How machine learning models can amplify inequities in medical diagnosis and treatment

August 17 2023, by Steve Nadis

Researchers investigated the fairness of machine learning models under subpopulation shift, which is critical for ensuring the safe and equitable deployment of ML models in high-stakes applications such as health care. Credit: Alex Shipps/MIT CSAIL via Midjourney

Prior to receiving a Ph.D. in computer science from MIT in 2017,

Marzyeh Ghassemi had already begun to wonder whether the use of AI techniques might enhance the biases that already existed in health care. She was one of the early researchers to take up this issue, and she's been exploring it ever since.

In a new paper, Ghassemi, now an assistant professor in MIT's Department of Electrical Science and Engineering (EECS), and three collaborators based at the Computer Science and Artificial Intelligence Laboratory, have probed the roots of the disparities that can arise in machine learning, often causing models that perform well overall to falter when it comes to subgroups for which relatively few data have been collected and utilized in the training process.

The paper—written by two MIT Ph.D. students, Yuzhe Yang and Haoran Zhang, EECS computer scientist Dina Katabi (the Thuan and Nicole Pham Professor), and Ghassemi—was presented last month at the 40th International Conference on Machine Learning in Honolulu, Hawaii.

In their analysis, the researchers focused on "subpopulation shifts"—differences in the way machine learning models perform for one subgroup as compared to another. "We want the models to be fair and work equally well for all groups, but instead we consistently observe the presence of shifts among different groups that can lead to inferior medical diagnosis and treatment," says Yang, who along with Zhang are the two lead authors on the paper.

The main point of their inquiry is to determine the kinds of subpopulation shifts that can occur and to uncover the mechanisms behind them so that, ultimately, more equitable models can be developed.

The new paper "significantly advances our understanding" of the

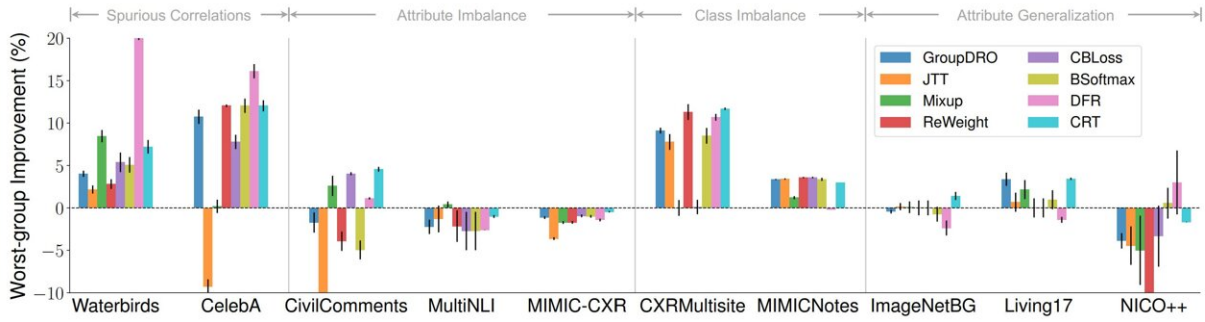
subpopulation shift phenomenon, claims Stanford University computer scientist Sanmi Koyejo. "This research contributes valuable insights for future advancements in machine learning models' performance on underrepresented subgroups."

Camels and cattle

The MIT group has identified four principal types of shifts—spurious correlations, attribute imbalance, class imbalance, and attribute generalization—which, according to Yang, "have never been put together into a coherent and unified framework. We've come up with a single equation that shows you where biases can come from."

Biases can, in fact, stem from what the researchers call the class, or from the attribute, or both. To pick a simple example, suppose the task assigned to the machine learning model is to sort images of objects—animals in this case—into two classes: cows and camels. Attributes are descriptors that don't specifically relate to the class itself. It might turn out, for instance, that all the images used in the analysis show cows standing on grass and camels on sand—grass and sand serving as the attributes here.

