Having looked at some of the positive aspects of the neural net, its ability to classify images and its relative simplicity of implementation, it is now time to look at some of the vulnerabilities and short comings of the neural net – and why fears of a future filled with Skynet or the machines from the Matrix are still many centuries off. Along the way, we may also gain some insight into how neural networks are, contrary to some claims, not at all how people and their brains work.

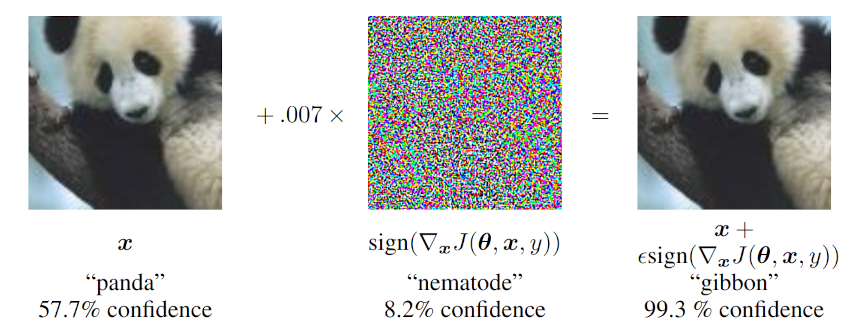
An excellent place to start our journey is with the succinct observation made by Dave Gershgorn in his article [*Fooling the machine*](https://www.popsci.com/byzantine-science-deceiving-artificial-intelligence/#:~:text=Researchers%20found%20that%20they%20could%20attack%20these%20systems,a%20gibbon%2C%20by%20applying%20light%20distortion.%20Christian%20Szegedy) for Popular Science in which he said

\*\*[a]n accelerating field of research suggests that most of the artificial intelligence we’ve created so far has learned enough to give a correct answer, but without truly understanding the information. And that means it’s easy to deceive.\*\*

Since the article dates from 2016, Gershgorn could only present the then state-of-the-art examples of adversarial attacks on neural nets. He draws heavily from the work of Christian Szegedy and others with two of the most relevant papers being [*Explaining and Harnessing Adversarial Examples*](https://arxiv.org/pdf/1412.6572.pdf) (Goodfellow, Shlens, and Szegedy) and [*Intriguing properties of neural networks*](https://arxiv.org/pdf/1312.6199.pdf) (Szegedy et al).

Both papers deal with the attempts to trick a trained neural net through the use of what looks to be speckle noise but which is really a small but deterministic perturbation to the image feature vector in a direction in the image space that turns the net’s classification from being waveringly correct to dogmatically wrong.

Perhaps the most famous example of this is the panda-to-gibbon example shown below (taken from the first paper)



To understand what is happening and how the addition of a ‘speckle noise’ image could persuade the net that the image it was 58% sure was a panda should now be interpreted, with nearly 100% certainty, as a gibbon, we need to remember just how the neural net does its classifications. Each of the above images, which we perceive as 2-dimensional (a bit more on this later) are actually serialized into a one-dimensional feature vector. In the case above, judging by the perturbation (i.e., a count of the colored pixels along an edge), each image is about 240x240 pixels. The relationship displayed in the figure above should really be interpreted as the linear algebra relationship

\[ {\vec P} + {\vec N} = {\vec G} \; , \]

where ${\vec P}$ is the displacement vector from the origin to the location of the original image in the panda-classification region, ${\vec G}$ is the location of the perturbed image, which is now somewhere in the gibbon-classification region, and ${\vec N} = 0.007 \nabla J$ is the displacement vector that connects the two.

Diagram

Description automatically generated

Since each of these vectors is in a $240^2 = 57,600$ dimensional space, our intuitions about concepts such as ‘near’ and ‘far’ poorly apply and the gibbon-region may be distinct and far away in some directions and quite close in others. Th structure of ${\vec N}$ (the middle image) was determined by experimenting with the net in order to find the “fastest” direction that nudges ${\vec P}$ out of the panda classification region and into the gibbon-classification region. The term ‘nudge’ is quite descriptive because the gain on the image was small (0.007) and the presence in the final image is not discernable by the human eye.

