

Bias and Fairness

Presentation by Chetan Parthiban




Algorithms can learn and enhance bias






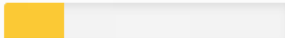











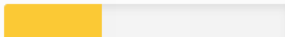
- Humans have conscious and subconscious bias – we make snap judgements about everything we see
- Data is made by humans
- Data is biased
- A good model learns the data well
- Therefore a good model should reflect the bias in data

Case 1: Facial Recognition

- Study of large commercial facial recognition software (Microsoft, IBM, Face+)
- Split a dataset of 1270 images into male/female as well as into 5 groups based on skin color



Gender Classifier	Overall Accuracy on all Subjects in Pilot Parliaments Benchmark (2017)
 Microsoft	93.7%
 FACE++	90.0%
 IBM	87.9%

Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
 Microsoft	94.0% 	79.2% 	100% 	98.3% 	20.8% 
 FACE++	99.3% 	65.5% 	99.2% 	94.0% 	33.8% 
 IBM	88.0% 	65.3% 	99.7% 	92.9% 	34.4% 

Case 1: Facial Recognition

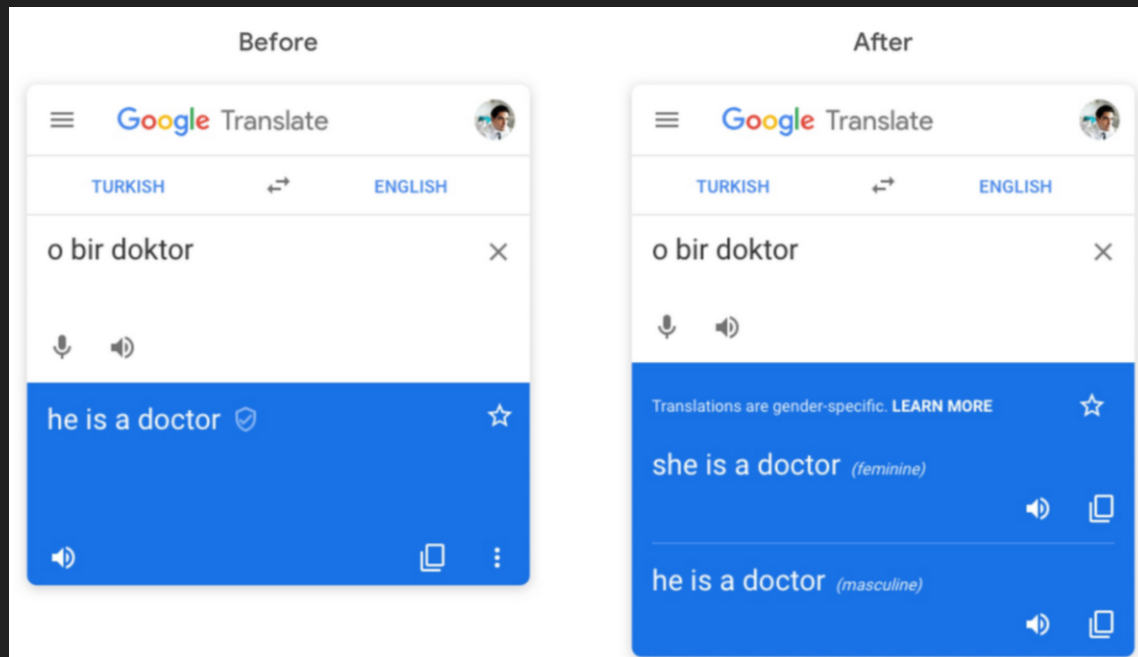
- Classification accuracy can be very different across groups
- Error analysis reveals 93.6% of faces misgendered by Microsoft were those of darker subjects.
- Error analysis reveals 95.9% of the faces misgendered by Face++ were those of female subjects.

