Social Bias in Machine Learning Image Classification

Luther Rice / GW Undergraduate Research Award Proposal

Ryan Steed
The George Washington University
ryansteed@gwu.edu

February 28, 2019

Abstract

The proliferating applications of machine learning are increasingly important to elements of society from facial recognition to criminal justice and hiring practices, but these techniques are susceptible to numerous forms of sociocultural bias. I propose to investigate the prejudices in machine learning applications involving human faces. Using both a traditional and a transfer learning approach, I will apply the popular Inception image classification model as a feature extractor to train a regression model on computer-generated faces to predict first-impression emotional responses, such as trustworthiness. I will compare the emotional response predictions for popular image labels, including race, sexuality, and gender, to test whether common facial image classification techniques can learn emotional prejudices about categories of people. What do these comparisons reveal about the sociocultural biases imprinted on images, annotations, and image collection methods? The conclusions of this investigation may be applied to discover and negate new sources of bias.

Project Description

Problem Statement

Machine learning techniques have been applied with great success to image classification and pattern recognition tasks for the purposes of remote sensing, face recognition, video screening for job applicants, and more [7]. While the performance of these models improves with the quantity and variance of data used to learn patterns, or training data, a technique called "transfer learning" uses the learned patterns from one model to improve the performance of a different model designed to classify similar but separate data [11].

But because the performance of these models depend on both the training data and the annotations used to label them, systematic biases in either source of data could result in biased predictions. Culturally, *a priori* bias often includes harmful stereotypes and introduces problems of unfairness or prejudice into subsequent conclusions. Often, these first-impression biases can be quantified with implicit association tests [3]. For example, Todorov and Willis measure the immediate judgments people make about others' faces on first sight, recording a spectrum of emotional responses - from trustworthiness to aggressiveness - after less than a second of exposure to computer-generated faces [12, 9]. These emotional labels can be used to compare faces on the basis of gender, ethnicity and other factors.

Previous applications of unsupervised machine learning methods demonstrated the existence of social and cultural biases in the statistical properties of language, but little research has been conducted with respect to the biases in transfer learning models or image classifiers for faces or people [1, 10]. Using facial impression data, I will investigate the propagation of bias through facial image classification by measuring the association of machine learning predictions and ground-truth emotional responses. What is the typical facial judgment bias of a machine learned image classifier, using industry-standard data and other data? Does the use of transfer learning to improve model performance augment sociocultural bias acquisition, or does a tuned deep learning approach exacerbate mimicry of stereotypes?

Research Plan

I hypothesize that off-the-shelf machine learning techniques, including pre-trained models, propagate first-impression stereotypes and prejudices from annotated image data. I will test this hypothesis by training a machine learning model to predict emotional ratings for a set of popular images of people used for image classification. I will measure the statistical association between predicted emotional ratings and image category to determine whether a machine learning model demonstrates human prejudices when analyzing human faces.

My methods are as follows: I will train an off-the-shelf convolutional neural network, Inception V3, on the 300 randomly computer-generated images measured by Todorov et. al's first impression association test [9]. I will conduct expreriments with two types of models: a transfer learning approach with a pre-trained model, which makes my results reproducible for many industrial applications; and a custom neural network, which provides more accurate and applicable estimates. Each of these images is labelled with a vector of emotional ratings, where each cell in the vector is a real number representing the strength of a particular emotion (e.g., "trustworthy," "calm" and "ambitious"). Using the bottleneck layer of the trained neural network (M1), I will extract features for each of the labelled images and use the features to train a regression model (M2) for predicting emotional responses to faces. My test data consists of images of people from the ImageNet "person" categories [8, 2]. These images are annotated with one of thousands of different classes, such as "homosexual," "black" and "female." I will choose labels according to their popularity so that the images tested accord with images commonly used to train image classifiers. I will use a face recognition algorithm to crop and normalize these images to match the computer-generated training set, removing hair and obstructions and standardizing backgrounds.

