USC Viterbi School of Engineering + amazon

USC Viterbi + amazon

Towards Trustworthy AI

Ninareh Mehrabi (PhD Fellow)

USC-Information Sciences Institute

September 24th

Presentation Outline

What is Trustworthy AI?

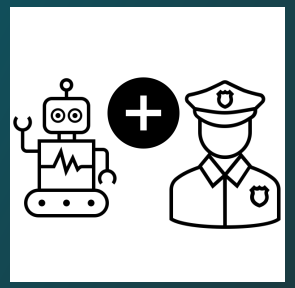
Adversarial Robustness through the Lens of Fairness

Attributing Fair Decisions

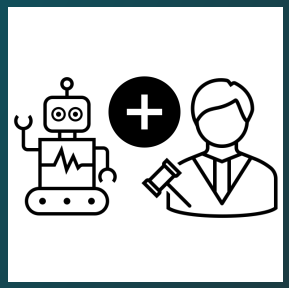
Future Plans

Conclusion and Remarks

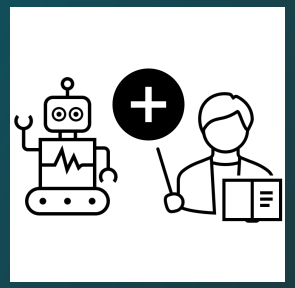
Trustworthy AI



Robustness.
Robust against adversarial attacks on privacy, security, safety, fairness.



Fairness.
Staying away from any favoritism or prejudice towards different individuals or demographic groups.



Explainability.
Understanding how AI makes decisions which can be beneficial for transparency and accountability as well.

Adversarial Robustness through the Lens of Fairness

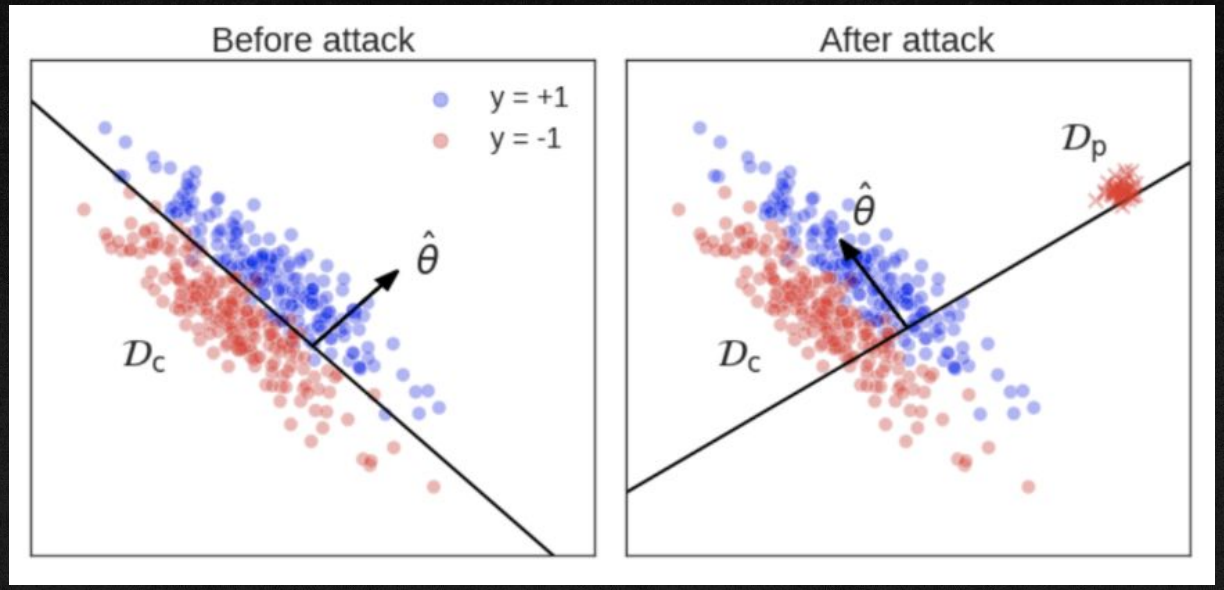
AAAI-2021

Introduction

- Previous work in adversarial machine learning mostly considered and analyzed vulnerability of models with regards to accuracy.
- In this work, we analyze vulnerability of machine learning models with regards to fairness metrics.
- We propose two families of data poisoning attacks that target fairness.
  - Anchoring Attack
    - Goal is to skew the decision boundary by placement of poisoned instances.
    - Two types of anchoring attack:
      - Random: Target points are chosen by random.
      - Non-random: Target points chosen based on popularity (near more similar instances).
  - Influence Attack on Fairness
    - Goal is to propose a loss function maximizing which can harm fairness.

Data Poisoning Attacks

- Poisoning Attacks: These types of attacks happen during the training process. The goal is to train a malicious model via some poisoned data instances.



