PHYS.ORG

When deep learning mistakes a coffee maker for a cobra

March 22 2017

Credit: Ecole Polytechnique Federale de Lausanne

Is this your sister?" That's the kind of question asked by image-recognition systems, which are becoming increasingly prevalent in our everyday devices. They may soon be used for tumor detection and genomics, too. These systems rely on what is known as "deep-learning" architectures – an exciting new development in artificial learning. But EPFL researchers have revealed just how sensitive these systems actually are: a tiny universal perturbation applied across an image can throw off even the most sophisticated algorithms.

Deep-learning systems, a major breakthrough in computer-based image recognition, are however surprisingly sensitive to minor changes in the data they analyze. Researchers at EPFL's Signal Processing Laboratory (LTS4), headed by Pascal Frossard, have shown that even the best deep-learning architectures can be fooled by introducing an almost invisible perturbation into digital images. Such a perturbation can cause a system to mistake a joystick for a Chihuahua, for example, or a coffee-maker for a cobra. Yet the human brain would have no problem correctly identifying the objects. The researchers' findings – which should help scientists better understand, and therefore improve, deep-learning systems – will be presented at the IEEE Computer Vision and Pattern Recognition 2017 conference, a major international academic event. We spoke with Alhussein Fawzi and Seyed Moosavi, the two lead authors of the research.

What is deep learning and what is the problem with today's systems?

Fawzi: Deep learning, or artificial neural networks, is an exciting new development in artificial intelligence. All the major tech firms are banking on this technology to develop systems that can accurately recognize objects, faces, text and speech. Various forms of these algorithms can be found in Google's search engine and in Apple's SIRI, for instance. Deep-learning systems can work exceptionally well; companies are considering using them to detect tumors from a CAT scan or to operate self-driving cars. The only problem is that they are often black boxes and we don't always have a good grasp of how they function. Should we just trust them blindly?

Is it really surprising that we can fool these systems?

Moosavi: Researchers had already shown two years ago that artificial neural networks could easily be tricked by small perturbations designed specifically to confuse them on a given image. But we found that a single, universal perturbation could cause a network to fail on almost all images. And this perturbation is so tiny that it is almost invisible to the naked eye. That's alarming and shows that these systems are not as robust as one might think.

How does "deep learning" work?

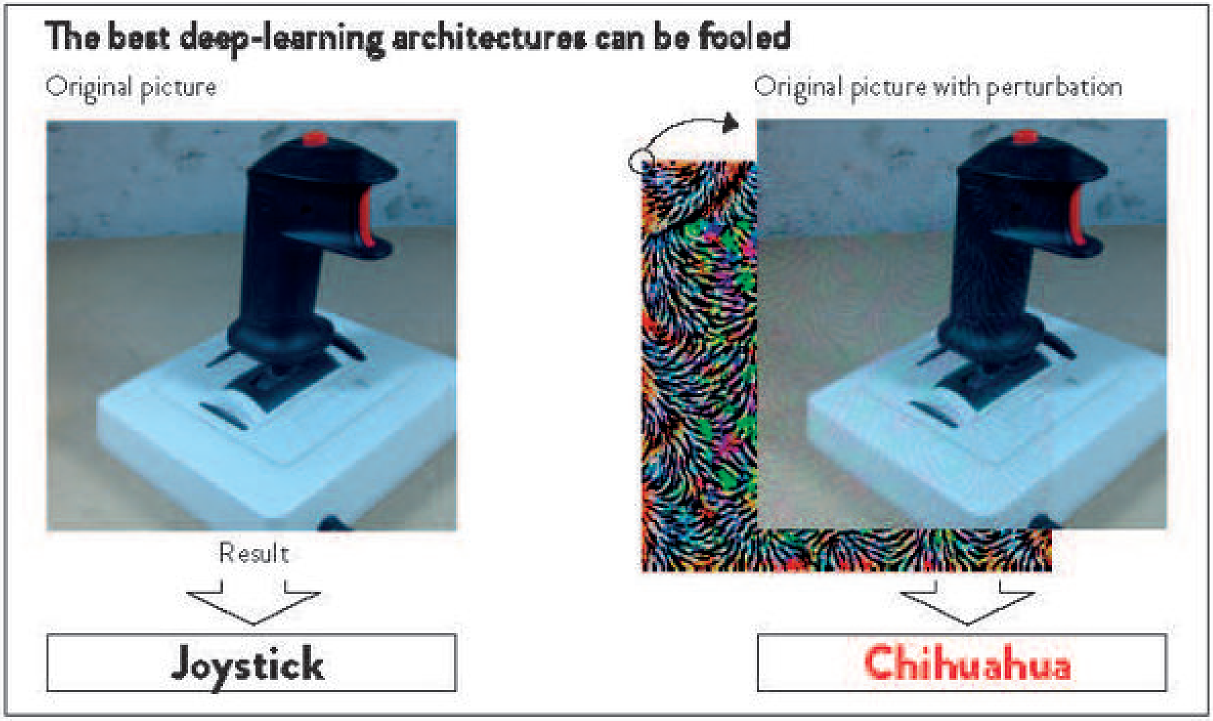
Moosavi: Deep-learning algorithms are learning algorithms designed to identify the important information in data sets. They consist of virtual "machines" built from artificial neural networks that are essentially neurons stacked in layers, with each layer performing a simple calculation.

