Supplementary Material to Learning Unbiased Representations via Mutual Information Backpropagation

Ruggero Ragonesi[1,2], Riccardo Volpi[3], Jacopo Cavazza[1], Vittorio Murino[1,4,5]

[1] PAVIS Department, Istituto Italiano di Tecnologia, Genova, Italy
[2] DITEN Department, University of Genova, Italy
[3] Naver Labs Europe, Grenoble, France
[4] Department of Computer Science, University of Verona, Italy
[5] Huawei Technologies Ltd., Ireland Research Center, Dublin, Ireland
name.lastname@iit.it

1. Implementation Details

In the following paragraphs, we provide the implementation details. We carried out all of our experiments using TensorFlow [1]. Concerning the architectures used, please refer to Figure 2.

To ease the discussion, we can divide the optimization problem presented in our work into the following two

$$\min_{\theta,\psi} \mathcal{L}_{task} + \lambda \mathcal{L}_{ne} \tag{1}$$

$$\max_{\phi} \mathcal{L}_{ne} \tag{2}$$

where the learning rates associated to (1) and (2) are α and η, respectively. We use the same notation of Algorithm 1.

Digit experiment. We train our models for 150 epochs, using mini-batches of size 1024. The learning rates α and η are both set to 10^{-4}. We use Adam [2] as optimizer for (1) and (2). For each gradient update to optimize (1) with respect to θ, ψ, we update MINE parameters 80 times ($K = 80$). That is, we perform 80 update steps to optimize (2), as to better train MINE (see Section 2 for a detailed discussion around this choice).

IMDB experiment. For both training splits (EB1 and EB2) we restrict the training set to 2000 samples (for each run we sampled different random images). This choice is motivated by the fact that using the whole training sets we observed higher baselines results then the ones published in previous art [1]. We trained each model for 6 epochs with mini-batch size set to 24. The learning rate α is set to 10^{-5}; the learning rate η is set to 10^{-1}. We use Adam [2] as optimizer for (1) and vanilla gradient descent for (2). We found a number $K = 20$ of MINE iterations to be sufficient in order to estimate the mutual information throughout training.

German experiment. We adopted the same settings as previous art that uses this benchmark [4]. The $1,000$ data samples available are split in 70% training and 30% test (randomly picked in each run). The model is trained for 500 epochs with mini-batch size set to 64. The learning rate α is set to 10^{-5}; the learning rate η is set to 10^{-1}. We use Adam [2] as optimizer for (1), and vanilla gradient descent for (2). We set to a number of MINE iterations $K = 30$.

Adult experiment. We adopted the same settings as Madras et al.[3]. The model is trained for $1,000$ epochs with mini-batch size set to 64. The learning rates α and η are set to 10^{-3}. We use Adam [2] as optimizer for both (1) and (2). We set to a number of MINE iterations $K = 30$.

2. Discussion on the Hyper-Parameters

In this section, we discuss the hyper-parameters that we adopted throughout the experiments reported in this work.

Choice of the number of iterations to update MINE. We found that increasing the number of iterations to estimate $I(Z, C)$ stabilizes the overall training procedure, as shown in Figure 3. As our intuition behind this fact, we posit that the better the estimation of the mutual information through MINE is, the more precise and effective the gradients $\nabla_\theta \mathcal{L}_{ne}$ are. The only drawback we observed is the increased computational cost, since the time increases

