Model Parallelism

The human brain provides the only known example of a generally intelligent piece of hardware, so there have naturally been comparisons drawn between the complexity of deep networks and the complexity of the brain. Simple arguments state the brain has roughly 100 billion neurons; would constructing deep networks with that many "neurons" suffice to achieve general intelligence? Unfortunately, such arguments miss the point that biological neurons are significantly more complex than "mathematical neurons." As a result, simple comparisons yield little value. Nonetheless, building larger deep networks has been a major research focus over the last few years.

The major difficulty with training very large deep networks is that GPUs tend to have limited memory (dozens of gigabytes typically). Even with careful encodings, neural networks with more than a few hundred million parameters are not feasible to train on single GPUs due to memory requirements. Model parallel training algorithms attempt to sidestep this limitation by storing large deep networks on the memories of multiple GPUs. A few teams have successfully implemented these ideas on arrays of GPUs to train deep networks with billions of parameters. Unfortunately, these models have not thus far shown performance improvements justifying the extra difficulty. For now, it seems that the increase in experimental ease from using smaller models outweighs the gains from model parallelism.

Hardware Memory Interconnects

Enabling model parallelism requires having very high bandwidth connections between compute nodes since each gradient update by necessity requires internode communication. Note that while data parallelism requires strong interconnects, sync operations need only be performed sporadically after multiple local gradient updates.

A few groups have used InfiniBand interconnects (InfiniBand is a high-throughput, low-latency networking standard), or Nvidia's proprietary NVLINK interconnects to attempt to build such large models. However, the results from such experiments have been mixed thus far, and the hardware requirements for such systems tend to be expensive.

Data Parallel Training with Multiple GPUs on Cifar10

In this section, we will give you an in-depth walkthrough of how to train a data-parallel convolutional network on the Cifar10 benchmark set. Cifar10 consists of 60,000 images of size 32 × 32. The Cifar10 dataset is often used to benchmark convolutional architectures. Figure 9-7 displays sample images from the Cifar10 dataset.

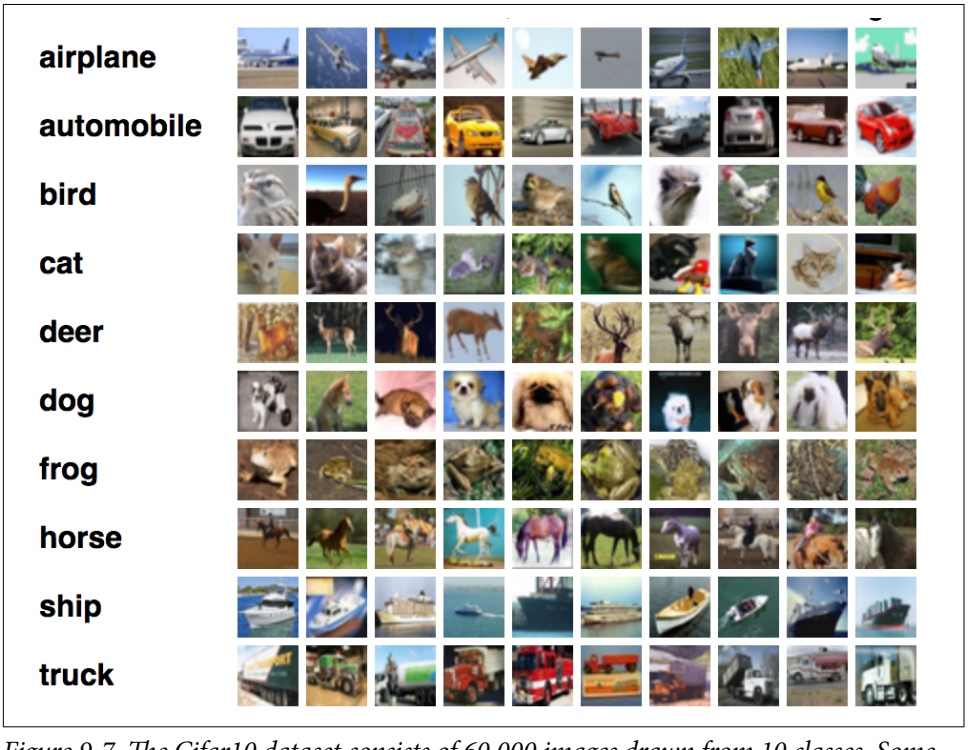

Figure 9-7. The Cifar10 dataset consists of 60,000 images drawn from 10 classes. Some sample images from various classes are displayed here.

The architecture we will use in this section loads separate copies of the model architecture on different GPUs and periodically syncs learned weights across cores, as Figure 9-8 illustrates.