to discover some internal representation for the data that is useful for solving other tasks, such as image classification.

Historically, unsupervised learning played an important role in enabling the first deep networks (apart from convolutional networks) to be successfully trained. Each layer of the network was first pre-trained using unsupervised learning and then the entire network was trained further using gradient-based supervised training. It was later discovered that the pre-training phase could be omitted and a deep network could be trained from scratch purely using supervised learning given appropriate conditions.

However, pre-training and representation learning remain central to deep learning in other contexts. The most notable example of pre-training is in natural language processing in which transformer models are trained on large quantities of text and are able to learn highly sophisticated internal representations of language that facilitates an impressive range of capabilities at human level and beyond.

Chapter 12

6.3.4 Transfer learning

The internal representation learned for one particular task might also be useful for related tasks. For example, a network trained on a large labelled data set of everyday objects can learn how to transform an image representation into one that is much better suited for classifying objects. Then, the final classification layer of the network can be retrained using a smaller labelled data set of skin lesion images to create a lesion classifier. This is an example of *transfer learning* (Hospedales *et al.*, 2021), which allows higher accuracy to be achieved than if only the lesion image data were used for training, because the network can exploit commonalities shared by natural images in general. Transfer learning is illustrated in Figure 6.13.

Section 1.1.1

In general, transfer learning can be used to improve performance on some task A, for which training data is in short supply, by using data from a related task B, for which data is more plentiful. The two tasks should have the same kind of inputs, and there should be some commonality between the tasks so that low-level features, or internal representations, learned from task B will be useful for task A. When we look at convolutional networks we will see that many image processing tasks require similar low-level features corresponding to the early layers of a deep neural network, whereas later layers are more specialized to a particular task, making such networks well suited to transfer learning applications.

Chapter 10

When data for task A is very scarce, we might simply re-train the final layer of the network. In contrast, if there are more data points, it is feasible to retrain several layers. The process of learning parameters using one task that are then applied to one or more other tasks is called *pre-training*. Note that for the new task, instead of applying stochastic gradient descent to the whole network, it is much more efficient to send the new training data once through the fixed pre-trained network so as to evaluate the training inputs in the new representation. Iterative gradient-based optimization can then be applied just to the smaller network consisting of the final layers. As well as using a pre-trained network as a fixed pre-processor for a different task, it is also possible to apply *fine-tuning* in which the whole network is adapted to the data for task A. This is generally done with a very small learning rate for a lim-