Debiasing Neural Networks using Differentiable Classification Parity Proxies

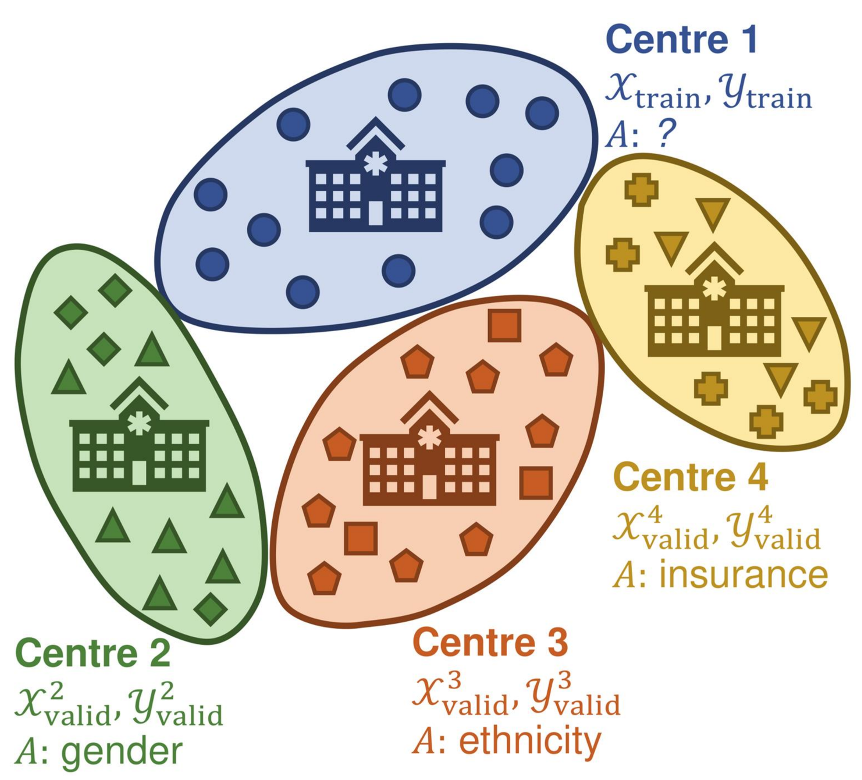
Ričards Marcinkevičs,✉ Ece Ozkan and Julia E. Vogt
Department of Computer Science, ETH Zürich ✉: ricards.marcinkevics@inf.ethz.ch

Motivation

Given:
- features $\boldsymbol{X} \in \mathbb{R}^p$, protected attribute $A \in \{0,1\}$, label $Y \in \{0,1\}$
- dataset $\mathcal{D} = \mathcal{D}_{\text{train}} \uplus \mathcal{D}_{\text{valid}} \uplus \mathcal{D}_{\text{test}} = \{(\boldsymbol{x}_i, y_i, a_i)\}_i$
- biased neural network $f_{\boldsymbol{\theta}}(\cdot)$ trained on $\{(\boldsymbol{x}_i, y_i)\}_i$ from $\mathcal{D}_{\text{train}}$

Goal: reduce the bias $\mu(\cdot)$ of the classifier $f_{\boldsymbol{\theta}}(\cdot)$, without considerably sacrificing its predictive performance $\rho(\cdot)$

Intra-processing Setting:
- $f_{\boldsymbol{\theta}}(\cdot)$ is debiased on the validation set $\mathcal{D}_{\text{valid}}$ *post hoc*
- the debiasing algorithm may edit the parameters $\boldsymbol{\theta}$
- A is not given at test time



Practical Example: a classifier was trained on data from the clinical centre **1**. When deployed in centres **2**, **3**, and **4**, it needs to be debiased according to the local considerations and constraints

Our Contribution:
1. Differentiable proxy functions for statistical parity (SPD) and equal opportunity difference (EOD)
2. Simple yet effective intra-processing debiasing techniques based on neural network pruning and fine-tuning
3. Experiments on tabular data and fully connected architectures