Finally, I will predict emotional responses on extracted features for each of the images in the test set with the regression model (M2). The predicted emotional vectors for each image form an association matrix that can be used to find statistically significant associations between emotions and popular person categories (e.g. "untrustworthy" and "black"). A modified version of the Word Embedding Association Test (WEAT) can be used to test the null hypothesis that there is no difference between two target categories (e.g. "homosexual" and "heterosexual") and an emotional trait [1]. Alternatively, since each category represents a binary label (e.g. "black" or "not black"), simpler statistical tests of correlation could also be used to provide evidence that the mean emotional prediction differs from uniformly random selection for each category. I could also train a classifier and perform statistical tests on binary associations, but using continuous emotional ratings is much more precise. In either case, this method answers the question: are the features extracted from labelled images of people in the ImageNet database using standard machine learning techniques associated with emotional first-impression prejudices?

Outcomes

If there are significant associations between images of people and emotional prejudices, I can claim that standard industry techniques for image classification are susceptible to first-impression bias injected by image annotators during the labelling process. Furthermore, any other associations I measure may reveal novel human prejudices which are also likely to appear in any application of similarly constructed artificially intelligent systems. My findings will be critical to the proliferating use of machine learning applications in societal settings where prejudice is harmful. There is a wealth of literature measuring the stereotypes perpetuated by image classifiers and other machine learning models, from search results to automated captioning [5, 4, 6]. Knowing the specific effects of bias in facial recognition not only opens the door to counteracting bias *a priori*, but also provides a key to new insights about human first-impression biases. As such, the results of this research will be particularly useful to AI and machine learning practitioners or statisticians wishing to avoid cultural stereotypes, psychologists studying prejudice and human interpretation of facial structures and expressions, and any public policy concerning fairness and bias in technology.

Working with Professor Caliskan, I expect to produce a publication-ready paper and software package by the end of the 2019 Fall semester, though I hope to finish the majority of my research in Spring 2019. The extent of my investigation (and exact duration of my project) will depend on the significance of the results from my exploratory experimental setup. I hope to extend my research to produce bias-countering algorithms and unsupervised applications that measure even more fundamental bias in image data in Spring 2020.

References

[1] A. Caliskan, J. J. Bryson, and A. Narayanan. Semantics Derived Automatically from Language Corpora Contain Human-like Biases. Technical report, Science, 2017.

[2] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

[3] A. G. Greenwald, D. E. McGhee, and J. L. Schwartz. Measuring Individual Differences in Implicit Cognition: The Implicit Association Test. *Journal of Personality and Social Psychology*, 74(6):1464–80, 6 1998.

[4] L. A. Hendricks, K. Burns, K. Saenko, T. Darrell, and A. Rohrbach. Women Also Snowboard: Overcoming Bias in Captioning Models. *CoRR*, 2018.

[5] M. Kay, C. Matuszek, and S. A. Munson. Unequal Representation and Gender Stereotypes in Image Search Results for Occupations. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems - CHI '15*, pages 3819–3828, New York, New York, USA, 2015. ACM Press.

[6] J. Kleinberg, H. Lakkaraju, J. Leskovec, J. Ludwig, and S. Mullainathan. Human Decisions and Machine Predictions. Technical Report 23180, National Bureau of Economic Research, 2 2017.

[7] D. Lu and Q. Weng. A Survey of Image Classification Methods and Techniques for Improving Classification Performance. *International Journal of Remote Sensing*, 28(5):823–870, 3 2007.

[8] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115(3):211–252, 12 2015.

[9] A. Todorov. *Face Value: The Irresistible Influence of First Impressions.* Princeton University Press, Princeton, 2017.

[10] A. Torralba and A. A. Efros. Unbiased Look at Dataset Bias. Technical report, CVPR, 2011.

[11] K. Weiss, T. M. Khoshgoftaar, and D. Wang. A Survey of Transfer Learning. *Journal of Big Data*, 3(1):9, 12 2016.

[12] J. Willis and A. Todorov. First Impressions. *Psychological Science*, 17(7):592–598, 7 2006.