FAT ML

Dominika Basaj

Fairness

- How should we define, measure, and deal with biases in training data sets? Can we design data collection practices that limit the effect of bias? How can we use additional sources of information to assess and correct for bias?
- What are meaningful formal fairness criteria? How do different criteria relate and trade-off? What are their limitations?
- Should we turn to the law for definitions of fairness? Are proposed formal fairness criteria reconcilable with the law?
- How can we use the tools of causal inference to reason about fairness in machine learning? Can causal inference lead to actionable recommendations and interventions? How can we design and evaluate the effect of interventions?
- Can we develop definitions of discrimination and disparate impact that move beyond distributional constraints such as demographic parity or the 80% rule?
- Who should decide what is fair when fairness becomes a machine learning objective?
- Are there any dangers in turning questions of fairness into computational problems?
- What are the societal implications of algorithmic experimentation and exploration? How can we manage the cost that such experimentation might pose to individuals?

Accountability

- What would human review entail if models were available for direct inspection?
- Are there practical methods to test existing algorithms for compliance with a policy?
- Can we prove that an algorithm behaves in some way without having to reveal the algorithm? Can we achieve accountability without transparency?
- How can we conduct reliable empirical black-box testing and/or reverse engineer algorithms to test for ethically salient differential treatment?
- Can we demonstrate the causal origins of the outcome predicted by a model?
- What constitutes sufficient evidence to someone other than the creator of a model that the model functions as intended? Can we describe the goals of modeling effectively?
- What are the societal implications of autonomous experimentation? How can we manage the risks that such experimentation might pose to users?

Transparency

- How can we develop interpretable machine learning methods that provide ways to manage the complexity of a model and/or generate meaningful explanations?
- Can we field interpretable methods in a way that does not reveal private information used in the construction of the model?
- Can we use adversarial conditions to learn about the inner workings of inscrutable algorithms? Can we learn from the ways they fail on edge cases?
- How can we use game theory and machine learning to build fully transparent, but robust models using signals that people would face severe costs in trying to manipulate?

Women also Snowboard: Overcoming Bias in Captioning Models

https://arxiv.org/pdf/1807.00517.pdf

Key issues:

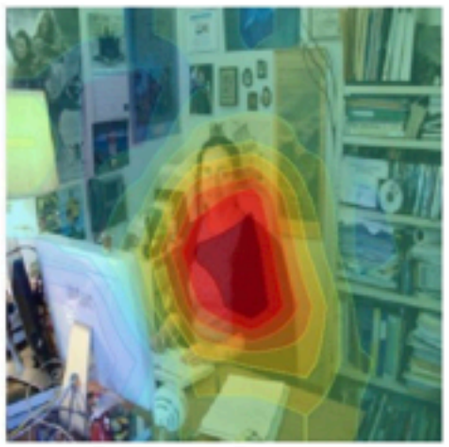
- Image captioning models tend to exaggerate biases present in training data
- Paper investigates generation of gender specific caption words
- It introduces a new Equalizer model that ensures equal gender probability when gender evidence is occluded

When description models predict gendered words such as “man” or “woman”, they should consider visual evidence associated with the described person, and not contextual cues like location (e.g., “kitchen”) or other objects in a scene (e.g., “snowboard”)

Women also Snowboard: Overcoming Bias in Captioning Models

Wrong

Baseline:
*A **man** sitting at a desk with a laptop computer.*

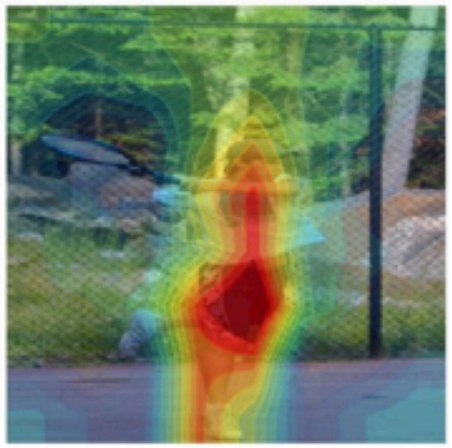


Right for the Right Reasons

Our Model:
*A **woman** sitting in front of a laptop computer.*

Right for the Wrong Reasons

Baseline:
*A **man** holding a tennis racquet on a tennis court.*



Right for the Right Reasons

Our Model:
*A **man** holding a tennis racquet on a tennis court.*

Women also Snowboard: Overcoming Bias in Captioning Models

Equalizer model

Appearance Confusion Loss (ACL)

