

Day 3

Speaker: Prof. Chetan Arora, IITD

Title: Fairness in Visual Recognition Tasks.

Success of DNN Models in Computer Vision

Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton
ImageNet classification with deep convolutional neural networks. NIPS 2012

Breakthrough in DL → Accuracy ↑ Even Human Indistinguishable

* Mite → Mite → Black widow → Cockroach

} Top 4 results are visually similar.
Eg. mite, leopard

Chetan Arora
Department of Computer Science and Engineering, IIT Delhi

DNNs: Generic Features!

Input → Column

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton
ImageNet Classification with Deep Convolutional Neural Networks. NIPS 2012

1000D \longrightarrow Next 6 {Nearest Image Based on Euclidean distance } [Embedding Vectors]
(Input)

* Input \xrightarrow{NN} Transformation / Embedding Vectors

All Good !



Chetan Arora
Department of Computer Science and Engineering, IIT Delhi

But How Little We Really “Understand” DNNs

- The effective capacity of neural networks is sufficient for memorizing the entire data set (in this case CIFAR 10, and ImageNet).
- State-of-the-art convolutional networks for image classification trained with stochastic gradient methods easily fit a random labeling of the training data.
- This phenomenon is qualitatively unaffected by explicit regularization and occurs even if we replace the true images by completely unstructured random noise.

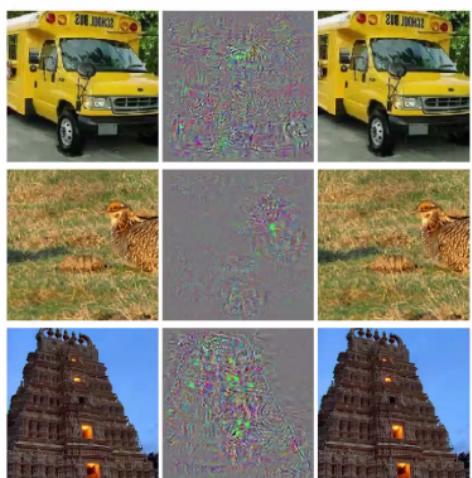
C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals.
Understanding deep learning requires rethinking generalization.

* Labelling had no significance. * Image memorization was taking place actually.
* Not affected by regularization .



Chetan Arora
Department of Computer Science and Engineering, IIT Delhi

But How Little We Really “Understand” DNNs



Ostrich

Average Distortion: 0.006

* Read Paper

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus
Intriguing properties of neural networks. ICLR 2014

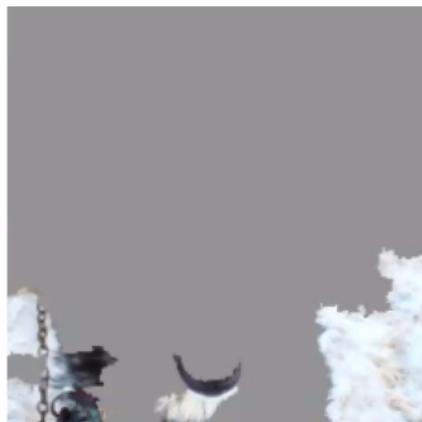
Image $\xrightarrow{\text{white Noise}}$? {All predicted Ostrich)
Prediction [Same Distribution Dataset \rightarrow Bad]



Can We Trust Deep Network Predictions?



Husky classified as wolf



M. T. Ribeiro, S. Singh, and C. Guestrin.
Why should I trust you?: Explaining the predictions of any classifier.

→ Which part the n/w thinks are important for classification?

Evaluation: Here, snow is considered a imp. part for determining a wolf.

∴ Embedding may not say, we learnt the semantic meaning of the data but the trends in the similarly distributed dataset.

For eg. Here all the wolf images had a snowy background, so maybe the snow was generalized as the feature of the snow



Can We Trust Deep Network Predictions?