Koh, Pang Wei, Jacob Steinhardt, and Percy Liang. "Stronger data poisoning attacks break data sanitization defenses." 2018.

Data Poisoning Objective

- The goal of the data poisoning attack as an optimization problem:

$$\max_{\mathcal{D}_p} L_{adv}(\hat{\theta}; \mathcal{D}_{test})$$

$$s.t. \quad |\mathcal{D}_p| = \epsilon |\mathcal{D}_c|$$

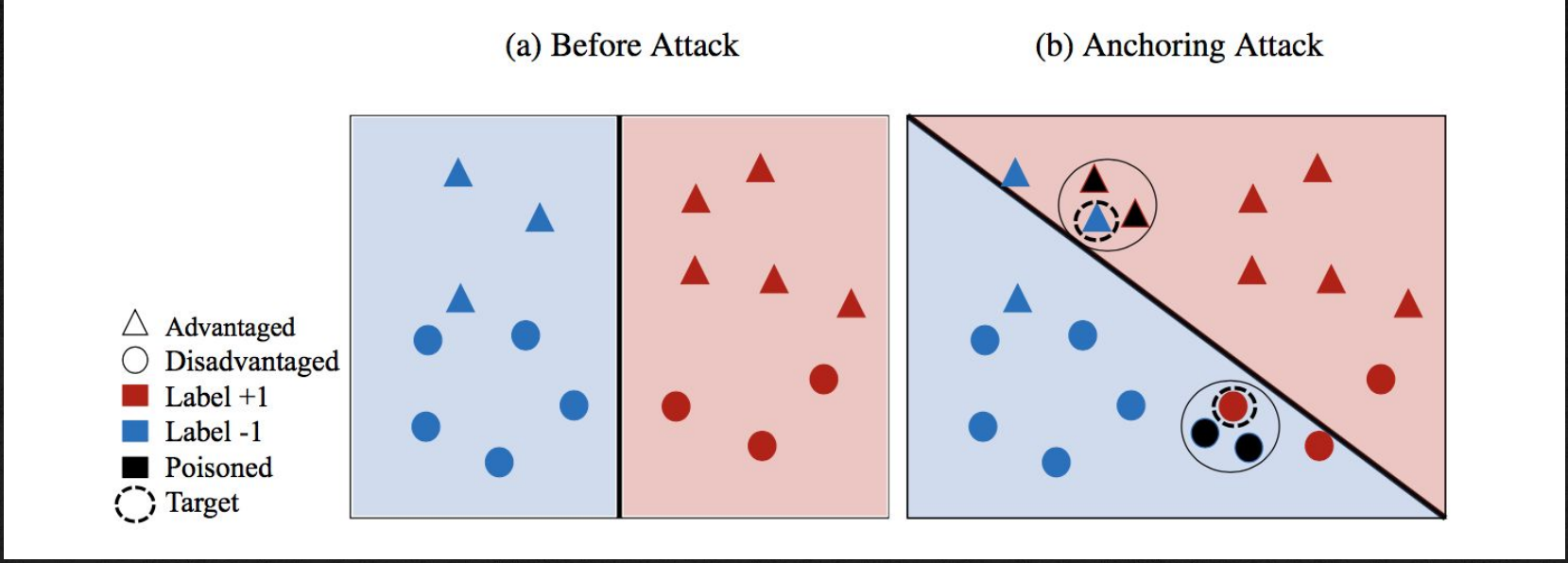
$$\mathcal{D}_p \subseteq \mathcal{F}_\beta$$

$$\text{where} \quad \hat{\theta} = \arg\min_{\theta} \mathcal{L}(\theta; \mathcal{D}_c \cup \mathcal{D}_p).$$

- We have:
  - The adversary who wants to maximize a targeted loss via some poisoned points.
  - A set of clean and poisoned data points.
  - A feasible set which the adversary wishes to put its poisoned instances there to avoid detection by sanitization techniques.
  - Finally, the defender who wants to minimizes its loss.

Anchoring Attack

- In this attack, we need to select some target instances.
- The goal is to cloud advantaged instances with label -1 with advantaged instances with label 1 and disadvantaged instances with label 1 with disadvantaged instances with label -1.
- This will skew and bias the decision boundary as shown below.



