Deep Networks

Pradeep Ravikumar

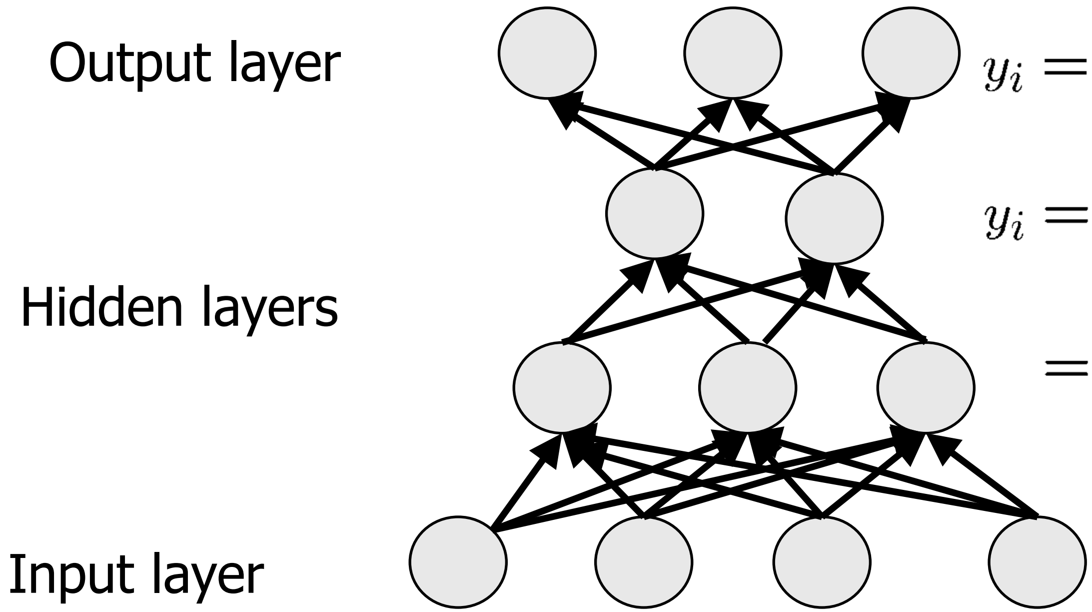
Co-instructor: Ziv Bar-Joseph

Machine Learning 10-701

Slides Courtesy: Barnabas Poczos, Ruslan Salakhutdinov, Yoshua Bengio, Geoffrey Hinton, Yann LeCun

Deep architectures

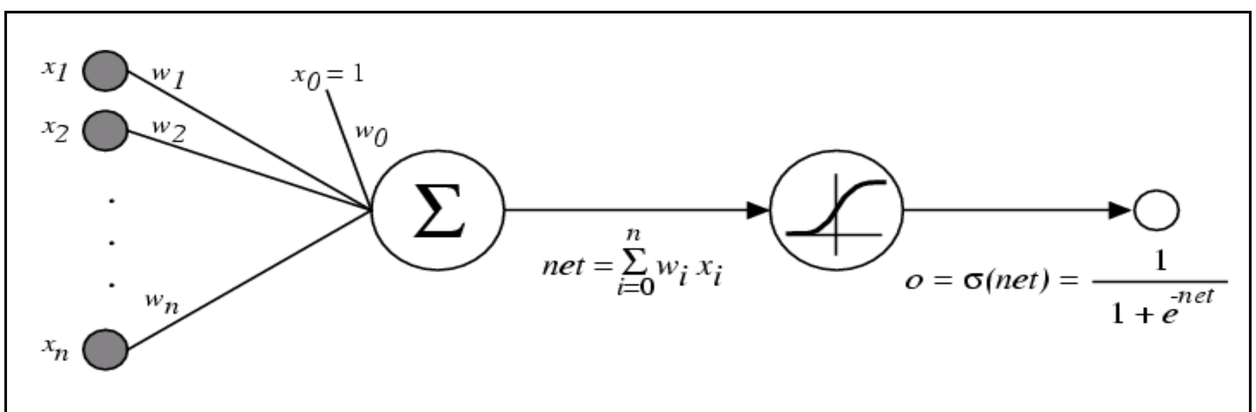
Definition: Deep architectures are composed of *multiple levels* of non-linear operations, such as neural nets with many hidden layers.



