Adversarial Learning: Evaluation the Impact of Input Strategies on Fairness and Accuracy

Ray Chen, Aayana Evanson, Sandy Shao

New York University

CSCI-GA 3033-112 Final Paper

# Abstract

Adversarial Learning is an essential approach in ensuring fairness and accuracy in machine learning models. This paper extends the framework of Zhang et al. by analysing the effectiveness of different inputs into the adversary; such as intermediate predictor layers, concatenated representations and original features. We evaluated these inputs on fairness and performance using the UCI Adult Dataset, Communities and Crime, German Credit, and COMPAS Recidivism.

Keywords: Adversarial Learning, Bias Mitigation, Fairness, Machine Learning Models.

E-mail address: [rc4000@nyu.edu](mailto:rc4000@nyu.edu), [ae2630@nyu.edu](mailto:ae2630@nyu.edu), [xs508@nyu.edu](mailto:xs508@nyu.edu)

# 1 Introduction

Machine Learning models are influenced by bias in their training data. Adversarial learning gives models solutions to detect and mitigate biases by using adversarial inputs. Our paper focuses on exploring some of these methods by building on Zhang et al. work by expanding adversarial inputs and analysing their impacts in the model.

# 2 Related Work

This study is built from the prior work on adversarial fairness by Zhang et al. where a method using single-layer predictor inputs was a focus.

# 3 Methods

Our approach modifies the adversarial debiasing framework using intermediate layers, concatenated layers and original input features. These changes will lead to an analysis of bias mitigation strategies.

# 4 Evaluation

We evaluate out approach using datasets such as UCI Adult, Communities and Crime, German Credit, and COMPAS Recidivism. Evaluation metrics include statistical parity, equality of odds, and accuracy. These metrics enable an analysis of fairness an predictive performance.

# 5 Results

The results show that intermediate an concatenate input provide superior bias detection without significantly impacting accuracy. Using original features potential for bias detection can be seen but performance is lost.

# 6 Discissions

From our findings.

# 7 Conclusion

# 8 Appendix

# 9 References

1. Zhang, B. H., Lemoine, B., & Mitchell, M. (2018). Mitigating Unwanted Biases with Adversarial Learning. AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society.

Title  
Abstract  
1. Introduction  
 - This section should motivate the importance of the research problem, give an overview of your approach to solving the problem, and explain why it is a novel and valuable contribution.  Introduction sections for ML conferences often end with a bulleted list of the specific contributions of the paper (both methodological and applied/impact).  
2. Related Work  
 - This section should describe the most closely related prior literature and why your work is different/better.  
3. Methods  
 - This section should descibe your methodological approach in detail.  (Usually, the novel contribution of the paper is methodological, and is described here.)  
4. Evaluation  
 - How did you evaluate your methodological approach?  Describe the datasets you used, evaluation metrics used to measure performance, other methods you compared to.  
5. Results  
 - Describe the results of the evaluation here. Tables and figures showing the results are common in this section.  
6. Discussion  
 - What are the take-home messages from your results?  Why should we care?  
7. Conclusions  
 - Briefly summarize what you did and why it matters.  
 - Limitations of the work often go here (or could be in Discussion)  
 - Possible directions for future work  
References