Classification Parity Proxies

Let $\mathcal{X} = \{\boldsymbol{x}_i\}_{i=1}^N$, $\mathcal{Y} = \{y_i\}_{i=1}^N$, $\mathcal{A} = \{a_i\}_{i=1}^N$. We propose differentiable proxies for the statistical parity difference:

$$\tilde{\mu}_{\text{SPD}}(f_{\boldsymbol{\theta}}, \mathcal{X}, \mathcal{Y}, \mathcal{A}) = \frac{\sum_{i=1}^N f_{\boldsymbol{\theta}}(\boldsymbol{x}_i)(1-a_i)}{\sum_{i=1}^N 1-a_i} - \frac{\sum_{i=1}^N f_{\boldsymbol{\theta}}(\boldsymbol{x}_i)a_i}{\sum_{i=1}^N a_i}$$

and equal opportunity difference:

$$\tilde{\mu}_{\text{EOD}}(f_{\boldsymbol{\theta}}, \mathcal{X}, \mathcal{Y}, \mathcal{A}) = \frac{\sum_{i=1}^N f_{\boldsymbol{\theta}}(\boldsymbol{x}_i)(1-a_i)y_i}{\sum_{i=1}^N (1-a_i)y_i} - \frac{\sum_{i=1}^N f_{\boldsymbol{\theta}}(\boldsymbol{x}_i)a_i y_i}{\sum_{i=1}^N a_i y_i}$$

Debiasing Methods

Intuition: directly minimise a differentiable bias proxy $\tilde{\mu}$ *without adversarial training*

Pruning for Debiasing: greedily prune individual units, *aka* neurons, in the neural network based on their contributions to the differentiable bias proxy. For unit j in layer l of the network $f_{\boldsymbol{\theta}}(\cdot)$:

$$S_{l,j} = \frac{1}{N}\sum_{i=1}^N \frac{\partial \tilde{\mu}(f_{\boldsymbol{\theta}}, \mathcal{X}, \mathcal{Y}, \mathcal{A})}{\partial h_j^l(\boldsymbol{x}_i)},$$

where $h_j^l(\cdot)$ is the activation of the j-th unit in layer l

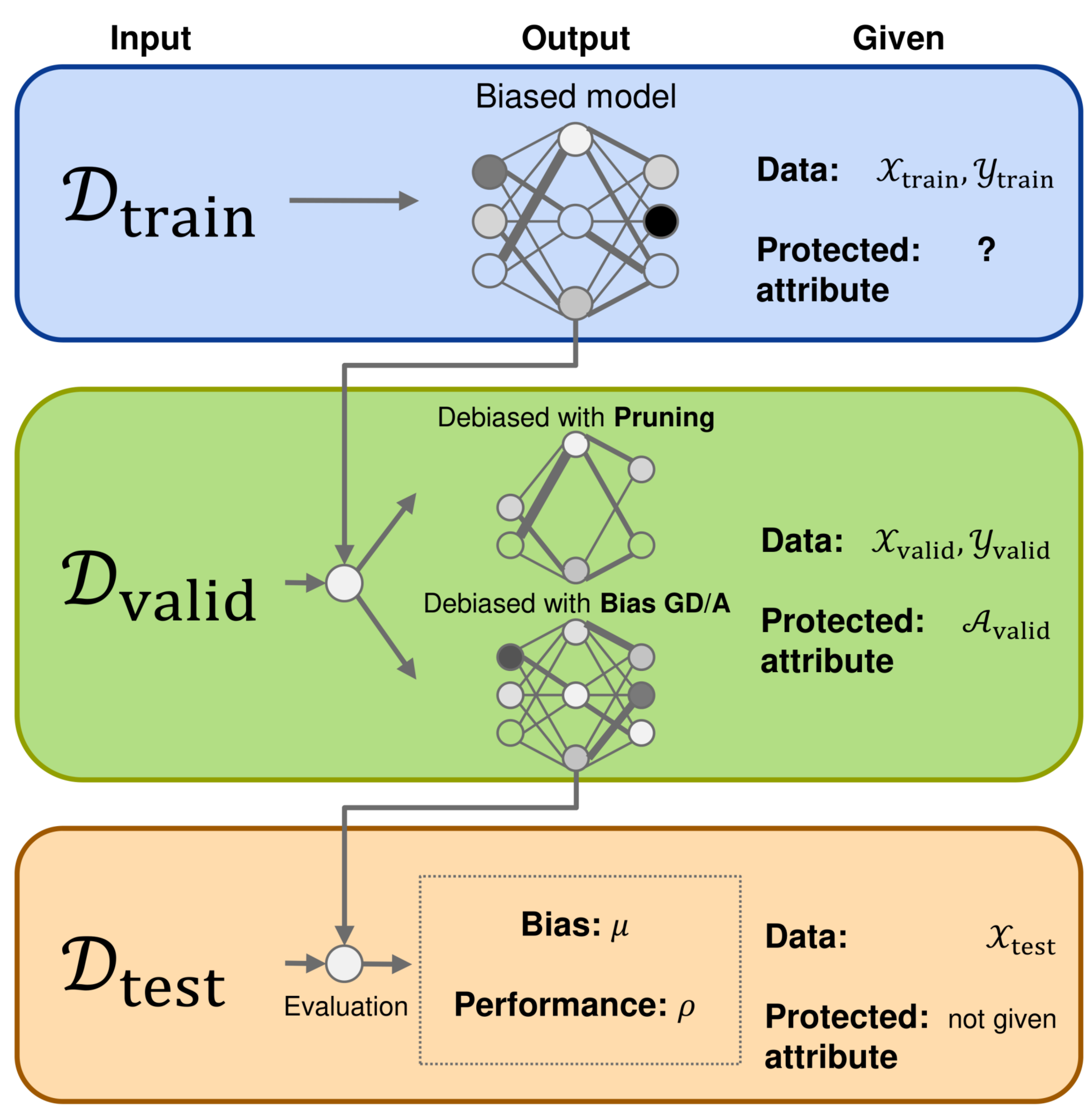
Bias Gradient Descent/Ascent (GD/A): fine-tune the network $f_{\boldsymbol{\theta}}(\cdot)$, minimising/maximising the proxy $\tilde{\mu}$ in the mini-batch gradient descent

