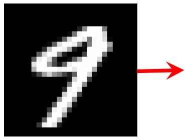
Learning handwritten digits with a neural net

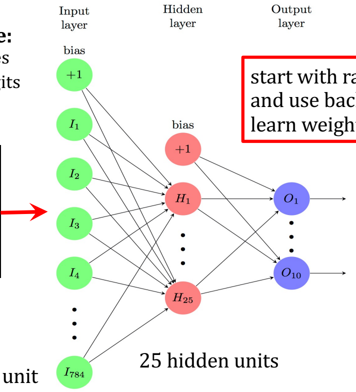
MNIST database:
3000 28x28 images
of handwritten digits

28

28



sample image



start with random initial weights and use back-propagation to learn weights to recognize digits

one output unit for each digit

select output unit with maximum response e.g. 9

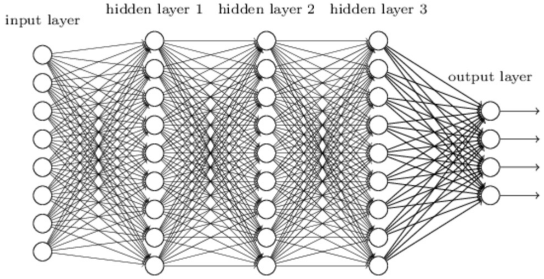
25 hidden units

one input unit for each pixel

1

"Deep" neural networks

- early work extended simple neural networks to have multiple, highly-connected hidden layers
- ***if*** such networks could be trained, they would be much more powerful than "shallow" neural nets
- ***but*** generic multi-layer networks are extremely hard to train!!



2

State-of-the-art recognition systems are based on *convolutional* neural networks

Public databases of face images serve as benchmarks:

Labeled Faces in the Wild (LFW, http://vis-www.cs.umass.edu/lfw)
> 13,000 images of celebrities, 5,749 different identities

YouTube Faces Database (YTF, http://www.cs.tau.ac.il/~wolf/ytfaces)
3,425 videos, 1,595 different identities

Private face image datasets:

(Facebook) Social Face Classification dataset
4.4 million face photos, 4,030 different identities

(Google) 100-200 million face images, ~ 8 million different identities

	LFW	YTF
Facebook DeepFace	97.4%	91.4%
Google FaceNet	99.6%	95.1%
Human performance	97.5%	89.7%

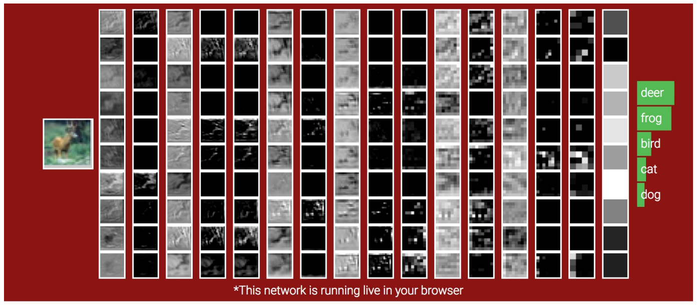
False accept

False reject

3

Convolutional Neural Networks (CNNs)

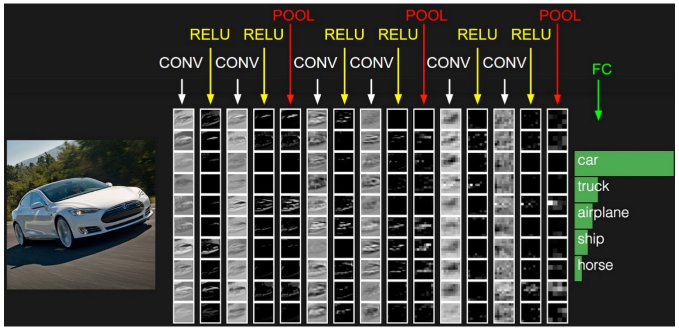
Fei-Fei Li, Justin Johnson, Serena Yeung (http://cs231n.stanford.edu/)