- based on the intuition that, given an image in which evidence of gender is absent, description models should be unable to accurately predict a gendered word
- to optimize the Appearance Confusion Loss, we require ground truth rationales indicating which evidence is appropriate

Confident Loss (Conf)

- increase the model's confidence when gender is in the image

If evidence to support a specific gender decision is not present in an image, the model should be confused about which gender to predict (enforced by an Appearance Confusion Loss term), and if evidence to support a gender decision is in an image, the model should be confident in its prediction (enforced by a Confident Loss term)

Women also Snowboard: Overcoming Bias in Captioning Models

Model	MSCOCO-Bias Error	MSCOCO-Bias Ratio	MSCOCO-Confident Error	MSCOCO-Confident Ratio	MSCOCO-Balanced Error	MSCOCO-Balanced Ratio	MSCOCO-Balanced Pointing Game
GT	-	0.466	-	0.548	-	1.000	-
Baseline-FT	0.129	0.265	0.143	0.384	0.203	0.597	40.7
Balanced	0.129	0.270	0.142	0.393	0.204	0.610	37.4
UpWeight	0.134	0.315	0.116	0.472	0.157	0.712	46.0
Equalizer w/o ACL	0.079	0.369	0.081	0.499	0.106	0.777	46.8
Equalizer w/o Conf	0.098	0.318	0.116	0.425	0.165	0.673	40.5
Equalizer	**0.070**	**0.437**	**0.071**	**0.563**	**0.081**	**0.973**	**48.7**

error rate = the number of man/woman misclassifications

ratio = gender ratio of sentences which belong to a “woman” set to sentences which belong to a “man” set

Achieving Fairness through Adversarial Learning: an Application to Recidivism Prediction

http://www.fatml.org/media/documents/achieving_fairness_through_adversearial_learning.pdf

- COMPAS - Correctional Offender Management Profiling for Alternative Sanctions a risk assessment and recidivism prediction score

- COMPAS performs similarly in terms of accuracy for white and black inmates, but the errors COMPAS makes indicate discrimination against black inmates

- For example, a black inmate who does not re-offend is more likely to be classified as “high risk” than a white inmate who does re-offend.

- Even if race is not an input feature, other features are correlated with race. For example, an inmate’s number of priors and previous time in jail are correlated with race because black inmates are more likely to be jailed for the same crimes as white inmates and are arrested at a higher rate for less serious crimes

Achieving Fairness through Adversarial Learning: an Application to Recidivism Prediction

Fairness definition

- **Demographic parity**

  A score S = S(x) satisfies parity if the proportion of individuals classified as high-risk is the same for each demographic.

