Inside the Mind of a Neural Network

In this lesson, we will take a peek inside a simple neural network to visualize what we have learned so far.

> Neural networks are useful for solving the kinds of problems that we don't really know how to solve with simple, crisp rules. Imagine writing a set of rules to apply to images of handwritten numbers to decide what the number was. You can imagine that wouldn't be easy, and our attempts probably not very successful either.

Mysterious Black Box

Once a neural network is trained and performs well enough on test data, you essentially have a mysterious *black box*. You don't really know how it works out the answer — it just does.

This isn't always a problem if you're just interested in answers, and don't really care how they're arrived at. But it is a disadvantage of these kinds of machine learning methods — the learning doesn't often translate into understanding or wisdom about the problem the black box has learned to solve.

Let's see if we can take a peek inside our simple neural network to see if we can understand what it has learned, to visualize the knowledge it has gathered through training.

We could look at the weights, which is after all what the neural network learns. But that's not likely to be that informative. Especially as the way neural networks work is to distribute their learning across different link weights. This gives them an advantage in that they are resilient to damage, just like biological brains are. It's unlikely that removing one node, or even quite a few nodes, will completely damage the ability of a neural network to work well.

Here's a crazy idea: Backwards Query!

Here's a crazy idea. Backwards Query!