EVALUATING AGE AND GENDER INFERENCE BIAS

Evaluating Age and Gender Inference Bias:

Microsoft Azure Cognitive Services

Kieran Kyllo Brooks

Department of Computer Science, University of Regina

CS 280: Risk and Reward in the Information Society

Dr. Trevor Tomesh

Abstract

Computer vision is a rapidly growing and increasingly mainstream usage of modern machine learning techniques. Inference on images of the human facial form is becoming increasingly common in applications such as law enforcement, intelligence gathering, marketing, sentiment analysis, lip reading, and biometrics. This study investigates inference bias present in the Microsoft Azure Cognitive Services Face API by exposing a sample of 67,228 facial raster images from the manually annotated Fairface dataset. Detection rates, gender inference error rates, and age inference error rates are compared across seven racial categories by chi square test at $p < 0.001$. Black faces were over represented in detection failure and gender inference error rates while white faces were most associated with age inference failure rates with a positive correlation to age overestimation. Azure Cognitive Services was most accurate when inferring the gender of White subjects and when inferring the age of East Asian facial images with inference errors positively correlated with age underestimation. These findings highlight and replicate the mounting evidence of computer vision biases present in the literature generally and may point to specific bias mechanics not yet well understood.

Keywords: computer vision, gender inference, age inference, ethics, facial recognition, machine learning, artificial intelligence.

Introduction

When applied to the human image, computer vision and machine learning inference tools are capable of increasingly accurate and potentially useful categorical inference from raster data. However, efforts to evaluate commercially available algorithm APIs have shown bias in gender discernment in sample sets representing POC populations (Buolamwini & Gebru, 2018). Emerging implementations of facial recognition and inference in law enforcement, security, intelligence gathering, and biometrics further underline the importance of gaining a complete understanding of how to measure, monitor, and eliminate racially motivated bias in emerging systems and implementations of this technology. More generally, understanding how bias may propagate through training data is an important avenue to explore and demonstrate experimentally. The intent of this study is to measure and compare error rates in facial inference for gender and age across racially diverse facial samples. For example, does the API infer a younger age for Southeast Asian faces? Is gender inference less accurate for darker skinned racial groups? etc. This study will use chi square and correlation analysis to determine if a difference exists between the racially categorized Fairface dataset (Karkkainen & Joo, 2021) and the inference values returned from the Microsoft Azure Cognitive Services Face API. Face detection failure rates, gender identification failure rates, and age category identification error rates will be compared across seven racial categories present in the Fairface annotations. Specifically, the chi square test will reveal if the difference in error rates across racial categories is due to random chance within a probability of 0.001 of less. The major hypotheses are as follows:

- H_O : The error rate will be equal across all race categories with respect to gender and age inference values at $p < 0.001$ (No evidence of difference).
- H_A : Statistically significant differences in error rates exist between race categories with respect to gender and age at $p < 0.001$. (Evidence of difference exists).

Fairface Dataset Background

The Fairface dataset defines 7 race groups: White, Black, Indian, East Asian, Southeast Asian, Middle Eastern, and Latino Hispanic. Race is defined as shared facial morphology generally characteristic of shared genetic ancestry. It is important to acknowledge the non-discrete nature of race and indeed gender in the wild as a caveat of the dataset in question which uses discrete categories for race and binary categories for gender. Fairface annotation values for age, gender, and race were generated using Amazon Mechanical Turk through a multi phased, model assisted 2/3 human inference worker consensus pipeline. Specifically, ground truth annotation in Fairface is not achieved through individual census of each image to validate factual ground truth, therefore the annotations in the dataset are themselves subject to human visual inference bias. However, Karkkainen & Joo achieved cross dataset performance accuracy superior or comparable to three other major benchmark datasets trained on a standardized model architecture, suggesting reliable annotation accuracy.

