Hardware for Deep Neural Networks

Originalartikel

Backup

<html> In case you didn’t make it to the ISCA (International Society for Computers and their Applications) session this year, you might be interested in a presentation by [Joel Emer] an MIT professor and scientist for NVIDIA. Along with another MIT professor and two PhD students ([Vivienne Sze], [Yu-Hsin Chen], and [Tien-Ju Yang]), [Emer’s] presentation covers hardware architectures for deep neural networks. The presentation covers the background on deep neural networks and basic theory. Then it progresses to deep learning specifics. One interesting graph shows how neural networks are getting better at identifying objects in images every year and as of 2015 can do a better job than a human over a set of test images. However, the real key is using hardware to accelerate the performance of networks. Hardware acceleration is important for several reasons. For one, many applications have lots of data associated. Also, training can involve many iterations which can take a long time. The presentation covers a lot more: a survey of tools, current hardware available, and the exploration of different kernels (algorithms) for use in different layers of the network. The real meat, though, is how to build hardware to best implement those kernels. Throwing parallel elements at the task is obvious, but the paper points out that memory access is the bottleneck. There are several strategies for reducing the cost of data access across the network. These strategies take advantage of data reuse and local accumulation of results. However, it is also possible to tune for lower power consumption by using a single memory pool at the expense of performance. Towards the end of the paper, different memory topologies and devices get scrutinized. This includes using memristors and stacks of memory cells at the IC level. If you are looking for a complete but accessible survey of the deep neural network landscape, the first half or so of this presentation will be of great interest to you. The back part is a lot more detailed and if you are just a curious hacker, may not be that useful to you. But if you are developing ASIC or even FPGA architectures for neural networks, it is great stuff. We’ve talked a lot about neural networks lately. There’s a host of classes and tutorials to get you started. </html>