Influence Attack on Fairness

- In the influence attack on fairness (IAF), we propose a loss function based on the decision boundary fairness measure* maximizing which using the influence attack will cause harm to fairness.
- The loss is composed of two parts one for controlling the accuracy and one controlling fairness regularized by a hyper-parameter as follows:

$$L_{adv}(\hat{\theta}; \mathcal{D}_{test}) = \ell_{acc} + \lambda \ell_{fairness}$$

$$where \ \ \ell_{fairness} = \frac{1}{N} \sum_{i=1}^{N} (z_i - \bar{z}) d_{\hat{\theta}}(x_i).$$

$$\max_{\mathcal{D}_p} L_{adv}(\hat{\theta}; \mathcal{D}_{test})$$

$$s.t. \ \ |\mathcal{D}_p| = \epsilon |\mathcal{D}_c|$$

$$\mathcal{D}_p \subseteq \mathcal{F}_\beta$$

$$\text{where} \ \ \hat{\theta} = \arg\min_{\theta} \mathcal{L}(\theta; \mathcal{D}_c \cup \mathcal{D}_p).$$

*Zafar, Muhammad Bilal, et al. "Fairness constraints: Mechanisms for fair classification." *Artificial Intelligence and Statistics*. PMLR, 2017.

Experimental Setup

- Baselines:
  - Basic Influence Attack on Accuracy*
    - This attack is merely targeting accuracy and is not optimized for fairness.
    - The goal for its inclusion is:
      - To show that attacks optimized for accuracy can not target fairness measures.
      - Good baseline to compare the performance with regards to accuracy.
  - Poisoning Attack on Fairness**
- Measures:
  - In addition to accuracy, we utilized two widely known fairness measures to report our results:
    - Statistical Parity Difference (SPD)
    - Equality of Opportunity Difference (EOD)
- Datasets:
  - We utilized three real world datasets in our experiments:
    - German
    - COMPAS
    - Drug Consumption

$$SPD = |p(\hat{Y} = +1 | x \in \mathcal{D}_a) - p(\hat{Y} = +1 | x \in \mathcal{D}_d)|$$

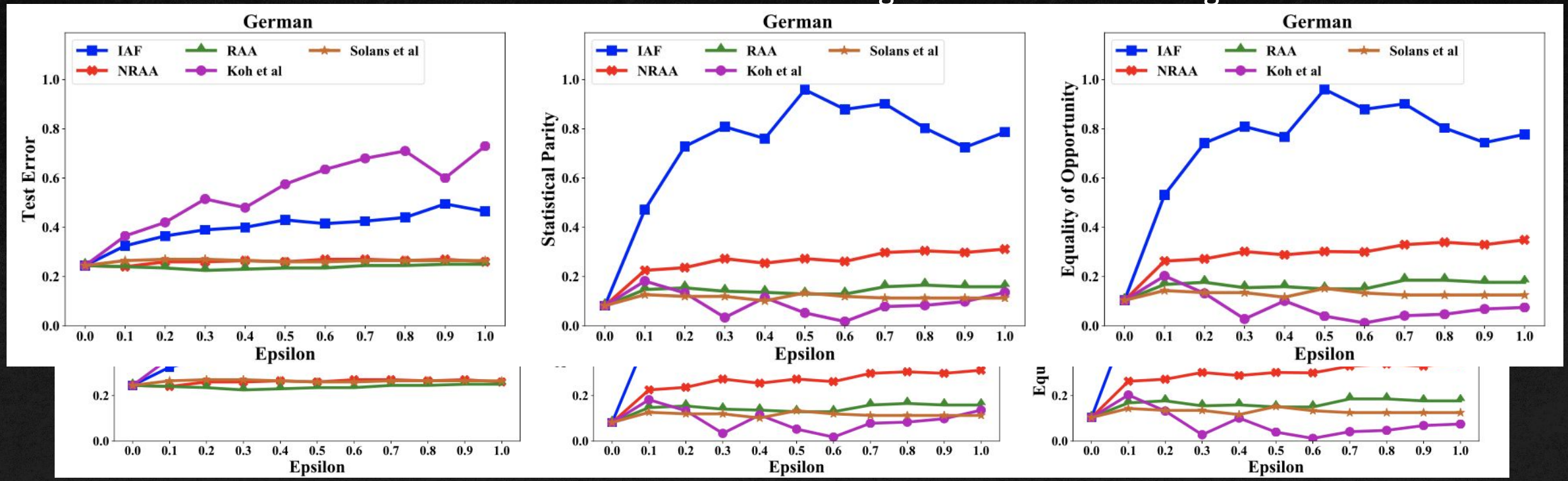
$$EOD = |p(\hat{Y} = +1 | x \in \mathcal{D}_a, Y = +1) - p(\hat{Y} = +1 | x \in \mathcal{D}_d, Y = +1)|$$

D_a: Advantaged Demographic Group
D_d: Disadvantaged Demographic Group.

*Koh, Pang Wei, Jacob Steinhardt, and Percy Liang. "Stronger data poisoning attacks break data sanitization defenses." 2018.
**Solans, David, Battista Biggio, and Carlos Castillo. "Poisoning Attacks on Algorithmic Fairness." 2020.

