One-network Adversarial Fairness

Tameem Adel[1], Isabel Valera[2], Zoubin Ghahramani[1,3] & Adrian Weller[1,4]

[1]University of Cambridge, UK. [2]MPI-IS, Germany. [3]Uber AI Labs, USA. [4]The Alan Turing Institute, UK.

tah47@cam.ac.uk

Introduction

Machine learning (ML) algorithms optimized:

- Not only for task performance, e.g. accuracy.
- But also other criteria, e.g. safety, interpretability, fairness.
- Here, our aim is to build an *accurate* as well as *fair* learner.
- Fairness: the outcome of a system should not discriminate between subgroups characterized by sensitive attributes such as gender or race.

Motivation

- Our Fair Adversarial Discriminative (FAD) learner adds a hidden layer, and an extra classifier at the network's top.
- This leads to a neural network (NN) that is:
  - maximally uninformative about the sensitive attributes; and
  - predictive of the class labels.
- The whole adversarial game happens in *one single NN*, leading to:
  - a much less tricky adversarial optimization; and
  - minimal overhead on the original model (slight modifications).

Contributions

- A fairness algorithm (FAD) that slightly modifies an unfair model's architecture to simultaneously optimize for accuracy and fairness.
- FAD also quantifies the tradeoff between accuracy and fairness.
- A variation of the algorithm in which diversity among minibatch elements is increased (FAD−MD).
- A novel generalization bound illustrating the theoretical relationship between the label classifier and the fair adversary.
- Experiments on two datasets demonstrate state-of-the-art effectiveness.

FAD with Minibatch diversity (FAD-MD)

We form minibatch elements as follows to make them as diverse as possible:

- Randomly choose few points to belong to the minibatch.
- From a pool of points, select the point via the score resulting from a one-class SVM. The class consists of the current minibatch elements.
- The next added data point is the point with the lowest score, i.e. the point least likely to be similar to the current minibatch elements.
- Continue this process until reaching the prespecified minibatch size.

Conclusion

- We introduced a fair adversarial framework applicable to any differentiable discriminative model.
- Instead of having to establish the architecture from scratch, we make slight adjustments to an existing differentiable classifier by:
  - adding a new hidden layer; and
  - adding a new classifier above it,

  to concurrently optimize for fairness and accuracy in one network.