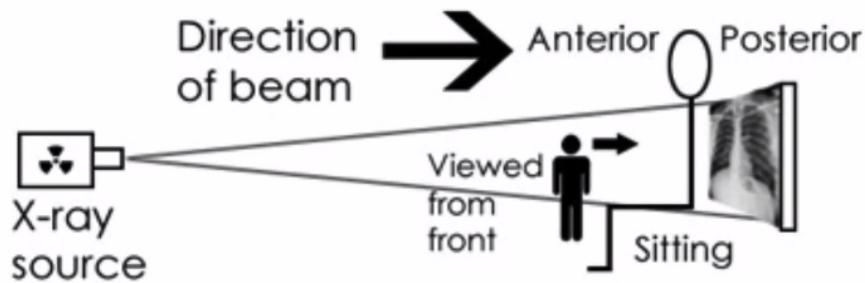
Classified Positive for Pneumonia



Classified Negative for Pneumonia



AP vs PA View



- * Usually Patient → PA View (Sitting with source in front)
If Patient can't stand → AP View

Here, N/w may get biased with AP & PA View rather than

Pneumonia or non-Pneumonia. (∴ Almost all AP View Patients are having Pneumonia)

- * Dataset always have the possibility of not learning its semantic meaning



Systematic Bias: Appearance

ImageNet

Chairs



Chairs by rotation



Chairs by background

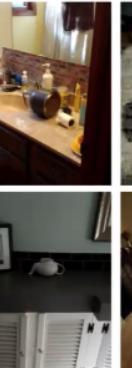


ObjectNet

Chairs by viewpoint



Teapots



T-shirts



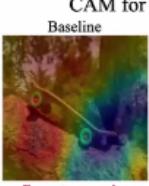
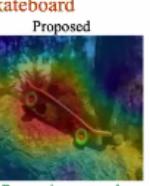
<https://objectnet.dev/>

ImageNet → Organised Data. { ObjectNet → Wild Images of data.

∴ Systematic Biases Occur.



Systematic Bias: Context

| Cause of bias |  | vs. |  | |
|---------------|--|--|--|--|
| Effect |  Learning from the wrong thing |  Learning from the right thing |  Cannot recognize when context is absent |  Recognize even when context is absent |

Krishna Kumar Singh, Dhruv Mahajan, Kristen Grauman, Yong Jae Lee, Matt Feiszli, Deepti Ghadiyaram
Don't Judge an Object by Its Context: Learning to Overcome Contextual Bias. CVPR 2020

- * ↑ Accuracy → doesn't mean it is right for the right reason.
- * Adversarial Attacks ^{* Intended} [To study the weakness of the NN]
- To read → One Pixel Attack Paper.
- * Kitchen from different countries



Systematic Bias → Fairness

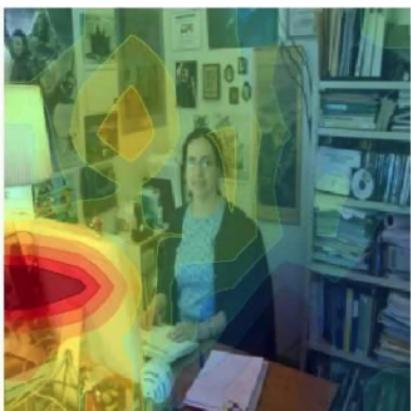


<https://www.gapminder.org/dollar-street>

Systematic Bias would occur if the n/w fails to recognise kitchens from a particular country. [Model Decisions are affecting some real people] → Question of Fairness arises



Systematic Bias → Fairness



→ Image Caption

A man sitting at a desk with a laptop computer



women??

* May be the dataset shows man-computer correlation.
Biased dataset Fairness??

Image Source: Women Also Snowboard: Overcoming Bias in Captioning Models. ECCV 2018

For ImageNet Experiment, { 256X256 Resolution → Both man/women differentiated
64X64 " → Only man.

* If the network is biased against a particular group or class → *

* Systematic Bias or Fairness Issue in CV.