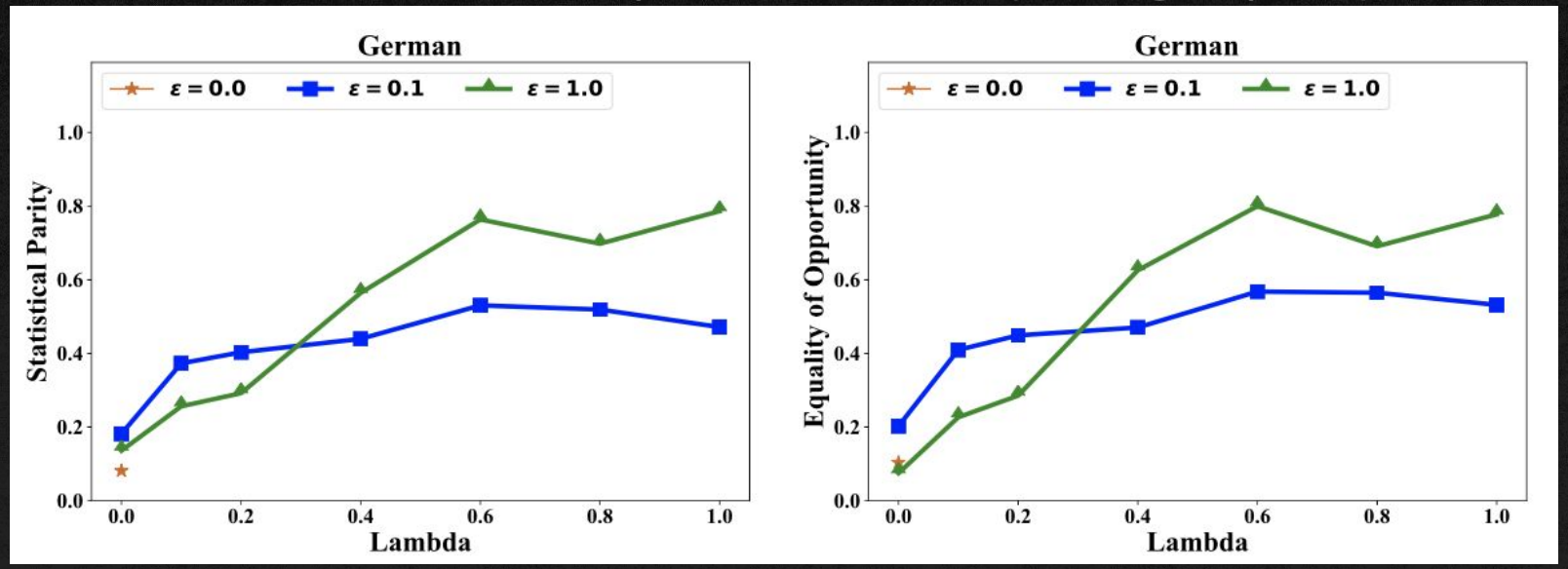
Results and Findings

- The influence attack *on accuracy* is not effective in affecting fairness measures, while it affects accuracy the strongest.
- The influence attack on fairness is shown to be the strongest attack that can target fairness

