The Flaw Lurking In Every Deep Neural Net

Written by Mike James

Tuesday, 27 May 2014

A recent paper with an innocent sounding title is probably the biggest news in neural networks since the invention of the backpropagation algorithm. But what exactly does it all mean?

Update: Also see - The Deep Flaw In All Neural Networks.

A recent paper "Intriguing properties of neural networks" by Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow and Rob Fergus, a team that includes authors from Google's deep learning research project outlines two pieces of news about the way neural networks behave that run counter to what we believed - and one of them is frankly astonishing.

I'm going to tell you about both, but it is the second that is the most amazing. So if you are in a hurry skip down the page.

The first concerns the way that we had long assumed that neural networks organized data. In a multi-layer network it has long been thought that neurons in each level learned useful features for the next level. At the very least it was supposed that in the last layer each neuron would learn some significant and usually meaningful feature.

The standard way of finding out if this is the case is to take a particular neuron and find the set of inputs that makes it produce a maximal output. Whatever it is that maximizes the neuron's output is assumed to be the feature that it responds to. For example, in a face recognizer a neuron might respond strongly to an image that has an eye or a nose - but notice there is no reason that the features should correspond to the neat labels that humans use.

What has been discovered is that a single neuron's feature is no more interpretable as a meaningful feature than a random set of neurons. That is, if you pick a random set of neurons and find the images that produce the maximum output on the set then these images are just as semantically similar as in the single neuron case.

This means that neural networks do not "unscramble" the data by mapping features to individual neurons in say the final layer. The information that the network extracts is just as much distributed across all of the neurons as it is localized in a single neuron.

This is an interesting finding and it leads on to the second and even more remarkable finding..

Every deep neural network has "blind spots" in the sense that there are inputs that are very close to correctly classified examples that are misclassified.

Since the very start of neural network research it has been assumed that networks had the power to generalize. That is, if you train a network to recognize a cat using a particular set of cat photos the network will, as long as it has been trained properly, have the ability to recognize a cat photo it hasn't seen before.

Within this assumption has been the even more "obvious" assumption that if the network correctly classifies the photo of a cat as a cat then it will correctly classify a slightly perturbed version of the same photo as a cat. To create the slightly perturbed version you would simply modify each pixel value, and as long as the amount was small, then the cat photo would look exactly the same to a human - and presumably to a neural network.

However, this isn't true.

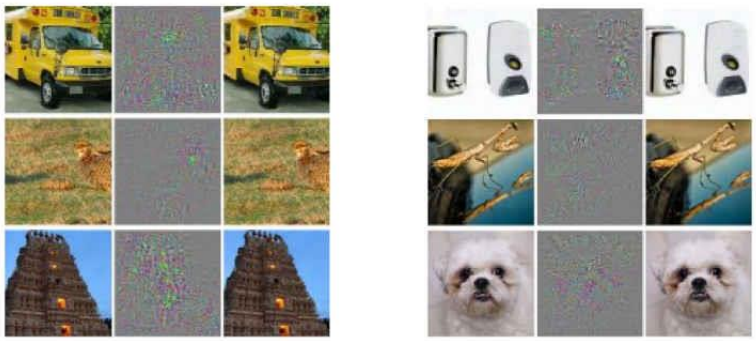
What the researchers did was to invent an optimization algorithm that starts from a correctly classified example and tries to find a small perturbation in the pixel values that drives the output of the network to another classification. Of course, there is no guarantee that such a perturbed incorrect version of the image exists - and if the continuity assumption mentioned earlier applied the search would fail.

However the search succeeds.

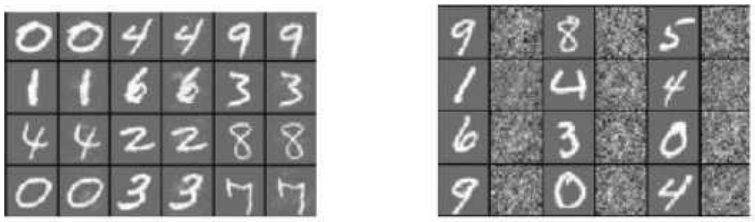
For a range of different neural networks and data sets it has proved very possible to find such "adversarial examples" from correctly classified data. To quote the paper:

"For all the networks we studied, for each sample, we always manage to generate very close, visually indistinguishable, adversarial examples that are misclassified by the original network."

To be clear, the adversarial examples looked to a human like the original, but the network misclassified them. You can have two photos that look not only like a cat but the same cat, indeed the same photo, to a human, but the machine gets one right and the other wrong.



The images to the right are correctly classified the ones to the far left are misclassified and the middle column gives the differences multiplied by ten to make them visible.



In the left-hand panel, the odd columns are correctly classified and the even columns are misclassified In the case of the right-hand panel everything is correctly classified and the even columns are random distortions of the originals. This demonstrates that the distortion has to be very specific - you have to move in a very specific direction to find an adversarial example.

What is even more shocking is that the adversarial examples seem to have some sort of universality. That is a large fraction were misclassified by different network architectures trained on the same data and by networks trained on a different data set.

"The above observations suggest that adversarial examples are somewhat universal and not just the results of overfitting to a particular model or to the specific selection of the training set"

This is perhaps the most remarkable part of the result. Right next to every correctly classified example there is an effectively indistinguishable example that is misclassified, no matter what network or training set was used.

So if you have a photo of a cat there is a set of small changes that can be made to it that makes the network classify it as a dog - irrespective of the network or its training.

What does this all mean?

The researchers take a positive approach and use the adversarial examples to train the network to get it right. They regard the adversarial examples as particularly tough training cases that can be used to improve the network and its generalization.

However, the discovery seems to be more than just a better training set.

The first thing to say is you might be thinking "so what if a cat photo that is clearly a photo a cat is recognized as a dog?" If you change the situation just a little and ask what does it matter if a self-driving car that uses a deep neural network misclassifies a view of a pedestrian standing in front of the car as a clear road?

The continuity and stability of deep neural networks matters for their practical application.

There is also the philosophical question raised by these blind spots. If a deep neural network is biologically inspired we can ask the question, does the same result apply to biological networks.

Put more bluntly "does the human brain have similar built-in errors?" If it doesn't, how is it so different from the neural networks that are trying to mimic it? In short, what is the brain's secret that makes it stable and continuous?

One possible explanation is that this is another manifestation of the curse of dimensionality. As the dimension of a space increases it is well known that the volume of a hypersphere becomes increasingly concentrated at its surface. (The volume that is not near the surface drops exponentially with increasing dimension.) Given that the decision boundaries of a deep neural network are in a very high dimensional space it seems reasonable that most correctly classified examples are going to be close to the decision boundary - hence the ability to find a misclassified example close to the correct one, you simply have to work out the direction to the closest boundary.

If this is part of the explanation, then it is clear that even the human brain cannot avoid the effect and must somehow cope with it; otherwise cats would morph into dogs with an alarming regularity.

The bottom line is that deep neural networks do not seem to be continuous with respect to the decisions they make and exhibit a new sort of instability. Rather than patch things up with adversarial training cases, research is needed to explore and eliminate the problem. Until this happens you cannot rely on a neural network in any safety critical system..

Car Not A Car!