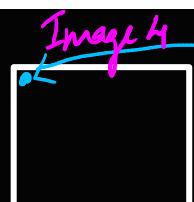
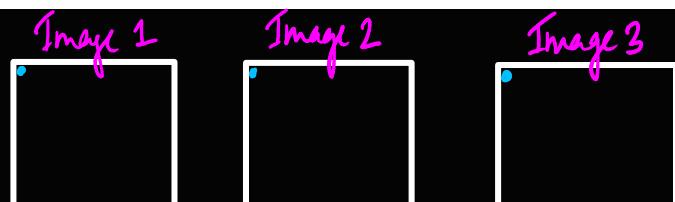
{ Size of the dataset is not a problem but the quality of the dataset. }
[To know if the dataset is having higher order correlation]



Fairness

- CV systems based on DNNs have been widely adopted in critical areas.
- Trained on large datasets that contain biases or spurious variations
- DNNs have been shown to inherit these biases and make decisions based on the sensitive, and protected attributes.

Unintended bias of large datasets.



in top left corner
→ Blue Pixel adds
Systematic Bias.

Example

Case
Study



Chetan Arora
Department of Computer Science and Engineering, IIT Delhi

Fairness in ML Systems

- Names that are associated with black people are found to be significantly more associated with unpleasant than with pleasant terms, compared to names associated with whites.
- The models learned on such text data for opinion or sentiment mining have a possibility of inheriting the prejudices reflected in the data.

A. Caliskan-Islam, J. J. Bryson, and A. Narayanan.

Semantics derived automatically from language corpora necessarily contain human biases.

Case Study:



Chetan Arora
Department of Computer Science and Engineering, IIT Delhi

Fairness in ML Systems

- Strong ethnic bias in COMPAS score: a predictive model for the “risk of crime recidivism”
- According to the score:
 - A black who did not re-offend were classified as high risk twice as much as whites who did not re-offend, and
 - White repeat offenders were classified as low risk twice as much as black repeat offenders

<http://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>

→ Not all Biases are Bad!



Inductive Bias → Helps with the model tasks.

Set of assumptions that the learner uses to predict outputs given inputs that it has not encountered

Check on multiple models & pick one with minimum cross validation error

Binary classification

- **Minimum cross-validation error:** When trying to choose among hypotheses, select the hypothesis with the lowest cross-validation error.
- **Maximum margin:** when drawing a boundary between two classes, attempt to maximize the width of the boundary.



Inductive Bias

Set of assumptions that the learner uses to predict outputs given inputs that it has not encountered

- **Minimum features:** Delete a feature unless there is good evidence that it is useful. This is the assumption behind feature selection algorithms.
- **Nearest neighbors:** Assume that most of the cases in a small neighborhood in feature space belong to the same class. Used in the k-nearest neighbors and many manifold learning algorithms.

The above two biases are desirable biases we chosen for our system.

But there are undesirable biases too.



Undesirable Biases in ML Systems

• Selection Bias

- Bias introduced by the selection of data for analysis in such a way that proper randomization is not achieved, thereby ensuring that the sample obtained is not representative of the population

• Reporting Bias

- Selective revealing or suppression of information. Authors under-reporting unexpected or undesirable experimental results

Bad { Unjustified dataset}



Good.
Just differentiate
2 classes

Differentiation Vs Discrimination

- Differentiation is the basis of prediction in any ML system.

- Unjustified basis for differentiation is Discrimination

- Practical irrelevance**

- Gender, Religion, Caste, Financial Status etc.

- Moral irrelevance**

- Caste, Disability etc.

Solon Barocas and Moritz Hardt
Fairness in Machine Learning. NIPS 2017 Tutorial



Doctrines of Discrimination

1. Disparate Treatment → Treating 2 groups differently.

- Formal**

- Using protected attributes explicitly

→ {
• **Intentional**

- Using proxies for the protected attributes

Eg. Indirect action.



Doctrines of Discrimination

2. Disparate Impact → Impact is different for 2 groups.

- **Unjustified**

- Predicted outcome is significantly different for two groups

- **Avoidable**

- There is way to achieve the same score with lesser disparate impact



Does ML Prevent Disparate Treatment?

- Automated decision leaves no scope of human discretion.
- Model learns what the data supports
- Protected attributes can be withheld

In favour of ML
i.e. no bias in the model.
??
No issue with ML



How Machines Learn to Discriminate

- **Tainted examples**

- Unreliable labels
- Group A are tax evaders.

- **Proxies**

- Considering features that are correlated with protected attribute

Biased
sample
Selection of
caused
discrimination.



