CHAPTER 9

Training Large Deep Networks

Thus far, you have seen how to train small models that can be completely trained on a good laptop computer. All of these models can be run fruitfully on GPU-equipped hardware with notable speed boosts (with the notable exception of reinforcement learning models for reasons discussed in the previous chapter). However, training larger models still requires considerable sophistication. In this chapter, we will discuss various types of hardware that can be used to train deep networks, including graphics processing units (GPUs), tensor processing units (TPUs), and neuromorphic chips. We will also briefly cover the principles of distributed training for larger deep learning models. We end the chapter with an in-depth case study, adapated from one of the TensorFlow tutorials, demonstrating how to train a CIFAR-10 convolutional neural network on a server with multiple GPUs. We recommend that you attempt to try running this code yourself, but readily acknowledge that gaining access to a multi-GPU server is trickier than finding a good laptop. Luckily, access to multi-GPU servers on the cloud is becoming possible and is likely the best solution for industrial users of TensorFlow seeking to train large models.

Custom Hardware for Deep Networks

As you've seen throughout the book, deep network training requires chains of tensorial operations performed repeatedly on minibatches of data. Tensorial operations are commonly transformed into matrix multiplication operations by software, so rapid training of deep networks fundamentally depends on the ability to perform matrix multiplication operations rapidly. While CPUs are perfectly capable of implementing matrix multiplications, the generality of CPU hardware means much effort will be wasted on overhead unneeded for mathematical operations.

Hardware engineers have noted this fact for years, and there exist a variety of alternative hardware for working with deep networks. Such hardware can be broadly divided into *inference only* or *training and inference*. Inference-only hardware cannot be used to train new deep networks, but can be used to deploy trained models in production, allowing for potentially orders-of-magnitude increases in performance. Training and inference hardware allows for models to be trained natively. Currently, Nvidia's GPU hardware holds a dominant position in the training and inference market due to significant investment in software and outreach by Nvidia's teams, but a number of other competitors are snapping at the GPU's heels. In this section, we will briefly cover some of these newer hardware alternatives. With the exception of GPUs and CPUs, most of these alternative forms of hardware are not yet widely available, so much of this section is forward looking.

CPU Training

Although CPU training is by no means state of the art for training deep networks, it often does quite well for smaller models (as you've seen firsthand in this book). For reinforcement learning problems, a multicore CPU machine can even outperform GPU training.

CPUs also see wide usage for inference-only applications of deep networks. Most companies have invested heavily in developing cloud servers built primarily on Intel server boxes. It's very likely that the first generation of deep networks deployed widely (outside tech companies) will be primarily deployed into production on such Intel servers. While such CPU-based deployment isn't sufficient for heavy-duty deployment of learning models, it is often plenty for first customer needs. Figure 9-1 illustrates a standard Intel CPU.