Bias Detection and Mitigation in CNN Image Classification

T2P4: Leanne Beltman, Seth Bonin, Justin Coffey, Connar Gibbon, Libby Stephan

Outline

- Introduction
  - Problem
  - Objective
  - Deliverables
- Current Progress
  - Attribute Correlations
  - Dimension Reductions
  - Computation Issues
- Future Steps
- Timeline

Introduction

Problem:

implicit bias in data → unfair/inaccurate predictions → devastating effects

Objective:

Determine methods to detect these bias and mitigation tactics.

- CNN image classification

Deliverables:

Written report with well-supported justification for proposed findings.

Current Progress

Exploratory Analysis

- **Goal**: find potential similarities and underrepresentation between classes to exploit in CNN training
- **Methods**:
  - 1. Frequently occurring attributes
  - 2. Correlations between attributes
  - 3. Plot images via dimension reduction

1. Frequently Occuring Attributes - CelebA

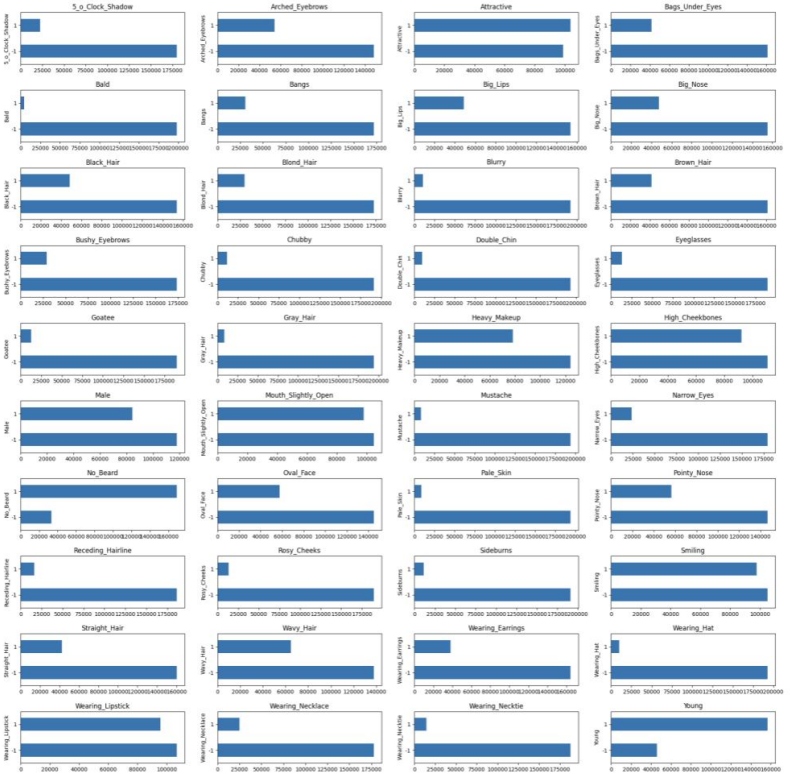
Goal: identify traits that may be underrepresented in the dataset

Even Representation:

- Attractive
- Smiling

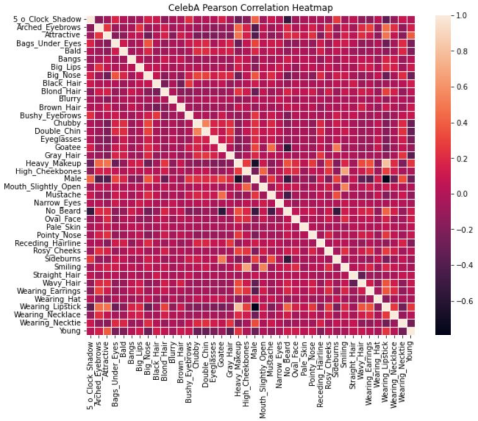
Uneven Representation:

