How sure is sure? Incorporating human error into machine learning

August 9 2023

Credit: Pixabay/CC0 Public Domain

Researchers are developing a way to incorporate one of the most human of characteristics—uncertainty—into machine learning systems.

Human error and uncertainty are concepts that many artificial intelligence systems fail to grasp, particularly in systems where a human provides feedback to a machine learning model. Many of these systems are programmed to assume that humans are always certain and correct, but real-world decision-making includes occasional mistakes and uncertainty.

Researchers from the University of Cambridge, along with The Alan Turing Institute, Princeton, and Google DeepMind, have been attempting to bridge the gap between human behavior and machine learning, so that uncertainty can be more fully accounted for in AI applications where humans and machines are working together. This could help reduce risk and improve trust and reliability of these applications, especially where safety is critical, such as medical diagnosis.

The team adapted a well-known image classification dataset so that humans could provide feedback and indicate their level of uncertainty when labeling a particular image. The researchers found that training with uncertain labels can improve these systems' performance in handling uncertain feedback, although humans also cause the overall performance of these hybrid systems to drop.

Their results will be reported at the AAAI/ACM Conference on Artificial Intelligence, Ethics and Society (AIES 2023) in Montréal.

'Human-in-the-loop' machine learning systems—a type of AI system that enables human feedback—are often framed as a promising way to reduce risks in settings where automated models cannot be relied upon to make decisions alone. But what if the humans are unsure?

"Uncertainty is central in how humans reason about the world but many AI models fail to take this into account," said first author Katherine Collins from Cambridge's Department of Engineering. "A lot of

developers are working to address model uncertainty, but less work has been done on addressing uncertainty from the person's point of view."

We are constantly making decisions based on the balance of probabilities, often without really thinking about it. Most of the time—for example, if we wave at someone who looks just like a friend but turns out to be a total stranger—there's no harm if we get things wrong. However, in certain applications, uncertainty comes with real safety risks.

"Many human-AI systems assume that humans are always certain of their decisions, which isn't how humans work—we all make mistakes," said Collins. "We wanted to look at what happens when people express uncertainty, which is especially important in safety-critical settings, like a clinician working with a medical AI system."

"We need better tools to recalibrate these models, so that the people working with them are empowered to say when they're uncertain," said co-author Matthew Barker, who recently completed his MEng degree at Gonville and Caius College, Cambridge. "Although machines can be trained with complete confidence, humans often can't provide this, and machine learning models struggle with that uncertainty."

For their study, the researchers used some of the benchmark machine learning datasets: one was for digit classification, another for classifying chest X-rays, and one for classifying images of birds. For the first two datasets, the researchers simulated uncertainty, but for the bird dataset, they had human participants indicate how certain they were of the images they were looking at: whether a bird was red or orange, for example.

These annotated 'soft labels' provided by the human participants allowed the researchers to determine how the final output was changed.

However, they found that performance degraded rapidly when machines were replaced with humans.

"We know from decades of behavioral research that humans are almost never 100% certain, but it's a challenge to incorporate this into machine learning," said Barker. "We're trying to bridge the two fields, so that machine learning can start to deal with human uncertainty where humans are part of the system."

The researchers say their results have identified several open challenges when incorporating humans into machine learning models. They are releasing their datasets so that further research can be carried out and uncertainty might be built into machine learning systems.

"As some of our colleagues so brilliantly put it, uncertainty is a form of transparency, and that's hugely important," said Collins. "We need to figure out when we can trust a model and when to trust a human and why. In certain applications, we're looking at a probability over possibilities. Especially with the rise of chatbots for example, we need models that better incorporate the language of possibility, which may lead to a more natural, safe experience."

"In some ways, this work raised more questions than it answered," said Barker. "But even though humans may be mis-calibrated in their uncertainty, we can improve the trustworthiness and reliability of these human-in-the-loop systems by accounting for human behavior."

Provided by University of Cambridge