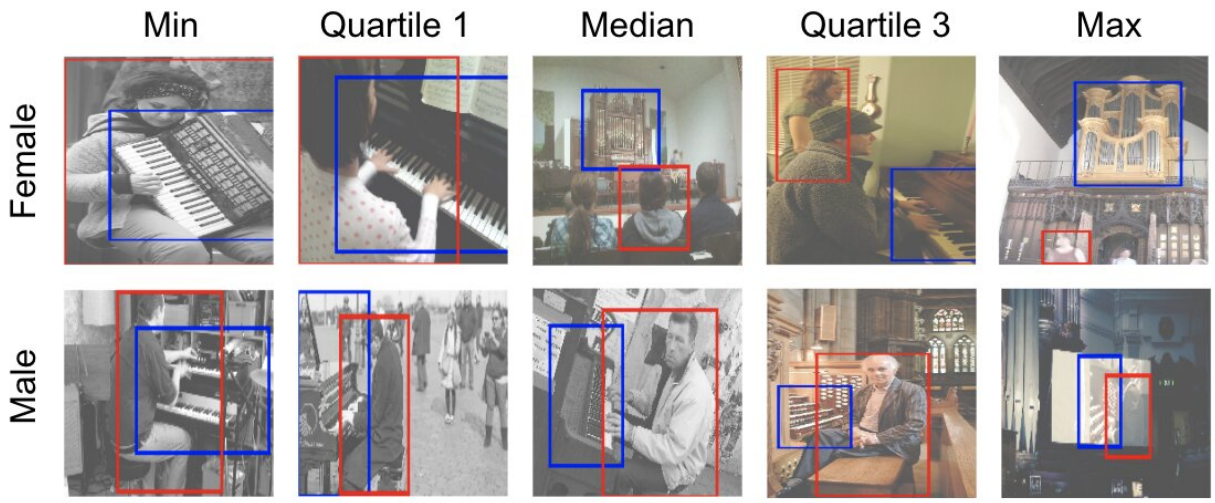
Tool helps clear biases from computer vision

October 2 2020, by Molly Sharlach



In one data set, REVISE uncovered a potential gender bias in images containing people (red boxes) and the musical instrument organ (blue boxes). Analyzing the distribution of inferred 3-D distances between the person and the organ showed that males tended to be featured as actually playing the instrument, whereas females were often merely in the same space as the instrument. Credit: Princeton Visual AI Lab

Researchers at Princeton University have developed a tool that flags potential biases in sets of images used to train artificial intelligence (AI) systems. The work is part of a larger effort to remedy and prevent the biases that have crept into AI systems that influence everything from credit services to courtroom sentencing programs.

Although the sources of bias in AI systems are varied, one major cause is stereotypical images contained in large sets of images collected from online sources that engineers use to develop computer vision, a branch of AI that allows computers to recognize people, objects and actions. Because the foundation of computer vision is built on these data sets, images that reflect societal stereotypes and biases can unintentionally influence computer vision models.

To help stem this problem at its source, researchers in the Princeton Visual AI Lab have developed an open-source tool that automatically uncovers potential biases in visual data sets. The tool allows data set creators and users to correct issues of underrepresentation or stereotypical portrayals before image collections are used to train computer vision models. In related work, members of the Visual AI Lab published a comparison of existing methods for preventing biases in computer vision models themselves, and proposed a new, more effective approach to bias mitigation.

The first tool, called REVISE (REvealing VIsual biaSEs), uses statistical methods to inspect a data set for potential biases or issues of underrepresentation along three dimensions: object-based, gender-based and geography-based. A fully automated tool, REVISE builds on earlier work that involved filtering and balancing a data set's images in a way that required more direction from the user. The study was presented Aug. 24 at the virtual European Conference on Computer Vision.

REVISE takes stock of a data set's content using existing image annotations and measurements such as object counts, the co-occurrence of objects and people, and images' countries of origin. Among these measurements, the tool exposes patterns that differ from median distributions.

