UVA DATA SCIENCE

Bias Evaluation in Open Model Platform

Becky Desrosiers, Abner Casillas-Colon, Naomi Ohashi, George Shoriz

1 May 2025

Presentation roadmap

- Background/purpose
- Data: source, assumptions, limitations
- Methods
- Findings
- Future work

Background

Sponsor background

- Research was conducted on behalf of Intel Labs for their Open Model Zoo Software
- OpenVINO Toolkit and Open Model Zoo are a set of open source AI Tools for public use

Purpose

- Evaluate various Bias Mitigation Techniques
- Identify a dataset to apply bias mitigation on an Open Model Zoo model

Data

FairFace Data

- Originally from the YFCC-100M Flickr
- 108,501 curated, balanced images
- Data Attributes:
  - Race, gender, and age groups
  - 7 Race groups: White, Black, Indian, East Asian, Southeast Asian, Middle Eastern, and Latino

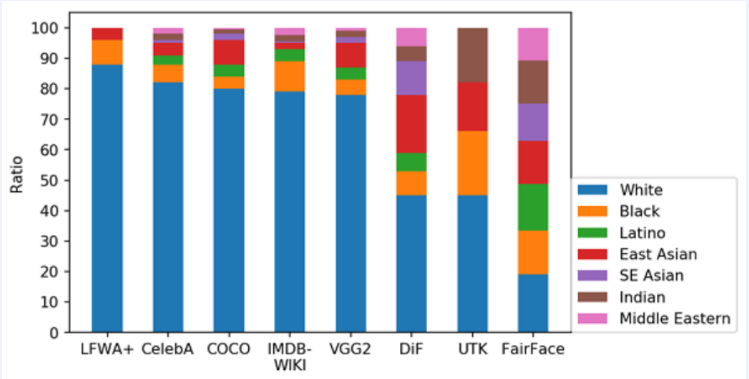
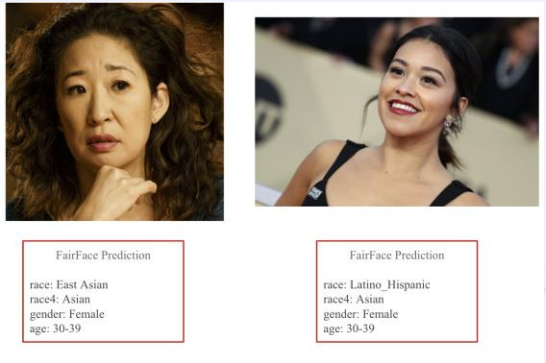


Fig 1. Racial compositions in face datasets



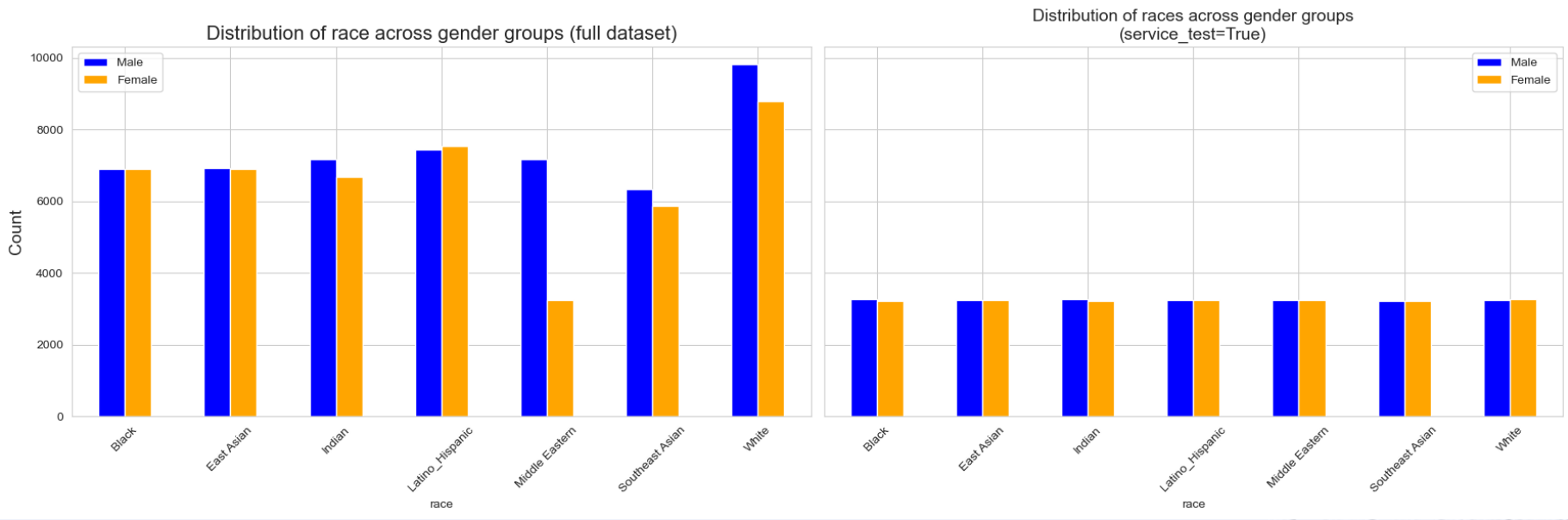
UVA DATA SCIENCE

Fig 1. Racial compositions in face datasets, Joo, et al (2021)

FairFace Data

- Service Test:
  - Amazon, Microsoft, Facebook, IBM tested FairFace for classification accuracy
  - Evaluated dataset: 40,252
- Challenges:
  - Representation does not guarantee fairness # ie. approvals
  - Intersectionality underrepresentation # ie. A few Indian women

FairFace Data



FairFace Data

- Service Test:
  - Amazon, Microsoft, Facebook, IBM tested FairFace for classification accuracy
  - Evaluated dataset: 40,252
- Challenges:
  - Representation does not guarantee fairness # ie. approvals
  - Intersectionality underrepresentation # ie. A few Indian women

Methods

Methodology

- Inputs as Batch Size, number of channels, image height, image width
- Face-Detection-0200
- Uses MobilenetV2 (CNN) architecture
- Output blob that contains predicted number of bounding boxes and confidence values

Findings

Top findings with Aequitas

- Dataset: FairFace (diverse gender, race, age)
- Tool: Aequitas Bias Metrics
- Thresholds evaluated:
  - 95% (strict)
  - 80% (general detection standard)

Top findings with Aequitas (Continued)

- At 95% threshold:
  - Missed faces more often for older adults and Black individuals
  - Southeast Asian, Latino/Hispanic, and children had better detection
- At 80% threshold:
  - Detection improved for all groups
  - Fairness gaps shrank significantly

Future work

Future work plan

- Dataset with non-faces
- AI Fairness 360 (AIF360)
- Explanations of each metric
- Qualitative bias analysis

Acknowledgements

Emanuel Moss
sponsor (Intel)

Elizabeth Watkins
sponsor (Intel)

Dawn Nafus
sponsor (Intel)

Philip Waggoner
mentor (UVA)

Questions?

Capstone Team

Naomi Ohashi

Becky Desrosiers

Abner Casillas Colon

George Shoriz