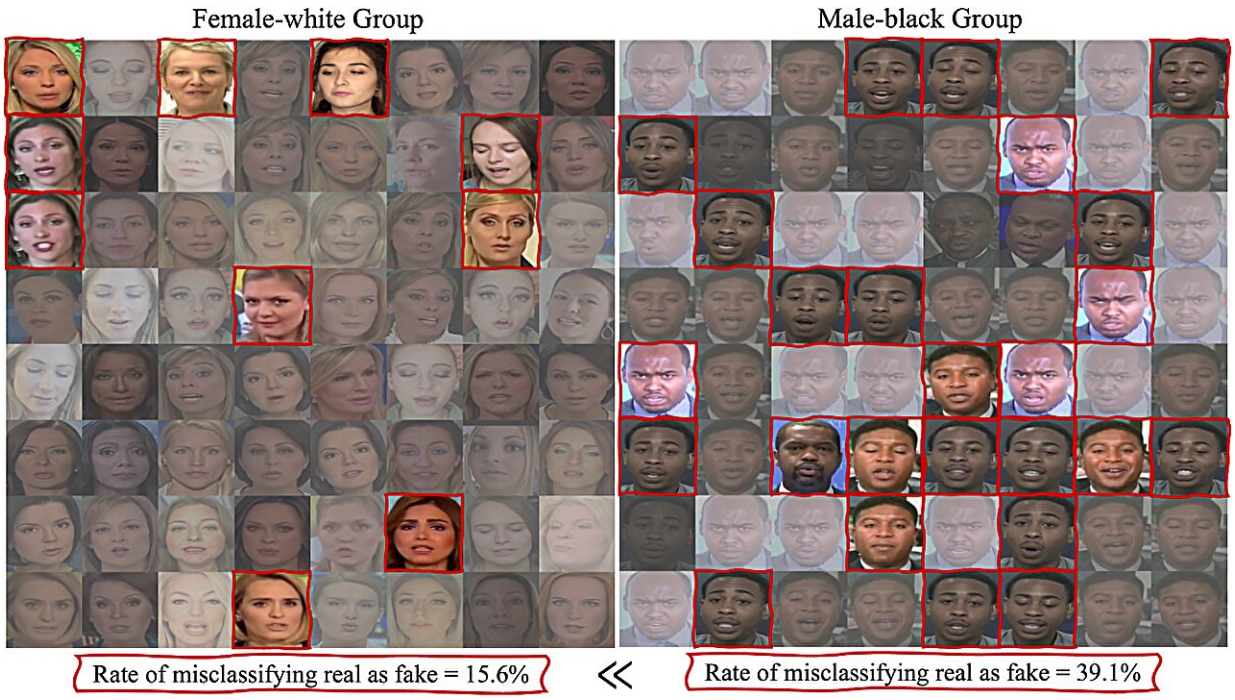
Team develops a new deepfake detector designed to be less biased

January 16 2024, by Tom Dinki



Deepfake detection algorithms often perform differently across races and genders, including a higher false positive rate on Black men than on white women. New algorithms developed at the University at Buffalo are designed to reduce such gaps. Credit: Siwei Lyu

The image spoke for itself. University at Buffalo computer scientist and deepfake expert Siwei Lyu created a photo collage out of the hundreds of faces that his detection algorithms had incorrectly classified as

fake—and the new composition clearly had a predominantly darker skin tone.

"A detection algorithm's accuracy should be statistically independent from factors like race," Lyu says, "but obviously many existing algorithms, including our own, inherit a bias."

Lyu, Ph.D., co-director of the UB Center for Information Integrity, and his team have now developed what they believe are the first-ever deepfake detection algorithms specifically designed to be less biased.

Their two machine learning methods—one that makes algorithms aware of demographics and one that leaves them blind to them—reduced disparities in accuracy across races and genders, while, in some cases, still improving overall accuracy.

The research, published on the *arXiv* preprint server, was presented at the Winter Conference on Applications of Computer Vision (WACV), held Jan. 4–8.

