This Ayn Random Number Generator you wrote claims to be fair, but the output is biased toward certain numbers.

Well, maybe those numbers are just intrinsically better!
Bias

Eric Wong
9/1/2022
Announcements

• Ed Discussion is up

• Sign up for paper presentations

• Materials from first lecture are posted
Paper presentations

• Readings posted by the end of Friday for next week (suggestions welcome)

• Presenters: prepare notes to lead seminar

• Everyone else: post in Ed Discussion before class
What is bias?

“A tendency for or against something”
3 Contextual biases

3.1 Biases in academia
   3.1.1 Academic bias
   3.1.2 Experimenter bias
   3.1.3 Funding bias
   3.1.4 Full text on net bias
   3.1.5 Publication bias

3.2 Biases in law enforcement
   3.2.1 Driving while black
   3.2.2 Racial profiling
   3.2.3 Victim blaming

3.3 Biases in media
   3.3.1 Agenda setting
   3.3.2 Gatekeeping
   3.3.3 Sensationalism

3.4 Other contexts
   3.4.1 Educational bias
   3.4.2 Inductive bias
   3.4.3 Insider trading
   3.4.4 Match fixing

3.5 Implicit bias
Types of bias

• Statistical bias (*board*)

• Learning bias (*board*)

• Data bias (*these slides*)
Harms of representation

1. Stereotyping
2. Recognition
3. Denigration
4. Under representation
1. Stereotyping

PC: Kate Crawford, “The Trouble with Bias” NeurIPS 2017 keynote
2. Recognition

Buolamwini & Gebru, “Gender Shaes: Intersectional Accuracy Disparities in Commercial Gender Classification”
3. Denigration

PC: Kate Crawford, “The Trouble with Bias” NeurIPS 2017 keynote
4. Under representation

Kay et al. 2015, “Unequal Representation and Gender Stereotypes in Image Search Results for Occupations”
PC: Kate Crawford, “The Trouble with Bias” NeurIPS 2017 keynote
Understand gender-specific translation order

Gender-specific translations show in alphabetical order by gender label. For example, in English, “feminine” comes alphabetically before “masculine.”

In languages without alphabets, gender-specific translations show in each language’s standard indexing order by gender label.
Bias part 2

Eric Wong
9/6/2022
Today

• Notebook on calculating bias

• Project proposal guidelines posted
  • Form teams and submit by Oct 4th

• Specific biases

• Paper presentation
Reminders

• Sign up for paper presentations

• Post/reply on Ed discussion

• Start thinking about projects and form groups
Feedback

• Temperature on DRL

• Lights in CHEM
Missingness bias

A bias that occurs when data is missing
Convolution layer

\[ f(x)_{m,n} = \sum_i \sum_j w_{ij} x_{m-i,n-j} \]
Case Study: ResNets

Blackening out pixels

Jain, et al. 2022 "Missingness Bias in Model Debugging."
Skewed class distribution

Jain, et al. 2022 "Missingness Bias in Model Debugging."
Attention layer

\[ f(Q, K, V) = \text{softmax} \left( \frac{QK^T}{\sqrt{d_k}} \right)V \]

Self attention: \( Q, K, V = W_Q x, W_K x, W_V x \)

PC: Lee et al. 2019 “Semantic Relation Classification via Bidirectional LSTM Networks with Entity-aware Attention using Latent Entity Typing”
Transformer

PC: https://towardsdatascience.com/bert-explained-state-of-the-art-language-model-for-nlp-f8b21a9b6270
Case Study: Vision transformer

Dosovitskiy et al. 2021 “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale”
“True” missingness

Jain, et al. 2022 "Missingness Bias in Model Debugging."
Mitigates learning bias

Original

ResNet-50: schooner
ViT-S: catamaran

Random

ResNet-50: crossword
ViT-S: catamaran

Least Salient

ResNet-50: wing
ViT-S: catamaran

Most Salient

ResNet-50: jigsaw puzzle
ViT-S: catamaran
Uniform class distribution
Missing words in text

Occupation templates:

“[MASK] is a doctor/nurse/dancer”

Compare probability scores for different pronouns

Liang et al. 2020 “Monolingual and multilingual reduction of gender bias in contextualized representations“
Measuring data bias in word embeddings

• Analogies

• Implicit Association Test

• Distances or alignment to group subspace
Analogies

Prediction of stereotypes as a measure of bias

doctor : man :: woman : nurse

Bolukbasi et al. 2016 “Man is to computer programmer as woman is to homemaker? Debiasing word embeddings”
Implicit Association

Cosine similarity between word embeddings of concepts and attributes

Examples:
- European vs American names (concepts)
- Pleasant vs Unpleasant (attributes)

Liang et al. 2020 “Monolingual and multilingual reduction of gender bias in contextualized representations“