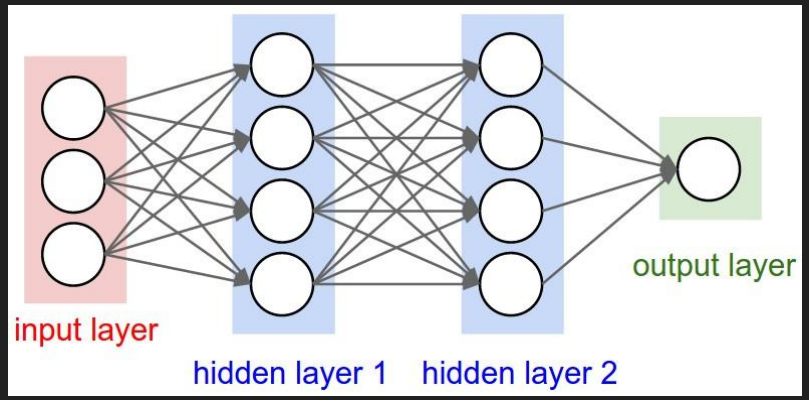
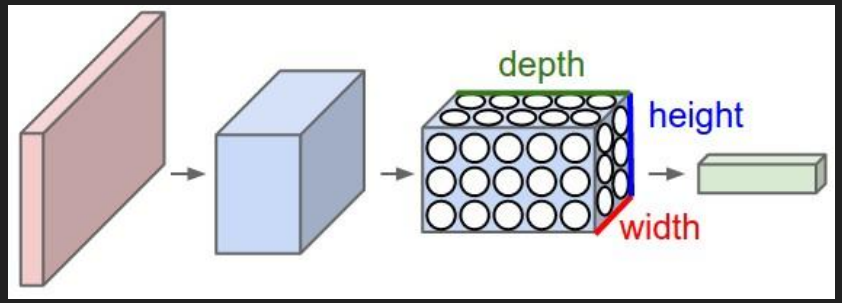
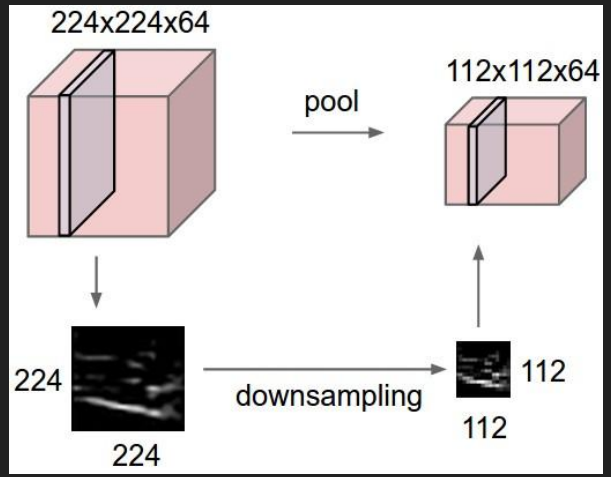
Scalable Deep Learning

Sammy Sidhu


input layer
hidden layer 1
hidden layer 2
output layer


depth
height
width


224x224x64
pool
112x112x64
224
224
downsampling
112
112

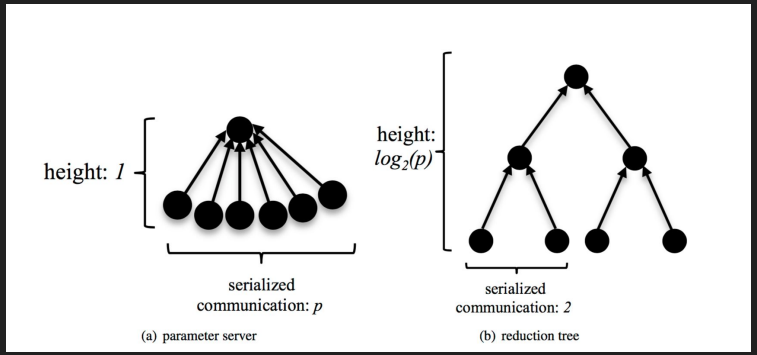
TL;DR

- Main Operations
  - Convolutions → Few Parameters, heavy compute
  - Matrix Multiplication → Tons of parameters, fast to compute given parameters size
  - Pooling, reduces number of parameters needed but lose resolution

- Typically when training on 1 GPU, convolutions dominate compute
- When you have many GPUs, Communication becomes an issue.

