ImageNet Classification with Deep Convolutional Neural Networks

Summary

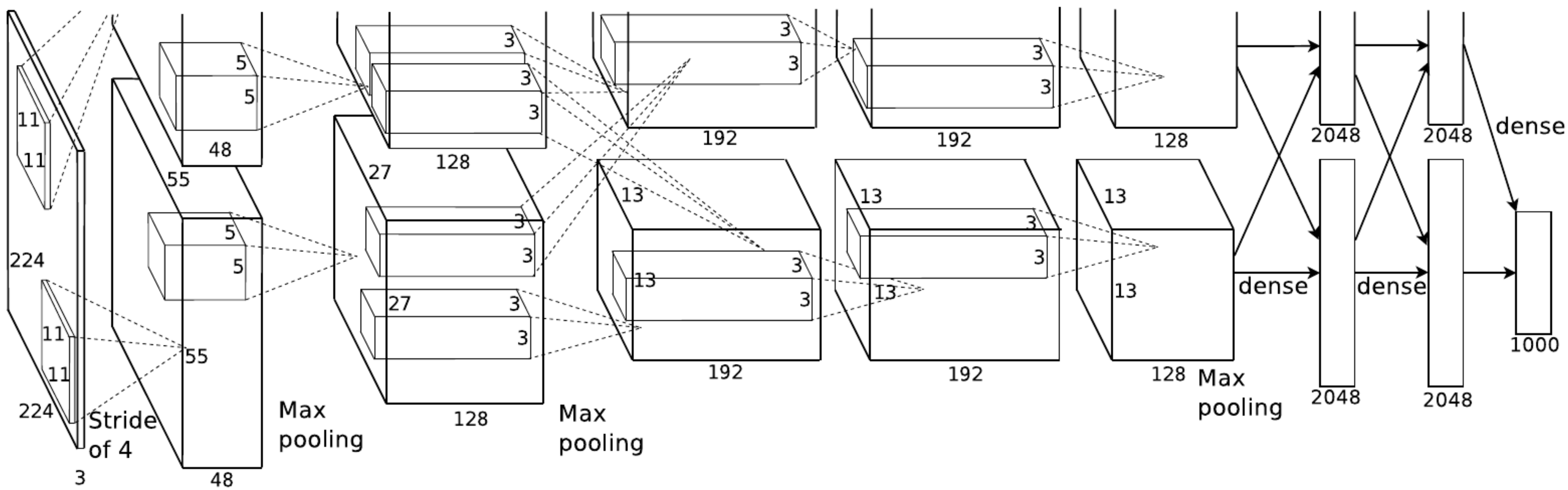
- Introduction
- Network Architecture
  - ReLU Nonlinearity
  - Training on Multiple GPUs
  - Overlapping Pooling
  - Overall Architecture
- Reducing overfitting
  - Data augmentation
  - Dropout
- Details of learning
- Results

Introduction

- ImageNet
  - Over 15 million high-quality labeled images
  - About 22,000 categories
  - Collected from the web, labeled by humans on Amazon's Mechanical Turk
  - Variable-resolution images

- ILSVRC
  - ImageNet Large-Scale Pascal Visual Object Challenge
  - Subset of ImageNet
  - 1,000 categories with about 1,000 images each
  - 1.2 million training images, 50k for validation, 150k for testing
  - Usually people report the top-1 and top-5 error rates