Given the data available to it, the machine could reach an erroneous conclusion—namely that cows can only be found on grass, not on sand, with the opposite being true for camels. Such a finding would be incorrect, however, giving rise to a spurious correlation, which, Yang explains, is a "special case" among subpopulation shifts—"one in which you have a bias in both the class and the attribute."



Worst-group improvements over ERM across different datasets when attributes are unknown in both training and validation set. SOTA algorithms only enhance subgroup robustness on certain types of shift (i.e., SC and CI). Complete results are in Appendix D.2. Credit: https://proceedings.mlr.press/v202/yang23s.html

In a medical setting, one could rely on machine learning models to determine whether a person has pneumonia or not based on an examination of X-ray images. There would be two classes in this situation, one consisting of people who have the lung ailment, another for those who are infection-free.

A relatively straightforward case would involve just two attributes: the people getting X-rayed are either female or male. If, in this particular dataset, there were 100 males diagnosed with pneumonia for every one female diagnosed with pneumonia, that could lead to an attribute imbalance, and the model would likely do a better job of correctly detecting pneumonia for a man than for a woman.

Similarly, having 1,000 times more healthy (pneumonia-free) subjects than sick ones would lead to a class imbalance, with the model biased toward healthy cases. Attribute generalization is the last shift highlighted in the new study. If your sample contained 100 male patients with pneumonia and zero female subjects with the same illness, you still

would like the model to be able to generalize and make predictions about female subjects even though there are no samples in the training data for females with pneumonia.

The team then took 20 advanced algorithms, designed to carry out classification tasks, and tested them on a dozen datasets to see how they performed across different population groups. They reached some unexpected conclusions: By improving the "classifier," which is the last layer of the neural network, they were able to reduce the occurrence of spurious correlations and class imbalance, but the other shifts were unaffected.

Improvements to the "encoder," one of the uppermost layers in the neural network, could reduce the problem of attribute imbalance. "However, no matter what we did to the encoder or classifier, we did not see any improvements in terms of attribute generalization," Yang says, "and we don't yet know how to address that."

Precisely accurate

There is also the question of assessing how well your model actually works in terms of evenhandedness among different population groups. The metric normally used, called worst-group accuracy or WGA, is based on the assumption that if you can improve the accuracy—of, say, medical diagnosis—for the group that has the worst model performance, you would have improved the model as a whole.

"The WGA is considered the gold standard in subpopulation evaluation," the authors contend, but they made a surprising discovery: boosting worst-group accuracy results in a decrease in what they call "worst-case precision." In medical decision-making of all sorts, one needs both accuracy—which speaks to the validity of the findings—and precision, which relates to the reliability of the methodology.

"Precision and accuracy are both very important metrics in classification tasks, and that is especially true in medical diagnostics," Yang explains. "You should never trade precision for accuracy. You always need to balance the two."

The MIT scientists are putting their theories into practice. In a study they're conducting with a medical center, they're looking at public datasets for tens of thousands of patients and hundreds of thousands of chest X-rays, trying to see whether it's possible for machine learning models to work in an unbiased manner for all populations. That's still far from the case, even though more awareness has been drawn to this problem, Yang says. "We are finding many disparities across different ages, gender, ethnicity, and intersectional groups."

He and his colleagues agree on the eventual goal, which is to achieve fairness in health care among all populations. But before we can reach that point, they maintain, we still need a better understanding of the sources of unfairness and how they permeate our current system. Reforming the system as a whole will not be easy, they acknowledge. In fact, the title of the paper they introduced at the Honolulu conference, "Change is Hard," gives some indications as to the challenges that they and like-minded researchers face.

More information: Paper: Change is Hard: A Closer Look at Subpopulation Shift

This story is republished courtesy of MIT News (web.mit.edu/newsoffice/), a popular site that covers news about MIT research, innovation and teaching.

Provided by Massachusetts Institute of Technology