Microsoft Azure Cognitive Services API Background

Microsoft exposes a web API as a paid service for inference on raster data returning facial attributes for age, gender, smile, facial hair, head pose, glasses, and emotion (Farley, P, n.d). Microsoft does not publish their model architecture or their training or validation data for this proprietary inference product. For the purposes of this study the API performed at around 250ms per call with performance mostly limited to request call image payload size. A joint 2018 investigation between MIT and Microsoft Research revealed significant differences in gender inference among several commercially available classifiers using their novel Pilot Parliaments Benchmark dataset. Buolamwini & Gebru revealed that gender inference error rates were higher among darker skinned subjects when compared to lighter skinned representing an error rate of 20.8% among darker skinned female subjects as their most significant result.

Study Methodology

A sample of 67228 annotated raster images were taken from the Fairface training dataset and their raster component exposed to the Cognitive Services Face API via python script (Appendix 1). API calls returned inferred gender and age or null if no face was detected. Results returned null were separated from results with gender and age and analyzed via chi square. Binary gender inference data were compared to ground truth and across race categories via chi square. Returned age inference values were represented as discrete integers whereas ground truth data exist as ranged bins. To account for this, inference was considered to be in error if the returned value was outside the range of the defined bin. Error rates were again analyzed via chi square with six degrees of freedom (BMJ, 2021). Correlation analysis was carried out for age and gender error as well as for age over and underestimation to determine if age inference errors correlate to directionality.

Results

H_O is rejected when comparing failure rates across race categories at $p < 0.001$. Black faces were significantly overrepresented in detection failure rates. Image detection favored Southeast Asian subjects with the lowest detection failure rate of 14.62% followed closely by Latin Hispanic and Indian. None of the race categories performed in line with the expected overall failure rate of 17.48% (Table 1).

Table 1

Observed Face Detection Failures

Category	Total Count	Observed	Expected	O-E	(O-E)^2	((O-E)^2)/E	Failure Rate
East Asian	9443.00	1503.00	1650.71	-147.71	21819.14	13.22	15.92%
Indian	9580.00	1456.00	1674.66	-218.66	47812.96	28.55	15.20%
Black	9481.00	2126.00	1657.36	468.64	219627.44	132.52	22.42%
White	12870.00	2450.00	2249.78	200.22	40087.87	17.82	19.04%
Middle Eastern	7165.00	1423.00	1252.50	170.50	29070.20	23.21	19.86%
Latin Hispanic	10340.00	1573.00	1807.52	-234.52	54997.71	30.43	15.21%
Southeast Asian	8349.00	1221.00	1459.47	-238.47	56869.35	38.97	14.62%
Totals	**67228**	**11752**	**11752**		**Test Statistic**	**284.7068211**	

Overall Failure Rate	17.481%
Overall Detection Rate	82.519%
Degrees Of Freedom	6
Alpha	0.001
Critical Value	22.458

Reject Null Hypothesis
Difference is unlikely to be due to chance.

Sig at 1 DOF p < 0.001

H_O is rejected when comparing gender inference error rates across race categories at p < 0.001. Black faces were again overwhelmingly represented in the gender inference error rates. Conversely, the API performed better than the expected overall error rate of 7.255% at p < 0.001 for White, Middle Eastern, and Latin Hispanic subjects. All other categories were not significantly different from expected values (Table 2).

Table 2

Observed Gender Inference Errors

Category	Total Count	Observed	Expected	O-E	(O-E)^2	((O-E)^2)/E	Error Rate
East Asian	7940.00	635.00	576.08	58.92	3471.81	6.03	8.00%
Indian	8124.00	596.00	589.43	6.57	43.19	0.07	7.34%
Black	7355.00	880.00	533.63	346.37	119969.47	224.82	11.96%
White	10420.00	534.00	756.01	-222.01	49289.15	65.20	5.12%
Middle Eastern	5742.00	292.00	416.60	-124.60	15526.28	37.27	5.09%
Latin Hispanic	8767.00	514.00	636.08	-122.08	14903.53	23.43	5.86%
Southeast Asian	7128.00	574.00	517.16	56.84	3230.31	6.25	8.05%
Totals	**55476**	**4025**	**4025**		**Test Statistic**	**363.0573472**	

Overall Error Rate	7.255%
Overall Positive Predictive Value	92.745%
Degrees Of Freedom	6
Alpha	0.001
Critical Value	22.458

Reject Null Hypothesis
Difference is unlikely to be due to chance.