Case 1: Facial Recognition

- None of the companies tested reported how well their computer vision products perform across gender, skin type, ethnicity, age or other attributes.
- Face recognition technology that has not been publicly tested for demographic accuracy is increasingly used by law enforcement and at airports.
- AI fueled automation now helps determine who is fired, hired, promoted, granted a loan or insurance, and even how long someone spends in prison.
- This technology is almost completely unregulated in the US

Case 2: Google Translate

- What happens when you translate from a language with a gender neutral pronoun (Turkish “o”) to English?



Case 2: Google Translate

Gender

by Google Translate

he is a soldier
she's a teacher
he is a doctor
she is a nurse

he is a president
he is an entrepreneur
she is a singer
he is a student
he is a translator

he is hard working
she is lazy

he is a painter
he is a hairdresser
he is a waiter
he is an engineer
he is an architect
he is an artist
he is a secretary
he is a dentist
he is a florist
he is an accountant
he is a baker
he is a lawyer

...

Machine Bias

Prediction Fails Differently for Black Defendants

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

Why is this so important?

- The areas where deep learning is starting to be applied have massive human impact
- It is a common misconception that machines are neutral
- Machines are good at discovering latent variables
- Systems are often not transparent
- *"The privileged are processed by people; the poor are processed by algorithms."* – Cathy O'Neil, Weapons of Math Destruction

Algorithms can be dangerous



Recommendation
algorithms are often
polarizing



Facebook leads to
genocides in
Myanmar by
promoting fake news

What does it mean to be fair

- **Group-Independent Predictions**

require that the decisions that are made are independent of group membership

- **Equal Metrics Across Groups**

require equal prediction metrics of some sort (this could be accuracy, true positive rates, false positive rates, and so on) across groups.

- **Individual Fairness**

requires that individuals who are similar with respect to the prediction task are treated similarly. The implicit assumption is that there exists an ideal feature space in which to compute similarity, that is reflected or recoverable in the available data.

- **Causal Fairness**

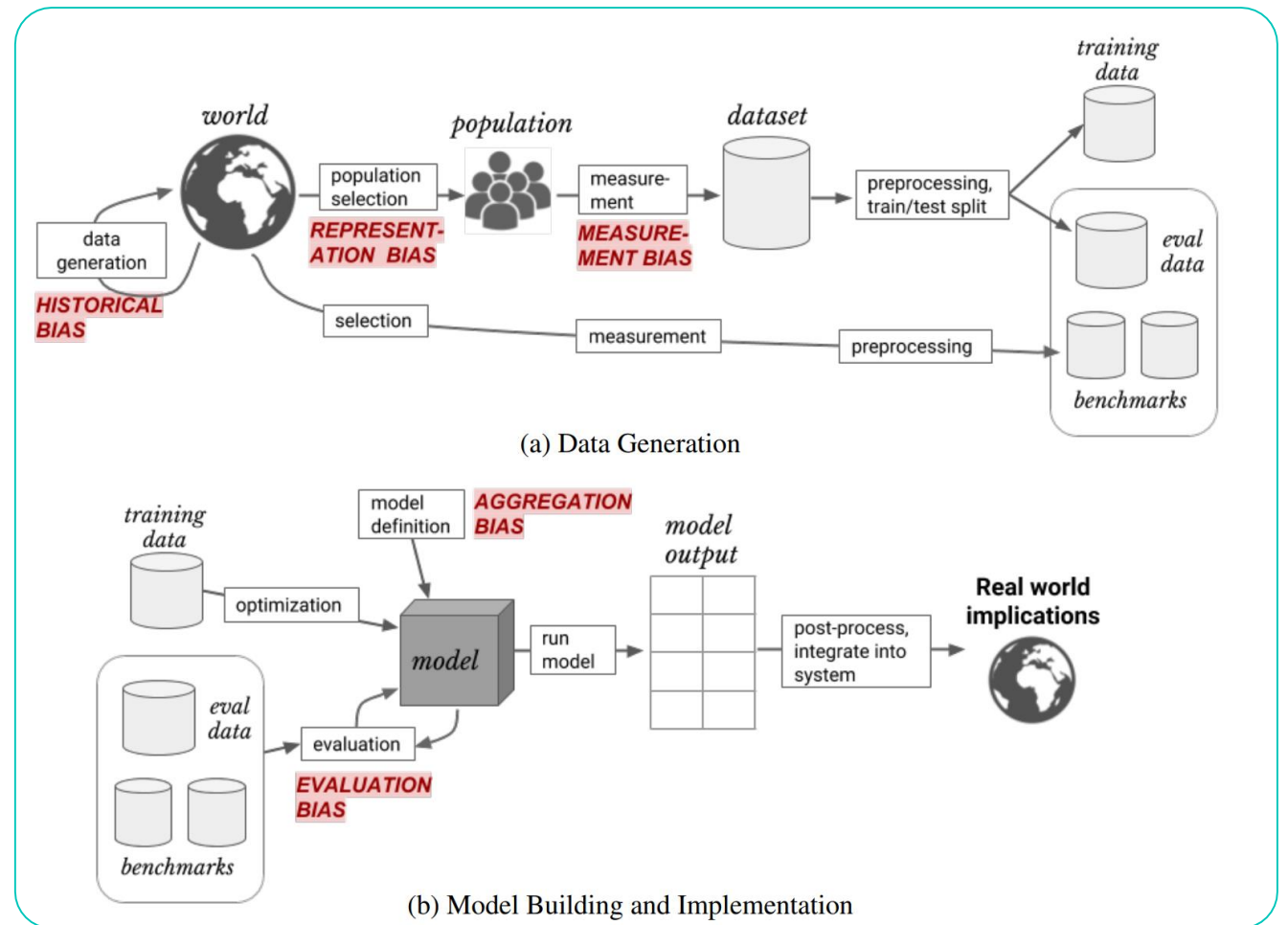
definitions place some requirement on the causal graph that generated the data and outcome. For example, *counterfactual fairness* requires that there is not a causal pathway from a sensitive attribute to the outcome decision