For example, in one of the tested data sets, REVISE showed that images

including both people and flowers differed between males and females: Males more often appeared with flowers in ceremonies or meetings, while females tended to appear in staged settings or paintings. (The analysis was limited to annotations reflecting the perceived binary gender of people appearing in images.)

Once the tool reveals these sorts of discrepancies, "then there's the question of whether this is a totally innocuous fact, or if something deeper is happening, and that's very hard to automate," said Olga Russakovsky, an assistant professor of computer science and principal investigator of the Visual AI Lab. Russakovsky co-authored the paper with graduate student Angelina Wang and Arvind Narayanan, an associate professor of computer science.

For example, REVISE revealed that objects including airplanes, beds and pizzas were more likely to be large in the images including them than a typical object in one of the data sets. Such an issue might not perpetuate societal stereotypes, but could be problematic for training computer vision models. As a remedy, the researchers suggest collecting images of airplanes that also include the labels mountain, desert or sky.

The underrepresentation of regions of the globe in computer vision data sets, however, is likely to lead to biases in AI algorithms. Consistent with previous analyses, the researchers found that for images' countries of origin (normalized by population), the United States and European countries were vastly overrepresented in data sets. Beyond this, REVISE showed that for images from other parts of the world, image captions were often not in the local language, suggesting that many of them were captured by tourists and potentially leading to a skewed view of a country.

In an example of a geographic discrepancy, REVISE found that images annotated as "dish" tended to refer to food items in Eastern Asia, rather than satellite dishes or plates, which were more common in images collected from other regions. Credit: Princeton Visual AI Lab

Researchers who focus on object detection may overlook issues of fairness in computer vision, said Russakovsky. "However, this geography analysis shows that object recognition can still can be quitebiased and exclusionary, and can affect different regions and people unequally," she said.

"Data set collection practices in computer science haven't been scrutinized that thoroughly until recently," said co-author Angelina Wang, a graduate student in computer science. She said images are mostly "scraped from the internet, and people don't always realize that their images are being used [in data sets]. We should collect images from more diverse groups of people, but when we do, we should be careful that we're getting the images in a way that is respectful."

"Tools and benchmarks are an important step ... they allow us to capture these biases earlier in the pipeline and rethink our problem setup and assumptions as well as data collection practices," said Vicente Ordonez-Roman, an assistant professor of computer science at the University of Virginia who was not involved in the studies. "In computer vision there are some specific challenges regarding representation and the propagation of stereotypes. Works such as those by the Princeton Visual

AI Lab help elucidate and bring to the attention of the computer vision community some of these issues and offer strategies to mitigate them."

A related study from the Visual AI Lab examined approaches to prevent computer vision models from learning spurious correlations that may reflect biases, such as overpredicting activities like cooking in images of women, or computer programming in images of men. Visual cues such as the fact that zebras are black and white, or basketball players often wear jerseys, contribute to the accuracy of the models, so developing effective models while avoiding problematic correlations is a significant challenge in the field.

In research presented in June at the virtual International Conference on Computer Vision and Pattern Recognition, electrical engineering graduate student Zeyu Wang and colleagues compared four different techniques for mitigating biases in computer vision models.

They found that a popular technique known as adversarial training, or "fairness through blindness," harmed the overall performance of image recognition models. In adversarial training, the model cannot consider information about the protected variable—in the study, the researchers used gender as a test case. A different approach, known as domain-independent training, or "fairness through awareness," performed much better in the team's analysis.

"Essentially, this says we're going to have different frequencies of activities for different genders, and yes, this prediction is going to be gender-dependent, so we're just going to embrace that," said Russakovsky.

The technique outlined in the paper mitigates potential biases by considering the protected attribute separately from other visual cues.

"How we really address the bias issue is a deeper problem, because of course we can see it's in the data itself," said Zeyu Wang. "But in in the real world, humans can still make good judgments while being aware of our biases"—and computer vision models can be set up to work in a similar way, he said.

Provided by Princeton University