$$y_i = \tanh\left(\sum_j W_{ij} x_j + W_{i0}\right)$$

$$y_i = \sigma\left(\sum_j W_{ij} x_j + W_{i0}\right)$$

$$= \frac{1}{1 + \exp[-(\sum_j W_{ij} x_j + W_{i0})]}$$



Goal of Deep architectures

Goal: Deep learning methods aim at

- learning *feature hierarchies*
- where features from higher levels of the hierarchy are formed by lower level features.

edges, local shapes, object parts

Low level representation

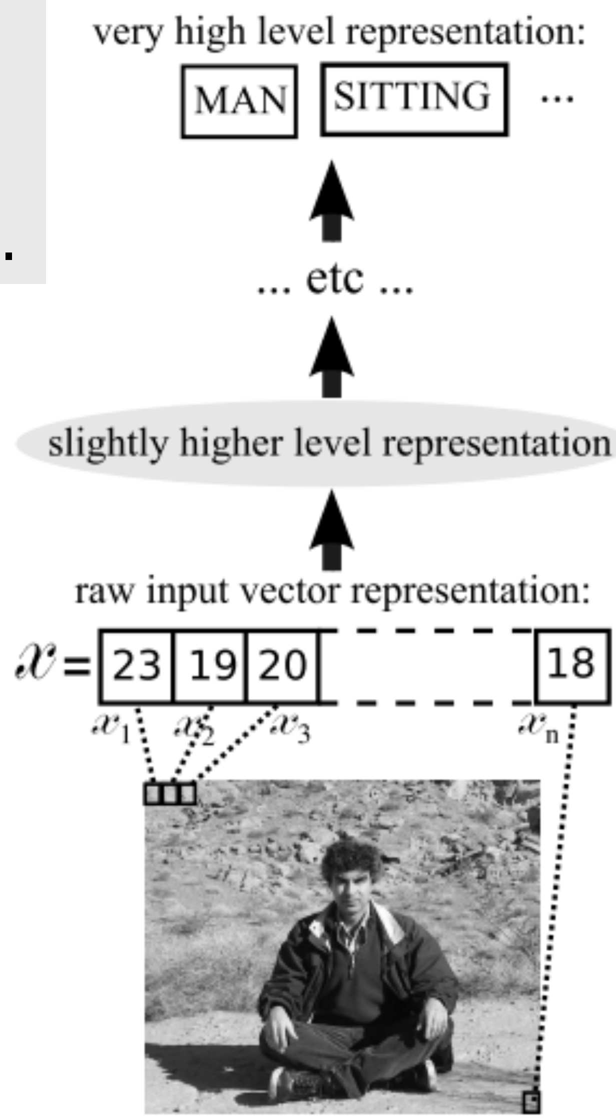
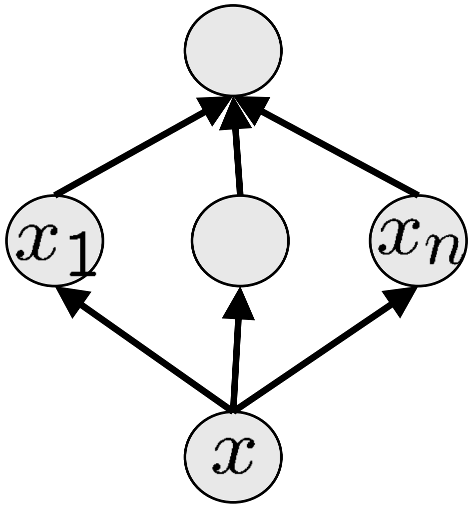


Figure is from Yoshua Bengio

Neurobiological Motivation

- Most current learning algorithms are shallow architectures (1-3 levels)
  (SVM, kNN, MoG, KDE, Parzen Kernel regression, PCA, Perceptron,…)

$$\text{SVM: } \hat{f}(\mathbf{x}) = \text{sign}\left(\sum_{i=1}^{n} \alpha_i y_i k(\mathbf{x}_i, \mathbf{x})\right)$$



- The mammal brain is organized in a deep architecture (Serre, Kreiman, Kouh, Cadieu, Knoblich, & Poggio, 2007)
  (E.g. visual system has 5 to 10 levels)

Breakthrough

Deep Belief Networks (DBN)

Hinton, G. E, Osindero, S., and Teh, Y. W. (2006).
A fast learning algorithm for deep belief nets.
Neural Computation, 18:1527-1554.

