ply an elaborate form of Lego stacking. Figure 6-7 demonstrates how a convolutional architecture might be built up out of constituent blocks.

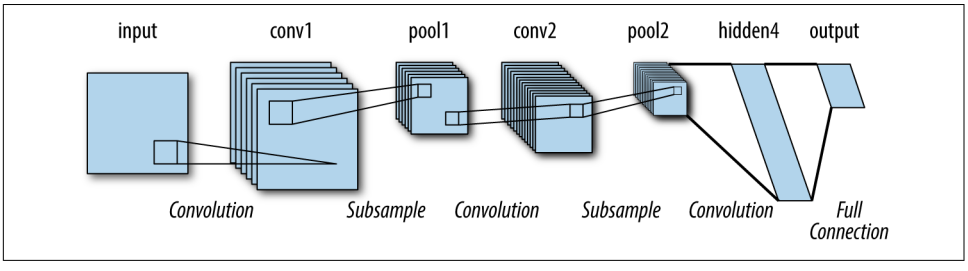

Figure 6-7. An illustration of a simple convolutional architecture constructed out of stacked convolutional and pooling layers.

Dilated Convolutions

Dilated or atrous convolutions are a newly popular form of convolutional layer. The insight here is to leave gaps in the local receptive field for each neuron (atrous means *a trous*, or "with holes" in French). The basic concept is an old one in signal processing that has recently found some good traction in the convolutional literature.

The core advantage to the atrous convolution is the increase in visible area for each neuron. Let's consider a convolution architecture whose first layer is a vanilla convolutional with 3×3 local receptive fields. Then a neuron one layer deeper in the architecture in a second vanilla convolutional layer has receptive depth 5×5 (each neuron in a local receptive field of the second layer itself has a local receptive field in the first layer). Then, a neuron two layers deeper has receptive view 7×7. In general, a neuron N layers within the convolutional architecture has receptive view of size $(2N + 1) \times (2N + 1)$. This linear growth in receptive view is fine for smaller images, but quickly becomes a liability for large images.

The atrous convolution enables exponential growth in the visible receptive field by leaving gaps in its local receptive fields. A "1-dilated" convolution leaves no gaps, while a "2-dilated" convolution leaves one gap between each local receptive field element. Stacking dilated layers leads to exponentially increasing local receptive field sizes. Figure 6-8 illustrates this exponential increase.

Dilated convolutions can be very useful for large images. For example, medical images can stretch thousands of pixels in every dimension. Creating vanilla convolutional networks that have global understanding could require unreasonably deep networks. Using dilated convolutions could enable networks to better understand the global structure of such images.

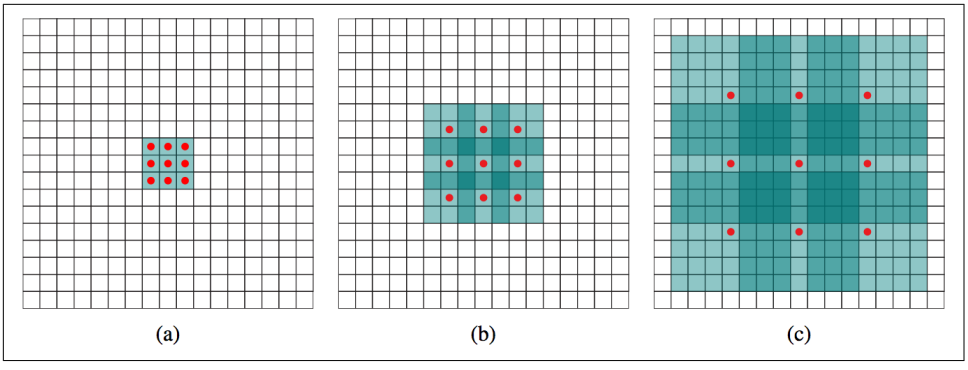

Figure 6-8. A dilated (or atrous) convolution. Gaps are left in the local receptive field for each neuron. Diagram (a) depicts a 1-dilated 3 × 3 convolution. Diagram (b) depicts the application of a 2-dilated 3 × 3 convolution to (a). Diagram (c) depicts the application of a 4-dilated 3 × 3 convolution to (b). Notice that the (a) layer has receptive field of width 3, the (b) layer has receptive field of width 7, and the (c) layer has receptive field of width 15.

Applications of Convolutional Networks

In the previous section, we covered the mechanics of convolutional networks and introduced you to many of the components that make up these networks. In this section, we describe some applications that convolutional architectures enable.

Object Detection and Localization

Object detection is the task of detecting the objects (or entities) present in a photograph. Object localization is the task of identifying where in the image the objects exist and drawing a "bounding box" around each occurrence. Figure 6-9 demonstrates what detection and localization on standard images looks like.

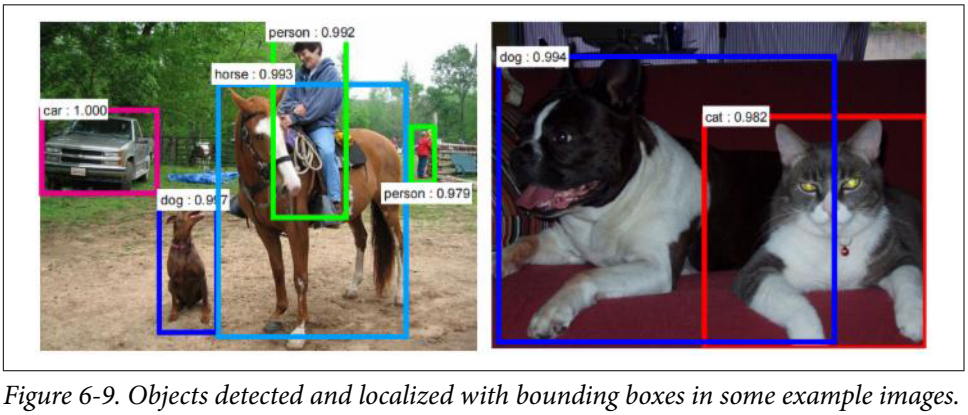

Figure 6-9. Objects detected and localized with bounding boxes in some example images.