Social Bias in Machine Learning Image Classification

Luther Rice / GW Undergraduate Research Award Proposal

Ryan Steed
The George Washington University
ryansteed@gwu.edu

November 27, 2018

Abstract

The proliferating applications of machine learning are susceptible to numerous forms of social and cultural bias. Machine learning is an extremely useful statistical technique to infer patterns from data, such as text and images. I propose to investigate the bias in machine learned image classifications of human faces. Using a transfer learning approach, I will re-train the popular Inception image classification model on images of people from ImageNet and compare the model's classifications to human subjects' emotional responses after a brief glimpse of computer generated facial features. 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

By comparing machine learning image classification predictions to split-second emotional responses to images of human faces, I would like to show that off-the-shelf machine learning techniques propagate stereotypes and prejudices from annotated image data, and investigate the transmission of these biases under several different training conditions.

Machine learning is the application of statistical techniques to infer patterns in data, especially where patterns are distributed across many weak signals and the relationships between the dependent variables are convoluted. Using "supervised" machine learning techniques, labelled data can be used to predict the classification of new data. For example, given the ImageNet database, a vast collection of images of labelled with short categorical descriptions, a machine learning model may be used to predict the correct category for a new image [9]. 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 [8]. 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 [13].

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. Bias includes any *a priori* information used to draw an intelligent conclusion [1]. Culturally, prior knowledge often includes harmful stereotypes and introduces problems of unfairness or prejudice into subsequent conclusions. 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 [2, 12].

Consequently, I wish to investigate the biases that result from learning conducted on annotated image data. But to demonstrate the existence of social and cultural bias in machine learned image classifiers, I must compare model predictions to data that has already been tagged with human responses, a ground-truth target set. Todorov and Willis measure the immediate judgments people make about others' faces, recording emotional responses - from trustworthiness to aggressiveness - after less than a second of exposure to a computer-generated faces [14, 11]. These emotional labels can be used to compare faces on the basis of gender, ethnicity and other factors.

Using facial impression data, I can investigate the propagation of bias through facial image classification by measuring the association of machine learning class 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? Do these associations match biases demonstrated by social psychological exams like the Implicit Association Test, and can these associations be inferred for objects other than faces [4]? Most importantly, do these associations reveal any previously unacknowledged and uncovered prejudices?

Outcomes

The answers to these questions are critical to the increasing 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 [6, 5, 7]. 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. If demonstrated implicit association biases appear in my artificial association measurements, I have grounds to claim that other associations I measure may reveal novel human prejudices and that these prejudices are likely to appear in any application of similarly constructed artificially intelligent systems. 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.

Research Plan

Working with Professor Caliskan, I expect to produce a publication-ready paper 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. Possible extensions include alteration of both the model type and training data used - for instance, how does using tailored convolutional neural network instead of a transfer learning model affect bias? What other people categories can be investigated, using other valid annotated images?

1. **Munge.** Scrape labelled images of people from the ImageNet "person..." category [3]. These images are all annotated with thousands of different classes, such as "homosexual," "black" and "female." Process each image to meet Inception image processing standards and facial recognition standards (removing hair, obstructions, etc.).

2. **Train.** Re-train Google's Inception model, a pre-trained image classifier trained on the general ImageNet database, on processed images of faces with specific people category annotations [10]. Later, try training a custom convolutional neural network instead of the more general Inception model to test the effect of transfer learning on associations. These tasks require a powerful GPU and desktop computer or remote virtual machine for scraping, feature extraction, processing, training, and classification.

3. **Test.** Using the trained image classifier, classify each of the annotated facial images from Todorov et. al [14], each tagged with an emotional response (e.g., "trustworthy," "calm" and "ambitious"), to produce a matrix of categories by emotional responses.

4. **Evaluate.** Regress emotional response on predicted category for each of the emotions and create novel methods for bias analysis to quantify the association. Are the machine-predicted people categories associated with split-second human emotional responses? (E.g., does a face classified as "black" have a low "trustworthy" score?) Do these assocations reveal any new insights about social, human, or evolutionary psychology?

References

[1] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

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

[3] 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.

[4] 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.

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

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going Deeper with Convolutions. Technical report, University of North Carolina at Chapel Hill, 2015.

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

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

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

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