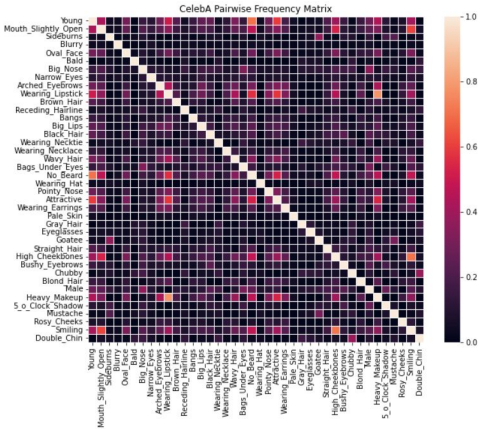
- Bald
- Wearing Hat



2. Correlations Between Attributes - CelebA

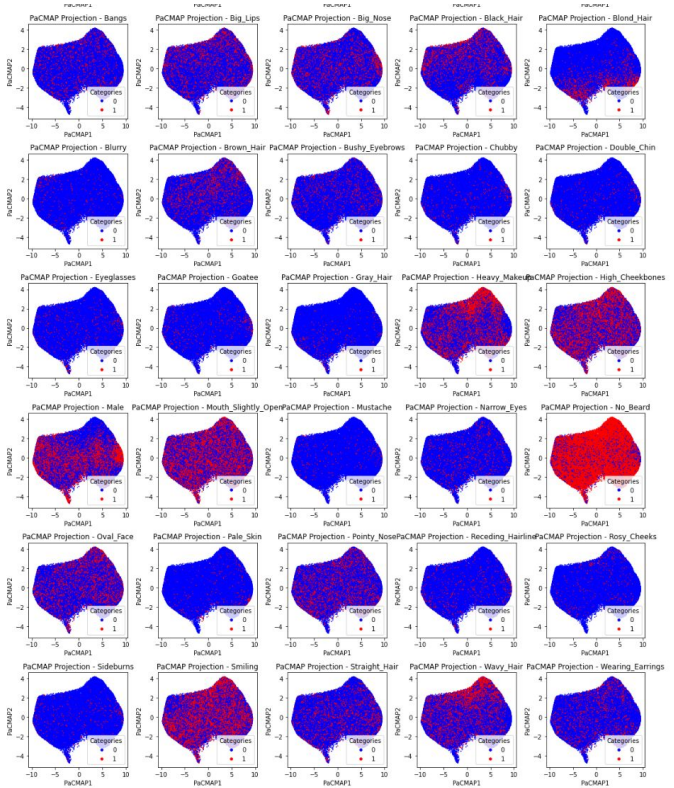
- Goal: find traits that typically appear together
- Two methods used: Pearson and frequency percentage
  - “Heavy Makeup” and “Lipstick”
  - “Smiling” and “High Cheekbones”





3. Dimension Reductions - PaCMAP CelebA

- Goal: find out which traits are or are not easily separable
- 64 x 64 resolution
- Training on subset (3,000)
- No clear separation of any class
  - Too low resolution
- No separation anywhere may mean unreliable PaCMAP results from poor resolution



3. Dimension Reductions - PaCMAP CelebA

CelebA Image

Full Resolution (178 x 128)

CelebA Image

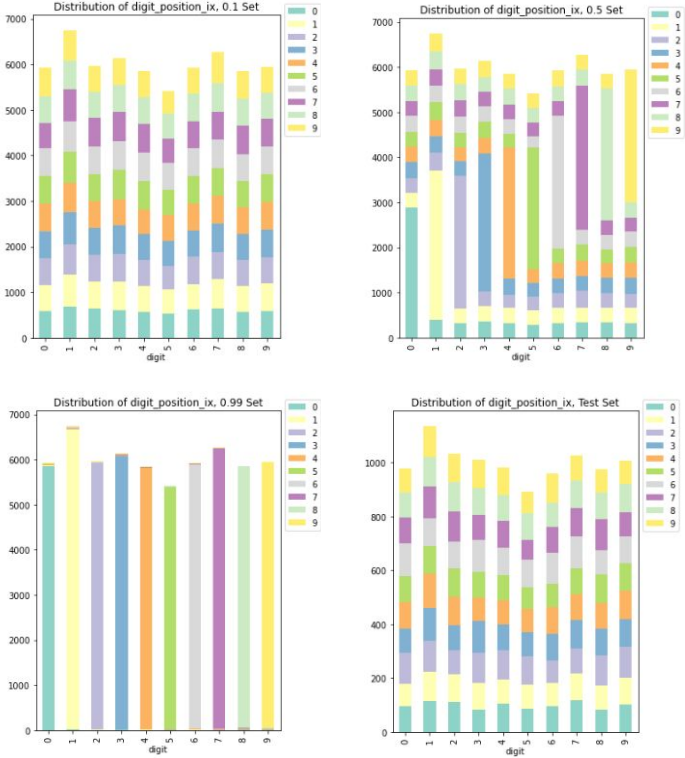
Reduced Resolution (64 x 64)

1 & 2. Evaluating Distributions – Biased MNIST

Datasets: Three correlation levels (0.1, 0.5, 0.99)

Goal: Understand how distributions of different variables change at different bias levels.