Bias is a type of Error

- **Statistical Bias** – difference between a statistic's expected value and true value
- **Unjust Bias** – disproportionate preference for a prejudice against a group
- We will be discussing unjust bias

A Framework for Understanding Unintended Consequences of Machine Learning

- **Historical Bias:** concern with the world as it is
- **Representation Bias:** issues when sampling
- **Measurement Bias:** issues with measuring features
- **Evaluation Bias:** error during model iteration
- **Aggregation Bias:** flawed assumptions affect the model definition



Historical Bias

- **Example: image search In 2018**
- 5% of Fortune 500 CEOs were women
- Should image search results for “CEO” reflect that number?

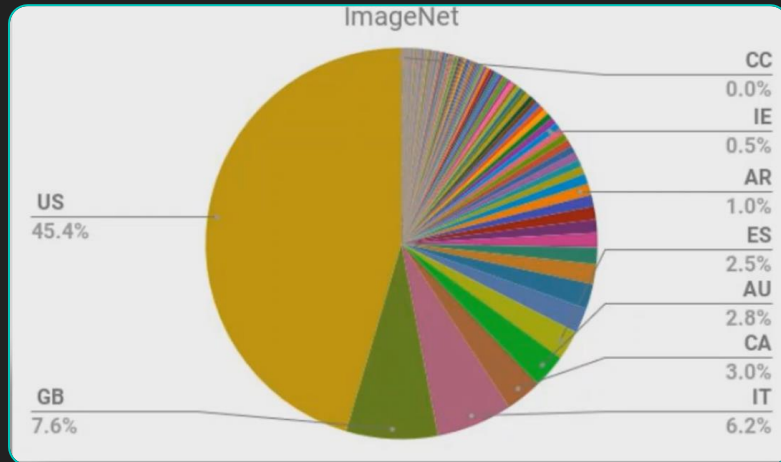
Representation Bias

- **The sampling methods only reach a portion of the population.**
- Data collected through smartphone apps can under-represent lower-income or older groups, who are less likely to own smartphones.
- **The population of interest has changed or is distinct from the population used during model training.**
- Data that is representative of Boston may not be representative if used to analyze the population of Indianapolis.
- Data representative of Boston 30 years ago will likely not reflect today's population

