FACT in AI

Reproducing and Extending *Mitigating Unwanted Biases with Adversarial Learning* (Zhang et al., 2018)

Vanessa Botha
Nithin Holla
Azamat Omuraliev
Leila Talha

University of Amsterdam

Overview

- Introduction
- Method
- Experiments & Results:
    - UCI Adult dataset
    - UCI Communities and Crime dataset
    - UTK Face dataset
- Discussion
- Conclusion
- Questions

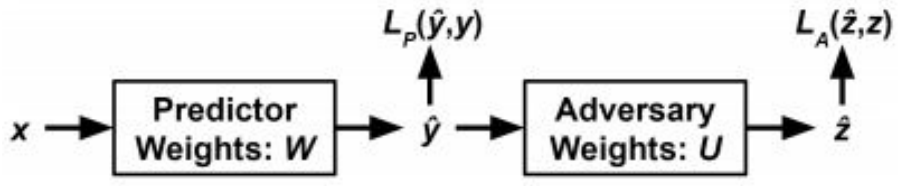
Introduction

Mitigating Unwanted Biases with Adversarial Learning (Zhang et al., 2018)

- Paper on fairness
- In-processing method
- Fairness measures:
    - Demographic Parity
    - Equality of Odds
    - *Equality of Opportunity*
- Predictor and adversary

Method

Adversarial Network



$$\nabla_W L_P - \mathrm{proj}_{\nabla_W L_A} \nabla_W L_P - \alpha \nabla_W L_A$$

- Predictor
- Adversary
  - Demographic parity
  - Equality of Odds
- Gradient based optimisation
- Custom update predictor
  - projection
  - alpha and adversary loss

Experiments

UCI Adult Dataset

Task

Predict if annual income is above $50k, based on 14 attributes. The protected variable is sex.

Predictor

$$\hat{y} = \sigma(w_1 \cdot x + b)$$

Adversary

$$s = \sigma\big((1 + |c|)\sigma^{-1}(\hat{y})\big)$$

$$\hat{z} = w_2 \cdot [s, sy, s(1 - y)] + b$$

Results UCI Adult Dataset

		Female		Male	
		Without	With	Without	With
Zhang *et al.* (2018)	FPR	**0.0248**	**0.0647**	**0.0917**	**0.0701**
	FNR	**0.4492**	**0.4458**	**0.3667**	**0.4349**
Faithful implementation	FPR	**0.0287** ± 0.0112	**0.1092** ± 0.0478	**0.1072** ± 0.0355	**0.1196** ± 0.1707
	FNR	**0.4491** ± 0.0709	**0.3334** ± 0.1007	**0.3803** ± 0.0775	**0.7030** ± 0.3852
Refined implementation	FPR	**0.0287** ± 0.0112	**0.0404** ± 0.0020	**0.1072** ± 0.0355	**0.0802** ± 0.0030
	FNR	**0.4491** ± 0.0709	**0.4313** ± 0.0120	**0.3803** ± 0.0775	**0.4542** ± 0.0080