- 0.1 correlation dataset has relatively equal frequencies of classes
- 0.5 dataset is noticeably unbalanced
- 0.99 dataset is extremely unbalanced
- Test dataset classes occur approximately equally



3. Dimension Reduction Analysis: Biased MNIST

Datasets: Three correlation levels (0.1, 0.5, 0.99)

Techniques Used:

- Principal Component Analysis (PCA)
- Uniform Manifold Approximation and Projection (UMAP)
- Pairwise Controlled Manifold Approximation (PaCMAP)

Key Question: What variables are highly correlated, and/or not easily separable via dimensionality reduction?

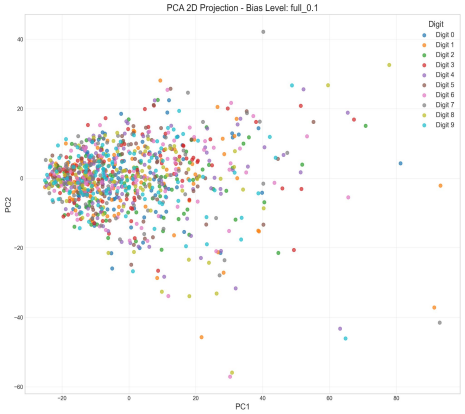
Goal: Understand how bias manifests in different correlation levels

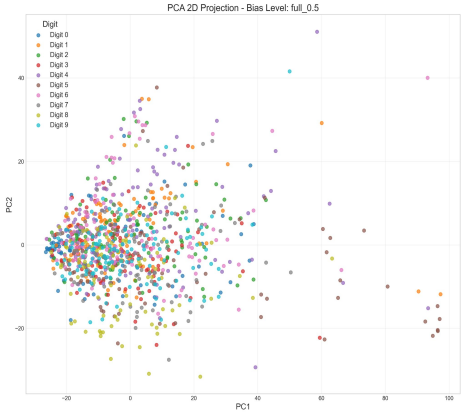
3a. Principal Component Analysis (PCA) - Biased MNIST

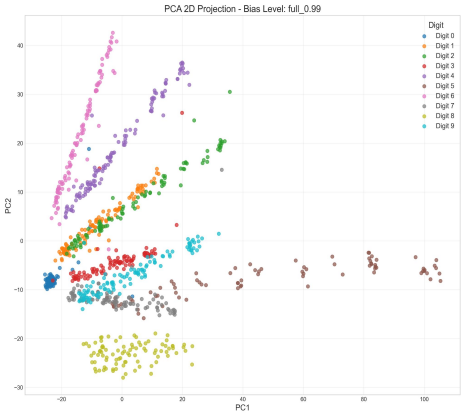
Key Findings - PCA

- **0.1 Correlation:** No clear separation between digits
- **0.5 Correlation:** Slight Improvement in separation, but still highly mixed
- **0.99 Correlation:** Distinct clusters for each digit

- Higher correlations create predictable patterns that models may exploit instead of learning genuine features





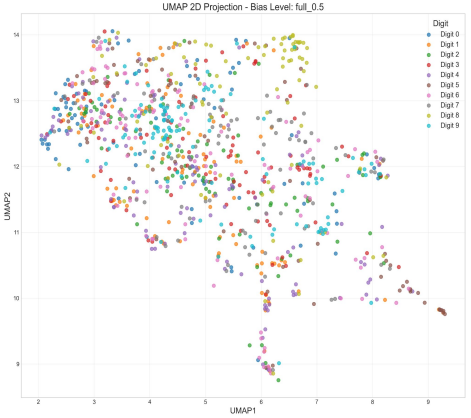


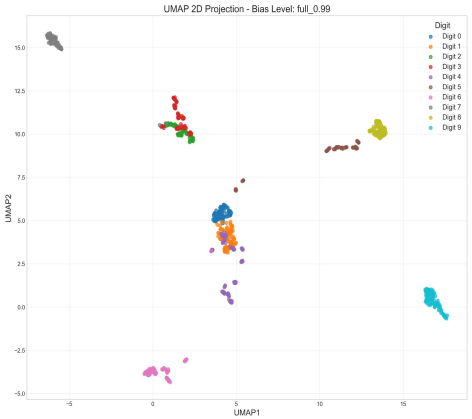
3b. Uniform Manifold Approx. and Projection (UMAP) - Biased MNIST

