive attributes distribution are shown in Figure 2. It shows the different distributions of attributes in each of three datasets, including gender, education level, job skill level, and race. The data instances are not imbalanced according to different sensitive types.

5.2 Evaluation Procedure

- We randomly divide the dataset into audit dataset for learning the autoencoder, and test dataset (for evaluating the autoencoder).
- We optimize the autoencoder model with grid search to find the best encoder to learn the representation on the audit dataset which randomly split into the training-validation dataset.
- We evaluate the group fairness of the learned autoencoder. We freeze the encoder weights and train a classifier to predict some task labels given the output on the test dataset. We use a multilayer perceptron (MLP) classifier to train the classifier with cross validation of audit dataset. Then the group fairness D_{dp} against different sensitive attribute pairs are evaluated on the test dataset.

We randomly split of datasets into audit dataset and test dataset to do the experiments 10 times and show the average results in the later sections.

5.3 Validation Loss for Learning Coupled Variational Autoencoder

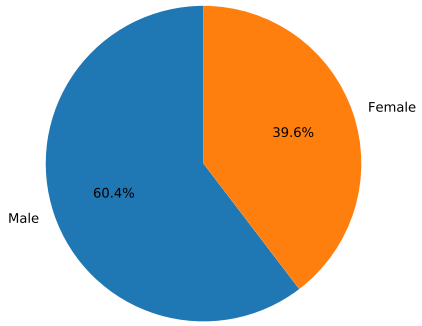
Autoencoder learns the representation z of latent dimension from the input data x. According to different input data sizes, we learn different dimensions of z with grid search on the validation dataset. We then select the best latent dimension z of encoder for representation learning on each dataset. If the input data size is D, then the latent dimension range is tried at $[\frac{D}{30}, \frac{D}{20}, \frac{D}{10}, \frac{D}{5}, \frac{D}{4}, \frac{D}{2}]$.

We show the validation loss result of learning representation with CVAE in Figure 3 (a), (b) and (c) for the three datasets.

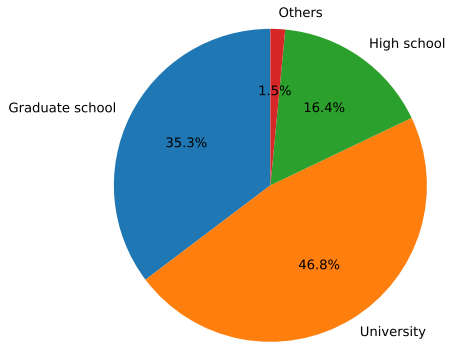
5.4 Classification Performance

In our training of MLP classification, we sampled balanced data instances of different classes to train the classification of each goal in the three datasets.

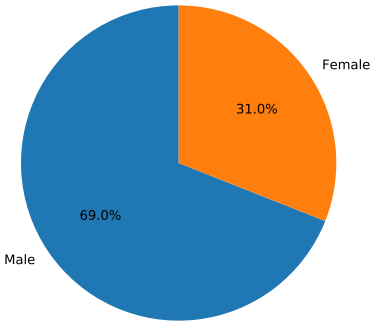
The classification performance of precision, recall and F1-score in three datasets with different prediction goals are shown in Table 2. In the UTKFace, the prediction goal is to predict gender or



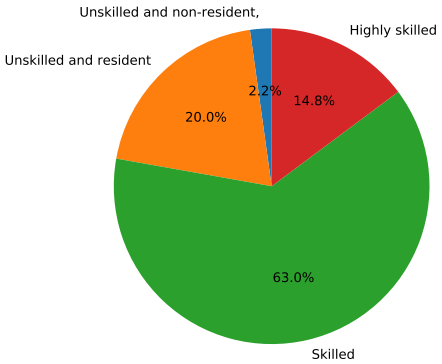
(a) Gender distribution on Default of Credit Card



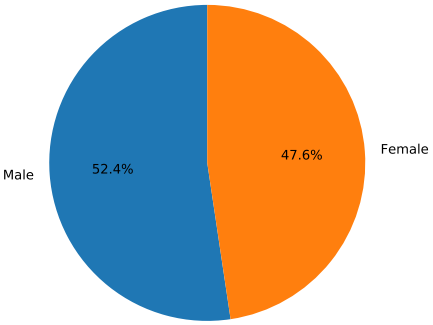
(b) Education distribution on Default of credit card



