How Does a Computer "See" Gender?

Demonstrating Uncertainty in Deep Learning Classification Models Though Data Visualization and Interactive Play

Christopher Baronavski
cbaronavski@pewresearch.org
Pew Research Center
Washington, District of Columbia

Peter Bell
pbell@pewresearch.org
Pew Research Center
Washington, District of Columbia

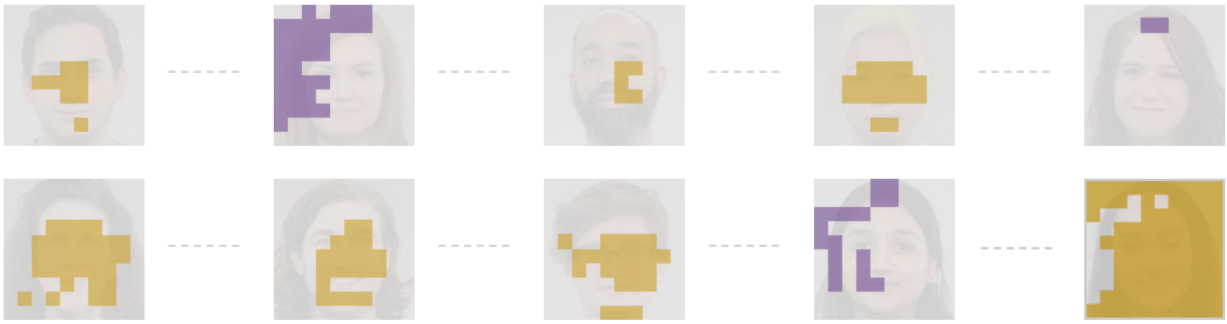

Figure 1: Progression through the interactive "Confound the Machine Vision System" challenge demonstrates that consistent, unambiguous patterns of facial features salient to the deep learning classification model are difficult to discern. See more at pewrsr.ch/cv.

ABSTRACT

A recent Pew Research Center survey found that a majority of Americans believe facial recognition technology can accurately identify individuals, as well as correctly determine gender and race. Additionally, a majority of Americans say that they trust law enforcement agencies to use facial recognition technology responsibly[3]. However, in practical use, machine learning applications are susceptible to biases[1] [6], and implementations of facial recognition and machine vision technologies may yield unpredictable, unexpected, or inaccurate results[9] [4].

Building on previous applications of deep learning gender classification of images, researchers at the Center sought to better understand how a concept such as "gender" is encoded in such a model. By systematically occluding regions of individual photographs of human faces and then submitting each occluded image for processing by a gender classification model, each region of the photo can be evaluated for its significance in the determination of the gender of the individual represented in that image. Indeed, when certain regions are occluded, it causes the classifier to completely switch its initial decision about the subject's gender.

A significant limitation of many gender classification systems (including the one utilized in the present study) is that they cannot account for subjects who do not identify as either a woman or a man, despite indications that about 1 in 5 adults in the U.S. know someone who identifies by a pronoun other than "he" or "she" [2]. Notwithstanding, there is limited evidence that marketers are already using public machine vision gender classification to target consumers, sparking concerns about privacy and misgendering of individuals [7] [8]. Additionally, unlike typical algorithms which tend to yield a predictable outcome after executing a predetermined set of steps, given the level of complexity inherent in deep learning classification systems, even the data scientists who design such systems often do not fully understand exactly how the classifier's decision is reached. As such, the information that has been encoded by training a deep learning classifier, essentially what has been "learned," has been described as a "black box"[5].

In light of these factors, we devised an interactive editorial feature that would expose users to the nitty-gritty decision making process of a machine learning system. This is how it works: After accepting the "Confound the Machine Vision System" challenge, users are shown a series of images of individual's faces over which a 10 by 10 square grid is laid. For each image, the user is presented 5 randomly selected square regions which correspond to regions of each subject's face, but only 1 of which is operant in the model's determination of the gender of the individual represented in that image. The user is prompted to find the square that will "confound" the machine vision system when it is occluded, thereby inverting the gender determination that the model has initially made. When rolling over an available square, that underlying region of the image is magnified to encourage careful consideration of the facial features corresponding to that region and how much these features (facial hair, eyebrows, ears, mouth, etc.) contribute to the model's determination of the subject's gender, often unexpectedly (Figure 2). Efforts were made to design an accessible interface, including both visual and auditory cues, as well as haptic feedback (when supported by the user's mobile device).

As users progress through the challenge, a data visualization which depicts the regions of each image that are most salient to the gender classification model accumulates below the grid, allowing users to directly compare each image (Figure 1). Notably, a consistent pattern across images fails to emerge. For example, a region of a particular facial image which seems trivial to a human observer, might be salient for the machine vision classifier. However, that corresponding region may not be salient for the image classifier as it processes subsequent images. These systems make their decisions in ways that are largely hidden from public view, and highly dependent on the data used to train them. As such, they can be prone to systematic biases and can fail in ways that are difficult to understand and hard to predict in advance.