Measurement Bias

- **The granularity of data varies across groups**
 - If a group of factory workers is more stringently or frequently monitored, more errors will be observed in that group.
- **The quality of data varies across groups**
 - Women are more likely to be misdiagnosed or not diagnosed for conditions where self-reported pain is a symptom (in this case “diagnosed with condition X” is a biased proxy for “has condition X”)
 - Using arrest rate as a proxy for crime
- **The defined classification task is an oversimplification**
 - Consider the prediction problem of deciding whether a student will be successful. Algorithms resort resort to some available label such as ‘GPA’ which ignores different indicators of success exhibited by parts of the population

Evaluation Bias



- Occurs when the evaluation and/or benchmark data for an algorithm doesn't represent the target population
- Quality is often measured on benchmarks e.g. ImageNet
- Can lead to overfitting to a particular benchmark or set of benchmarks
- This is especially problematic if the benchmark is not representative

Aggregation Bias

- Arises when a one-size-fit-all model is used for groups with different conditional distributions
- Diabetes patients have known differences in associated complications across ethnicities
- Because these factors have different meanings and importance within different subpopulations, a single model is unlikely to be best-suited for any group in the population even if they are equally represented in the training data



Bias in NLP

Implicit-Association Test (IAT)

- Psychology test used to demonstrate subconscious human bias
- Creates pairings between two targets (e.g. male/female) with two attributes (pleasant/unpleasant)
- Uses reaction time to check associations
- If the categories under study are associated with the presented attributes to differing degrees, the pairing reflecting the stronger association should be easier for the participant

Task 1 (practice):

Black White

Aaliyah

Task 2 (practice):

Pleasant Unpleasant

Suffering

Press E to classify as Pleasant
or I to classify as Unpleasant

Tasks 3 and 4 (data collection):

Black/ White/

Pleasant Unpleasant

Happiness

Press E to classify as Black or Pleasant
or I to classify as White or Unpleasant

Task 5 (practice):

White Black

Eminem

Press E to classify as White
or I to classify as Black

Tasks 6 and 7 (data collection):

White/ Black/

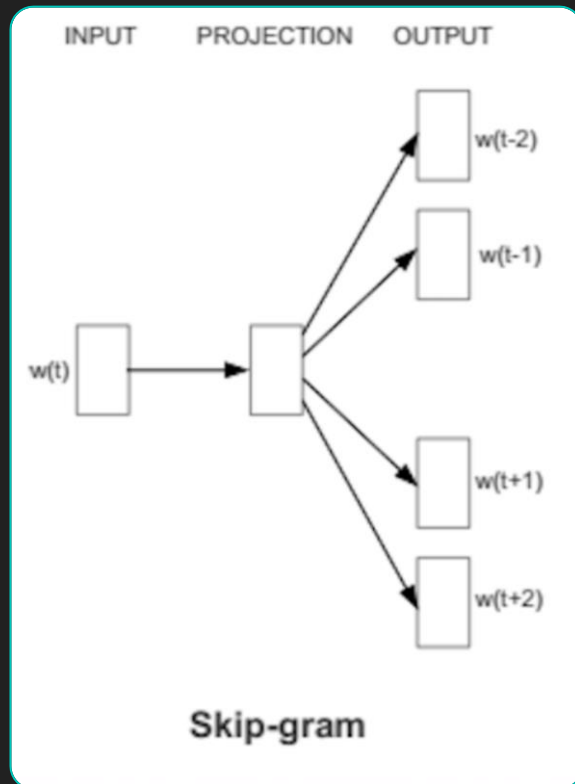
Pleasant Unpleasant

Shanice

Press E to classify as White or Pleasant
or I to classify as Black or Unpleasant

Example of a typical IAT procedure

Word2Vec



- Language Modeling = Unsupervised Pretraining
- Difficult to obtain task specific datasets big enough to learn language
- Word2Vec architecture is a perceptron, trained with the skip-gram objective

