#h1 Chapter 29 - Visualizing cDBNNs

#h2 Understanding what has been learned

#h3 29.1 - Introduction

#pg In the previous chapter, we have seen convolutional neural networks achieve remarkable accuracy on classifying images as containing one of many different visual objects the neural networks have been trained on. By adding more and more convolutional layers, greater accuracies were achieved. The same neural networks were now capable of moving beyond simply classifying an image as containing a visual object, newer generations of cDBNNs can now segment the object from its background, reason about the relative position of one visual object versus another (the pizza is in the box which itself is on the table) and even compute the position and orientation of an object (e.g., the pizza) in world coordinates. A lot of these models start off by training on the large and annotated image database ImageNet, which has been the gold standard to benchmark neural networks for machine vision for some years. We used a subset of its images (100 categories with roughly 1000 images per category) in the previous chapter to train a full version of AlexNet. The complete ImageNet dataset contains 1000 categories, with thousands of images per category. Even with that many categories, neural networks such as AlexNet, InceptionV3 and VGG19 perform exceptional well after rigorous training. What adds to the remarkable feat of these networks is that the training data isn't all that carefully curated. If we plot a handful of images of 'power drill', we notice that the images differ greatly in resolution, contrast, and brightness. Moreover, the visual object in it for which the image has been labeled 'power drill' is often small compared to the entire image. It is not centered, something occluded or partial visible, and in a wide range of different orientations. Not to mention that power drills themselves vary quite a bit in their appearance, even if we images of just the drill, at fixed distance and orientation relative to the camera, and posing in front of a white backdrop. It is generally agreed on that such data sets are preferable to the highly curated ones. If these neural networks need to operate 'in the wild' they need not expect these pristine conditions. Usually object are partially occluded, poorly lit and at a weird angle relative to us. In other words, a highly curated image database doesn't generalize well to the real world. Second, the large amount in variation between images for a single category (like a power drill, or baboon, or jumbo jet) forces the neural network to really get the thing object the visual object that is invariant to position, location, orientation, color, etc. If you paint a blue 747 jet green, it still is a jumbo jet. And imagine standing in front of a 747, on the runway looking up to its cockpit, or image seeing the jet at an angle through the terminal window as its taxis to the gate. The images are widely different from a pixel-by-pixel perspective. But they are still both images of a 747. So, what exactly is learned by a convolutional neural network that allows it to pick up on these invariant properties, while at the same time not being thrown off by the things that vary in the image that do not change the fact our image is still an image of a 747?

#h3 28.2 - Representations

#pg In neuroscience, one of the active lines of research focusses on the neural representation of sensory inputs. As we have seen in the chapter on biological vision, and as we will see later for artificial neural networks as well, vision is best characterized as a sequence of layers of neurons (or nodes) that increase both in the amount of the image they can respond to and well as the complexity of their response. In the first cortical area processing the inputs coming from the photoreceptors in our eye, named V1, each neuron 'sees' only a very small portion of the entire visual field (or image if it is an artificial neural network). When we say 'sees only a small portion' we mean that a neuron in V1 only receives inputs from a small, localized patch within the visual field. Within this patch, which we refer to as a receptive field, the neuron also expressed a preference to a particular kind of visual input within that patch. We refer to that as its selectivity or tuning. In V1, most neurons detect boundaries between dark and bright, in other words: edges. But they also typically want that boundary to have a certain orientation. Some cells require yet another thing, they want to have this edge move across their receptive field at a certain speed at a certain orientation. These are the building blocks of our ability to detect movement. After V1, neurons that have receptive fields that overlap or are at least nearby in visual space all project to a higher layer. This means that the neurons in this next layer (called V2) has larger receptive fields, since we now combine many nearby smaller receptive fields together. In addition, V2 receives outputs from many differently tuned V1 neurons, tuned to different orientations and spatial frequencies. This allows V2 to create a more complex tuning, usually imagined as textures. This process is repeated several times in the human visual system through various cortical areas. At each subsequent layer, the neuron sees a little bit more of the visual field and has a more complex tuning. At a certain stage, this tuning becomes very specific. Neurons now seem to respond to complete man-made objects, animals, plants, etc. All the way up to an area called the fusiform face area, which only really responds to images to human faces. And it gets a bit weirder. In a famous study where electrodes were implanted directly into the brain of a human subject, they found what has been dubbed the 'Jennifer Aniston' cell. It is important to point out that these electrodes were planted by the neurosurgeon to precisely pinpoint to brain area that caused the intractable epilepsy in these patients, so the doctors could spare as much tissue as possible and only remove the area that caused the seizures. The Jennifer Aniston cell was found by accident. The research team was quite surprised when they found a cell only responding and becoming active when a picture of Jennifer Aniston was shown to the subject. Interestingly, when Brad Pitt was in the same picture (both actors were married for some time) the cell stopped firing! This reignited an old debate: the existence of a 'grandmother cell'. If we take the process of ever-increasing complexity and the experimental findings at face value, it seems to confirm that at the very end of processing visual information, each cell is so specialized, it encodes a single unique object (Jennifer Aniston) rather than a class of objects (human faces in general). In the same way I should have a cell or an only a few cells truly encoding one thing and one thing only: my grandmother. I don't buy into that framework. It is a rigid system, there are too many different unique objects to begin with, and what happens if that one cell dies (which is not uncommon, neurons do die off with age). Do I simply forget everything about my grandmother but not about anyone else? Rather, I favor a distributed system, where the activity many neurons (or nodes) at many different layers together form a code. Between visual objects of a very different class (jumbo jets and grandmothers) these codes are widely different. At the same time, the codes for two objects of the same class (my grandmother on one side, versus my grandmother on the other side of my family) are far more similar, but they are far from the same. Since artificial neural networks use architectures that are inspired by biological vision, perhaps studying what has been learned by them sheds light on this question. But even without comparing biological and artificial neural networks, it is still informative to try and interpret what has been learned in cDBNNs.

#h3 29.3 - Visualizing Neural Networks using TensorFlow Lucid

#h3 29.4 - Demos

#h3 29.5 - References