How Machines Learn to Discriminate

MacBook → Rich
Correlation
may cause Bias.
Data discrimination

- Limited features
 - Features may be less informative or less reliably collected for certain parts of the population
 - A feature set that supports accurate predictions for the majority group may not for a minority group
- Sample size disparity
 - Objective functions are biased against the minority classes



Problem Formulation

Example: A company wants to hire a software engineer (SWE) and is going to advertise for the same. An ML system needs to predict which persons to show the advertisement based upon if he/she is currently a SWE.

- Subset [
- X : features of an individual (browsing history etc.)
 - A : sensitive or protected attributes (gender etc.) \rightsquigarrow Subset of X , No division bet^n $X \& A$
 - $C = c(X, A)$: predicted score/class (show ad or not). Also denoted as \hat{Y}
 - Y : target variable (whether the person is a SWE)

Notation: $\mathbb{P}_a\{E\} = \mathbb{P}\{E|A = a\} \rightarrow$ Prob of an event E , given a protected attribute, $A = a = \mathbb{P}_a\{E\}$ (May be biased)



Fundamental Criterion for Fairness

- **Independence:** \hat{Y} independent of $A \rightarrow$ Prediction is independent of the protected attribute
- **Separation:** \hat{Y} independent of A conditional on Y
- **Sufficiency:** Y independent of A conditional on \hat{Y}



First criterion: Independence

- Require \hat{Y} and A to be independent. Denoted as $\hat{Y} \perp A$

- That is, for all groups g_1, g_2 and all predictions c :

Prob. of Predicting a class c of group g_1 [from A] $\rightarrow \mathbb{P}_{g_1}\{\hat{Y} = c\} = \mathbb{P}_{g_2}\{\hat{Y} = c\}$ \rightarrow Prob. of Predicting the same class c of group g_2

- When \hat{Y} is a binary 0/1-variable:

↳ Same for Binary. $\mathbb{P}_{g_1}\{\hat{Y} = 1\} = \mathbb{P}_{g_2}\{\hat{Y} = 1\}$,

- Also called **demographic parity**, or **statistical parity**

* \hat{Y} = Score for reoffending | If model is independent, the Prob. of reoffending for Blacks & White must be same.



First criterion: Independence

Approximate versions

- $\frac{\mathbb{P}_{g_1}\{\hat{Y}=1\}}{\mathbb{P}_{g_2}\{\hat{Y}=1\}} \geq 1 - \epsilon$

- Also known in the legal context as "80% Rule". Declare unfair if the impact differs by more than 20%

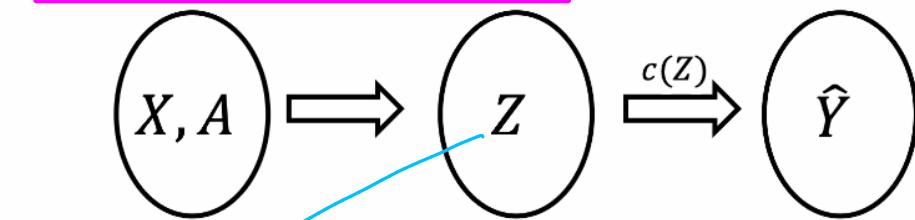
- Additive version: $|\mathbb{P}_{g_1}\{\hat{Y} = 1\} - \mathbb{P}_{g_2}\{\hat{Y} = 1\}| \leq \epsilon$



Achieving Fairness by Independence

Representation Learning Approach

- Maximize Mutual Information between X and latent representation Z while minimizing the same between A and Z



Latent Representation

$$I(X, Y) = \sum_{x \in X} \sum_{y \in Y} \mathbb{P}(x, y) \log \frac{\mathbb{P}(x, y)}{\mathbb{P}(x)\mathbb{P}(y)}$$