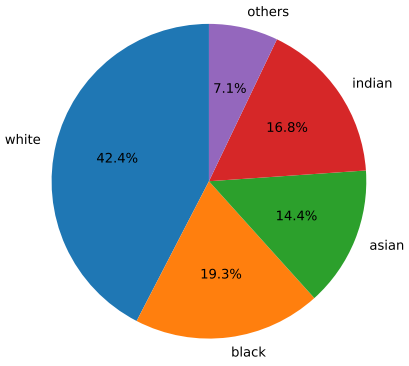
(c) Gender distribution on German Credit Risk



(d) Job distribution on German credit risk

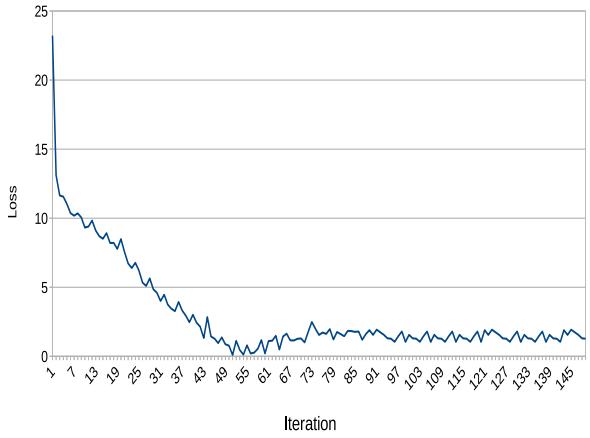


(e) Gender distribution on UTKFace

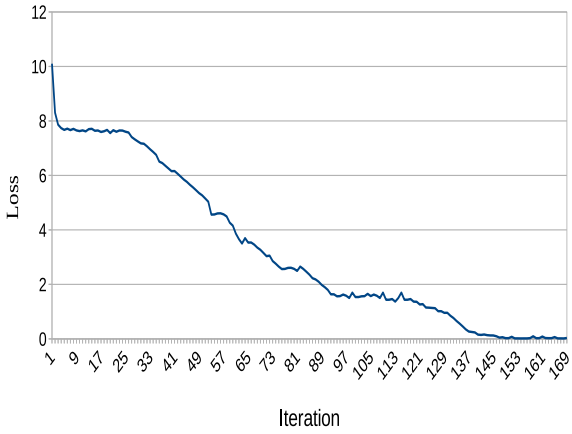


(f) Race distribution in UTKFace

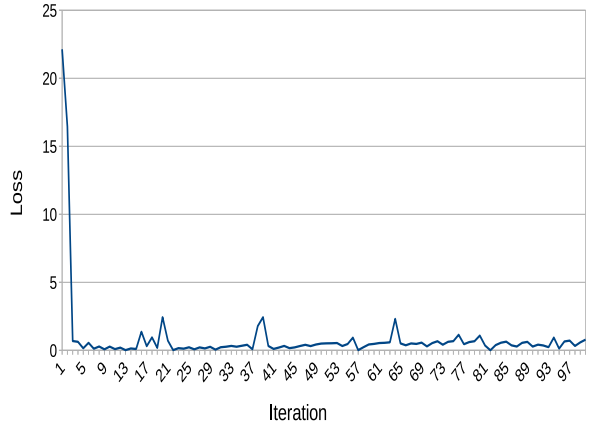
Figure 2: Examples of attribute distributions on the three datasets



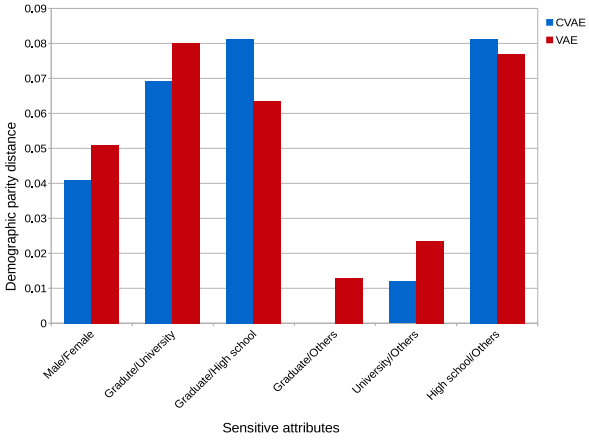
(a) CVAE Validation loss on Default of Credit Card