[1] https://www.tensorflow.org/

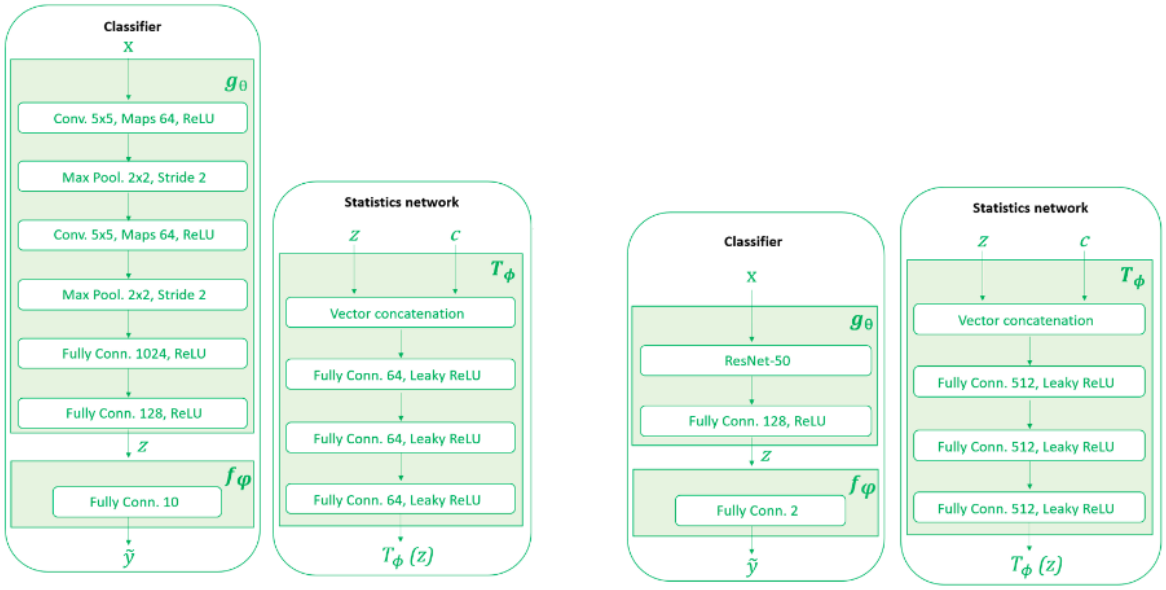

Figure 1. Description of the architectures (classifiers and statistics networks) for the experiments on Digits *(left)*, IMDB *(right)*.

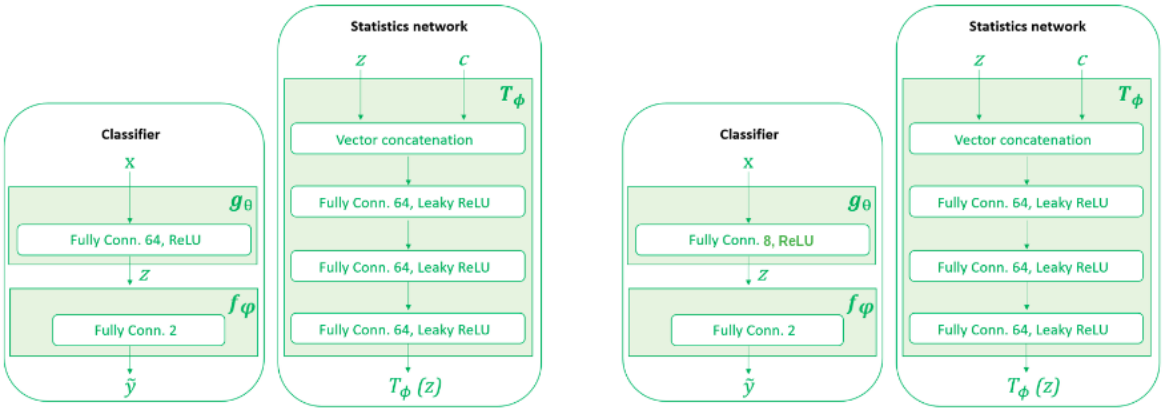

Figure 2. Description of the architectures (classifiers and statistics networks) for the experiments on Gernan *(left)*, Adult *(right)*.

linearly with the number of iterations employed to estimate the mutual information.

Choice of the hyper-parameter λ. The hyper-parameter λ regulates the trade-off between minimizing the task loss and reducing the mutual information between the biased attribute and the learned representation in (1). In Section 5 of the paper, we describe how to properly tune it. We report in Figure 4 the complete version of the analysis reported in the manuscript for the Digit experiment. We report the evolution of mutual information, test accuracy and training accuracy for different values of the hyper-parameter λ.

References

[1] Byungju Kim, Hyunwoo Kim, Kyungsu Kim, Sungjin Kim, and Junmo Kim. Learning not to learn: Training deep neural networks with biased data. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[2] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.

[3] David Madras, Elliot Creager, Toniann Pitassi, and Richard S. Zemel. Learning adversarially fair and transferable representations. *CoRR*, abs/1802.06309, 2018.

[4] Daniel Moyer, Shuyang Gao, Rob Brekelmans, Aram Galstyan, and Greg Ver Steeg. Invariant representations without adversarial training. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 9084–9093. Curran Associates, Inc., 2018.