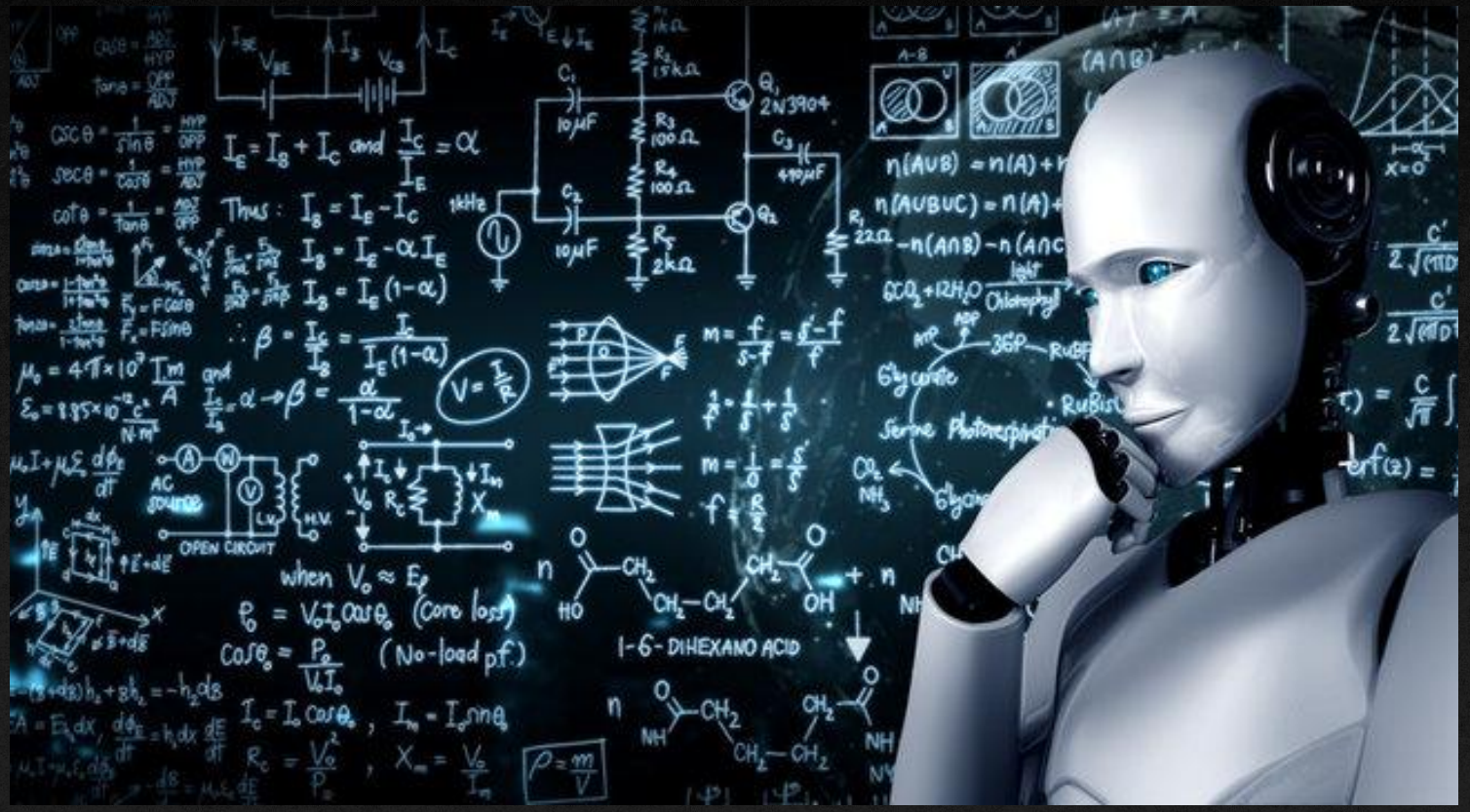


Effect of λ in IAF

- The goal of this experiment was to demonstrate the effect of λ in our influence attack on fairness loss function.
- By increasing λ in influence attack on fairness loss:
  - Harms on fairness becomes more noticeable.
  - This harm is more noticeable with more poisoned instances (i.e., larger epsilon).



Attributing Fair Decisions with Attention Interventions

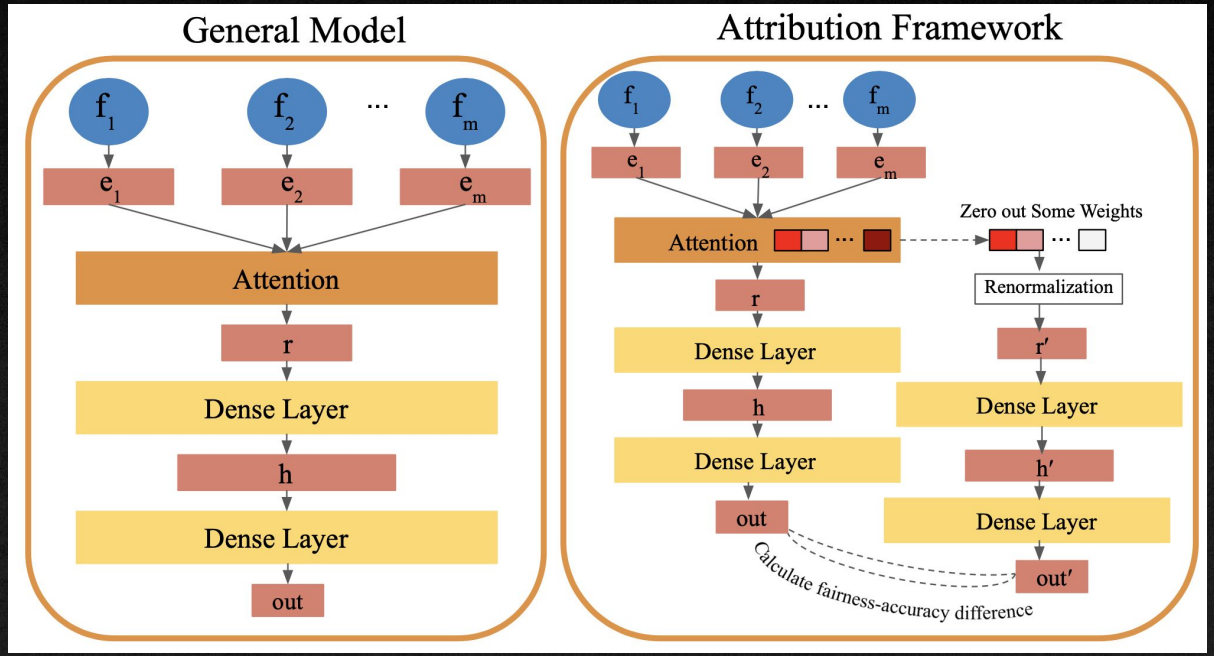


Introduction

- We propose an attribution framework for attention-based classification in tabular data, which is interpretable in the sense that it allows to quantify the effect of each attribute on the outcomes.
- We then use these attributions to study the effect of different input features on the fairness and accuracy of the models.
- Using this attribution framework, we propose a post-processing bias mitigation technique that can reduce unfairness and provide competitive accuracy vs. fairness trade-offs.
- Lastly, we show the versatility of our framework by applying it to non-tabular data such as text.