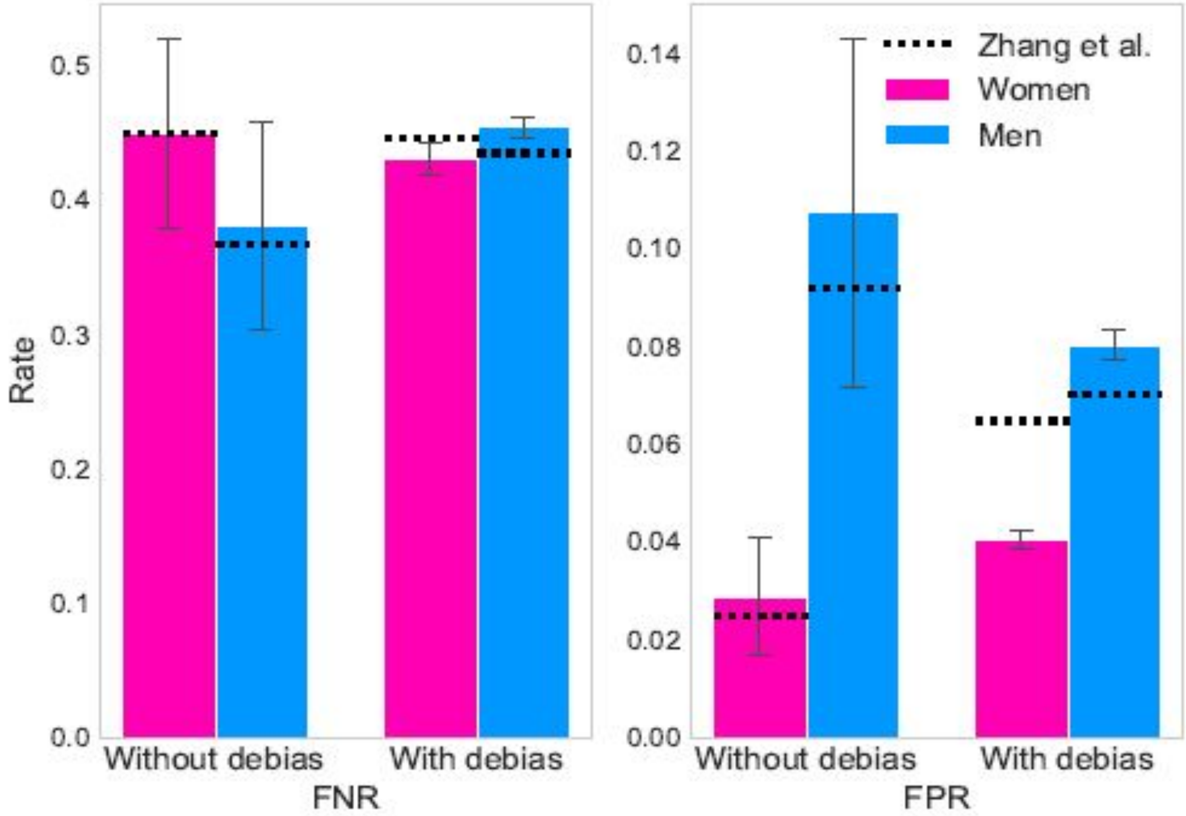
Results UCI Adult Dataset

Accuracy
Zhang et al.

- biased: 86.0%
- debiased: 84.5%

Accuracy
refined implementation

- biased: 84.95%
- debiased: 84.45%



Experiments

UCI Communities and Crime Dataset

Task

Predict rate of violent crimes within a community, based on 114 continuous attributes. The protected variable is the percentage of white citizens.