More 2 quantities are diff higher the Mutual Information bet'n them

* Must be
↳ $Z \perp A$
to
 $\hat{Y} \perp A$



Learning Not to Learn

→ To read

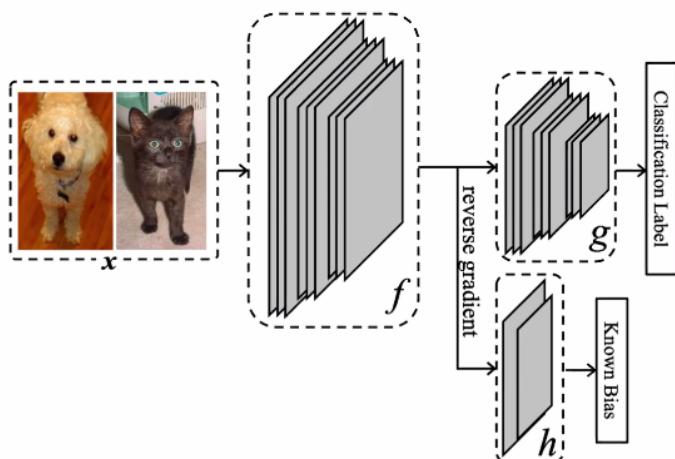


Image Source: Learning Not to Learn: Training Deep Neural Networks with Biased Data. CVPR 2019.

* To remove the class bias ↑ [To Be Updated]



Problems with Independence based Fairness

- • Ignores possible correlation between Y and A :
 - Recall our example problem: predict which persons to show the advertisement based upon if he/she is currently a SWE.
 - There could be less woman software engineers.
- Rules out perfect predictor $\hat{Y} = Y$.
- Permits laziness: Accept the qualified in one group, random people in other.
 - Allows to trade false negatives for false positives: Can create random false positives to make $\mathbb{P}_{g_1}\{\hat{Y} = 1\} = \mathbb{P}_{g_2}\{\hat{Y} = 1\}$.

} Dataset may likely have some correlation.

} Eg. 30% correct pred & other 20% random.



Fundamental Criterion for Fairness

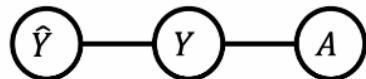
- **Independence:** \hat{Y} independent of A
- **Separation:** \hat{Y} independent of A conditional on Y
- **Sufficiency:** Y independent of A conditional on \hat{Y}



Second criterion: Separation

- Require \hat{Y} and A to be independent, conditional on target variable Y
- Denoted as $\hat{Y} \perp A | Y$
- For all the groups g_1, g_2 and all values c and y :

$$\mathbb{P}_{g_1}\{\hat{Y} = c | Y = y\} = \mathbb{P}_{g_2}\{\hat{Y} = c | Y = y\}$$
- PGMs: Random variable \hat{Y} separated from A if $\hat{Y} \perp A | Y$



Second criterion: Separation

$$\mathbb{P}_{g_1}\{\hat{Y} = c | Y = y\} = \mathbb{P}_{g_2}\{\hat{Y} = c | Y = y\}$$

- Also known as **Equalized Odds OR Equal Opportunity**
- For the outcome $Y = 1$, the constraint requires that \hat{Y} has equal true positive rates across the two demographics $A = a$ and $A = b$.
- For $Y = 0$, the constraint equalizes false positive rates.

Let $Y = 1, g_1 = \text{women}$
 $Y = \text{SWE}, \hat{Y} = \text{SWE or}$
 not SWE

$$\begin{aligned} & P(\text{no. of men / SWE}) \\ &= P(\text{no. of women / SWE}) \end{aligned}$$

