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.