Predictor

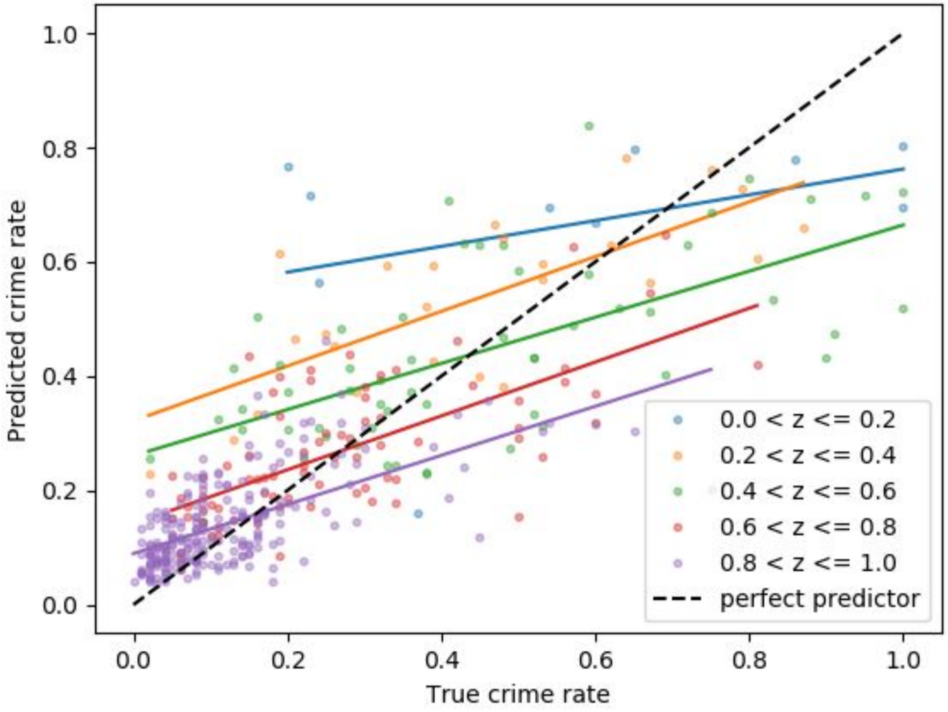
$$\hat{y} = \sigma(w_1 \cdot x + b)$$

Adversary

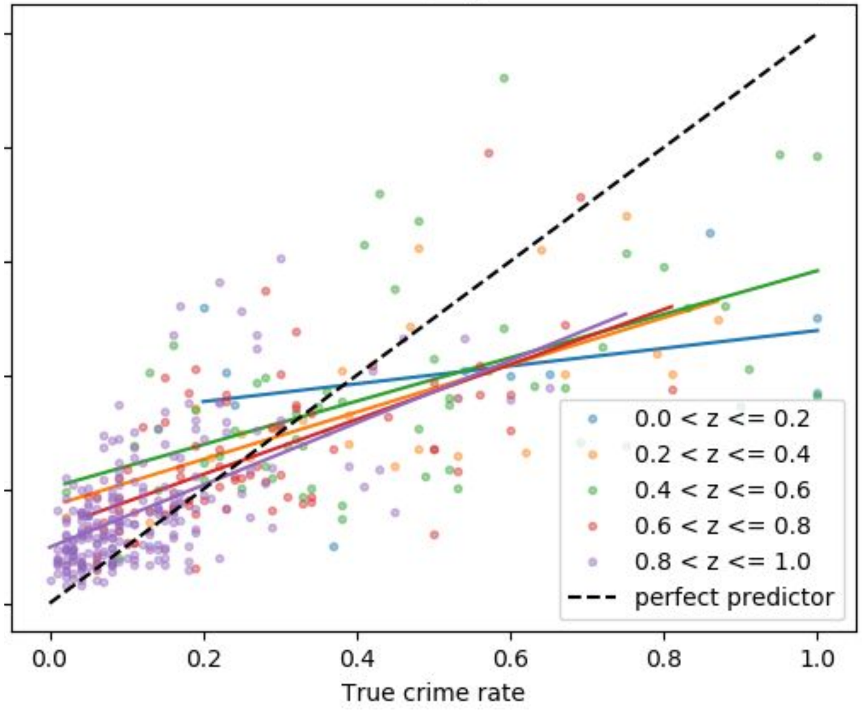
$$s = \sigma\big((1 + |c|)\sigma^{-1}(\hat{y})\big)$$

$$\hat{z} = w_2 \cdot [s, sy, s(1-y)] + b$$

Results UCI Communities and Crime Dataset



Predictor MSE: 0.017



Predictor MSE: 0.022, Adversary MSE: 0.043

Experiments

Age estimation on UTKFace Dataset

Task:

Predict age, based on images annotated with gender (protected variable) and race.

Predictor:

Deep convolutional neural network, regularized with Dropout

Adversary:

Satisfying demographic parity

$$s = \sigma\Big((1 + |c|)\sigma^{-1}(\hat{y})\Big)$$

$$\hat{z} = w_2 \cdot s + b.$$

Results

UTKFace Dataset

	Female		Male	
	Without	With	Without	With
AUC mean	**0.8632**	**0.8732**	**0.8702**	**0.8727**
AUC std	± 0.0013	± 0.0040	± 0.0003	± 0.0045

- Difference before debiasing
  >
  Difference after debiasing
- Improved performance

Discussion

Reproducing scientific research

- Reproduced to certain degree
- Original research lacks specifications
    - Hyperparameters
    - Data splits
    - Unclear reporting metric
- Adversarial setup unstable

Discussion

Extending adversarial debiasing

- Model-agnostic debiasing method
- Works
  - on a different domain
  - for different debiasing scheme
  - for complex predictor
  - for continuous targets
- Tackling instability and sensitivity of adversarial learning

Conclusion

- Universal but not robust solution
- Results difficult to reproduce
- No standard for metrics capturing bias under different scenarios

Questions

References:

Dheeru Dua and Casey Graff. 2017. UCI Machine Learning Repository. http://archive.ics.uci.edu/ml

Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. 2018. Mitigating Unwanted Biases with Adversarial Learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (New Orleans, LA, USA) (AIES '18). Association for Computing Machinery, New York, NY, USA, 335–340. https://doi.org/10.1145/3278721.3278779

Song Yang Zhang, Zhifei and Hairong Qi. 2017. Age Progression/Regression by Conditional Adversarial Autoencoder. In *IEEE Conference on Computer Visionand Pattern Recognition (CVPR)*. IEEE.

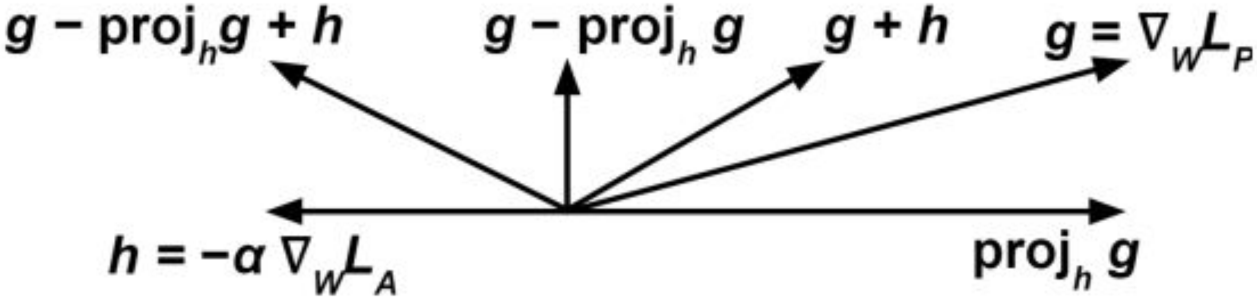


Figure 2: Diagram illustrating the gradients in Eqn. 1 and the relevance of the projection term $\text{proj}_h g$. Without the projection term, in the pictured scenario, the predictor would move in the direction labelled $g + h$ in the diagram, which actually *helps* the adversary. With the projection term, the predictor will never move in a direction that helps the adversary.