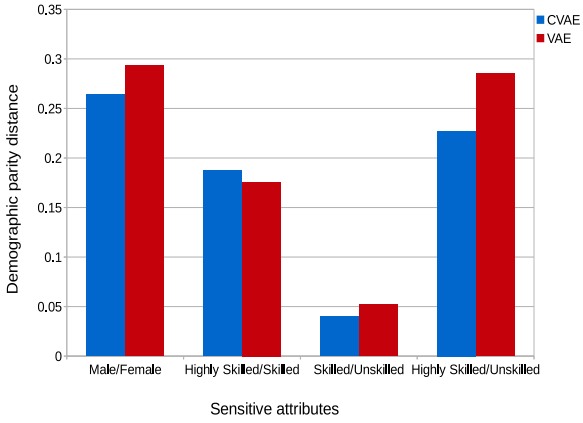
(b) CVAE Validation loss on German Credit Risk



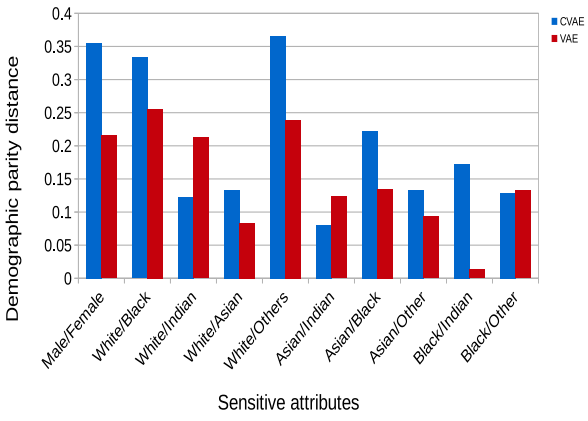
(c) CVAE Validation loss on UTK Faces



(d) Parity differences of the sensitive attributes on Default of Credit Card



(e) Parity differences of the sensitive attributes on German Credit Risk



(f) Parity differences of the sensitive attributes on UTK Faces

Figure 3: (a)–(c) Validation loss of the CVAE learning representation, (d)–(f) The demographic parity distances of different sensitive attributes on the three datasets

race based on other attributes, so there are two types of classification involved. We can see that, the prediction F1-score is around 0.68–0.84. Probably we use MLP classification, which is not a simple model, so the F1-score is not high. However, it does not affect the fairness evaluation of the sensitive attributes.

Table 2: Classification performance

Datasets prediction	**Precision**	**Recall**	**F1**
Predict Default of Credit Card	0.5213	1.0000	0.6853
Predict German Credit Risk	0.7639	0.9333	0.8401
Predict UTKFace gender	0.7103	0.8800	0.7861
Predict UTKFace race	0.5533	0.8667	0.6754

5.5 Group Fairness Evaluation of Autoencoder

As we said, we use MLP classification result and calculate the demographic parity differences to evaluate the group fairness of Coupled Variational Autoencoder (CVAE).

We also run the same procedure to get the best learning representation model for the baseline-VAE, and compare the fairness with CVAE.

Here we consider different pairs of sensitive attributes for each different datasets. The group of the attributes considered are among genders, education level, job skill level, and race in these datasets.

Evaluation on the Default of Credit Card In this dataset, we predict a person has default of next month payment or not. We evaluate the demographic parity given the male and female, and the different education level. Figure 3 (d) shows the demographic parity distance Δ_{dp} on the Default of Credit Card dataset. It shows the CVAE model has good result for fairness on the sensitive attributive with the demographic parity difference values around 0.01–0.08. Compared with VAE model, it has good fairness on gender and most of the sensitive attributes of education levels.

Evaluation on the German Credit Risk In this dataset, we predict a person to have credit risk or not according to the person information and the credit record. We evaluate the demographic parity given the male and female, and the different job skill level. Figure 3 (e) shows the demographic parity distance Δ_{dp} on the German Credit Risk dataset. It shows the fairness with demographic parity difference values is around 0.05-0.3, which is much larger than that on the Default of Credit Card dataset. Similarly to the Default of Credit Card dataset, the CVAE performs a little bit better than VAE for these attritubes on the demographic parity distance metric for most of the attribute pairs.

Evaluation on UTKFace In this dataset, there are two types of classifications are applied. First is to predict a person is female or male given the face information and other attributes in the dataset. The second is to predict a person's different races. We evaluate the demographic parity difference given the male and female, and the different education level. Figure 3 (f) shows the demographic parity distance D_{dp} on the UTKFace dataset. It shows the fairness evaluation with D_{dp} values around 0.13-0.36. However, the CVAE model performs worse than the VAE model on most of the sensitive attributes pairs based on the demographic parity distance measurement. This is probably the robust loss function of learning representation for capture the high accuracy which drives the bias to the race classification.

