form of the model or the way it is trained.

The development of backpropagation and gradient-based optimization dramatically increased the capability of neural networks to solve practical problems. However, it was also observed that in networks with many layers, it was only weights in the final two layers that would learn useful values. With a few exceptions, notably models used for image analysis known as convolutional neural networks (LeCun *et al.*, 1998), there were very few successful applications of networks having more than two layers. Again, this constrained the complexity of the problems that could be addressed effectively with these kinds of network. To achieve reasonable performance on many applications, it was necessary to use hand-crafted *pre-processing* to transform the input variables into some new space where, it was hoped, the machine learning problem would be easier to solve. This pre-processing stage is sometimes also called *feature extraction*. Although this approach was sometimes effective, it would clearly be much better if features could be learned from the data rather than being hand-crafted.

Chapter 10

By the start of the new millennium, the available neural network methods were once again reaching the limits of their capability. Researchers began to explore a raft of alternatives to neural networks, such as kernel methods, support vector machines, Gaussian processes, and many others. Neural networks fell into disfavour once again, although a core of enthusiastic researchers continued to pursue the goal of a truly effective approach to training networks with many layers.

1.3.3 Deep networks

The third, and current, phase in the development of neural networks began during the second decade of the 21st century. A series of developments allowed neural networks with many layers of weights to be trained effectively, thereby removing previous limitations on the capabilities of these techniques. Networks with many layers of weights are called *deep neural networks* and the sub-field of machine learning that focuses on such networks is called *deep learning* (LeCun, Bengio, and Hinton, 2015).

One important theme in the origins of deep learning was a significant increase in the scale of neural networks, measured in terms of the number of parameters. Although networks with a few hundred or a few thousand parameters were common in the 1980s, this steadily rose to the millions, and then billions, whereas current state-of-the-art models can have in the region of one trillion (10^{12}) parameters. Networks with many parameters require commensurately large data sets so that the training signals can produced good values for those parameters. The combination of massive models and massive data sets in turn requires computation on a massive scale when training the model. Specialist processors called *graphics processing units*, or GPUs, which had been developed for very fast rendering of graphical data for applications such as video games, proved to be well suited to the training of neural networks because the functions computed by the units in one layer of a network can be evaluated in parallel, and this maps well onto the massive parallelism of GPUs (Krizhevsky, Sutskever, and Hinton, 2012). Today, training for the largest models is performed on large arrays of thousands of GPUs linked by specialist high-speed interconnections.