In the end, return a debiased network $f_{\tilde{\boldsymbol{\theta}}}(\cdot)$ maximising the **bias-constrained objective**:

$$\varphi_{\rho,\mu,\varepsilon}(f_{\boldsymbol{\theta}}, \mathcal{X}, \mathcal{Y}, \mathcal{A}) = \begin{cases} \rho(f_{\boldsymbol{\theta}}, \mathcal{X}, \mathcal{Y}, \mathcal{A}), & \text{if } |\mu(f_{\boldsymbol{\theta}}, \mathcal{X}, \mathcal{Y}, \mathcal{A})| < \varepsilon \\ 0, & \text{otherwise} \end{cases},$$

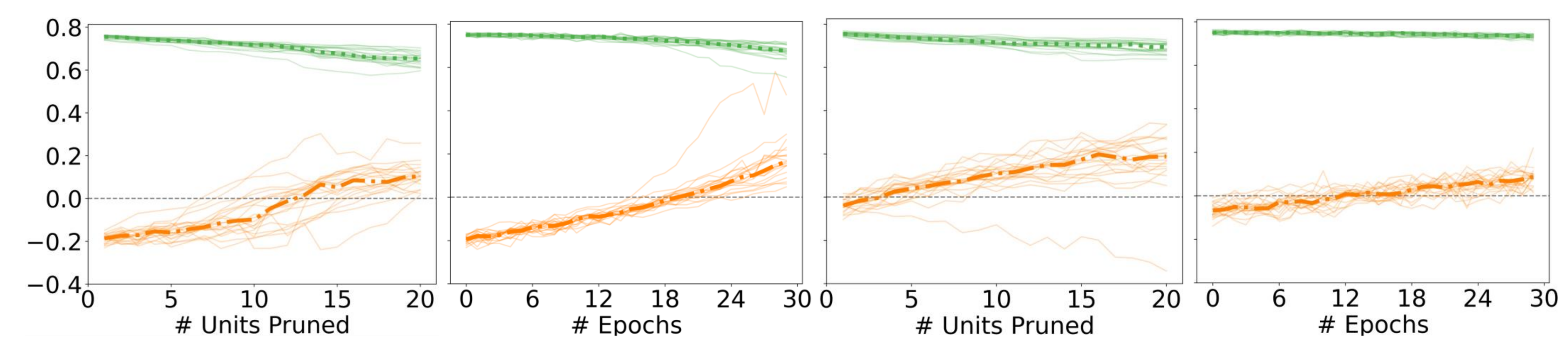
where $\varepsilon > 0$ is an upper/lower bound on bias

Debiasing Procedure:



Results

Changes in the bias, given by the SPD (*a*,*b*) and EOD (*c*,*d*), and balanced accuracy (BA) of the neural network during pruning (*a*,*c*) and bias GD/A (*b*,*d*). The results were obtained on the MIMIC-III dataset, using insurance type as the protected attribute:



