Why deep-learning methods confidently recognize images that are nonsense

December 16 2021, by Rachel Gordon

A deep-image classifier can determine image classes with over 90 percent confidence using primarily image borders, rather than an object itself. Credit: Rachel Gordon

For all that neural networks can accomplish, we still don't really understand how they operate. Sure, we can program them to learn, but making sense of a machine's decision-making process remains much like a fancy puzzle with a dizzying, complex pattern where plenty of integral pieces have yet to be fitted.

If a model was trying to classify an image of said puzzle, for example, it could encounter well-known, but annoying adversarial attacks, or even more run-of-the-mill data or processing issues. But a new, more subtle type of failure recently identified by MIT scientists is another cause for concern: "overinterpretation," where algorithms make confident predictions based on details that don't make sense to humans, like random patterns or image borders.

This could be particularly worrisome for high-stakes environments, like split-second decisions for self-driving cars, and medical diagnostics for diseases that need more immediate attention. Autonomous vehicles in particular rely heavily on systems that can accurately understand surroundings and then make quick, safe decisions. The network used specific backgrounds, edges, or particular patterns of the sky to classify traffic lights and street signs—irrespective of what else was in the image.

The team found that neural networks trained on popular datasets like CIFAR-10 and ImageNet suffered from overinterpretation. Models trained on CIFAR-10, for example, made confident predictions even when 95 percent of input images were missing, and the remainder is senseless to humans.

"Overinterpretation is a dataset problem that's caused by these nonsensical signals in datasets. Not only are these high-confidence images unrecognizable, but they contain less than 10 percent of the original image in unimportant areas, such as borders. We found that these images were meaningless to humans, yet models can still classify them with high confidence," says Brandon Carter, MIT Computer Science and Artificial Intelligence Laboratory Ph.D. student and lead author on a paper about the research.

Deep-image classifiers are widely used. In addition to medical diagnosis and boosting autonomous vehicle technology, there are use cases in

security, gaming, and even an app that tells you if something is or isn't a hot dog, because sometimes we need reassurance. The tech in discussion works by processing individual pixels from tons of pre-labeled images for the network to "learn."

Image classification is hard, because machine-learning models have the ability to latch onto these nonsensical subtle signals. Then, when image classifiers are trained on datasets such as ImageNet, they can make seemingly reliable predictions based on those signals.

Although these nonsensical signals can lead to model fragility in the real world, the signals are actually valid in the datasets, meaning overinterpretation can't be diagnosed using typical evaluation methods based on that accuracy.

To find the rationale for the model's prediction on a particular input, the methods in the present study start with the full image and repeatedly ask, what can I remove from this image? Essentially, it keeps covering up the image, until you're left with the smallest piece that still makes a confident decision.

To that end, it could also be possible to use these methods as a type of validation criteria. For example, if you have an autonomously driving car that uses a trained machine-learning method for recognizing stop signs, you could test that method by identifying the smallest input subset that constitutes a stop sign. If that consists of a tree branch, a particular time of day, or something that's not a stop sign, you could be concerned that the car might come to a stop at a place it's not supposed to.

While it may seem that the model is the likely culprit here, the datasets are more likely to blame. "There's the question of how we can modify the datasets in a way that would enable models to be trained to more closely mimic how a human would think about classifying images and

therefore, hopefully, generalize better in these real-world scenarios, like autonomous driving and medical diagnosis, so that the models don't have this nonsensical behavior," says Carter.

This may mean creating datasets in more controlled environments. Currently, it's just pictures that are extracted from public domains that are then classified. But if you want to do object identification, for example, it might be necessary to train models with objects with an uninformative background.

More information: Brandon Carter, Siddhartha Jain, Jonas Mueller, David Gifford, Overinterpretation reveals image classification model pathologies. arXiv:2003.08907v3 [cs.LG], arxiv.org/abs/2003.08907

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