Of course, a normal person would still classify the perturbed image as a panda even if the gain on the perturbation were large enough for the human eye to perceive since we have other faculties at play. But the neural net is, put frankly, stupid. Gershgorn quotes one of Szegedy’s coauthors, Ian Goodfellow, as saying

\*\*these neural networks aren’t really learning certain ideas, but just how to recognize when they [find] the right idea\*\*

which is a subtle but important point. No neural network can know that it is correct or understand why its answer is the right one, it simply responds to the metaphoric ‘atta-boys’ it has gotten to return with what the majority of us tell it as right.

When these results were published, the danger of this type of adversarial attack was downplayed because it depends on the attacker being to play with the neural net to find the fastest direction via gradient descent (the name of this technique explains why Goodfellow, Shlens, and Szegedy denoted the perturbation as $\nabla J$). If the net is public, that is to say one can give multiple inputs to the net in order to see how each is classified, then it is always possible to find the gradient direction that fools the network into misclassifying, even if the probabilities that the network assigns to the classification are hidden. The article [*Attacking machine learning with adversarial examples*](https://openai.com/research/attacking-machine-learning-with-adversarial-examples) by the staff at OpenAI discusses in great detail how this type of gradient attack works and how people try, and often fail, to defend against it.

The conventional thinking at the time was that if the network’s classification model were completely hidden from public scrutiny then a network could be trustworthy and therefore be reliably depended upon. But even this conclusion was shown, shortly thereafter, to be on shakier ground than first believed. In an article entitled [*Slight Street Sign Modifications Can Completely Fool Machine Learning Algorithms*](https://na01.safelinks.protection.outlook.com/?url=https%3A%2F%2Fspectrum.ieee.org%2Fslight-street-sign-modifications-can-fool-machine-learning-algorithms&data=05%7C01%7C%7C1030232ecbd94c9bf87408db1c1c75c3%7C84df9e7fe9f640afb435aaaaaaaaaaaa%7C1%7C0%7C638134680794263535%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C3000%7C%7C%7C&sdata=c4QIGspanMYTmwS%2BwkIwczzAsEEt0yJ2ZCyeClA%2BBFc%3D&reserved=0)*,* Evan Ackerman points out that image classifying neural networks can be easily and completely fooled by the interplays of light and shadow on real-world traffic signs into making what would be disastrous conclusions should the network be behind the wheel of a car.

Citing the paper, [*Robust Physical-World Attacks on Deep Learning Visual Classification*](https://arxiv.org/pdf/1707.08945.pdf), by Eykholt et al, Ackerman offers this image of 5 frames of a stop sign at different distances and orientations with respect to the camera

A picture containing text, outdoor, sign

Description automatically generated

The stop signs aren’t real, but rather printed posters that have a specially chosen dappling of lighter and darker regions. These posters fooled the neural network into believing that each frame the camera was actually looking at a Speed Limit 45 sign. The key point here is that these authors didn’t inject the network with a feature vector tuned to the original input image. Rather they decorated an object with a perturbation that carried out a successful attack regardless of how much of the feature vector that object consumed. This is a real-world possible scenario and a frightening one.

This last point brings us back to an observation alluded to earlier. Clearly, people of all mental capacities, backgrounds, and circumstances perceive the visual world quite differently than do neural networks. We don’t serialize an image nor are we only able to receive images of a particular size; we can have our eyes wide open or half-closed changing the ‘number of pixels’ we are receiving and not just what percentage any given object consumes. We also can change our certainty in our classification and segmentation based on self-awareness and belief and not just on ‘atta-boys’ received from the world around us. So, sleep well knowing that the days of Skynet are far off and that you remain safe – unless you happen to find yourself in a self-driving car.