Do trucks mean Trump? AI shows how humans misjudge images

September 20 2022, by Patricia Waldron

Credit: Unsplash/CC0 Public Domain

A study on the types of mistakes that humans make when evaluating

images may enable computer algorithms that help us make better decisions about visual information, such as while reading an X-ray or moderating online content.

Researchers from Cornell and partner institutions analyzed more than 16 million human predictions of whether a neighborhood voted for Joe Biden or Donald Trump in the 2020 presidential election based on a single Google Street View image. They found that humans as a group performed well at the task, but a computer algorithm was better at distinguishing between Trump and Biden country.

The study also classified common ways that people mess up, and identified objects—such as pickup trucks and American flags—that led people astray.

"We're trying to understand, where an algorithm has a more effective prediction than a human, can we use that to help the human, or make a better hybrid human-machine system that gives you the best of both worlds?" said first author J.D. Zamfirescu-Pereira, a graduate student at the University of California at Berkeley.

He presented the work, entitled "Trucks Don't Mean Trump: Diagnosing Human Error in Image Analysis," at the *2022 Association for Computing Machinery (ACM) Conference on Fairness, Accountability, and Transparency* (FAccT).

Recently, researchers have given a lot of attention to the issue of algorithmic bias, which is when algorithms make errors that systematically disadvantage women, racial minorities, and other historically marginalized populations.

"Algorithms can screw up in any one of a myriad of ways and that's very important," said senior author Emma Pierson, assistant professor of

computer science at the Jacobs Technion-Cornell Institute at Cornell Tech and the Technion with the Cornell Ann S. Bowers College of Computing and Information Science. "But humans are themselves biased and error-prone, and algorithms can provide very useful diagnostics for how people screw up."

The researchers used anonymized data from a New York Times interactive quiz that showed readers snapshots from 10,000 locations across the country and asked them to guess how the neighborhood voted. They trained a machine learning algorithm to make the same prediction by giving it a subset of Google Street View images and supplying it with real-world voting results. Then they compared the performance of the algorithm on the remaining images with that of the readers.

Overall, the machine learning algorithm predicted the correct answer about 74% of the time. When averaged together to reveal "the wisdom of the crowd," humans were right 71% of the time, but individual humans scored only about 63%.

People often incorrectly chose Trump when the street view showed pickup trucks or wide-open skies. In a New York Times article, participants noted that American flags also made them more likely to predict Trump, even though neighborhoods with flags were evenly split between the candidates.

The researchers classified the human mistakes as the result of bias, variance, or noise—three categories commonly used to evaluate errors from machine learning algorithms. Bias represents errors in the wisdom of the crowd—for example, always associating pickup trucks with Trump. Variance encompasses individual wrong judgments—when one person makes a bad call, even though the crowd was right, on average. Noise is when the image doesn't provide useful information, such as a house with a Trump sign in a primarily Biden-voting neighborhood.

Being able to break down human errors into categories may help improve human decision-making. Take radiologists reading X-rays to diagnose a disease, for example. If there are many errors due to bias, then doctors may need retraining. If, on average, diagnosis is successful but there is variance between radiologists, then a second opinion might be warranted. And if there is a lot of misleading noise in the X-rays, then a different diagnostic test may be necessary.

Ultimately, this work can lead to a better understanding of how to combine human and machine decision-making for human-in-the-loop systems, where humans give input into otherwise automated processes.

"You want to study the performance of the whole system together—humans plus the algorithm, because they can interact in unexpected ways," Pierson said.

More information: J.D. Zamfirescu-Pereira et al, Trucks Don't Mean Trump: Diagnosing Human Error in Image Analysis, *2022 ACM Conference on Fairness, Accountability, and Transparency* (2022). DOI: 10.1145/3531146.3533145

Provided by Cornell University