challenge with error rates half that of the nearest competitors. This victory dramatically galvanized the (already nascent) trend toward deep learning architectures in computer vision. The AlexNet architecture is illustrated in Figure 1-6.

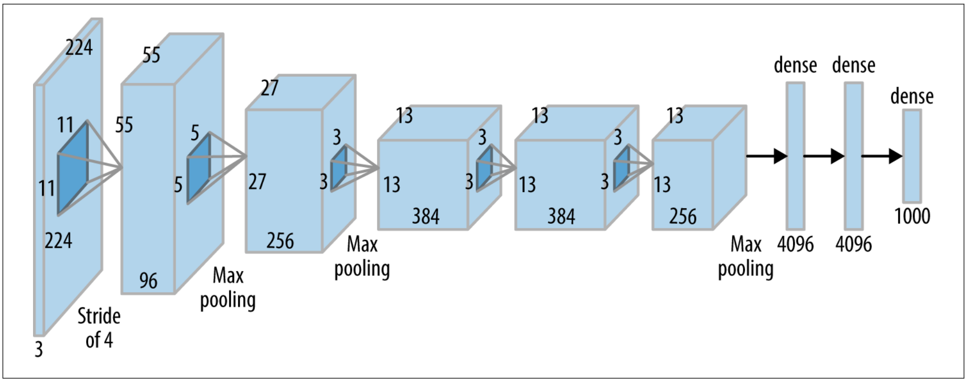

Figure 1-6. The AlexNet architecture for image processing. This architecture was the winning entry in the ILSVRC 2012 challenge and galvanized a resurgence of interest in convolutional architectures.

ResNet

Since 2012, convolutional architectures consistently won the ILSVRC challenge (along with many other computer vision challenges). Each year the contest was held, the winning architecture increased in depth and complexity. The ResNet architecture, winner of the ILSVRC 2015 challenge, was particularly notable; ResNet architectures extended up to 130 layers deep, in contrast to the 8-layer AlexNet architecture.

Very deep networks historically were challenging to learn; when networks grow this deep, they run into the vanishing gradients problem. Signals are attenuated as they progress through the network, leading to diminished learning. This attenuation can be explained mathematically, but the effect is that each additional layer multiplicatively reduces the strength of the signal, leading to caps on the effective depth of networks.

The ResNet introduced an innovation that controlled this attenuation: the bypass connection. These connections allow part of the signal from deeper layers to pass through undiminished, enabling significantly deeper networks to be trained effectively. The ResNet bypass connection is illustrated in Figure 1-7.

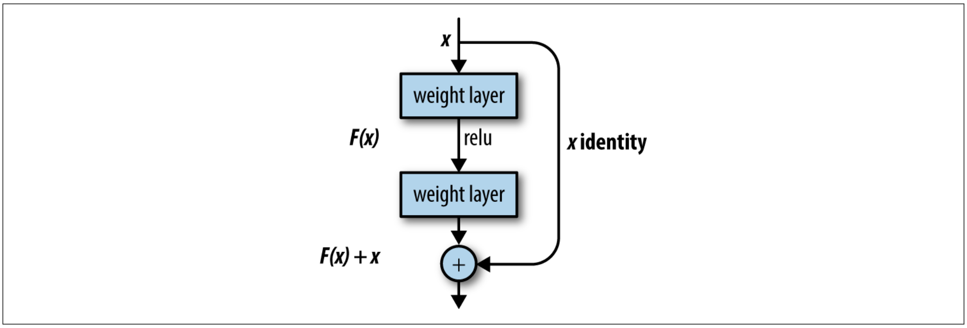

Figure 1-7. The ResNet cell. The identity connection on the righthand side permits an unmodified version of the input to pass through the cell. This modification allows for the effective training of very deep convolutional architectures.

Neural Captioning Model

As practitioners became more comfortable with the use of deep learning primitives, they experimented with mixing and matching primitive modules to create higher-order systems that could perform more complex tasks than basic object detection. Neural captioning systems automatically generate captions for the contents of images. They do so by combining a convolutional network, which extracts information from images, with an LSTM layer that generates a descriptive sentence for the image. The entire system is trained *end-to-end*. That is, the convolutional network and the LSTM network are trained together to achieve the desired goal of generating descriptive sentences for provided images.

This end-to-end training is one of the key innovations powering modern deep learning systems since it lessens the need for complicated preprocessing of inputs. Image captioning models that don't use deep learning would have to use complicated image featurization methods such as SIFT, which can't be trained alongside the caption generator.

A neural captioning model is illustrated in Figure 1-8.