Attributing Fairness with Attention

- In order to observe the effect of the k^{th} feature:
  - Get the output of the classifier using the model on the left-hand side on the figure with all the attention weights intact.
  - Zero out the attention weight corresponding to the k^{th} feature and obtain the outcome.
  - Observe the outcome differences by calculating difference in the fairness metrics of the original outcome compared to the new outcome as shown on the right-hand side of the figure.
- Notice this can be done for the accuracy as well, as inspired form the work in NLP*.



*Serrano, Sofia, and Noah A. Smith. "Is Attention Interpretable?." *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 2019.

Mitigating Bias w Attention

- The post-processing bias mitigation algorithm with attention:
  - Use the original attention model and obtain all the attention weights for all the k features in all the i samples.
  - Then observe the effect of the k^{th} feature on fairness by zeroing out its attention weight.
  - If the feature contributes to unfairness, add it to the unfair feature set.
  - For all the features in unfair feature set:
    - Use the decay-rate to decrease their weights for all the samples.
  - Use new attention weights and obtain the final outcome.

Algorithm 1: Bias Mitigation with Attention

Input: decay rate d_r $(0 \leq d_r < 1)$, n test samples, indexed by variable i.
Output: final predictions, unfair features.
Calculate the attention weights α_{ki} for k^{th} feature in sample i using the attention layer as in Eq. 1.
unfair_feature_set = $\{\}$
for *each feature (index)* k **do**
  if $SPD(\hat{\mathbf{y}}_o, \mathbf{a}) - SPD(\hat{\mathbf{y}}_z^k, \mathbf{a}) \geq 0$ **then**
    unfair_feature_set = unfair_feature_set $\cup \{k\}$
  end
end
for *each feature (index)* k **do**
  if k *in unfair_feature_set* **then**
    Set $\alpha_{ki} \leftarrow (d_r \times \alpha_{ki})$ for all n samples
  end
end
Use new attention weights to obtain the final predictions $\hat{Y}$.
return $\hat{Y}$, unfair_feature_set

