Debiasing Deep Chest X-Ray Classifiers using Intra-and Post-processing Methods

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

1 Motivation

Given

- $\boldsymbol{X}$, $A \in \{0,1\}$, $Y \in \{0,1\}$
- $\mathcal{D} = \{(\boldsymbol{x}_i, y_i, a_i)\}_i = \mathcal{D}_{\text{train}} \biguplus \mathcal{D}_{\text{valid}} \biguplus \mathcal{D}_{\text{test}}$
- neural network $f_{\boldsymbol{\theta}}(\cdot)$ trained on $\{(\boldsymbol{x}_i, y_i)\}_i$ from $\mathcal{D}_{\text{train}}$

Goal: reduce the bias μ of $f_{\boldsymbol{\theta}}(\cdot)$, sacrificing the performance ρ *as little as possible*

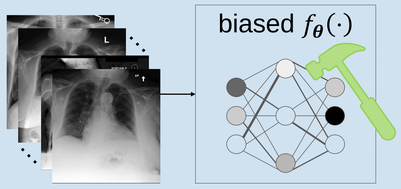
Bias Measures

$$\mu_{\text{SPD}} = \mathbb{P}_{X,A}\big(\hat{Y} = 1 | A = 0\big) - \mathbb{P}_{X,A}\big(\hat{Y} = 1 | A = 1\big)$$

$$\mu_{\text{EOD}} = \mathbb{P}_{X,Y,A}\big(\hat{Y} = 1 | Y = 1, A = 0\big) - \mathbb{P}_{X,Y,A}\big(\hat{Y} = 1 | Y = 1, A = 1\big)$$

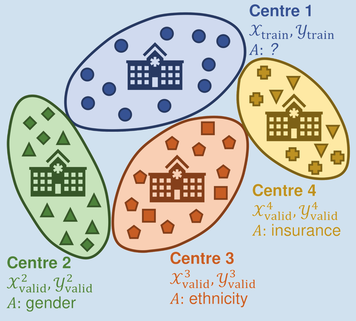
Intra-processing Setting

- debias *post hoc* on $\mathcal{D}_{\text{valid}}$
- edit the parameters $\boldsymbol{\theta}$
- A is not given at test time



Example

- train on centre **1**
- deploy in centres **2**, **3**, **4** with *local* fairness constraints
- no access to $\mathcal{D}_{\text{train}}$



Contributions

- *differentiable* proxy functions for the SPD and EOD
- debiasing algorithms based on pruning and fine-tuning
- experiments on MIMIC-III and MIMIC-CXR

2 Debiasing Algorithms

Classification Parity Proxies

Let $\mathcal{X} = \{\boldsymbol{x}_i\}_{i=1}^N$, $\mathcal{Y} = \{y_i\}_{i=1}^N$, $\mathcal{A} = \{a_i\}_{i=1}^N$. *Differentiable* proxies for classification disparity:

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

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

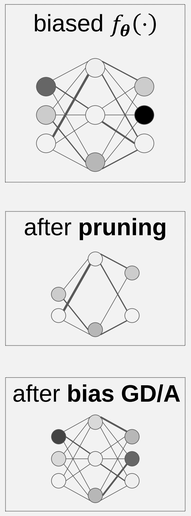
Observation: $\tilde{\mu}_{\text{SPD}}(f_{\boldsymbol{\theta}}(\cdot), \mathcal{X}, \mathcal{Y}, \mathcal{A}) \propto \widehat{\text{Cov}}\big(A, f_{\boldsymbol{\theta}}(\boldsymbol{X})\big)$ and $\tilde{\mu}_{\text{EOD}}(f_{\boldsymbol{\theta}}(\cdot), \mathcal{X}, \mathcal{Y}, \mathcal{A}) \propto \widehat{\text{Cov}}\big(A, f_{\boldsymbol{\theta}}(\boldsymbol{X}) | Y = 1\big)$