Sig at 1 DOF p < 0.001

Table 3

Observed Age Inference Errors

Category	Total Count	Misaged	Expected	O-E	(O-E)^2	((O-E)^2)/E	Error Rate
East Asian	7940.00	2956.00	3299.75	-343.75	118161.20	35.81	37.23%
Indian	8124.00	3410.00	3376.21	33.79	1141.53	0.34	41.97%
Black	7355.00	3159.00	3056.63	102.37	10479.92	3.43	42.95%
White	10420.00	4589.00	4330.40	258.60	66875.55	15.44	44.04%
Middle Eastern	5742.00	2371.00	2386.29	-15.29	233.78	0.10	41.29%
Latin Hispanic	8767.00	3619.00	3643.43	-24.43	597.06	0.16	41.28%
Southeast Asian	7128.00	2951.00	2962.29	-11.29	127.48	0.04	41.40%
Totals	**55476**	**23055**	**23055**		**Test Statistic**	**55.32403392**	

Overall Error Rate	41.56%
Overall Positive Predictive Value	58.44%
Degrees Of Freedom	6
Alpha	0.001
Critical Value	22.458

Reject Null Hypothesis
Difference is unlikely to be due to chance.

Sig at 1 DOF p < 0.001

H_O is rejected when comparing age inference error rates across race categories at $p < 0.001$. White subjects were significantly overrepresented in the error rates while conversely East Asian subjects were significantly underrepresented.

Table 4

Correlation Coef Matrix

	Gender Error	Age Error	Age Underestimate	Age Overestimate
East Asian	1.17%	-3.59%	3.38%	-6.82%
Indian	0.13%	0.35%	0.90%	-0.35%
Black	7.10%	1.10%	2.01%	-0.42%
White	-3.95%	2.42%	-4.05%	6.07%
Middle Eastern	-2.84%	-0.18%	-2.18%	1.59%
Latin Hispanic	-2.33%	-0.25%	-1.28%	0.78%
Southeast Asian	1.18%	-0.12%	1.58%	-1.44%

Sig at 1 DOF p < 0.001

Correlation analysis was carried out to tabulate the overall correlation to error across the data. Analysis reflects the positive correlation of Black subjects to gender error and negative correlation to gender error among White, Middle Eastern and Latin Hispanic subjects. Interestingly a significant negative correlation exists between errors and age overestimation among East Asian subjects. Meaning, when the API incorrectly inferred the age it was less

likely to overestimate. Conversely inference on White subjects was positively correlated with age overestimation. Meaning, when the API incorrectly inferred the age of White subjects it was more likely to overestimate.

Discussion

All three studied inference API return types showed significant evidence of difference from expected values. Facial inference for gender and age values are significantly better than chance however they are not equally accurate across discrete race categories. Controlling for the likely influence of the manually annotated ground truth of Fairface is problematic in forming a conclusion that definitively points to bias in either Fairface or Azure CS. It is clear that bias is a characteristic of computer vision implementations when applied to facial inference tasks. This study focused on racial categorizations however other sets of non-discrete human facial inference categories could conceivably exist and research in this area could reveal avenues for a deeper understanding of these as yet hidden non-discrete sets.