This study describes a novel technique that incorporates audio-visual cues, data visualization, and elements of randomness and play to help users more fully appreciate the complexities and limitations inherent in machine vision and deep learning systems. By engaging users in hands-on "intervention" into a classification model's decision making process, they not only gain practical insight into the functioning of an image classifier and the uncertainty associated with its decisions, but are also prompted to think more critically about the often brittle and unpredictable nature of machine vision as this technology continues to encroach into their daily lives. Furthermore, as artificial intelligence, deep learning, and machine vision become increasingly prevalent topics in public discourse, this interactive challenge provides a model for other journalists seeking to explicate the complex and sometimes indeterminate inner workings of machine vision technologies.

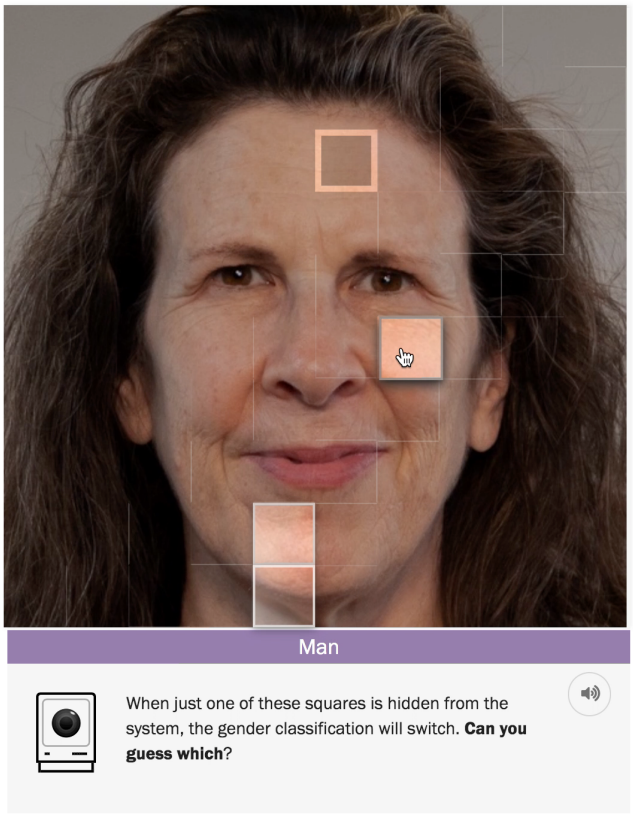

Figure 2: Choices are presented to the user within the interface of the interactive challenge.

KEYWORDS

machine vision, deep learning, data visualization, play

REFERENCES

[1] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. *Machine Bias: There's software used across the country to predict future criminals. And it's biased against blacks.* Retrieved December 10, 2019 from https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[2] Pew Research Center. 2019. *About one-in-five U.S. adults know someone who goes by a gender-neutral pronoun.* Retrieved February 12, 2020 from https://www.pewresearch.org/fact-tank/2019/09/05/gender-neutral-pronouns/

[3] Pew Research Center. 2019. *More Than Half of U.S. Adults Trust Law Enforcement to Use Facial Recognition Responsibly.* Retrieved December 10, 2019 from https://www.pewresearch.org/internet/2019/09/05/more-than-half-of-u-s-adults-trust-law-enforcement-to-use-facial-recognition-responsibly/

[4] Anita Chabria. 2019. *Facial recognition software mistook 1 in 5 California lawmakers for criminals, says ACLU.* Retrieved December 10, 2019 from https://www.latimes.com/california/story/2019-08-12/facial-

recognition-software-mistook-1-in-5-california-lawmakers-for-criminals-says-aclu

[5] Kalev Leetaru. 2019. *We Must Recognize Just How Brittle And Unpredictable Today's Correlative Deep Learning AI Is.* Retrieved December 10, 2019 from https://www.forbes.com/sites/kalevleetaru/2019/06/24/we-must-recognize-just-how-brittle-and-unpredictable-todays-correlative-deep-learning-ai-is/#412f7b035bb1

[6] Steve Lohr. 2018. *Facial Recognition Is Accurate, if You're a White Guy.* Retrieved December 10, 2019 from https://www.nytimes.com/2018/02/09/technology/facial-recognition-race-artificial-intelligence.html

[7] Tomoyoshi Otsu and Yu Miyaji. 2019. *Taxi face recognition, Google officials introduce criticism gushing out [In Japanese].* Retrieved February 13, 2020 from https://www.asahi.com/articles/ASMBJ5H29MBJULFA025.html

[8] Annie Palmer. 2019. *Privacy fears as Tokyo taxis use facial recognition cameras to guess riders' age and gender for targeted advertisements.* https://www.dailymail.co.uk/sciencetech/article-6951727/Tokyo-taxis-use-facial-recognition-guess-riders-age-gender-targeted-advertisements.html

[9] Roman V. Yampolskiy. 2019. Unpredictability of AI. arXiv:cs.AI/1905.13053