6 Conclusion

In this project, we explore the group fairness of the recently proposed coupled variational autoencoder and compare it with the general variational autoencoder. It shows the coupled variational autoencoder has good fairness on the gender, education level, but not as good as the general variational autoencoder in other attributes especially gender and race attributes in the evaluated UTKFace dataset according to the demographic parity difference metric. It indicates the proposed coupled variational autoencoder has been claimed robust and accuracy, but the fairness is not still a problem and needs to be considered and improved. In the future, we could optimize the autoencoder loss function against sensitive attribute types to reduce the bias. We could explore other types of definition of fairness, such as equal opportunity, fair subgroup accuracy and further investigate the fairness of this autoencoder model.

References

[1] Alexander Amini, Ava P Soleimany, Wilko Schwarting, Sangeeta N Bhatia, and Daniela Rus. Uncovering and mitigating algorithmic bias through learned latent structure. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 289–295, 2019.

[2] Arthur Asuncion and David Newman. Uci machine learning repository, 2007.

[3] Solon Barocas, Moritz Hardt, and Arvind Narayanan. Fairness in machine learning. *NIPS Tutorial*, 1, 2017.

[4] Rachel KE Bellamy, Kuntal Dey, Michael Hind, Samuel C Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, et al. Ai fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *IBM Journal of Research and Development*, Vol 63, 2018.

[5] Richard Berk, Hoda Heidari, Shahin Jabbari, Matthew Joseph, Michael Kearns, Jamie Morgenstern, Seth Neel, and Aaron Roth. A convex framework for fair regression. *ACM Conference on Fairness, Accountability, and Transparency, ACM FAccT*, 2017.

[6] Reuben Binns. Fairness in machine learning: Lessons from political philosophy. In *Conference on Fairness, Accountability and Transparency*, pages 149–159, 2018.

[7] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91, 2018.

[8] Shichen Cao, Jingjing Li, Kenric P Nelson, and Mark A Kon. Coupled vae: Improved accuracy and robustness of a variational autoencoder. *arXiv preprint arXiv:1906.00536*, 2019.

[9] Elliot Creager, David Madras, Jörn-Henrik Jacobsen, Marissa A Weis, Kevin Swersky, Toniann Pitassi, and Richard Zemel. Flexibly fair representation learning by disentanglement. *Proceedings of Machine Learning Research*, 2019.

[10] Brian d'Alessandro, Cathy O'Neil, and Tom LaGatta. Conscientious classification: A data scientist's guide to discrimination-aware classification. *Big data*, 5(2):120–134, 2017.

[11] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *International Conference on Learning Representations, ICLR*, 2015.

[12] Ayanna Howard and Jason Borenstein. The ugly truth about ourselves and our robot creations: the problem of bias and social inequity. *Science and engineering ethics*, 24(5):1521–1536, 2018.

[13] Apurv Jain. Weapons of math destruction: how big data increases inequality and threatens democracy, 2017.

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

[15] Preethi Lahoti, Alex Beutel, Jilin Chen, Kang Lee, Flavien Prost, Nithum Thain, Xuezhi Wang, and Ed Chi. Fairness without demographics through adversarially reweighted learning. *Advances in Neural Information Processing Systems, NeurIPS*, 33, 2020.

[16] Christos Louizos, Kevin Swersky, Yujia Li, Max Welling, and Richard Zemel. The variational fair autoencoder. *International Conference on Learning Representations, ICLR*, 2016.

[17] Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, Ian Goodfellow, and Brendan Frey. Adversarial autoencoders. *International Conference on Learning Representations, ICLR*, 2016.

[18] Kenric P Nelson, Sabir R Umarov, and Mark A Kon. On the average uncertainty for systems with nonlinear coupling. *Physica A: Statistical Mechanics and its Applications*, 468:30–43, 2017.

[19] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. On fairness and calibration. *Advances in Neural Information Processing Systems, NeurIPS*, 30:5680–5689, 2017.

[20] Antonio Torralba and Alexei A Efros. Unbiased look at dataset bias. In *CVPR 2011*, pages 1521–1528. IEEE, 2011.

[21] I-Cheng Yeh and Che-hui Lien. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems with Applications*, 36(2):2473–2480, 2009.

[22] Song Yang Zhang, Zhifei and Hairong Qi. Age progression/regression by conditional adversarial autoencoder. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017.

[23] Junbo Zhao, Yoon Kim, Kelly Zhang, Alexander Rush, and Yann LeCun. Adversarially regularized autoencoders. In *International conference on machine learning*, pages 5902–5911. PMLR, 2018.