Key Findings - UMAP

- **0.1 Correlation:** No clear separation between digits
- **0.5 Correlation:** Slight Improvement in separation, but still highly mixed
- **0.99 Correlation:** Distinct clusters for each digit

- Higher correlations create predictable patterns that models may exploit instead of learning genuine features



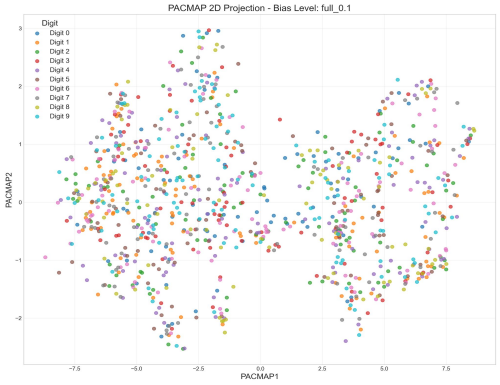


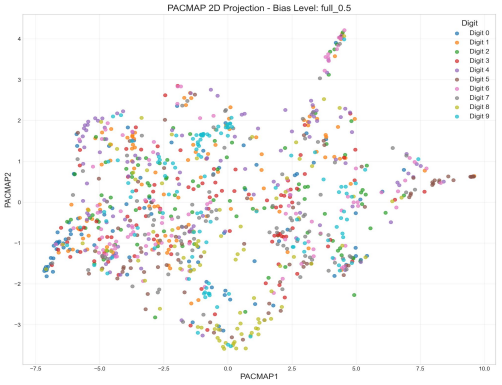
3c. Pairwise Controlled Manifold Approx. (PaCMAP) - Biased MNIST

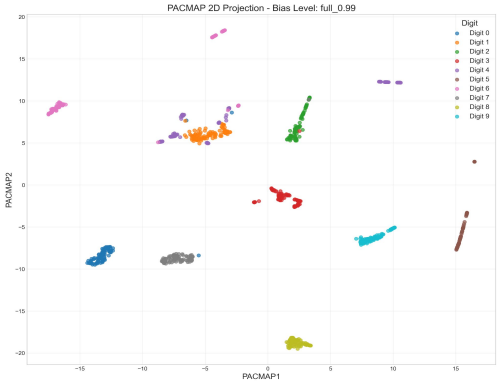
Key Findings - PaCMAP

- **0.1 Correlation:** No clear separation between digits
- **0.5 Correlation:** Slight Improvement in separation, but still highly mixed
- **0.99 Correlation:** Distinct clusters for each digit

- Higher correlations create predictable patterns that models may exploit instead of learning genuine features







Implications for Future Use

- There appears to be a threshold between 0.5 and 0.99 where the bias becomes the dominant signal for learning
- Next step would be to plot PCA components needed to explain 95% variance (0.1, 0.5, 0.75, 0.9, 0.95, 0.99) to see exactly where the transition occurs

Model Performance:

- Struggle to find consistent patterns at 0.1 and 0.5, potentially leading to lower accuracy
- Overfit to the bias attribute at 0.99
- Models trained on the 0.99 level likely to fail when tested on unbiased data

Computation Issues

Challenges Encountered:

- CelebA dataset size required image resizing (to 64x64)
- Unable to perform some dimensionality reduction techniques (PCA, UMAP) on full dataset on full resolution
- Computation power restrictions limited analysis options

Solutions Applied

- Created separate notebooks for dimensionality reduction
- Saved results as CSV files for easier visualization and analysis
- Modified approaches to work within computational constraints

Future Steps

Next Phase

- Build neural network models using insights from data exploration
- Focus on target variables identified as challenging:
  - Highly correlated pairs (Heavy_Makeup & Wearing_Lipstick)
  - Underrepresented attributes (Bald, Gray Hair)
  - Variables from 0.5 or 0.99 correlation MNIST datasets
- Test model performance on balanced test dataset
- Implement bias mitigation techniques

Potential Challenges

- Extremely imbalanced classes in 0.99 dataset may make resampling difficult
- Need to balance intentional bias creation with ability to mitigate it
- Computation time

Timeline

Task	Date
Develop the neural network	March 28
Initial bias calculation	April 4
Implement mitigation strategies	April 11
Re-calculate bias in new model	April 18
Create final project report	April 25