Training with multiple GPUs

- Data Vs Model Parallelism
- Single Node (Typically up to 8 GPUs)
  - 40 PCI-e lanes per CPU
  - PCI + QPI is typically fast enough
- Multiple Nodes (4-8 GPUs per Machine)
  - Sync vs Async
  - Faster Networking (Infiniband + MPI)
  - RDMA (Direct GPU vs Memory)
  - Merging Gradients (Tree Merge -- FireCaffe style)



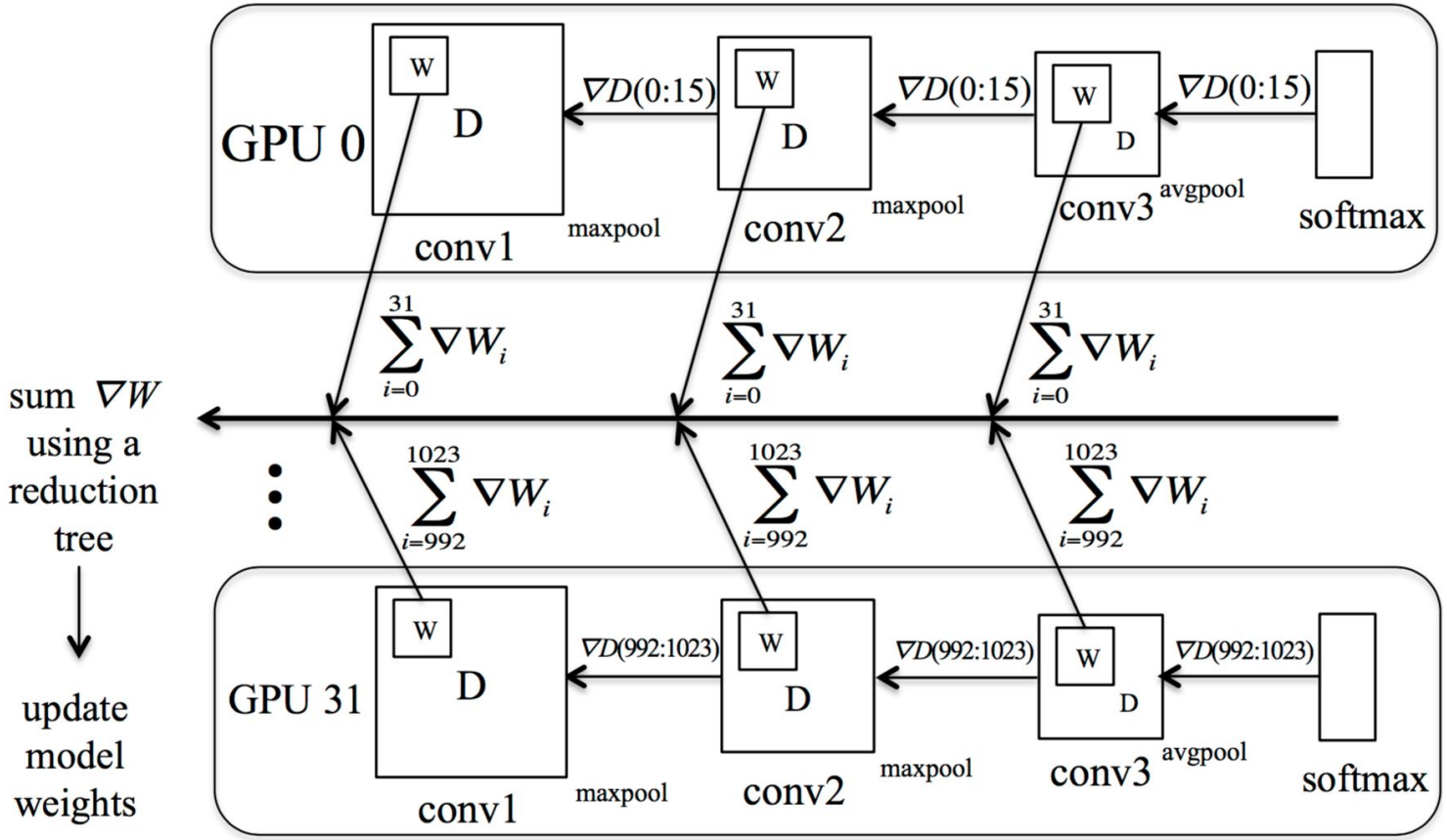
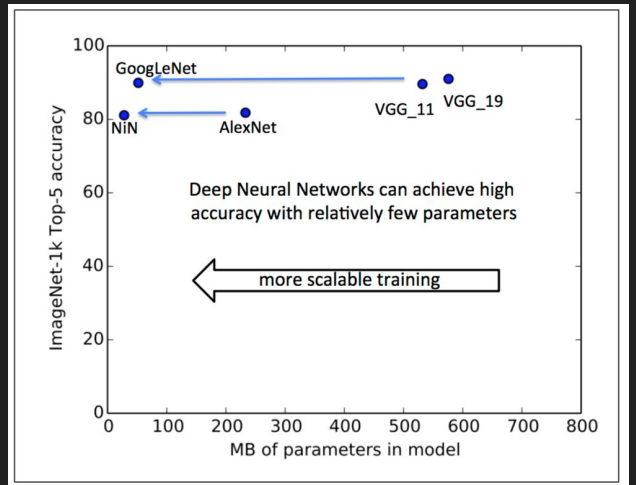
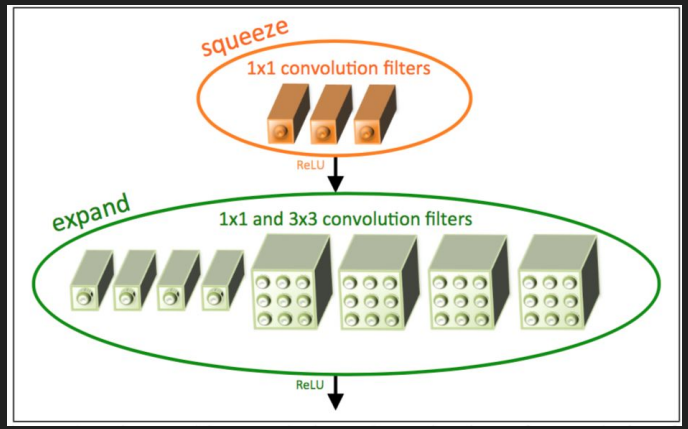