Lyu, the study's senior author, collaborated with his former student, Shu Hu, Ph.D., now an assistant professor of computer and information technology at Indiana University-Purdue University Indianapolis, as well as George Chen, Ph.D., assistant professor of information systems at Carnegie Mellon University. Other contributors include Yan Ju, a Ph.D. student in Lyu's Media Forensic Lab at UB, and postdoctoral researcher Shan Jia.

Ju, the study's first author, says detection tools are often less scrutinized than the artificial intelligence tools they keep in check, but that doesn't mean they don't need to be held accountable, too.

"Deepfakes have been so disruptive to society that the research

community was in a hurry to find a solution," she says, "but even though these algorithms were made for a good cause, we still need to be aware of their collateral consequences."

Demographic aware vs. demographic agnostic

Recent studies have found large disparities in deepfake detection algorithms' error rates—up to a 10.7% difference in one study—among different races. In particular, it's been shown that some are better at guessing the authenticity of lighter-skinned subjects than darker-skinned ones.

This can result in certain groups being more at risk of having their real image pegged as a fake, or perhaps even more damaging, a doctored image of them pegged as real.

The problem is not necessarily the algorithms themselves, but the data they've been trained on. Middle-aged white men are often overly represented in such photo and video datasets, so the algorithms are better at analyzing them than they are underrepresented groups, says Lyu, SUNY Empire Professor in the UB Department of Computer Science and Engineering, within the School of Engineering and Applied Sciences.

"Say one demographic group has 10,000 samples in the dataset and the other only has 100. The algorithm will sacrifice accuracy on the smaller group in order to minimize errors on the larger group," he adds. "So it reduces overall errors, but at the expense of the smaller group."

While other studies have attempted to make databases more demographically balanced—a time-consuming process—Lyu says his team's study is the first attempt to actually improve the fairness of the algorithms themselves.

To explain their method, Lyu uses an analogy of a teacher being evaluated by student test scores.

"If a teacher has 80 students do well and 20 students do poorly, they'll still end up with a pretty good average," he says. "So instead we want to give a weighted average to the students around the middle, forcing them to focus more on everyone instead of the dominating group."

First, their demographic-aware method supplied algorithms with datasets that labeled subjects' gender—male or female—and race—white, Black, Asian or other—and instructed it to minimize errors on the less represented groups.

"We're essentially telling the algorithms that we care about overall performance, but we also want to guarantee that the performance of every group meets certain thresholds, or at least is only so much below the overall performance," Lyu says.

However, datasets typically aren't labeled for race and gender. Thus, the team's demographic-agnostic method classifies deepfake videos not based on the subjects' demographics—but on features in the video not immediately visible to the human eye.

"Maybe a group of videos in the dataset corresponds to a particular demographic group or maybe it corresponds with some other feature of the video, but we don't need demographic information to identify them," Lyu says. "This way, we do not have to handpick which groups should be emphasized. It's all automated based on which groups make up that middle slice of data."

Improving fairness—and accuracy

The team tested their methods using the popular FaceForensic++ dataset

and state-of-the-art Xception detection algorithm. This improved all of the algorithm's fairness metrics, such as equal false positive rate among races, with the demographic-aware method performing best of all.

Most importantly, Lyu says, their methods actually increased the overall detection accuracy of the algorithm—from 91.49% to as high as 94.17%.

However, when using the Xception algorithm with different datasets and the FF+ dataset with different algorithms, the methods—while still improving most fairness metrics—slightly reduced overall detection accuracy.

"There can be a small tradeoff between performance and fairness, but we can guarantee that the performance degradation is limited," Lyu says. "Of course, the fundamental solution to the bias problem is improving the quality of the datasets, but for now, we should incorporate fairness into the algorithms themselves."

More information: Yan Ju et al, Improving Fairness in Deepfake Detection, *arXiv* (2023). DOI: 10.48550/arxiv.2306.16635

Provided by University at Buffalo