- **Equality of odds**

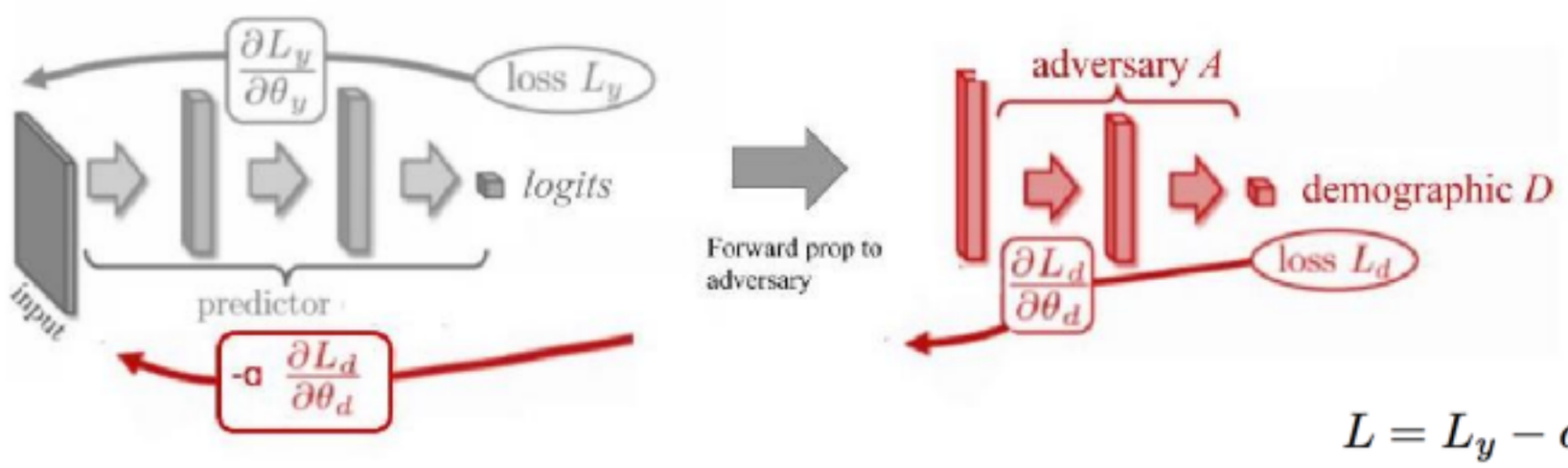
A score S = S(x) satisfies equality of odds if the proportion of individuals classified as high-risk is the same for each demographic, when true future recidivism is held constant. White and black inmates that do recidivate should have the same proportion of high risk classification

- **Calibration**

A score S = S(x) is calibrated if it reflects the same likelihood of recidivism irrespective of the individual's demographic. In this application, black inmates who are classified as high risk should have the same probability of true recidivism as white inmates classified as high risk.

Achieving Fairness through Adversarial Learning: an Application to Recidivism Prediction

Our goal is for neural network N to predict Yˆ accurately and for A to predict D poorly



We input the logit from N (the unnormalized predicted recidivism probability, i.e. just before the sigmoid) to an adversarial neural network A that learns to classify demographic D. If Yˆ is biased for demographic D, A should learn to have a high accuracy because the logit will be highly predictive of D. Our goal is for neural network N to predict Yˆ accurately and for A to predict D poorly

Achieving Fairness through Adversarial Learning: an Application to Recidivism Prediction

High Risk Gap: $|HighRisk_{white} - HighRisk_{black}|$

False Positive Gap: $|FP_{white} - FP_{black}|$

False Negative Gap: $|FN_{white} - FN_{black}|$

Model	High Risk Gap	FN Gap	FP Gap
COMPAS Scores (Our Test Set)	0.18	0.22	0.17
Our Recidivism Model	0.21	0.27	0.15
Our Chosen Adversarial Model	**0.02**	**0.02**	**0.01**

Model	AUC
COMPAS Scores (Our Test Set)	0.66
Our Recidivism Model	0.72
Our Chosen Adversarial Model	0.70

Model	Accuracy	FP Gap	FN Gap
COMPAS Scores (Our Test Set)	0.68	0.17	0.22
Our Recidivism Model	0.70	0.15	0.27
Our Chosen Adversarial Model	0.70	**0.01**	**0.02**
Bechavod et al. AVD Penalizers (2017)	0.65	**0.02**	**0.04**
Bechavod et al. SD Penalizers (2017)	0.66	**0.02**	**0.03**
Bechavod et al. Vanilla Regularized (2017)	0.67	0.20	0.30
Zafar et al. (2017)	0.66	**0.03**	0.11
Zafar et al. Baseline (2017)	0.66	**0.01**	0.09
Hardt et al. (2016)	0.65	**0.01**	**0.01**