Google Unveils Neural Network with “Superhuman” Ability to Determine the Location of Almost Any Image

Here’s a tricky task. Pick a photograph from the Web at random. Now try to work out where it was taken using only the image itself. If the image shows a famous building or landmark, such as the Eiffel Tower or Niagara Falls, the task is straightforward. But the job becomes significantly harder when the image lacks specific location cues or is taken indoors or shows a pet or food or some other detail.

Nevertheless, humans are surprisingly good at this task. To help, they bring to bear all kinds of knowledge about the world such as the type and language of signs on display, the types of vegetation, architectural styles, the direction of traffic, and so on. Humans spend a lifetime picking up these kinds of geolocation cues.

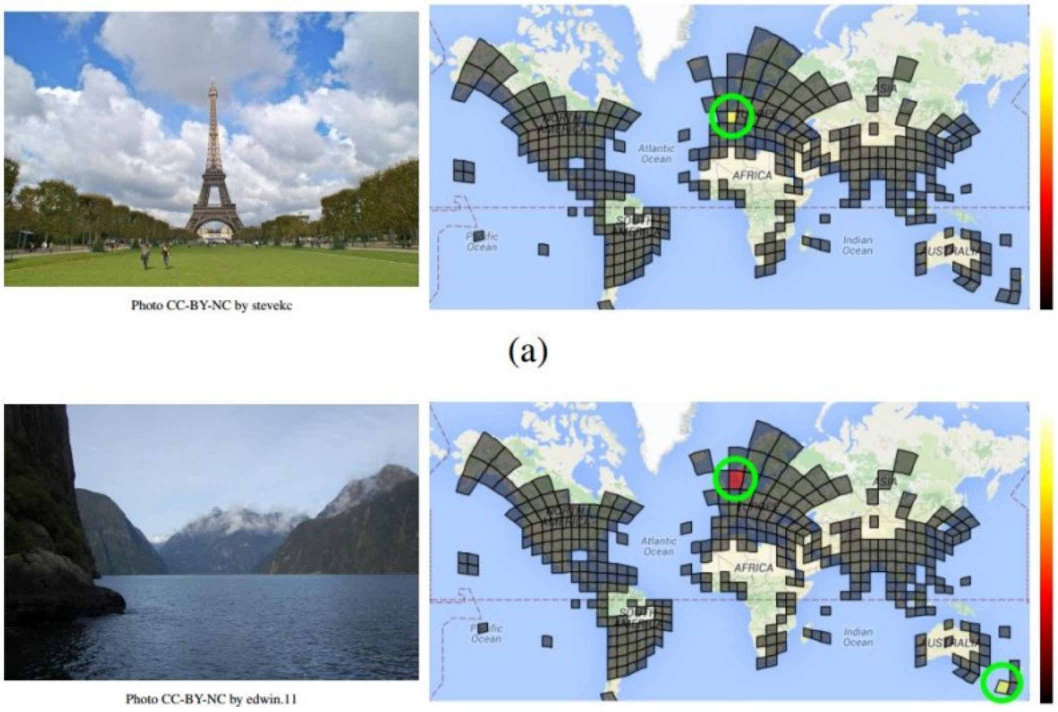
So it’s easy to think that machines would struggle with this task. And indeed, they have.

Today, that changes thanks to the work of Tobias Weyand, a computer vision specialist at Google, and a couple of pals. These guys have trained a deep-learning machine to work out the location of almost any photo using only the pixels it contains.

Their new machine significantly outperforms humans and can even use a clever trick to determine the location of indoor images and pictures of specific things such as pets, food, and so on that have no location cues.

Their approach is straightforward, at least in the world of machine learning. Weyand and co begin by dividing the world into a grid consisting of over 26,000 squares of varying size that depend on the number of images taken in that location.

So big cities, which are the subjects of many images, have a more fine-grained grid structure than more remote regions where photographs are less common. Indeed, the Google team ignored areas like oceans and the polar regions, where few photographs have been taken.



(a)

Next, the team created a database of geolocated images from the Web and used the location data to determine the grid square in which each image was taken. This data set is huge, consisting of 126 million images along with their accompanying Exif location data.

Weyand and co used 91 million of these images to teach a powerful neural network to work out the grid location using only the image itself. Their idea is to input an image into this neural net and get as the output a particular grid location or a set of likely candidates.

They then validated the neural network using the remaining 34 million images in the data set. Finally they tested the network—which they call PlaNet—in a number of different ways to see how well it works.

The results make for interesting reading. To measure the accuracy of their machine, they fed it 2.3 million geotagged images from Flickr to see whether it could correctly determine their location. "PlaNet is able to localize 3.6 percent of the images at street-level accuracy and 10.1 percent at city-level accuracy," say Weyand and co. What's more, the machine determines the country of origin in a further 28.4 percent of the photos and the continent in 48.0 percent of them.

That's pretty good. But to show just how good, Weyand and co put PlaNet through its paces in a test against 10 well-traveled humans. For the test, they used an online game that presents a player with a random view taken from Google Street View and asks him or her to pinpoint its location on a map of the world.

Anyone can play at www.geoguessr.com. Give it a try—it's a lot of fun and more tricky than it sounds.

Needless to say, PlaNet trounced the humans. "In total, PlaNet won 28 of the 50 rounds with a median localization error of 1131.7 km, while the median human localization error was 2320.75 km," say Weyand and co. "[This] small-scale experiment shows that PlaNet reaches superhuman performance at the task of geolocating Street View scenes."

An interesting question is how PlaNet performs so well without being able to use the cues that humans rely on, such as vegetation, architectural style, and so on. But Weyand and co say they know why: "We think PlaNet has an advantage over humans because it has seen many more places than any human can ever visit and has learned subtle cues of different scenes that are even hard for a well-traveled human to distinguish."

They go further and use the machine to locate images that do not have location cues, such as those taken indoors or of specific items. This is possible when images are part of albums that have all been taken at the same place. The machine simply looks through other images in the album to work out where they were taken and assumes the more specific image was taken in the same place.

That's impressive work that shows deep neural nets flexing their muscles once again. Perhaps more impressive still is that the model uses a relatively small amount of memory unlike other approaches that use gigabytes of the stuff. "Our model uses only 377 MB, which even fits into the memory of a smartphone," say Weyand and co.

That's a tantalizing idea—the power of a superhuman neural network on a smartphone. It surely won't be long now!

Ref: arxiv.org/abs/1602.05314 : PlaNet—Photo Geolocation with Convolutional Neural Networks