We can show a machine a thousand images of cats, for example, and tell the machine that these are "cats." Roughly speaking, the first layer of neurons will analyze the low level features, such as the edges and corners, the second the basic shapes, and so on, until the machine pieces together the entire shape of a cat. The machine thus learns on its own from many images; each layer builds on the results of the previous one to perform its own calculations. The name "deep learning" comes from this progressive, layer-by-layer approach.

Fawzi: Once the machine has learned to recognize a cat, it can identify cats in images it has never seen. It is "smart" enough to recognize objects in new contexts – it doesn't just work by memorization.

How did you carry out your study?

Fawzi: We calculated the smallest possible perturbation that could throw off the image-recognition algorithms of the best learning systems. This perturbation prevented the systems from correctly recognizing most natural images. The perturbation was basically just a slight change in the value of an image's pixels. To the human eye, there was only a very minor difference between the original and modified images. But for the deep-learning systems, the difference was huge. A sock was identified as an elephant; a green plant as a macaw. What was really astonishing was that the same perturbation could fool many different types of systems.


The best deep-learning architectures can be fooled
Original picture
Original picture with perturbation
Result
Joystick
Chihuahua

Credit: Ecole Polytechnique Federale de Lausanne

Moosavi: Going back to the cat example, a human being – even a two-year-old – will have learned what a "cat" is after seeing five images, even if the images are combinations of pictures and drawings. As humans, we have an inherent ability to look at the big picture and ignore small perturbations since they don't change the underlying concept displayed in an image. But deep-learning systems are not capable of such abstraction. They don't identify a concept but rather sort through a logical series of clues. And that's what can lead to classification errors.

What problems would this cause in real-world applications?

Moosavi: Our goal is to better understand deep-learning systems so that we can improve their performance. Right now there is a lot of interest in deep-learning applications for medical imaging, such as to identify certain proteins or tumors. But when it comes to healthcare, having fool-proof technology is essential. So we must have a good understanding of the limitations of existing systems, so that we can make them more robust and capable of delivering guaranteed results.

Since it was fairly easy to find our perturbation, it isn't a stretch to imagine that people with bad intentions could also come up with ways to trick deep-learning systems. That could create potential security threats, for instance.

So what's the next step?

Moosavi: We think more theoretical research needs to be done on deep learning. We would like to see scientists look more closely at how artificial neural networks work, their properties, their considerable potential – and the associated risks. The code we used in our research is available to the public, so anyone can test the perturbation and study the problem further. In terms of our next steps, we plan to learn more about how complicated architectures of artificial neural networks work, so that we can make them more robust. We are already in contact with some potential partners that seem interested in our approach.

More information: Universal adversarial perturbations. arxiv.org/pdf/1610.08401.pdf

Provided by Ecole Polytechnique Federale de Lausanne

Citation: When deep learning mistakes a coffee maker for a cobra (2017, March 22) retrieved 23 April 2024 from https://phys.org/news/2017-03-deep-coffee-maker-cobra.html