(*a*) Pruning, SPD (*b*) Bias GD/A, SPD (*c*) Pruning, EOD (*d*) Bias GD/A, EOD

Bias BA

We compared proposed methods to other debiasing algorithms on a range of tabular benchmarks. Table below reports (*a*) bias and (*b*) bias-constrained objective ($\varepsilon = 0.05$) before and after debiasing:

	Bias Measure	Method	Adult: *Sex*	Bank: *Age*	COMPAS: *Race*	MIMIC-III: *Insurance*
(*a*)	SPD	Standard	-0.32±0.02	0.18±0.04	0.19±0.03	-0.19±0.03
		Random	*-0.04±0.01*	*0.03±0.04*	0.09±0.04	*-0.04±0.01*
		ROC	*-0.04±0.02*	0.08±0.04	**-0.01±0.01**	-0.05±0.01
		Eq. Odds	-0.09±0.01	0.06±0.03	*0.03±0.06*	**-0.01±0.00**
		Pruning	*-0.04±0.07*	**0.02±0.02**	*0.03±0.04*	**-0.01±0.03**
		Bias GD/A	**-0.01±0.04**	*0.03±0.05*	**0.01±0.04**	**-0.01±0.02**
	EOD	Standard	-0.14±0.02	*0.01±0.04*	0.20±0.05	-0.05±0.04
		Random	-0.07±0.03	0.02±0.04	0.09±0.04	-0.04±0.04
		ROC	-0.05±0.03	0.04±0.04	**-0.01±0.01**	-0.04±0.04
		Eq. Odds	**-0.01±0.04**	0.04±0.10	0.03±0.06	**0.01±0.04**
		Pruning	*-0.03±0.03*	*0.01±0.05*	*0.02±0.06*	**-0.01±0.04**
		Bias GD/A	-0.04±0.03	**0.00±0.06**	*0.02±0.06*	*0.03±0.04*
(*b*)	SPD	Standard	0.00; [0.00, 0.00]	0.00; [0.00, 0.00]	0.00; [0.00, 0.00]	0.00; [0.00, 0.00]
		Random	0.59; [0.59, 0.60]	0.52; [0.00, 0.55]	0.00; [0.00, 0.00]	0.67; [0.66, 0.68]
		ROC	**0.78; [0.00, 0.80]**	0.00; [0.00, 0.56]	0.50; [0.50, 0.50]	0.66; [0.00, 0.67]
		Eq. Odds	0.00; [0.00, 0.00]	0.00; [0.00, 0.69]	0.59; [0.00, 0.60]	0.57; [0.56, 0.58]
		Pruning	0.54; [0.52, 0.57]	*0.83; [0.80, 0.85]*	*0.62; [0.41, 0.64]*	*0.70; [0.69, 0.71]*
		Bias GD/A	*0.67; [0.64, 0.69]*	**0.85; [0.00, 0.87]**	**0.63; [0.46, 0.64]**	**0.73; [0.73, 0.74]**
	EOD	Standard	0.00; [0.00, 0.00]	**0.86; [0.00, 0.87]**	0.00; [0.00, 0.00]	0.37; [0.00, 0.75]
		Random	0.00; [0.00, 0.00]	**0.86; [0.00, 0.87]**	0.00; [0.00, 0.00]	*0.74; [0.00, 0.76]*
		ROC	**0.81; [0.00, 0.82]**	**0.86; [0.00, 0.87]**	0.50; [0.50, 0.50]	0.72; [0.00, 0.74]
		Eq. Odds	*0.72; [0.53, 0.74]*	*0.00; [0.00, 0.68]*	*0.59; [0.00, 0.60]*	0.57; [0.55, 0.57]
		Pruning	**0.81; [0.79, 0.81]**	**0.86; [0.00, 0.87]**	0.56; [0.00, 0.62]	**0.75; [0.55, 0.75]**
		Bias GD/A	**0.81; [0.00, 0.82]**	**0.86; [0.00, 0.87]**	**0.62; [0.00, 0.64]**	**0.75; [0.55, 0.75]**

Conclusion

- Differentiable proxy functions for the SPD and EOD
- Novel debiasing algorithms based on pruning and fine-tuning
- Promising preliminary results on tabular benchmarks and FCNNs

Future Work:
- other architectures, e.g. CNNs
- comparison with adversarial approaches
- application to medical imaging data