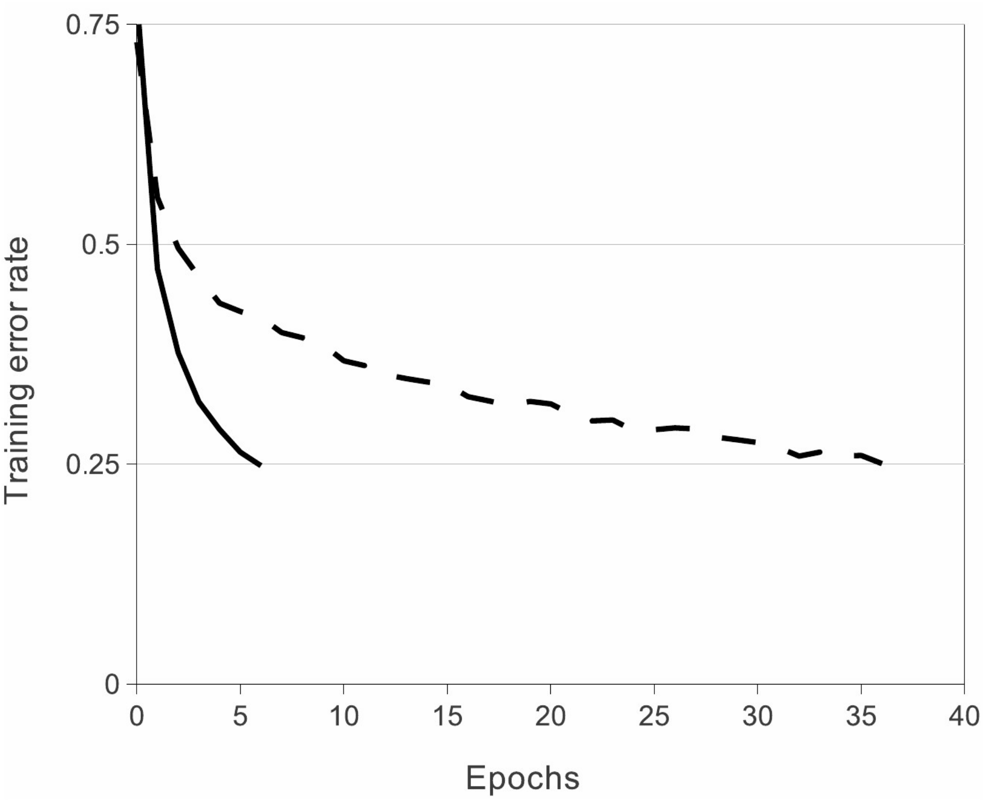
Network architecture



- ReLU Nonlinearity
- Training on Multiple GPUs
- Overlapping Pooling
- Overall Architecture

ReLU Nonlinearity (Rectified Linear Units)

- Standard ways to model a neuron's output: tanh or sigmoid
- ReLU: f(x) = max(0, x)
- Train several times faster than tanh
- On CIFAR-10, a convolutional neural network with ReLUs reaches a 25% error rate six times faster than with tanh



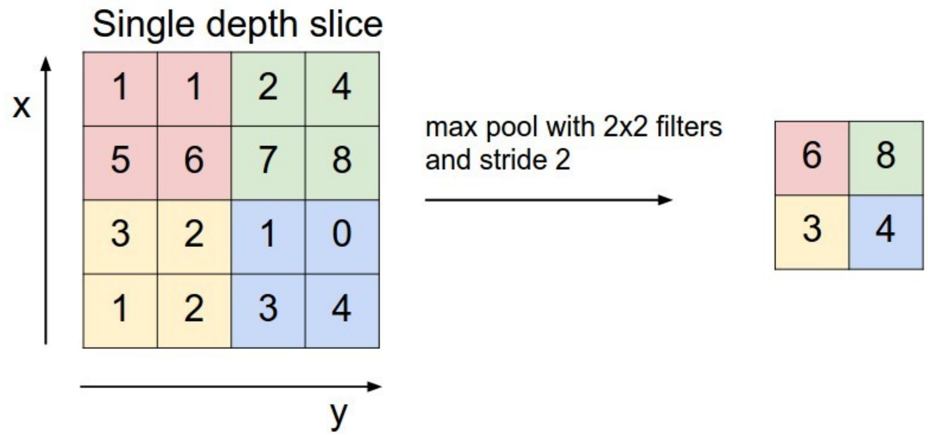
Training on Multiple GPUs

- Network trained using GTX 580 GPU (only 3GB memory)
- GPU are well suited for cross-GPU parallelization
- Parallelization scheme: put half of the kernels on each GPU
- GPUs only communicate in certain layers
  - Layer 3 takes all inputs from layer 2
  - Layer 4 takes all inputs from layer 3 that reside on the same GPU
- The two-GPU net is slightly faster than the one-GPU net
- Reduces the top-1 and top-5 errors of 1.7% and 1.2%

