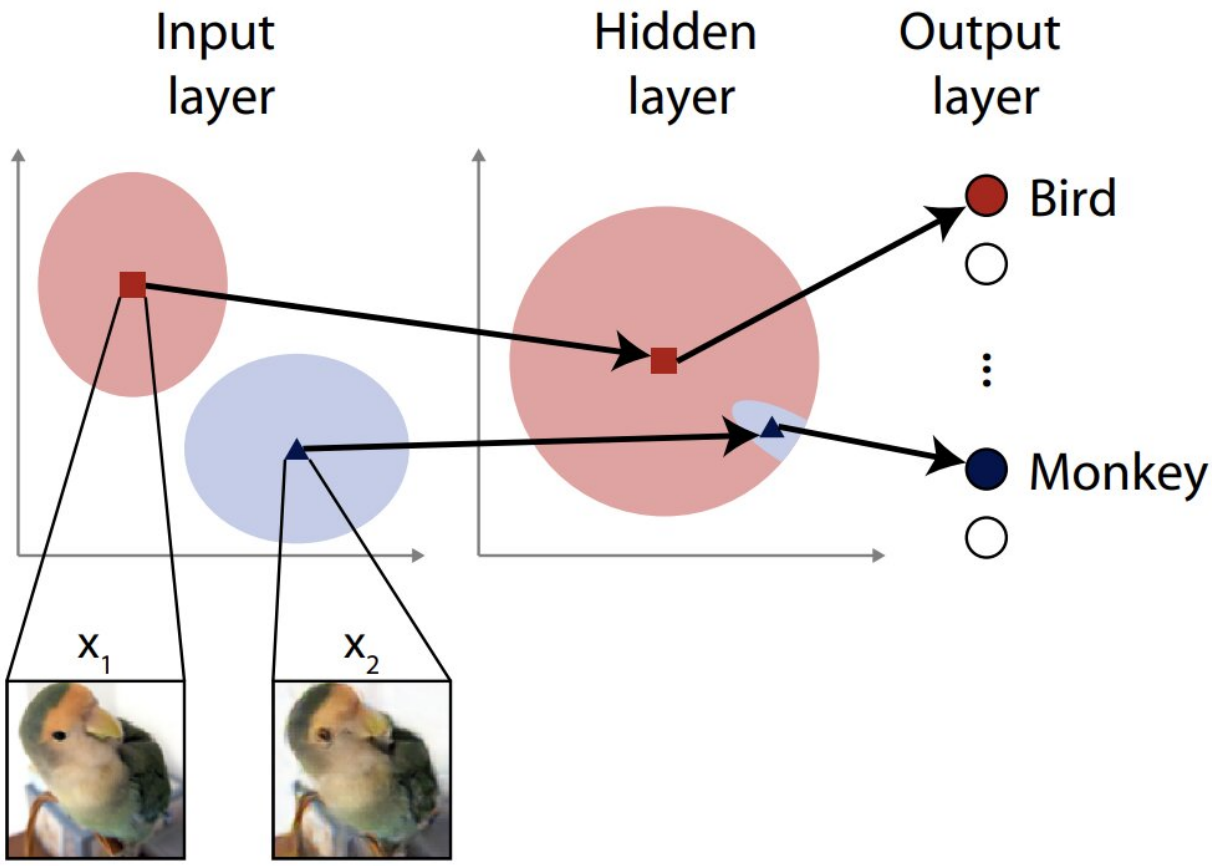
Using the brain as a model inspires a more robust AI

September 15 2023



Bird or monkey? To our eyes the input images x1 and x2 look the same, but hidden features nudge a typical neural network to classify this bird image as a monkey by mistake. It's said the images are distant at the input space, but close in the hidden-layer space. The researchers aimed to close this exploit. Credit: 2023 Ohki & Ukita CC-BY

Most artificially intelligent systems are based on neural networks, algorithms inspired by biological neurons found in the brain. These networks can consist of multiple layers, with inputs coming in one side and outputs going out of the other. The outputs can be used to make automatic decisions, for example, in driverless cars.