Semantics derived automatically from language corpora necessarily contain human biases

- Creates an analogue to IAT for use with language models – the Word Embedding Association Test (WEAT)
- Uses the same notions of targets and attributes
- Computes similarities using cosine similarity of the word embeddings
- Statistical significance computed with a permutation test

- The test statistic is

$$s(X, Y, A, B) = \sum_{x \in X} s(x, A, B) - \sum_{y \in Y} s(y, A, B)$$

where

$$s(w, A, B) = \text{mean}_{a \in A} \cos(\vec{w}, \vec{a}) - \text{mean}_{b \in B} \cos(\vec{w}, \vec{b})$$

In other words, $s(w, A, B)$ measures the association of the word w with the attribute, and $s(X, Y, A, B)$ measures the differential association of the two sets of target words with the attribute.

- Let $\{(X_i, Y_i)\}_i$ denote all the partitions of $X \cup Y$ into two sets of equal size. The one-sided p -value of the permutation test is

$$\Pr_i[s(X_i, Y_i, A, B) > s(X, Y, A, B)]$$

WEAT used on Word2Vec

Replicating IAT experimental results

- **IAT Finding**
- With 38,797 interpretable subjects, female names were found to be more associated with family than career words with an effect size of 0.72 and $p\text{-value} < 10^{-2}$
- **WEAT Finding**
- Females are more associated with family and males with career w/ effect size of 1.81 and $p\text{-value} < 10^{-3}$.

Stimuli: We use the same stimuli found in [Nosek et al. \(2002a, p. 114\)](#).

- **Male names:** John, Paul, Mike, Kevin, Steve, Greg, Jeff, Bill.
- **Female names:** Amy, Joan, Lisa, Sarah, Diana, Kate, Ann, Donna.
- **Career words :** executive, management, professional, corporation, salary, office, business, career.
- **Family words :** home, parents, children, family, cousins, marriage, wedding, relatives.