Intuition: minimise/maximise a differentiable proxy directly

Pruning for Debiasing

i. for layer $1 \leq l \leq L$, evaluate influence
$$S_{l,j} = \frac{1}{N} \sum_{i=1}^N \frac{\partial \tilde{\mu}(f_{\boldsymbol{\theta}}(\cdot), \mathcal{X}, \mathcal{Y}, \mathcal{A})}{\partial z_j^l(\boldsymbol{x}_i)}$$
ii. prune the most influential units
iii. evaluate bias and performance on $\mathcal{D}_{\text{valid}}$
iv. repeat steps i-iii

Return the model with the minimal bias and performance $> \varrho$

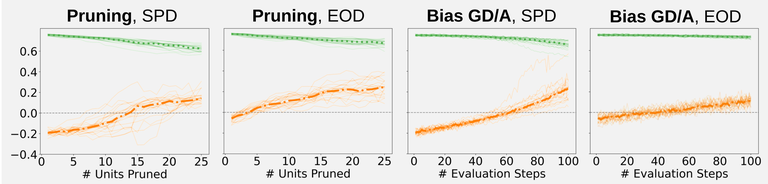


Bias Gradient Descent/Ascent (GD/A)

. fine-tune $f_{\boldsymbol{\theta}}(\cdot)$ on $\mathcal{D}_{\text{valid}}$
– few epochs, small LR
. mini-batch gradient decent/ascent
. stop early: performance $> \varrho$

3 Results

Bias and performance when debiasing on MIMIC-III:



Results for the VGG-16 and ResNet-18 trained on MIMIC-CXR:

Enlarged cardiomediastenum, *Sex*, **VGG-16**

Method	EOD	BA
Standard	-0.05±0.02	0.77±0.01
Random	-0.03±0.03	*0.75±0.01*
ROC	-0.05±0.02	*0.75±0.03*
Eq. Odds	*0.01±0.03*	*0.75±0.01*
Adv. Intra	-0.04±0.03	0.73±0.01
Pruning	**0.00±0.02**	**0.76±0.02**
Bias GD/A	*-0.01±0.04*	**0.76±0.01**

Pneumonia, *Ethnicity*, **VGG-16**

Method	EOD	BA
Standard	-0.14±0.04	0.73±0.02
Random	-0.11±0.06	**0.71±0.02**
ROC	*-0.07±0.06*	0.65±0.06
Eq. Odds	**0.00±0.06**	*0.70±0.01*
Adv. Intra	-0.13±0.05	*0.70±0.02*
Pruning	-0.09±0.05	**0.71±0.03**
Bias GD/A	-0.08±0.06	**0.71±0.02**

Enlarged cardiomediastenum, *Sex*, **ResNet-18**

Method	EOD	BA
Standard	-0.05±0.04	0.76±0.01
Random	**0.00±0.03**	0.73±0.02
ROC	-0.05±0.03	*0.74±0.04*
Eq. Odds	*0.01±0.03*	*0.74±0.01*
Adv. Intra	-0.04±0.04	0.73±0.02
Pruning	*-0.01±0.03*	*0.74±0.02*
Bias GD/A	**0.00±0.03**	**0.76±0.01**

Pneumonia, *Ethnicity*, **ResNet-18**

Method	EOD	BA
Standard	-0.14±0.05	0.73±0.02
Random	*-0.06±0.06*	0.65±0.04
ROC	-0.07±0.04	0.65±0.05
Eq. Odds	**-0.01±0.06**	0.70±0.01
Adv. Intra	-0.14±0.03	*0.71±0.02*
Pruning	-0.11±0.05	0.70±0.02
Bias GD/A	-0.11±0.05	**0.73±0.02**

4 Outlook

- Other diseases and protected attributes in MIMIC-CXR
- Other architectures, e.g. SqueezeNet, DenseNet
- Beyond SPD and EOD
- Multiple and multicategorical protected attributes and labels
- Beyond gradient-based criteria for pruning