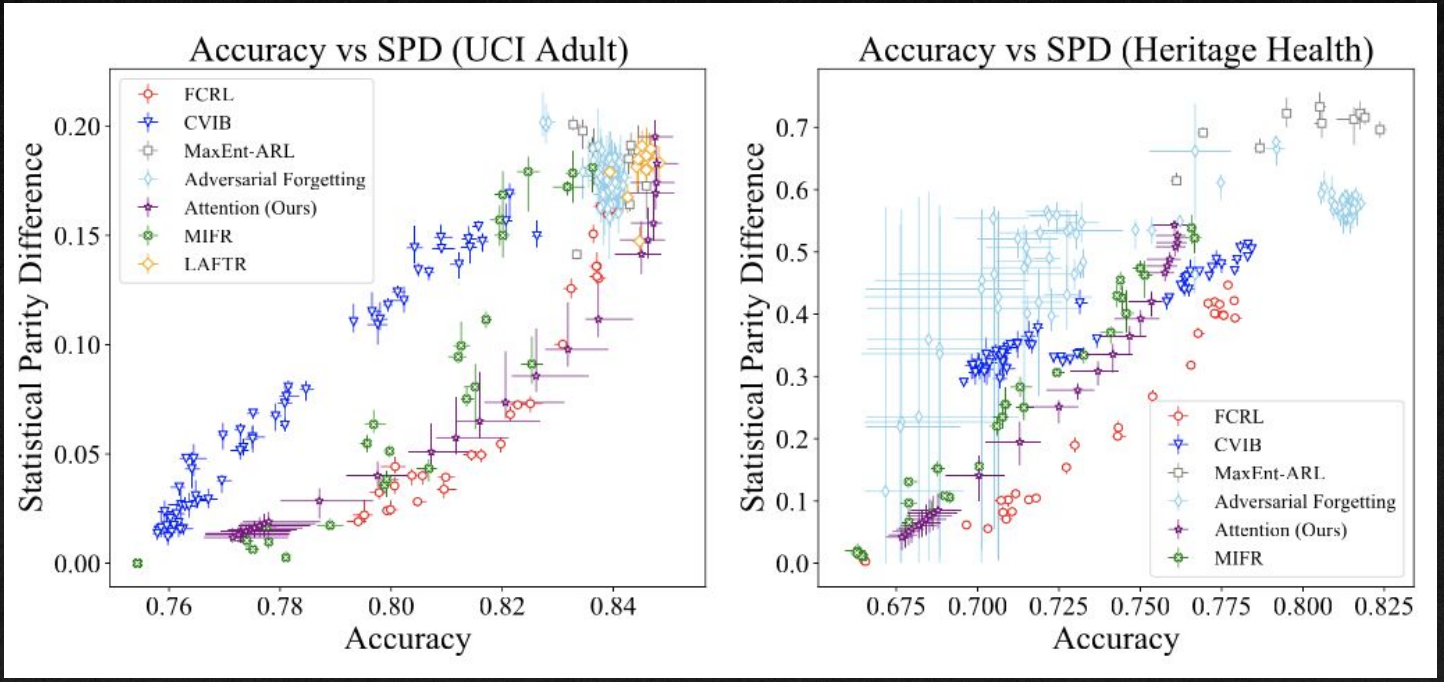
Experimental Setup

- Tabular data:
  - Adult dataset:
    - Prediction task of whether individual's income exceeds 50k or not.
    - Gender as the sensitive attribute.
  - Heritage Health dataset:
    - Prediction task of patient survival.
    - Age as the sensitive attribute.
- Non-Tabular (text) data:
  - Contains bios* of people with the prediction task of whether person's occupation is nurse or dentist.

*De-Arteaga, Maria, et al. "Bias in Bios: A Case Study of Semantic Representation Bias in a High-Stakes Setting." *FAT*. 2019.

Results

- We show that the attention attribution framework is able to identify problematic and unfair features and reduce the bias by the proposed post-processing mitigation technique by having a comparable results to the SOTA.



Advantages

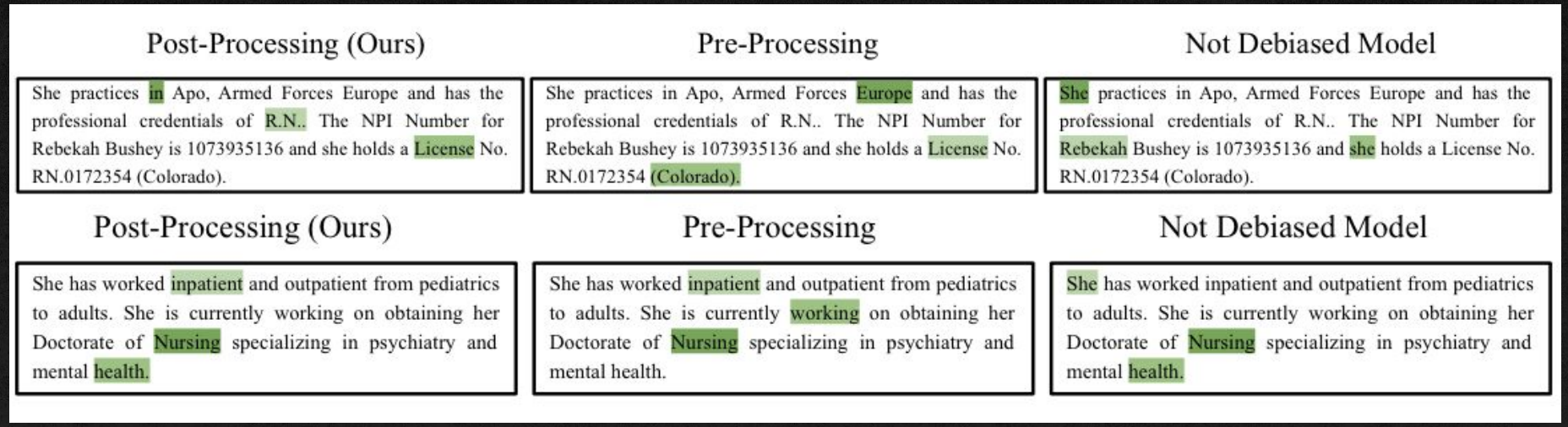
- In addition, our technique has the following advantages:
  - Able to provide explanation such that the users exactly know what feature and by how much it was manipulated to get the corresponding outcome.
  - It needs only one round of training.
  - The adjustments to attention weights are made post-training; thus, it is possible to achieve different trade-offs.
  - Our approach does not need to know sensitive attributes while training; thus, it could work with other sensitive attributes not known beforehand or during training.
  - Is scalable for large scale datasets.

Results on Non-Tabular Data

- Better results compared to the pre-processing baseline as well as the original model.