WEAT used on Word2Vec

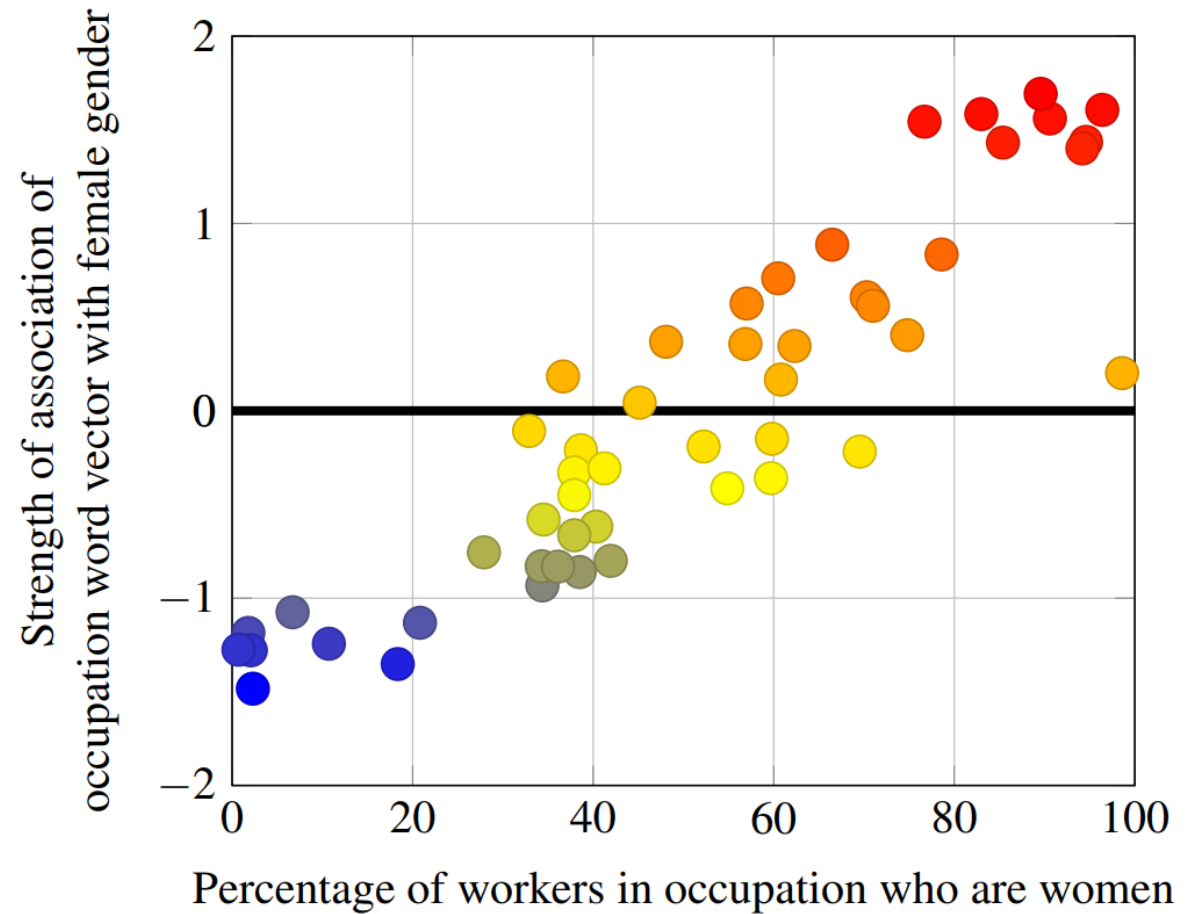
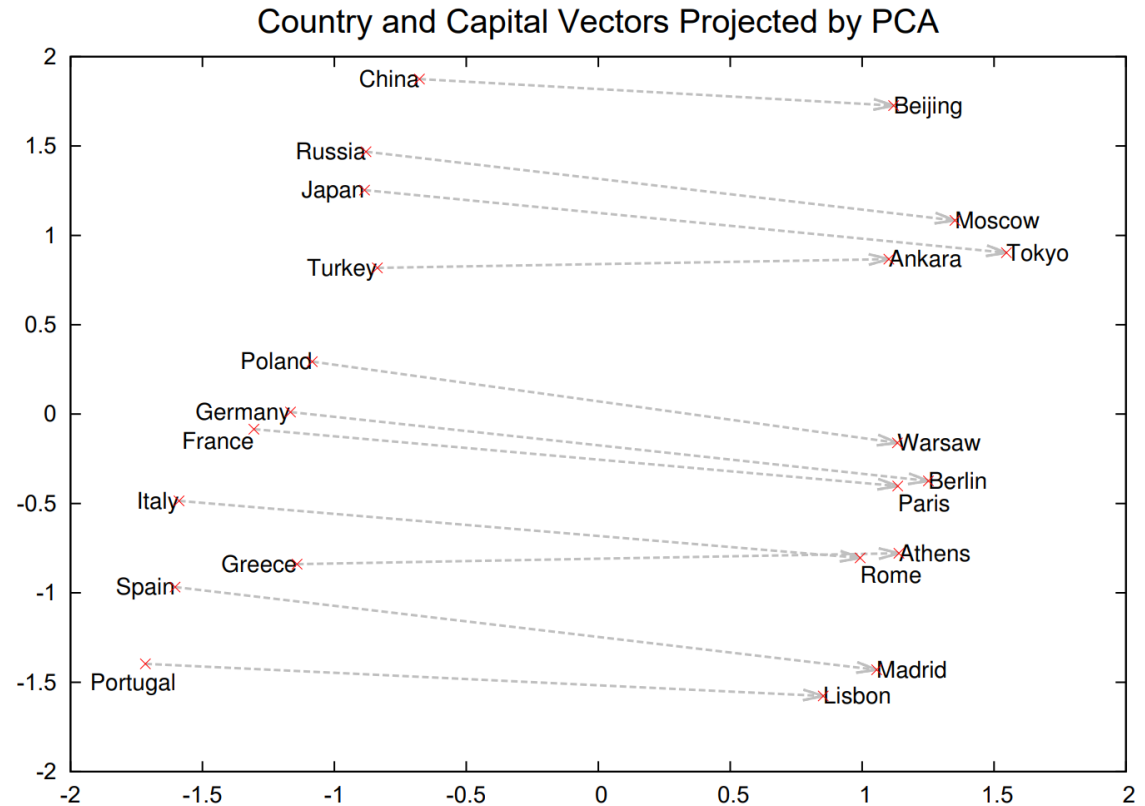