model is unfair b/w
 many women in the
 same 1. Optimal
 ↓

* Equalizing the error rates for all classes



Desirable Properties of Separation

$$\mathbb{P}_a\{\hat{Y} = c | Y = y\} = \mathbb{P}_b\{\hat{Y} = c | Y = y\}$$

Bad but
 fair model

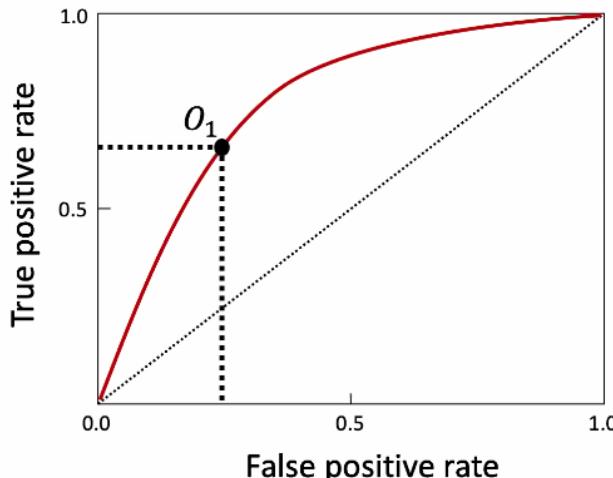
- Optimality compatibility: $\hat{Y} = Y$ is allowed
- Penalizes laziness: Incentive to reduce errors uniformly in all groups
 - Recall, none of the above is achieved by Independence.
- Assume Y is unbiased: Problematic with tainted and skewed samples

To read
→

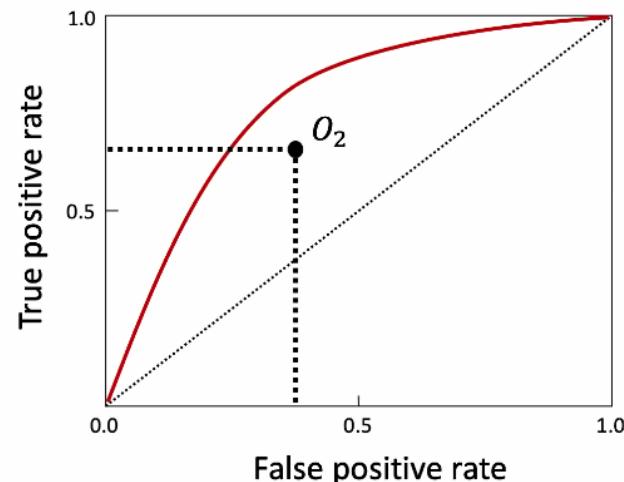
[To Be Updated]



Achieving Fairness by Separation



Achieving Fairness by Separation

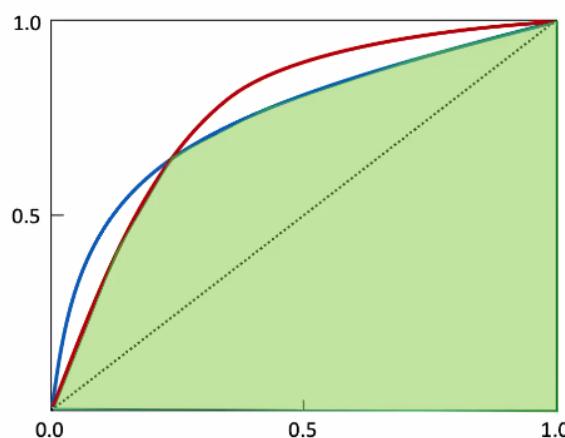


Create Random False Positives

Operating point O_1 can always be converted to O_2 by creating random false positives



Achieving Fairness by Separation



If Red and Blue are the curves corresponding to the two groups, then any point in the green region is feasible and satisfies equal opportunity



Fundamental Criterion for Fairness

- **Independence:** \hat{Y} independent of A
- **Separation:** \hat{Y} independent of A conditional on Y
- **Sufficiency:** Y independent of A conditional on \hat{Y}



Third criterion: Sufficiency

- Random variable \hat{Y} is sufficient for A , if $Y \perp A | \hat{Y}$



- For the purpose of predicting Y , we don't need to see A when we have \hat{Y}

$$\mathbb{P}_{g_1}\{Y = y | \hat{Y} = c\} = \mathbb{P}_{g_2}\{Y = y | \hat{Y} = c\}$$

- Also called **Predictive Rate Parity**

* Computer Company had PRP Passed but the model failed

If a person is offending then the prob. of person offending will be same across both classes (Black & White)



Third criterion: Sufficiency

$$Y \perp A | \hat{Y}$$

- For the purpose of predicting Y , we don't need to see A when we have \hat{Y}

$$\mathbb{P}_a\{Y = y | \hat{Y} = c\} = \mathbb{P}_b\{Y = y | \hat{Y} = c\}$$

Examples:

- For the purpose of showing the SWE advertisement, I don't need to know gender, if I already know the programming score.
- For the purpose of approving loan, I don't need to know the race, when I know the credit rating