FAD

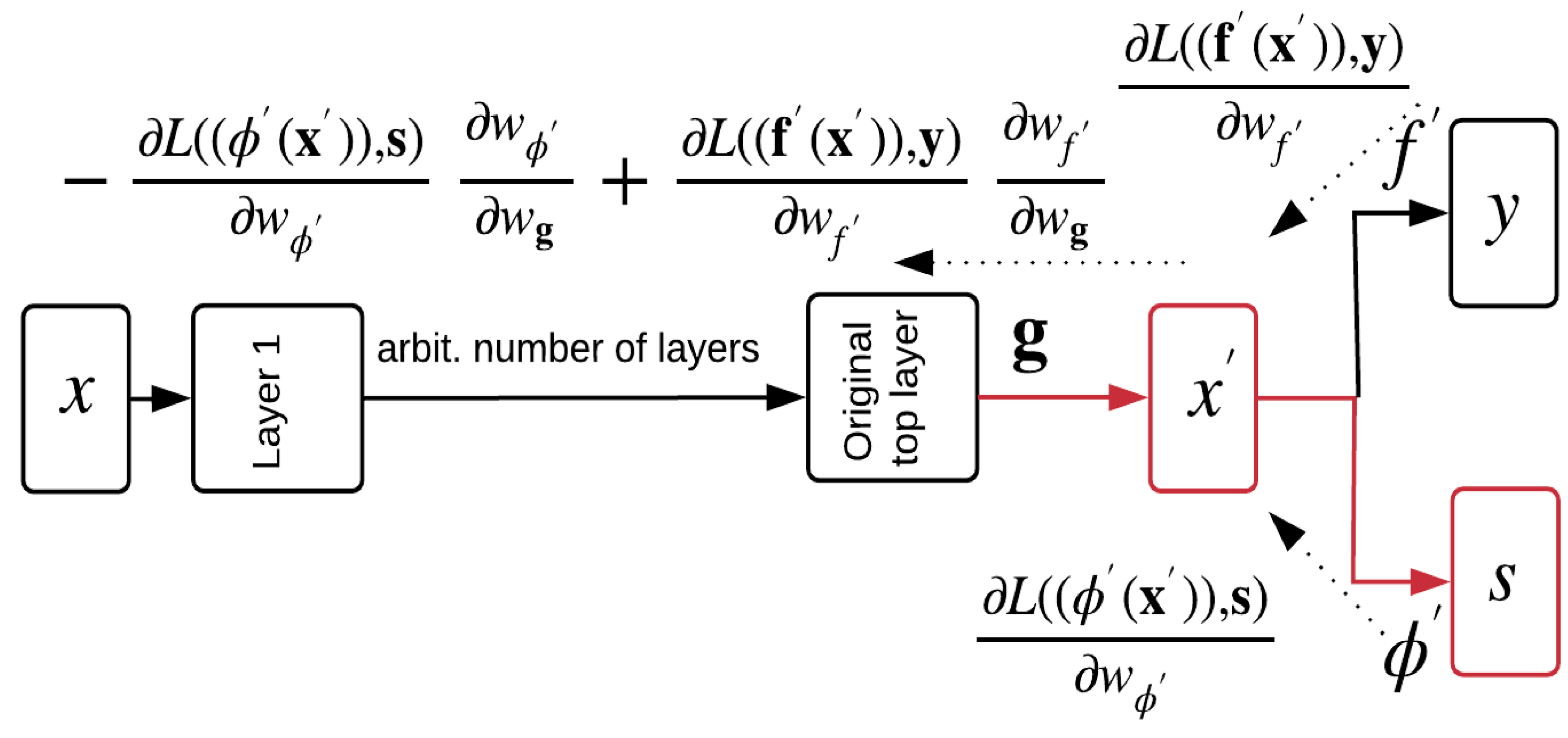

Figure 1: Architecture of the proposed FAD. The parts added, due to FAD, to an unfair deep architecture with input x are (shown in red): i) the layer g where x' is learned and; 2) the sensitive attribute s predictor ϕ' at the network's top.

Experiments

Classification Accuracy

COMPAS	Unfair ($\beta = 0$) 89.3%	Unfair (Zafar et al. 2017b) 66.8%	Unfair (Zafar et al. 2017a) 69.0%	FAD 88.4%
	FAD−MD **88.7**%	Zafar et al. (2017b) 66.2%	Zafar et al. (2017a) 67.5%	Hardt et al. (2016) 64.4%
	Feldman et al. (2015) 86.8%	Kamishima et al. (2012) 72.4%	Fish et al. (2016) 81.2%	Bechavod (2017) 66.4%
	Komiyama et al. (2018) 86.6%	Agarwal et al. (2018) 71.2%	Narasimhan (2018) 77.7%	
Adult	Unfair ($\beta = 0$) 90.1%	Unfair (Zafar et al. 2017b) 85.8%	Unfair (Zafar et al. 2017a) 87.0%	FAD 88.6%
	FAD−MD **89**%	Zafar et al. (2017b) 83.1%	Zafar et al. (2017a) 84.0%	Hardt et al. (2016) 84.6%
	Feldman et al. (2015) 82.1%	Kamishima et al. (2012) 84.3%	Fish et al. (2016) 84.0%	Bechavod (2017) 78.3%
	Komiyama et al. (2018) 85.7%	Agarwal et al. (2018) 86.2%	Narasimhan (2018) 81.5%	

Experiments

(Un)fairness

COMPAS	Unfair ($\beta = 0$) Disp_{DI}: 0.6 Disp_{FPR}: 0.21 Disp_{FNR}: 0.29	Unfair (Zafar et al. 2017b) – 0.18 0.3	Unfair (Zafar et al. 2017a) 0.62 – –	FAD **0.08** **0.01** **0.01**	FAD−MD 0.11 **0.01** 0.02	Zafar et al. (2017b) – 0.03 0.1	Zafar et al. (2017a) 0.38 – –	Hardt et al. (2016) – **0.01** **0.01**
	Feldman et al. (2015) 0.95 0.4 0.45	Kamishima et al. (2012) 0.9 0.2 0.15	Fish et al. (2016) 0.15 0.03 0.03	Bechavod (2017) – **0.01** 0.03	Komiyama et al. (2018) 0.2 – –	Agarwal et al. (2018) 0.09 0.05 0.05	Narasimhan (2018) 0.1 0.09 0.11	
Adult	Unfair ($\beta = 0$) Disp_{DI}: 0.71 Disp_{FPR}: 0.36 Disp_{FNR}: 0.32	Unfair (Zafar et al. 2017b) – 0.35 0.4	Unfair (Zafar et al. 2017a) 0.68 – –	FAD 0.14 0.02 **0.01**	FAD−MD **0.13** 0.01 0.02	Zafar et al. (2017b) – 0.12 0.09	Zafar et al. (2017a) 0.29 – –	Hardt et al. (2016) – 0.04 0.03
	Feldman et al. (2015) 0.25 0.3 0.4	Kamishima et al. (2012) 0.3 0.07 0.08	Fish et al. (2016) 0.16 0.02 0.03	Bechavod (2017) – **0.0** 0.04	Komiyama et al. (2018) 0.28 – –	Agarwal et al. (2018) **0.13** 0.04 0.05	Narasimhan (2018) 0.19 0.14 0.08	

Conclusion (contd.)

- We analyzed and evaluated the tradeoff between fairness and accuracy.
- We proposed a minibatch diversity variation of the learning procedure which is of independent interest for adversarial frameworks in general.
- We provided a theoretical interpretation of the two classifiers (adversaries) constituting the model.
- We demonstrated strong empirical performance of our methods compared to previous leading approaches.