Figure 1. Occupation-gender association
Pearson's correlation coefficient $\rho = 0.90$ with $p\text{-value} < 10^{-18}$.

Analogies in Word2Vec

- Additive relationships exist in word2vec
- King:He::Queen:She
- King – He + She ~ Queen
- He:Doctor::She:Nurse
- Maybe there is a direction corresponding to gender?



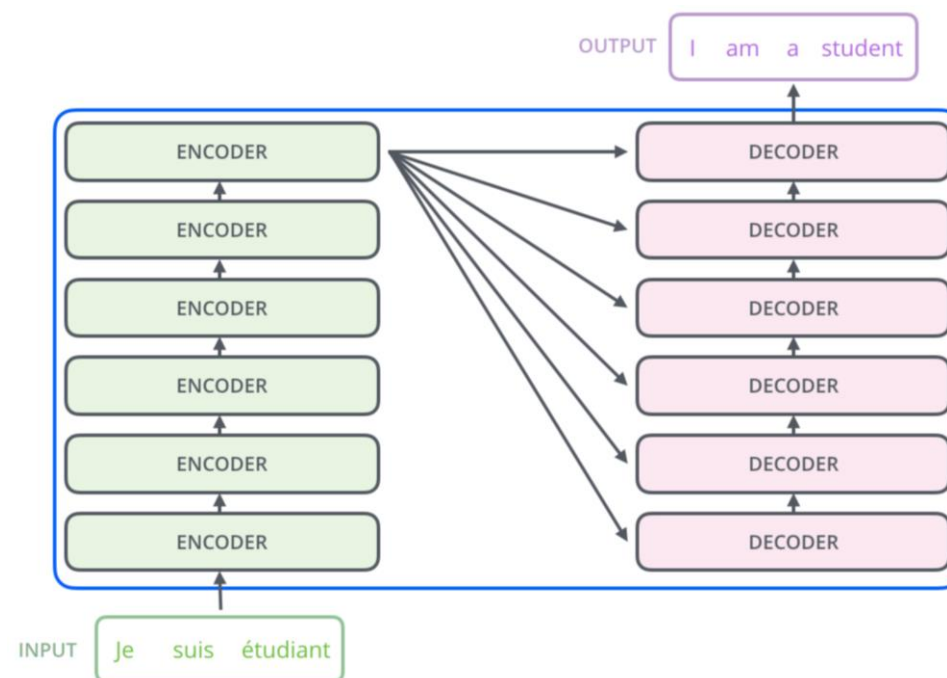
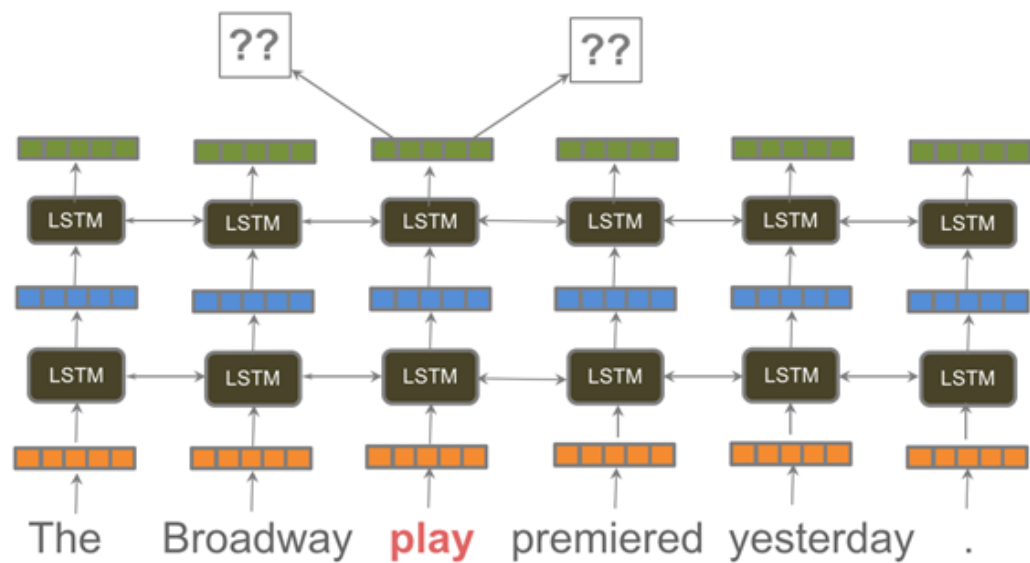
Fixing(?) Bias

