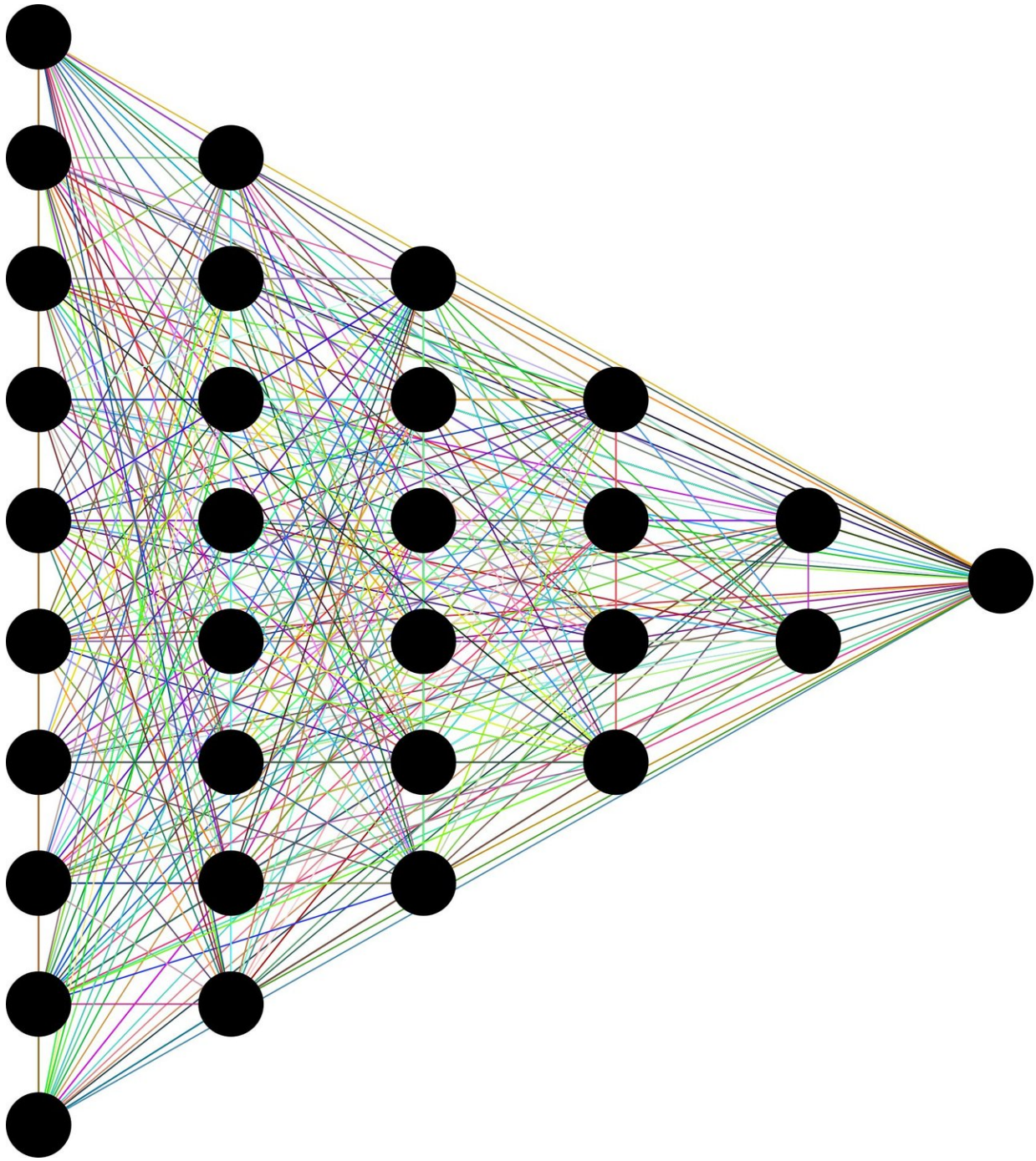
A neural network learns when it should not be trusted

November 19 2020, by Daniel Ackerman



Credit: CC0 Public Domain

Increasingly, artificial intelligence systems known as deep learning

neural networks are used to inform decisions vital to human health and safety, such as in autonomous driving or medical diagnosis. These networks are good at recognizing patterns in large, complex datasets to aid in decision-making. But how do we know they're correct? Alexander Amini and his colleagues at MIT and Harvard University wanted to find out.

They've developed a quick way for a neural network to crunch data, and output not just a prediction but also the model's confidence level based on the quality of the available data. The advance might save lives, as deep learning is already being deployed in the real world today. A network's level of certainty can be the difference between an autonomous vehicle determining that "it's all clear to proceed through the intersection" and "it's probably clear, so stop just in case."