The Convolutional Neural Network in this example is classifying images live in your browser using Javascript, at about 10 milliseconds per image. It takes an input image and transforms it through a series of functions into class probabilities at the end. The transformed representations in this visualization can be losely thought of as the activations of the neurons along the way. The parameters of this function are learned with backpropagation on a dataset of (image, label) pairs. This particular network is classifying CIFAR-10 images into one of 10 classes and was trained with ConvNetJS. Its exact architecture is [conv-relu-conv-relu-pool]x3-fc-softmax, for a total of 17 layers and 7000 parameters. It uses 3x3 convolutions and 2x2 pooling regions. By the end of the class, you will know exactly what all these numbers mean.

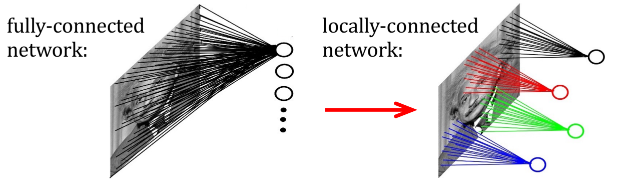
4

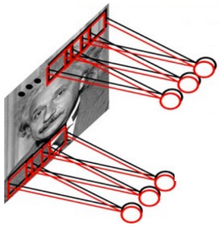
Sample stages of a CNN



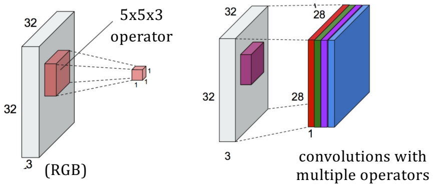
CONV: "convolution" layer with weights that are learned
RELU: "rectified linear unit" applies an activation function
POOL: "pooling" selects maximum value in small neighborhoods
FC: "fully-connected" neural network

Convolutional layer

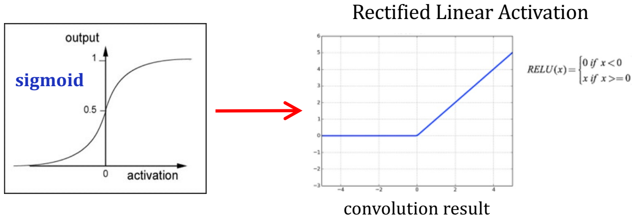




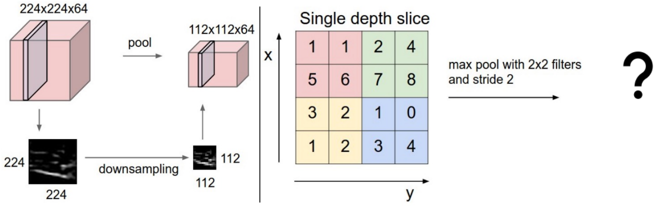
- early layers perform a *convolution* of their inputs
- multiple convolution operators (e.g. red & black)
- weights in convolution operators are *learned*
- convolution operators are typically small (e.g. 5x5)



ReLU & max pooling layers



$$RELU(x)=\begin{cases}0 \text{ if } x<0\\ x \text{ if } x>=0\end{cases}$$



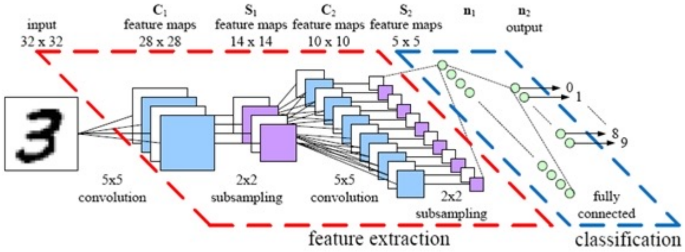
Single depth slice

1	1	2	4
5	6	7	8
3	2	1	0
1	2	3	4

max pool with 2x2 filters and stride 2

Adding a fully-connected neural net layer

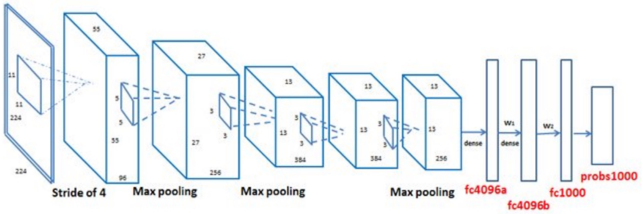
Recognizing digits from the MNIST database with a CNN:



LeNet

LeCun, Bottou, Bengio, Haffner (1998)

AlexNet, ZF Net, GoogLeNet, VGGNet, ResNet, …

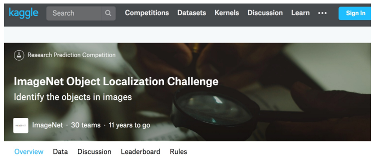


AlexNet: Krizhevsky, Sutskever, Hinton (2012)

ImageNet Large Scale Visual Recognition Challenge (ILSVRC)

IMAGENET

Annually since 2010



Research Prediction Competition

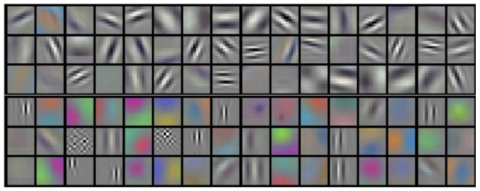
ImageNet Object Localization Challenge

Identify the objects in images

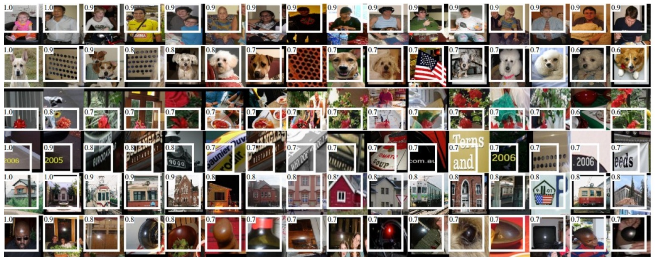
ImageNet · 30 teams · 11 years to go

Overview Data Discussion Leaderboard Rules

9



Filters generated in first convolution layer of AlexNet



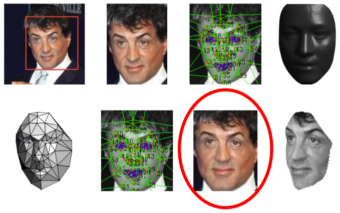
Maximally activating images from some POOL5 neurons of AlexNet (Girshick et al., 2014)

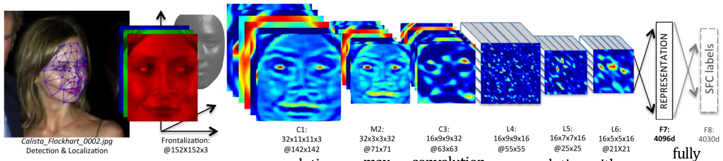
10

Facebook's DeepFace system

Taigman et al., 2014

- detect face
- 2D align face in crop window using 6 fiducial points
- align to 3D shape model using 67 fiducial points
- use 3D model + image to generate frontal view





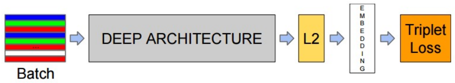
convolution | max pooling | convolution | convolution with locally varying weights | fully connected network

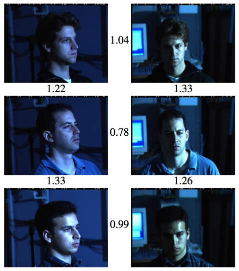
11

Google's FaceNet system

Schroff et al., 2015

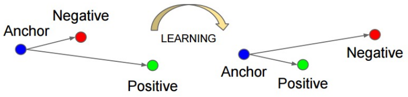
FaceNet also uses a deep convolutional network





threshold = 1.1 classifies pairs correctly (smaller value means more similar)

- learns mapping from images to a space where distance between images captures similarity
- training data: triplets of face thumbnails
  - two same ID, one different ID
- learning process: minimize distance between anchor & positive images (same ID), maximize distance between anchor & negative images



12