Achieving Fairness by Sufficiency

$$Y \perp A \mid \hat{Y}$$

- Sufficiency satisfied by Bayes optimal score:

$$\hat{Y} = r(X, A) = \mathbb{E}[Y \mid X = x, A = a]$$

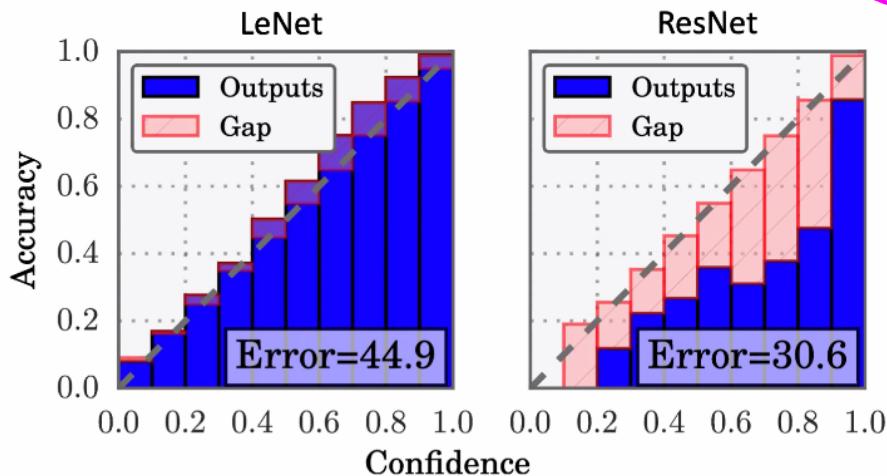
- Can be achieved using **calibration by group**: $\mathbb{P}\{Y = 1 \mid \hat{Y} = c, A = a\} = r$

Solon Barocas and Moritz Hardt

Fairness in Machine Learning. NIPS 2017 Tutorial



Neural Network Calibration



* NN Networks are overconfident / underconfident
 → Accuracy Scores can't actually say we are accurate

Chuan Guo, Geoff Pleiss, Yu Sun, Kilian Q. Weinberger
On Calibration of Modern Neural Networks. ICML 2017



Achieving Fairness by Sufficiency

$$\mathbb{P}_{g_1}\{Y = 1 \mid \hat{Y} = r\} = \mathbb{P}_{g_2}\{Y = 1 \mid \hat{Y} = r\}$$

- Formally, we say that a score R is calibrated if for all score values r in the support of R we have

$$\mathbb{P}\{Y = 1 \mid \hat{Y} = r\} = r$$

- Fairness by sufficiency can be achieved by group-wise calibrating the scores

* Achieving Fairness by Sufficiency

$$\mathbb{P}_g\{Y = 1 | \hat{Y} = r\} = r$$

→ Plat Scaling:

- Given uncalibrated score \hat{r} , fit a sigmoid function: $S = \frac{1}{1+\exp(\alpha\hat{r}+\beta)}$
- Minimize log loss: $-\mathbb{E}[r \log S + (1-r)\log(1-S)]$

Expectation

We want to make \hat{r} closer to r

Pass \hat{Y}
by another
sigmoid
layer



Achieving Fairness by Sufficiency

$$\mathbb{P}_g\{Y = 1 | \hat{Y} = r\} = r$$

Plat Scaling:

- Minimize log loss: $-\mathbb{E}[r \log S + (1-r)\log(1-S)]$, where $S = \frac{1}{1+\exp(\alpha\hat{r}+\beta)}$
- Objective function nudges S to take similar value as r (differentiate w.r.t. S)

$$\bullet \frac{r}{S} - \frac{1-r}{1-S} = 0. \Rightarrow \frac{r}{S} = \frac{1-r}{1-S} \Rightarrow r(1-S) = S(1-r). \Rightarrow r = S$$



Fairness Tradeoffs

**Any two of the three criteria we saw are
mutually exclusive except in degenerate cases.**

For example:

- If $A \not\perp Y$ and $\hat{Y} \not\perp Y$, then either independence or separation holds, not both.

* Any two of the three fairness criteria cannot be true at the same time.