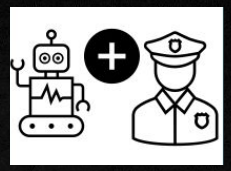
Method	Dentist TPRD (stdev)	Nurse TPRD (stdev)	Accuracy (stdev)
Post-Processing (Ours)	**0.0202 (0.010)**	**0.0251 (0.020)**	0.951 (0.013)
Pre-Processing	0.0380 (0.016)	0.0616 (0.025)	0.946 (0.011)
Not Debiased Model	0.0474 (0.025)	0.1905 (0.059)	**0.958 (0.011)**

- Better qualitative results.

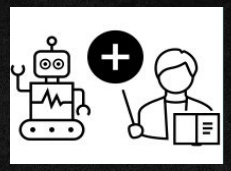


Future Plans

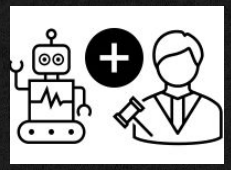
- Continue work on the intersection of fairness and adversarial machine learning with newer types of attacks and techniques.
  - Some work done in the past:
    - Poisoning Attack on Fairness (AAAI 2021).



- Continue work on the intersection of fairness and interpretability. Using explanations to combat unfairness.
  - Some work done in the past:
    - Attributing Fair Decisions (Under Review).



- Auditing different models and resources with regards to fairness.
  - Some work done in the past:
    - Commonsense Reasoning Resources (EMNLP 2021).
    - Community detection algorithms (ASONAM 2019).
    - Named Entity Recognition models (ACM HT 2020).
    - A survey on Bias and Fairness (ACM Computing Surveys 2021).



Conclusion

Covered some of my work done towards trustworthy AI in the intersection of:

Adversarial ML and Fairness

- Data poisoning attacks on fairness

Interpretability and Fairness

- Attributing fair decisions using attention interventions

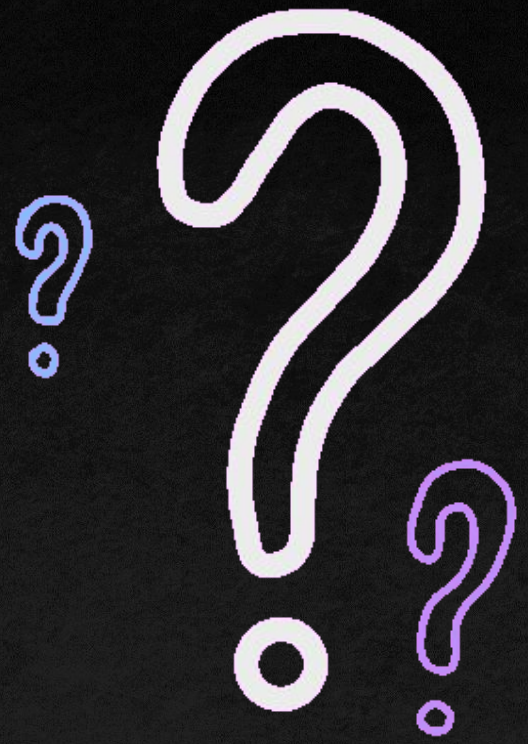
Discussed some future directions and plans

Acknowledgments

USC Viterbi + amazon

References

1. Mehrabi, Ninareh, et al. "Exacerbating Algorithmic Bias through Fairness Attacks." *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 35. No. 10. 2021.
2. Mehrabi, Ninareh, et al. "Attributing Fair Decisions with Attention Interventions." *arXiv preprint arXiv:2109.03952*(2021).
3. Koh, Pang Wei, Jacob Steinhardt, and Percy Liang. "Stronger data poisoning attacks break data sanitization defenses." 2018.
4. Zafar, Muhammad Bilal, et al. "Fairness constraints: Mechanisms for fair classification." *Artificial Intelligence and Statistics*. PMLR, 2017.
5. Solans, David, Battista Biggio, and Carlos Castillo. "Poisoning Attacks on Algorithmic Fairness." 2020.
6. Serrano, Sofia, and Noah A. Smith. "Is Attention Interpretable?." *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 2019.
7. De-Arteaga, Maria, et al. "Bias in Bios: A Case Study of Semantic Representation Bias in a High-Stakes Setting." *FAT*. 2019.
8. Mehrabi, Ninareh, et al. "A survey on bias and fairness in machine learning." *ACM Computing Surveys (CSUR)* 54.6 (2021): 1-35.
9. Mehrabi, Ninareh, et al. "Lawyers are Dishonest? Quantifying Representational Harms in Commonsense Knowledge Resources." *EMNLP* 2021.
10. Mehrabi, Ninareh, et al. "Debiasing community detection: The importance of lowly connected nodes." *2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE, 2019.
11. Mehrabi, Ninareh, et al. "Man is to person as woman is to location: Measuring gender bias in named entity recognition." *Proceedings of the 31st ACM Conference on Hypertext and Social Media*. 2020.



Questions