Current “SOTA” debiasing methods work effectively as follows:

1. Create a list of words that vary along a parameter of interest
e.g. [He, She, His, Her, Boy, Girl...]
 2. Embed all words of the list using your language model
 3. Use PCA to find the directions of highest variance
 4. Shrink these directions
 - Hard PCA: shrinks direction to 0
 - Soft PCA: shrinks direction by a percentage
 - Conceptors: shrink all directions proportionally to its current variance
- Empirically works well at reducing WEAT scores for Word2Vec (and ELMo, a contextualized language model)

Confirmation Bias, Lipstick on a Pig, and More

- **Confirmation bias**
 - Just because WEAT score goes down, it does not mean that the model is no longer biased
 - Danger of using a substitute measurement
- **Lipstick on a Pig**
 - Models down the line may reamplify the bias anyways – especially clear in the linear case
- **Generating word lists is not trivial**
 - Consider race
- **Fails for SOTA language models**
 - WEAT and PCA debiasing completely fails on transformer based architectures (BERT & co)



Transformers

BERT Passes WEAT

- BERT does not show significant associations in WEAT
 - Debiasing methods do not affect BERT WEAT scores
 - This does not mean BERT is unbiased
-
- Log-probability Bias Score
 - Use a template sentence “[Target] is an [Attribute]”
 - Compute $\log[P(\text{Target} = T \mid \text{Attribute} = A) / P(\text{Target} = T)]$

Why? (speculative)

01

BERT has high variance for a single token w.r.t context so PCA based methods do not capture the correct direction

02

It is unclear the cosine similarity is a sufficient measure of similarity

03

BERT is much more complicated than previous language models, and even behaves quite differently in different layers

Final Remarks

AI Safety/Ethics may not be as flashy as going for new SOTA results, but it is incredibly important

The field is still nascent so there is a lot of room for innovation

It is important for algorithm designers and people training models to consider this and to help educate other so that biased (or dangerous) models don't make it into production

References

- A Framework for Understanding Unintended Consequences of Machine Learning:
<https://arxiv.org/abs/1901.10002>
- Conceptor Debiasing of Word Representations Evaluated on WEAT:
<https://arxiv.org/abs/1906.05993>
- Distributed Representations of Words and Phrases and Their Compositionality:
<https://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf>
- Semantics derived automatically from language corpora necessarily contain human biases: <https://arxiv.org/abs/1608.07187>
- Gendershades.org