Attacks to mislead a neural network can involve exploiting vulnerabilities in the input layers, but typically only the initial input layer is considered when engineering a defense. For the first time, researchers augmented a neural network's inner layers with a process involving random noise to improve its resilience.

Artificial intelligence (AI) has become a relatively common thing; chances are you have a smartphone with an AI assistant or you use a search engine powered by AI. While it's a broad term that can include many different ways to essentially process information and sometimes make decisions, AI systems are often built using artificial neural networks (ANN) analogous to those of the brain.

And like the brain, ANNs can sometimes get confused, either by accident or by the deliberate actions of a third party. Think of something like an optical illusion—it might make you feel like you are looking at one thing when you are really looking at another.

The difference between things that confuse an ANN and things that might confuse us, however, is that some visual input could appear perfectly normal, or at least might be understandable to us, but may nevertheless be interpreted as something completely different by an ANN.

A trivial example might be an image-classifying system mistaking a cat for a dog, but a more serious example could be a driverless car mistaking a stop signal for a right-of-way sign. And it's not just the already

controversial example of driverless cars; there are medical diagnostic systems, and many other sensitive applications that take inputs and inform, or even make, decisions that can affect people.

As inputs aren't necessarily visual, it's not always easy to analyze why a system might have made a mistake at a glance. Attackers trying to disrupt a system based on ANNs can take advantage of this, subtly altering an anticipated input pattern so that it will be misinterpreted, and the system will behave wrongly, perhaps even problematically.

There are some defense techniques for attacks like these, but they have limitations. Recent graduate Jumpei Ukita and Professor Kenichi Ohki from the Department of Physiology at the University of Tokyo Graduate School of Medicine devised and tested a new way to improve ANN defense.

Is it a bird? Is it a plane? This is a sample of images the researchers generated for their simulated attack prior to running their new defense method. The x1 images were classified correctly, the x2 images are the adversarial examples that tricked an undefended network into classifying them wrongly. Credit: 2023 Ohki & Ukita CC-BY

"Neural networks typically comprise layers of virtual neurons. The first layers will often be responsible for analyzing inputs by identifying the elements that correspond to a certain input," said Ohki.

"An attacker might supply an image with artifacts that trick the network into misclassifying it. A typical defense for such an attack might be to deliberately introduce some noise into this first layer. This sounds counterintuitive that it might help, but by doing so, it allows for greater adaptations to a visual scene or other set of inputs. However, this method is not always so effective and we thought we could improve the matter by looking beyond the input layer to further inside the network."

Ukita and Ohki aren't just computer scientists. They have also studied the human brain, and this inspired them to use a phenomenon they knew about there in an ANN. This was to add noise not only to the input layer, but to deeper layers as well. This is typically avoided as it's feared that it will impact the effectiveness of the network under normal conditions. But the duo found this not to be the case, and instead the noise promoted greater adaptability in their test ANN, which reduced its susceptibility to simulated adversarial attacks.

"Our first step was to devise a hypothetical method of attack that strikes deeper than the input layer. Such an attack would need to withstand the

resilience of a network with a standard noise defense on its input layer. We call these feature-space adversarial examples," said Ukita.

"These attacks work by supplying an input intentionally far from, rather than near to, the input that an ANN can correctly classify. But the trick is to present subtly misleading artifacts to the deeper layers instead. Once we demonstrated the danger from such an attack, we injected random noise into the deeper hidden layers of the network to boost their adaptability and therefore defensive capability. We are happy to report it works."

While the new idea does prove robust, the team wishes to develop it further to make it even more effective against anticipated attacks, as well as other kinds of attacks they have not yet tested it against. At present, the defense only works on this specific kind of attack.

"Future attackers might try to consider attacks that can escape the feature-space noise we considered in this research," said Ukita. "Indeed, attack and defense are two sides of the same coin; it's an arms race that neither side will back down from, so we need to continually iterate, improve and innovate new ideas in order to protect the systems we use every day."

The research is published in the journal *Neural Networks*.

More information: Jumpei Ukita et al, Adversarial attacks and defenses using feature-space stochasticity, *Neural Networks* (2023). DOI: 10.1016/j.neunet.2023.08.022

Provided by University of Tokyo