Figure 1. **Data parallel DNN training in FireCaffe:** Each worker (GPU) gets a subset of each batch.

What kinds of models are good for parallel training?

- Less parameters → Less communication
- More Convolutions → less parameters
- We can see this progression in NN architecture as well
- AlexNet (138M) → googLeNet (~12M) → NiN (~8M) → ResNet-152 (~2M)
- SqueezeNet(1.2M, ~400k pruned)





FireCaffe with googLeNet

Table 3. Accelerating the training of ultra-deep, computationally intensive models on ImageNet-1k.

	Hardware	Net	Epochs	Batch size	Initial Learning Rate	Train time	Speedup	Top-1 Accuracy	Top-5 Accuracy
Caffe	1 NVIDIA K20	GoogLeNet [41]	64	32	0.01	21 days	1x	68.3%	88.7%
FireCaffe (ours)	32 NVIDIA K20s (Titan supercomputer)	GoogLeNet	72	1024	0.08	23.4 hours	20x	68.3%	88.7%
FireCaffe (ours)	128 NVIDIA K20s (Titan supercomputer)	GoogLeNet	72	1024	0.08	10.5 hours	**47x**	68.3%	88.7%

SqueezeNet Results

CNN architecture	Compression Approach	Data Type	Original → Compressed Model Size	Reduction in Model Size vs. AlexNet	Top-1 ImageNet Accuracy	Top-5 ImageNet Accuracy
AlexNet	None (baseline)	32 bit	240MB	1x	57.2%	80.3%
AlexNet	SVD [5]	32 bit	240MB → 48MB	5x	56.0%	79.4%
AlexNet	Network Pruning [11]	32 bit	240MB → 27MB	9x	57.2%	80.3%
AlexNet	Deep Compression [10]	5-8 bit	240MB → 6.9MB	35x	57.2%	80.3%
SqueezeNet (ours)	None	32 bit	4.8MB	**50x**	57.5%	80.3%
SqueezeNet (ours)	Deep Compression	8 bit	4.8MB → 0.66MB	**363x**	57.5%	80.3%
SqueezeNet (ours)	Deep Compression	6 bit	4.8MB → 0.47MB	**510x**	57.5%	80.3%