Fairness by Learning Orthogonal Disentangled Representations (supplementary material)

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

1 Results in numbers

Table 1: Results on all datasets. Bold face numbers on target task are the highest. Bold numbers in sensitive task are the closest to the majority label classifier.

	CIFAR-10		CIFAR-100		YaleB		Adult		German	
	T	S	T	S	T	S	T	S	T	S
Majority	0.6000	0.1000	0.0500	0.0100	0.0260	0.5056	0.7500	0.6700	0.7100	0.6900
Baseline	0.9775	0.2344	**0.7199**	0.3069	0.7800	0.9600	0.8500	0.8400	0.8700	0.8000
LFR [7]	-	-	-	-	-	-	0.8230	**0.6700**	0.7230	0.8050
VAE [3]	**-**	-	-	-	-	-	0.8190	0.6600	0.7250	0.7950
VFAE [4]	-	-	-	-	0.8500	0.5700	0.8130	**0.6700**	0.7270	0.7970
Xie et al. [6] (trade-off #1)	0.9752	0.2083	0.7132	0.1543	0.8900	0.5700	0.8440	0.6770	0.7440	0.8020
Roy et al. [5] (trade-off #1)	**0.9778**	0.2344	0.7117	0.1688	0.8900	0.4000	0.8460	0.6550	**0.8633**	0.7270
Xie et al. [6] (trade-off #2)	0.9735	0.2064	0.704	0.1484	-	-	-	-	-	-
Roy et al. [5] (trade-off #2)	0.9679	0.2114	0.705	0.1643	-	-	-	-	-	-
Ours	0.9725	**0.1907**	0.7074	**0.1447**	**0.8923**	**0.5292**	**0.8520**	0.6826	0.7700	**0.7100**

2 Biased categories accuracy

Biased categories accuracy reported in [6] which represents the target accuracy of the biased sensitive group. It is used as an indicator of how well the model performs on minority classes. The closer the biased accuracy to the overall accuracy, the better the model is in terms of fairness. The results are shown in Table. 2

3 Implementation details

Extended YaleB dataset: For the Extended YaleB dataset, we use an experimental setup similar to Xie *et al.* [6] and Louizos *et al.* [4] by using the same

Table 2: Biased categories accuracy on three datasets

	Entropy + KL Orth.		**w/o Entropy w/o KL**	
Dataset	**Overall**	**Biased**	**Overall**	**Biased**
Adult	0.8520	0.8157	0.8494	0.8108
German	0.77	0.7241	0.7433	0.6897
Yale	0.8923	0.5867	0.8650	0.4933

train/test split strategy. We used $38 \times 5 = 190$ samples for training and 1096 for testing. The model setup is similar to [6, 5], 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}. For the extended YaleB dataset we set $\lambda_{OD} = 0.037$, $\lambda_E = 1.0$, $\gamma_{OD} = 1.1$, $\gamma_E = 2.0$, $d_T = d_S = 100$.

CIFAR datasets: Similar to [5], we employed ResNet-18 [1] 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 [2] is used in all experiments. For the CIFAR-10 dataset we set $\lambda_{OD} = 0.063$, $\lambda_E = 1.0$, $\gamma_{OD} = 1.7$, $\gamma_E = 1$, $d_T = d_S = 128$. For the CIFAR-100 dataset we set $\lambda_{OD} = 0.0325$, $\lambda_E = 0.1$, $\gamma_{OD} = 1.2$, $\gamma_E = 1.67$, $d_T = d_S = 128$.

Tabular datasets: For the Adult and German datasets, we follow the setup appeared in [5] 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 Adult dataset we set $\lambda_{OD} = 0.037$, $\lambda_E = 0.55$, $\gamma_{OD} = 0.8$, $\gamma_E = 1.66$, $d_T = d_S = 2$. For the German dataset we set $\lambda_{OD} = 0.01$, $\lambda_E = 1.0$, $\gamma_{OD} = 1.4$, $\gamma_E = 2.0$, $d_T = d_S = 2$.

4 Sensitivity analysis heatmaps

References

1. 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)
2. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)

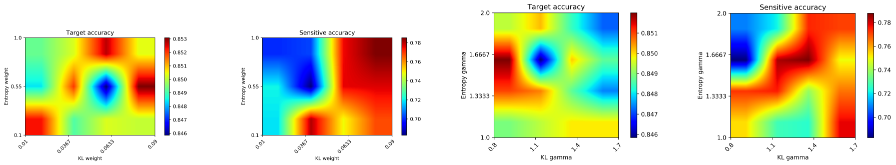

Fig. 1: Sensitivity analysis on the Adult dataset

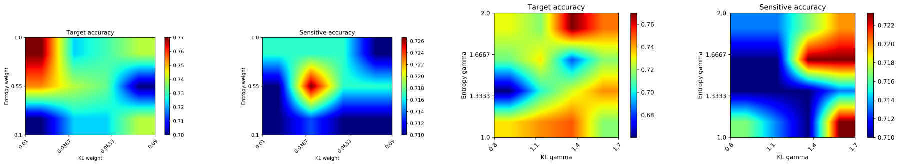

Fig. 2: Sensitivity analysis on the German dataset

3. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
4. Louizos, C., Swersky, K., Li, Y., Welling, M., Zemel, R.: The variational fair autoencoder. arXiv preprint arXiv:1511.00830 (2015)
5. 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)
6. 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)
7. Zemel, R., Wu, Y., Swersky, K., Pitassi, T., Dwork, C.: Learning fair representations. In: International Conference on Machine Learning. pp. 325–333 (2013)

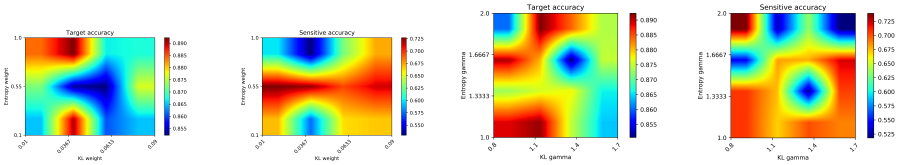

Fig. 3: Sensitivity analysis on the extended YaleB dataset