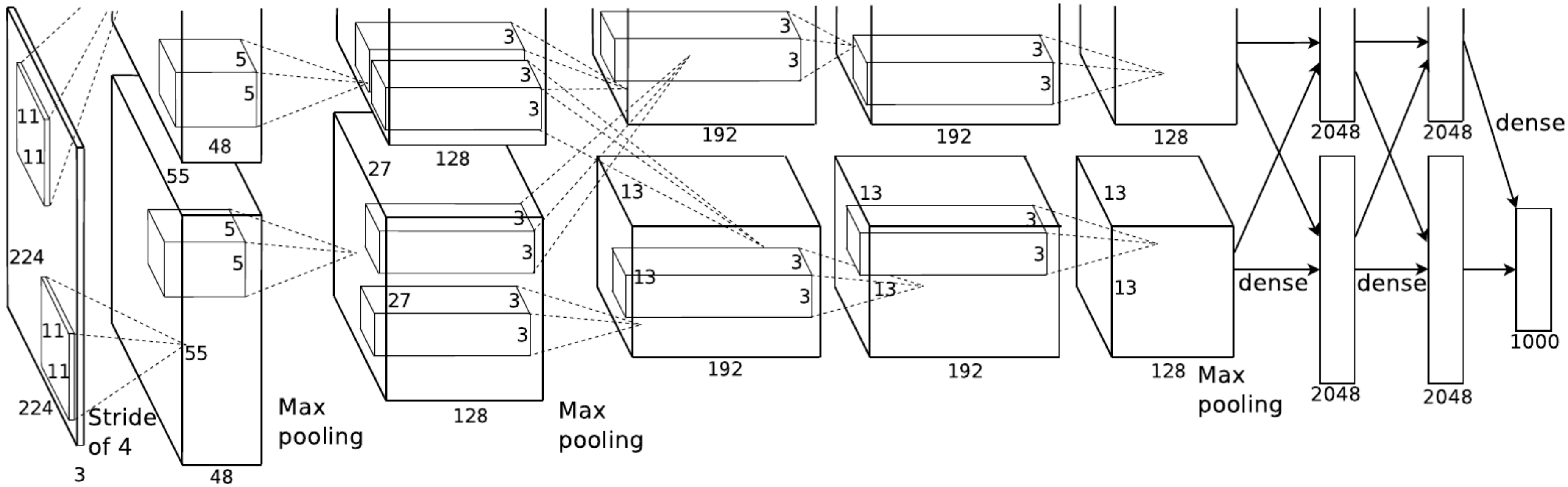
Overlapping Pooling

- Max Pooling Layer
  - A grid of pooling units summarizing neighborhoods of size z x z, separated by a stride s
  - Traditionally, s = z (no overlapping)

- Overlapping Pooling
  - s < z (a same unit can be selected several times as a the max)
  - In the described method authors took s = 2 and z = 3 in the whole network
  - Reduced top-1 and top-5 error rates by 0.4% and 0.3%, compared to s = z = 2
  - Makes the model more robust to overfitting



Overall Architecture



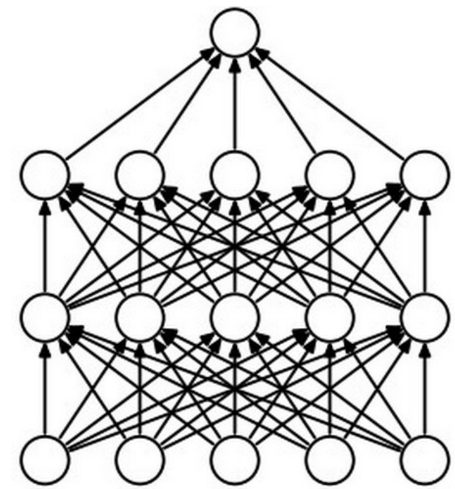
- 8 layers (5 convolutional, 3 fully-connected)
- Last layer: 1000-way softmax
- Cost function: categorical cross-entropy
- The kernels in the layers 2, 4 and 5 only take kernels maps from the previous layers which reside on the same GPU
- Neurons in a fully-connected layer are connected to all neurons in the previous layer
- Apart from the softmax, ReLU is the only non-linearity function used

Data augmentation

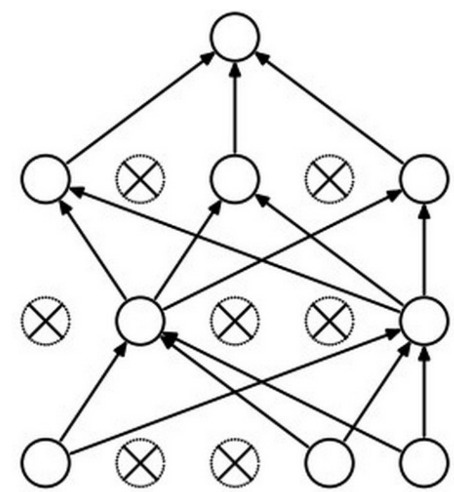
- Two types of image transformation that involve minimal computation
- Transformations are computed on CPU while GPU is training the previous batch
- Image extraction with reflection
  - Randomly select 224 x 224 patches from the 256 x 256 images
  - Take horizontal reflections
- Altering the intensities of the RGB channels in training images (reduces the top-1 error rate by over 1%)

Dropout

- During training, set to zero the output of randomly selected hidden neurons with probability 0.5.
- Reduces co-adaptations of neurons
  - A neuron cannot rely on the presence of particular other neurons
  - A neuron is encouraged to learn more robust features (reduces overfitting)
- Simulates the combination of several neural networks
- Requires more iterations to converge
- Much more efficient than combining many different models



(a) Standard Neural Net



(b) After applying dropout.

Details of learning

- SGD + momentum (0.9) + decay (0.0005)
- Parameters initialization
  - Each layer: zero-mean Gaussian distribution with standard deviation of 0.01
  - Bias in convolutional layers are set to 1 (accelerates early stages of the training)
  - Remaining bias are set 0 otherwise
- Same learning rate in all layers. Initialized to 0.01, decrease by a factor 10 when the validation error stops improving
- 90 epochs over 1.2 million images (~5 days of training on 2 GPUs)

Results

- Results on ILSVRC-2010

Model	Top-1	Top-5
Sparse coding	47.1%	28.2%
SIFT + FVs	45.7%	25.7%
CNN	37.5%	17.0%

- Results on ILSVRC-2012

Model	Top-5
SIFT + FVs	26.2%
1 CNN	16.4%
5 CNNs	15.3%

Questions?