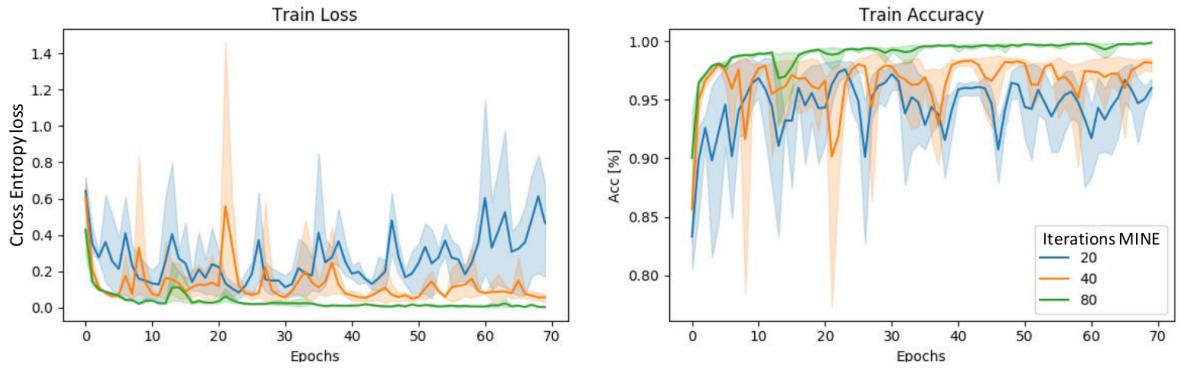

Figure 3. Training (cross-entropy) loss *(left)* and training accuracy *(right)* with $\lambda = 1.0$ for different number of iterations of MINE (K) on the digit recognition task (setting $\sigma = 0.02$). An increased number of iterations ($K = 20, 40, 80$ in *blue*, *orange* and *green*, respectively) has the effect of stabilizing the training procedure, *i.e.* it allows the model minimizing the loss function and fitting the training data. The charts report the average of 3 runs.

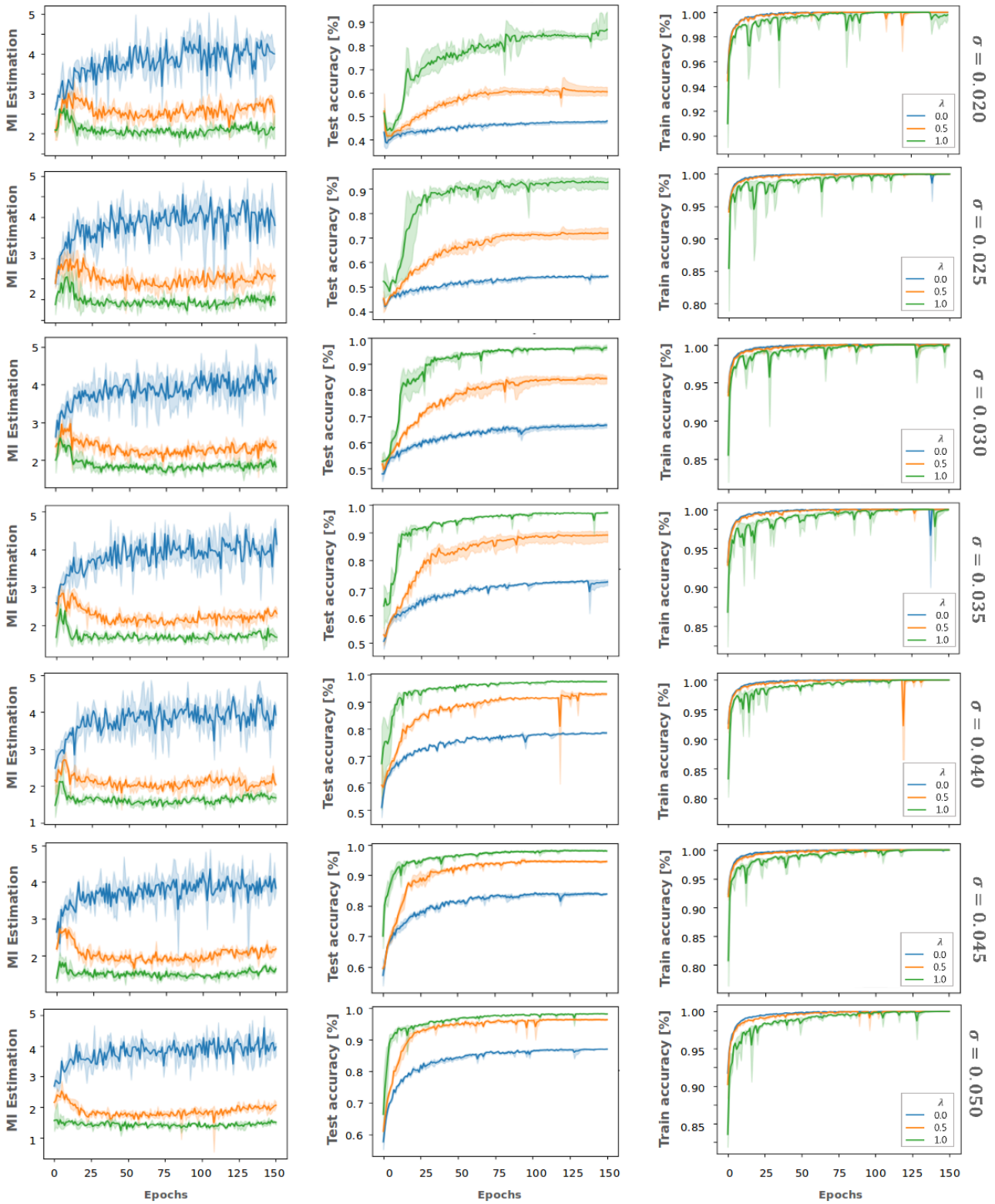

Figure 4. Values for mutual information (left column), test accuracy (middle column) and train accuracy (right column). We accounted for the different color, modelled by different σ (check Section 5 of the paper), and here represented by different rows. It is visible how a decrease in the (estimated) mutual information correlates with an improved performance.