This Neural Net Hallucinates Sheep

**POSTED BY JANELLE SHANE ON MAR 07, 2018**

**I**f you’ve been on the internet today, you’ve probably interacted with a neural network. They’re a type of machine learning algorithm that’s used for everything from language translation to finance modeling. One of their specialties is image recognition. Several companies—including Google, Microsoft, IBM, and Facebook—have their own algorithms for labeling photos. But image recognition algorithms can make really bizarre mistakes.

Janelle Shane

Microsoft Azure’s computer vision API [added](https://t.umblr.com/redirect?z=https://azure.microsoft.com/en-us/services/cognitive-services/computer-vision/&t=YzZkZWU0NThlNjQ2ZDhkODNiNzI1NDhmNDY0MTljNWYyZGZhMTVmNSxIekoxamdxTw==&b=t:IinBpZOgiMHH_bphj7hrOQ&p=http://aiweirdness.com/post/171451900302/do-neural-nets-dream-of-electric-sheep&m=1) the above caption and tags. But there are no sheep in the image of above. None. I zoomed all the way in and inspected every speck.

Janelle Shane

It also tagged sheep in this image. I happen to know there were sheep nearby. But none actually present.

*It hasn’t realized that “sheep” means the actual animal, not just a sort of treeless grassiness.*

Janelle Shane

Here’s one more example. In fact, the neural network hallucinated sheep every time it saw a landscape of this type. What’s going on here?

The way neural networks learn is by looking at lots of examples. In this case, its trainers gave it lots of images that humans had labeled by hand—and lots of those images contained sheep. Starting with no knowledge at all of what it was seeing, the neural network had to make up rules about which images should be labeled “sheep.” And it looks like it hasn’t realized that “sheep” means the actual animal, not just a sort of treeless grassiness. (Similarly, it labeled the second image with “rainbow” likely because it was wet and rainy, not realizing that the band of colors is essential.)

Are neural networks just hyper-vigilant, finding sheep everywhere? No, as it turns out. They only see sheep where they expect to see them. They can find sheep easily in fields and mountainsides, but as soon as sheep start showing up in weird places, it becomes obvious how much the algorithms rely on guessing and probabilities.

Bring sheep indoors, and they’re labeled as cats. Pick up a sheep (or a goat) in your arms, and they’re labeled as dogs.



Paint them orange, and they become flowers.



Put the sheep on leashes, and they’re labeled as dogs. Put them in cars, and they’re dogs or cats. If they’re in the water, they could end up being labeled as birds or even polar bears.

And if goats climb trees, they become birds. Or possibly giraffes. (It turns out that Microsoft Azure is somewhat notorious for seeing giraffes everywhere due to a rumored overabundance of giraffes in the original dataset.)



The thing is, neural networks match patterns. They see patches of fur-like texture, a bunch of green, and conclude that there are sheep. If they see fur and kitchen shapes, it may conclude instead that there are cats.

If life plays by the rules, image recognition works well. But as soon as people—or sheep—do something unexpected, the algorithms show their weaknesses.

Want to sneak something past a neural network? In a delightfully cyberpunk twist, surrealism might be the answer. Maybe future secret agents will dress in chicken costumes, or drive cow-spotted cars.

There are lots, lots more examples of hilarious mistakes in a Twitter [thread](https://twitter.com/JanelleCShane/status/969239712190746624) I started with the simple question:

Janelle Shane

And you can [test](https://t.umblr.com/redirect?z=https://azure.microsoft.com/en-us/services/cognitive-services/computer-vision/&t=YzZkZWU0NThlNjQ2ZDhkODNiNzI1NDhmNDY0MTljNWYyZGZhMTVmNSxIekoxamdxTw==&b=t:IinBpZOgiMHH_bphj7hrOQ&p=http://aiweirdness.com/post/171451900302/do-neural-nets-dream-of-electric-sheep&m=1) Microsoft Azure’s image recognition API and see for yourself that even top-notch algorithms are relying on probability and luck. Another algorithm, [NeuralTalk2](https://t.umblr.com/redirect?z=https://github.com/karpathy/neuraltalk2&t=MzM1ODA4M2RjMGM1ODI3ODc0ZGFkOTQ4NzY4ZmZiMzMwZmI5NmVlYixIekoxamdxTw==&b=t:IinBpZOgiMHH_bphj7hrOQ&p=http://aiweirdness.com/post/171451900302/do-neural-nets-dream-of-electric-sheep&m=1), is the one I mostly used for the Twitter thread.

Want to know when I post another experiment? You can sign up [here](https://t.umblr.com/redirect?z=https://janelleshane.activehosted.com/f/1&t=M2M0MzU0NGYwMjg3M2IwNDNkMzliZmM1MTk4OGQ4MTRjNWFiNWIxMSxIekoxamdxTw==&b=t:IinBpZOgiMHH_bphj7hrOQ&p=http://aiweirdness.com/post/171451900302/do-neural-nets-dream-of-electric-sheep&m=1).

*Janelle Shane is a research scientist in optics. She also trains neural networks, a type of machine learning algorithm, to write unintentional humor as they struggle to imitate human datasets. This*[*post*](http://aiweirdness.com/post/171451900302/do-neural-nets-dream-of-electric-sheep)*is reprinted with permission from her blog, AIWeirdness.com, which she started as a PhD student in electrical engineering at U.C. San Diego. Follow her on Twitter*[*@JanelleCShane*](https://twitter.com/JanelleCShane)*.*