Current methods of uncertainty estimation for neural networks tend to be computationally expensive and relatively slow for split-second decisions. But Amini's approach, dubbed "deep evidential regression," accelerates the process and could lead to safer outcomes. "We need the ability to not only have high-performance models, but also to understand when we cannot trust those models," says Amini, a Ph.D. student in Professor Daniela Rus' group at the MIT Computer Science and Artificial Intelligence Laboratory (CSAIL).

"This idea is important and applicable broadly. It can be used to assess products that rely on learned models. By estimating the uncertainty of a learned model, we also learn how much error to expect from the model, and what missing data could improve the model," says Rus.

Amini will present the research at next month's NeurIPS conference, along with Rus, who is the Andrew and Erna Viterbi Professor of Electrical Engineering and Computer Science, director of CSAIL, and deputy dean of research for the MIT Stephen A. Schwarzman College of

Computing; and graduate students Wilko Schwarting of MIT and Ava Soleimany of MIT and Harvard.

Efficient uncertainty

After an up-and-down history, deep learning has demonstrated remarkable performance on a variety of tasks, in some cases even surpassing human accuracy. And nowadays, deep learning seems to go wherever computers go. It fuels search engine results, social media feeds, and facial recognition. "We've had huge successes using deep learning," says Amini. "Neural networks are really good at knowing the right answer 99 percent of the time." But 99 percent won't cut it when lives are on the line.

"One thing that has eluded researchers is the ability of these models to know and tell us when they might be wrong," says Amini. "We really care about that 1 percent of the time, and how we can detect those situations reliably and efficiently."

Neural networks can be massive, sometimes brimming with billions of parameters. So it can be a heavy computational lift just to get an answer, let alone a confidence level. Uncertainty analysis in neural networks isn't new. But previous approaches, stemming from Bayesian deep learning, have relied on running, or sampling, a neural network many times over to understand its confidence. That process takes time and memory, a luxury that might not exist in high-speed traffic.

The researchers devised a way to estimate uncertainty from only a single run of the neural network. They designed the network with bulked up output, producing not only a decision but also a new probabilistic distribution capturing the evidence in support of that decision. These distributions, termed evidential distributions, directly capture the model's confidence in its prediction. This includes any uncertainty

present in the underlying input data, as well as in the model's final decision. This distinction can signal whether uncertainty can be reduced by tweaking the neural network itself, or whether the input data are just noisy.

Confidence check

To put their approach to the test, the researchers started with a challenging computer vision task. They trained their neural network to analyze a monocular color image and estimate a depth value (i.e. distance from the camera lens) for each pixel. An autonomous vehicle might use similar calculations to estimate its proximity to a pedestrian or to another vehicle, which is no simple task.

Their network's performance was on par with previous state-of-the-art models, but it also gained the ability to estimate its own uncertainty. As the researchers had hoped, the network projected high uncertainty for pixels where it predicted the wrong depth. "It was very calibrated to the errors that the network makes, which we believe was one of the most important things in judging the quality of a new uncertainty estimator," Amini says.

To stress-test their calibration, the team also showed that the network projected higher uncertainty for "out-of-distribution" data—completely new types of images never encountered during training. After they trained the network on indoor home scenes, they fed it a batch of outdoor driving scenes. The network consistently warned that its responses to the novel outdoor scenes were uncertain. The test highlighted the network's ability to flag when users should not place full trust in its decisions. In these cases, "if this is a health care application, maybe we don't trust the diagnosis that the model is giving, and instead seek a second opinion," says Amini.

The network even knew when photos had been doctored, potentially hedging against data-manipulation attacks. In another trial, the researchers boosted adversarial noise levels in a batch of images they fed to the network. The effect was subtle—barely perceptible to the human eye—but the network sniffed out those images, tagging its output with high levels of uncertainty. This ability to sound the alarm on falsified data could help detect and deter adversarial attacks, a growing concern in the age of deepfakes.

Deep evidential regression is "a simple and elegant approach that advances the field of uncertainty estimation, which is important for robotics and other real-world control systems," says Raia Hadsell, an artificial intelligence researcher at DeepMind who was not involved with the work. "This is done in a novel way that avoids some of the messy aspects of other approaches—e.g. sampling or ensembles—which makes it not only elegant but also computationally more efficient—a winning combination."

Deep evidential regression could enhance safety in AI-assisted decision making. "We're starting to see a lot more of these [neural network] models trickle out of the research lab and into the real world, into situations that are touching humans with potentially life-threatening consequences," says Amini. "Any user of the method, whether it's a doctor or a person in the passenger seat of a vehicle, needs to be aware of any risk or uncertainty associated with that decision." He envisions the system not only quickly flagging uncertainty, but also using it to make more conservative decision making in risky scenarios like an autonomous vehicle approaching an intersection.

"Any field that is going to have deployable machine learning ultimately needs to have reliable uncertainty awareness," he says.