Conclusion

Without development of bias balancing technologies or techniques in training dataset development, widespread deployment of facial inference technologies will see artifacts of systemic bias propagate into their implementation results. Bias mitigation technologies will benefit from the establishment of a large verified discrimination aware dataset of human facial morphology types free of the artifacts of human annotation bias to serve as a common standard for training and validation of emerging facial inference applications and technologies. Further study into the sensitivity of emerging facial inference model architectures against a gold standard training and validation dataset for accurate and bias free inference (if it is possible) may reveal novel bias propagation pathways that are as yet not well understood.

References

Buolamwini, J., & Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. *Proceedings of Machine Learning Research*, *81*, 77–91. https://doi.org/http://proceedings.mlr.press/v81/buolamwini18a.html

Farley, P. (n.d.). *Call the detect API - face - azure cognitive services*. Face - Azure Cognitive Services | Microsoft Docs. Retrieved October 26, 2021, from https://docs.microsoft.com/en-us/azure/cognitive-services/face/face-api-how-to-topics/howtodetectfacesinimage.

Karkkainen, K., & Joo, J. (2021). Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*. https://doi.org/10.1109/wacv48630.2021.00159

The chi squared tests: The BMJ. The BMJ | The BMJ: leading general medical journal. Research. Education. Comment. (2021, April 12). Retrieved October 26, 2021, from https://www.bmj.com/about-bmj/resources-readers/publications/statistics-square-one/8-chi-squared-tests.

Brooks, K. (2021). Evaluating Age and Gender Inference Bias: Microsoft Azure Cognitive Services (Version 1) [Computer software]. https://github.com/kierankyllo/CS280RP

Appendix 1 – Python Script

```
import time
from urllib.parse import urlparse
from io import BytesIO
from PIL import Image, ImageDraw
from azure.cognitiveservices.vision.face import FaceClient
from msrest.authentication import CognitiveServicesCredentials
from azure.cognitiveservices.vision.face.models import TrainingStatusType, Person
import pandas as pd

# Defines a function to expose the raster to the API given a path and the client object
# returns unknown or an (age, gender) tuple
def getFaceDetails(path_string, face_client):

    detected_faces = face_client.face.detect_with_stream(
                open(path_string, 'rb'),
                # You can use enum from FaceAttributeType, or direct string
                return_face_attributes=[
                    'age',
                    'gender'
                ]
            )
    if not detected_faces:
        unknown = 'undetected'
        return unknown, unknown

    age = int(detected_faces[0].face_attributes.age)
    gender = str(detected_faces[0].face_attributes.gender)
    gender = gender[7:]

    return age, gender

# Define sample size
sample_size = 67228

# This key will serve all examples in this document.
KEY = "ENDPOINT KEY RECIEVED FROM Azure"

# This is the endpoint provided to you from Azure
ENDPOINT = "https://fairface.cognitiveservices.azure.com/"

# Define the path for the fairface raster data, available at https://github.com/joojs/fairface
face_file_path = 'FACE FILE ROOT PATH NOT INCLUDING /VAL'

# Create an authenticated FaceClient object
face_client = FaceClient(ENDPOINT, CognitiveServicesCredentials(KEY))

# Import the training image labels into a dataframe, file must be in root of script
df = pd.read_csv('fairface_label_train.csv', delimiter= ',', skiprows=0, nrows=sample_size, usecols=[0,1,2,3] )

# Add new labels for inference values
df['age_infer'] = ''
df['gender_infer'] = ''

# Timer start
tick = time.perf_counter()

# Iterate through the dataframe and execute inference on test images and record the results in the dataframe
for index, row in df.iterrows():
    age, gender = getFaceDetails(face_file_path+row['file'], face_client)
    #write data to df
    row['age_infer'] = age
    row['gender_infer'] = gender
    #print activity to terminal
    print(row[0], row['age_infer'], row['gender_infer'])

# Timer end
tock = time.perf_counter()

# Output results dataframe to csv
df.to_csv('results.csv')

# Report the experiment completion time
print("Completed", sample_size, f"trials in {tock - tick:0.2f} seconds")
```