Adversarial Examples Study

Hello!
Computers often get confused by things called ‘adversarial examples’ which are carefully crafted images that contain changes which cause the computer to misclassify the image. For example the right hand image below is a slightly altered version of the dog on the left. Humans still see a dog, while the computer sees an ostrich.

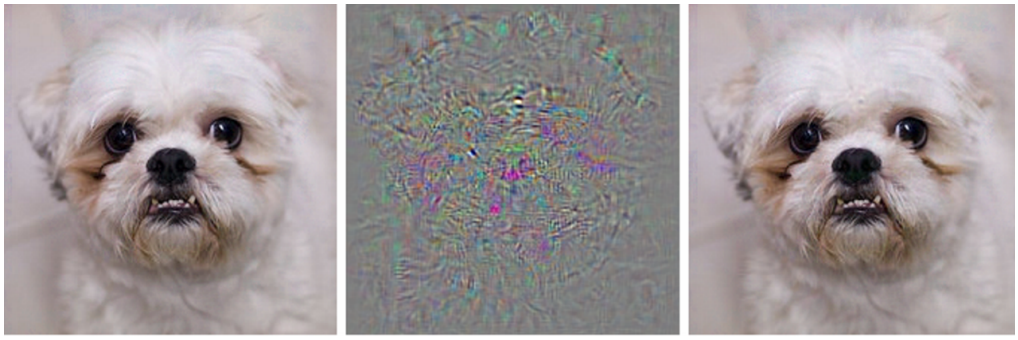

Image credit: Christian Szegedy (Google)

I’m interested in whether we can flip this experiment? We first take an image of one digit, and change it to make it look like another digit, to a human. We then ask a computer, trained on digit images to classify the new digit: Does it see the new number the humans see, or the original number? In effect can a human be more susceptible in this case to adversarial examples?

Tasks There are three tasks that need doing (the screen will be showing one of them).
1) Drawing digits: Simply drag the circles around to make the two lines into the shape of the digit it asked for. Please use both of the lines to make the digit!
2) Decide which of the symbols look more like a particular digit.
3) Decide which digit a symbol looks most like.

Use of the data The digit images from task 1 will be used to train a computer classifier to recognise digits. The decisions you make (from tasks 2 and 3) will be used to make the adversarial examples later tested on the computer classifier. The data is anonymously collected and may be released for other researchers to use. If we find interesting results we may publish a paper.

Taking part in this experiment is optional! If you want your responses to be removed from the data at a future date (prior to its release), please email me (details below) with the dates and times that you took part, and I’ll remove those times from the data.

By taking part you agree to the analysis of the data as above, and its release for use by other researchers.

This project isn’t funded, and is just something I’m interested in. The project has been approved by the Computer Science ethical review procedure. If you’ve any questions about the project, please email me: m.t.smith@sheffield.ac.uk or visit me in office COM136. Or if I’m not available please speak to Wil Ward, w.ward@sheffield.ac.uk. A copy of this information sheet can be downloaded from www.michaeltsmith.org.uk/other/informationsheet.pdf.

Thank you for taking part!

- Dr Mike Smith