Mathematical Proof:



Fairness Tradeoffs

- If $A \not\perp Y$ and $\hat{Y} \not\perp Y$, then either independence or separation holds, not both.
- ✓ • Recall that Separation $\Rightarrow \hat{Y} \perp A | Y$, and Independence $\Rightarrow \hat{Y} \perp A$
- ✓ • If separation holds then $\frac{\mathbb{P}\{\hat{Y}, A\}}{\mathbb{P}\{Y\}} = \frac{\mathbb{P}\{\hat{Y}\}}{\mathbb{P}\{Y\}} \times \frac{\mathbb{P}\{A\}}{\mathbb{P}\{Y\}}$
- ✓ • If independence holds then $\frac{\mathbb{P}\{\hat{Y}, A\}}{\mathbb{P}\{Y\}} = \frac{\mathbb{P}\{\hat{Y}\}\mathbb{P}\{A\}}{\mathbb{P}\{Y\}}$.
- The two can not be true at the same time, unless $A \perp Y$ or $\hat{Y} \perp Y$

Case Study:



The COMPAS Debate



Bernard Parker, left, was rated high risk;
Dylan Fugett, right, was rated low risk.
(Josh Ritchie for ProPublica)

Machine Bias

- There's software used across the country to predict future criminals. And it's biased against blacks.

Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica May 23, 2016



Essence of COMPAS debate

ProPublica's main charge: Black defendants face higher false positive rate.

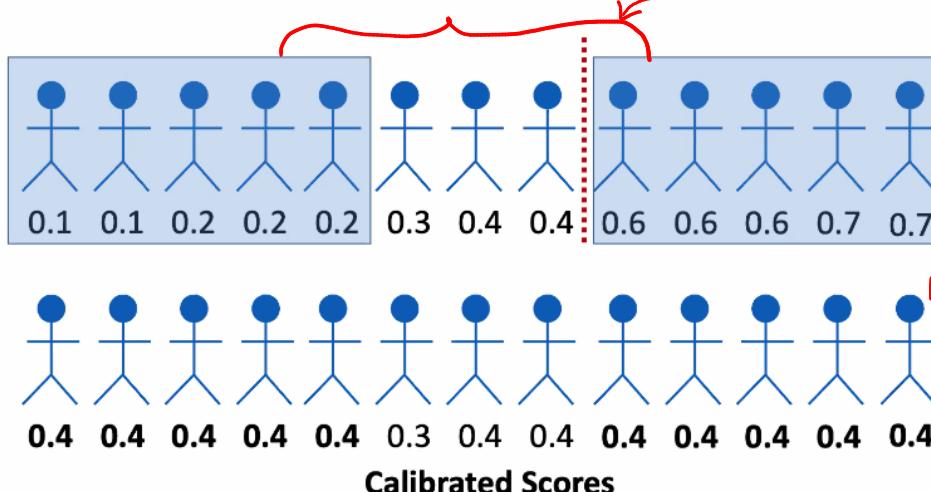
Northpointe's main defence: Scores are calibrated by group.

- Recall that sufficiency $\Rightarrow Y \perp A | R$, and can be achieved using **calibration by group**: $\mathbb{P}\{Y = 1 | R = r, A = a\} = r$ implemented by Platt Scaling

Though we discuss only calibration here, but neither calibration nor equality of false positive rates rule out blatantly unfair practices



Calibration is Insufficient



Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, Aziz Huq
Algorithmic decision making and the cost of fairness. KDD 2017



Calibration is Insufficient: Another Example

| | White | Black |
|--------------------------|-------|-------|
| 0 previous convictions | 5% | 5% |
| 1-2 previous convictions | 20% | 20% |
| 3+ previous convictions | 40% | 40% |
| Average recidivism | 20% | 20% |

Sam Corbett-Davies, Stanford University
Optimization and Fairness Symposium



Calibration is Insufficient: Another Example

| | White | Black |
|--------------------------|-------|-------|
| 0 previous convictions | | 5% |
| 1-2 previous convictions | | 20% |
| 3+ previous convictions | | 40% |
| Average recidivism | 20% | |

Calibration by group
Model hides true data.
Calibration ✓
Fairness X