Autoencoders

Bengio, Y., Lamblin, P., Popovici, P., Larochelle, H. (2007).
Greedy Layer-Wise Training of Deep Networks,
Advances in Neural Information Processing Systems 19

Convolutional neural networks running on GPUs (2012)

Alex Krizhevsky, Ilya Sutskever, Geoffrey Hinton, Advances in Neural Information Processing Systems 2012

Deep Convolutional Networks

Deep Convolutional Networks

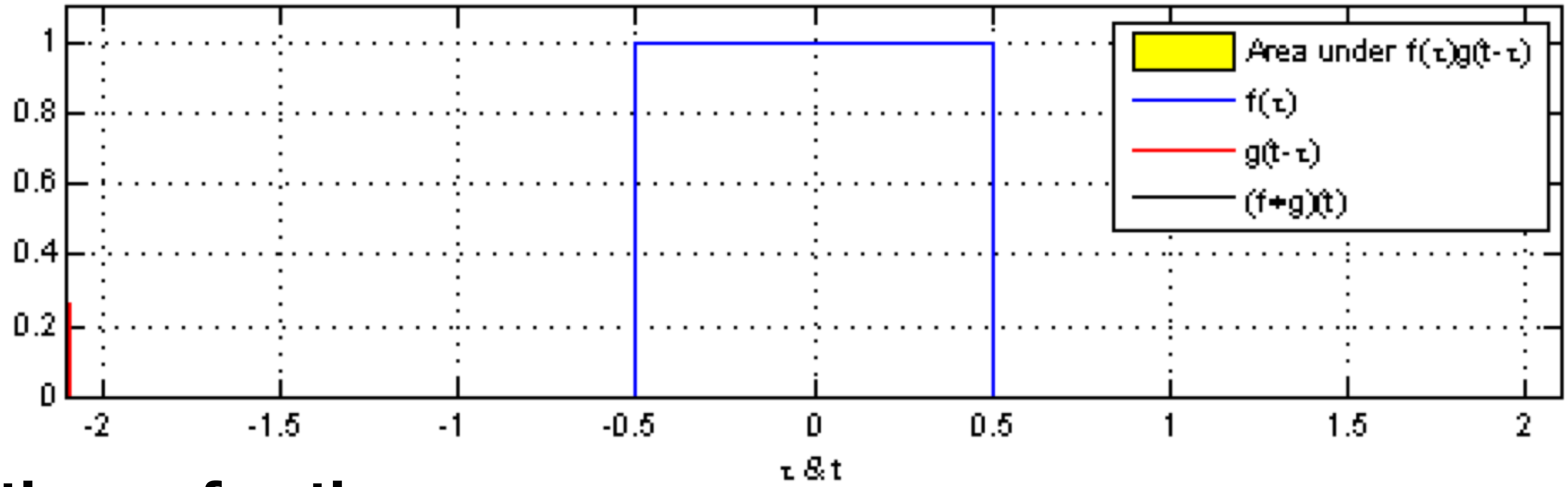
Compared to standard feedforward neural networks with similarly-sized layers,

- CNNs have much fewer connections and parameters
- and so they are easier to train,
- while their theoretically-best performance is likely to be only slightly worse.

LeNet 5

Y. LeCun, L. Bottou, Y. Bengio and P. Haffner: **Gradient-Based Learning Applied to Document Recognition**, *Proceedings of the IEEE, 86(11):2278-2324, November **1998***

Convolution



Continuous functions:

$$(f * g)(t) = \int_{-\infty}^{\infty} f(\tau)\, g(t-\tau)\, d\tau - = \int_{-\infty}^{\infty} f(t-\tau)\, g(\tau)\, d\tau.$$

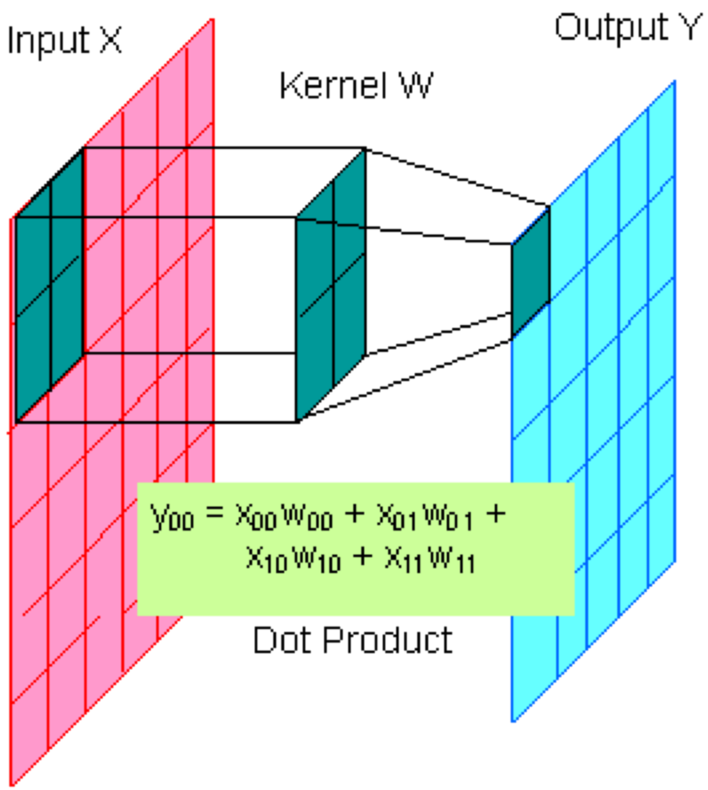
Discrete functions:

$$(f * g)[n] = \sum_{m=-\infty}^{\infty} f[m]\, g[n-m] = \sum_{m=-\infty}^{\infty} f[n-m]\, g[m]$$

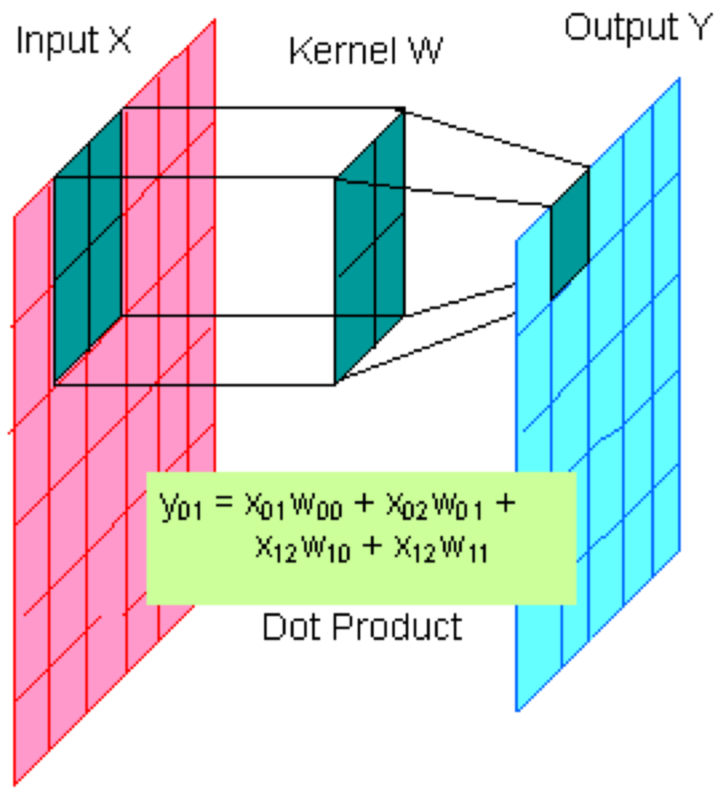
If discrete g has support on {-M,…,M} :

$$(f * g)[n] = \sum_{m=-M}^{M} f[n-m] g[m]$$

2-Dimensional Convolution



2-Dimensional Convolution



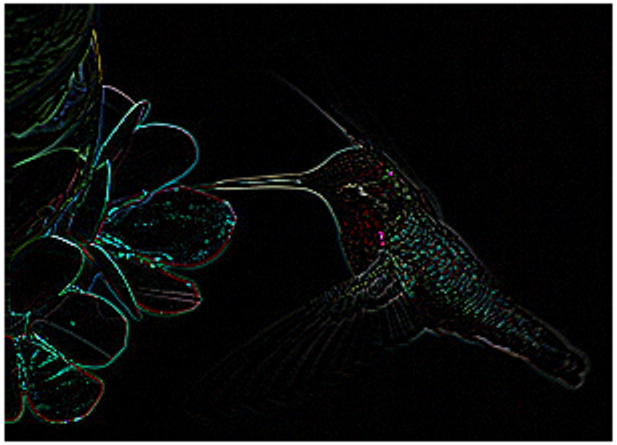
2-Dimensional Convolution

$$f[x,y] * g[x,y] = \sum_{n_1=-\infty}^{\infty} \sum_{n_2=-\infty}^{\infty} f[n_1,n_2] \cdot g[x-n_1, y-n_2]$$

https://graphics.stanford.edu/courses/cs178/applets/convolution.html

Filter (=kernel)

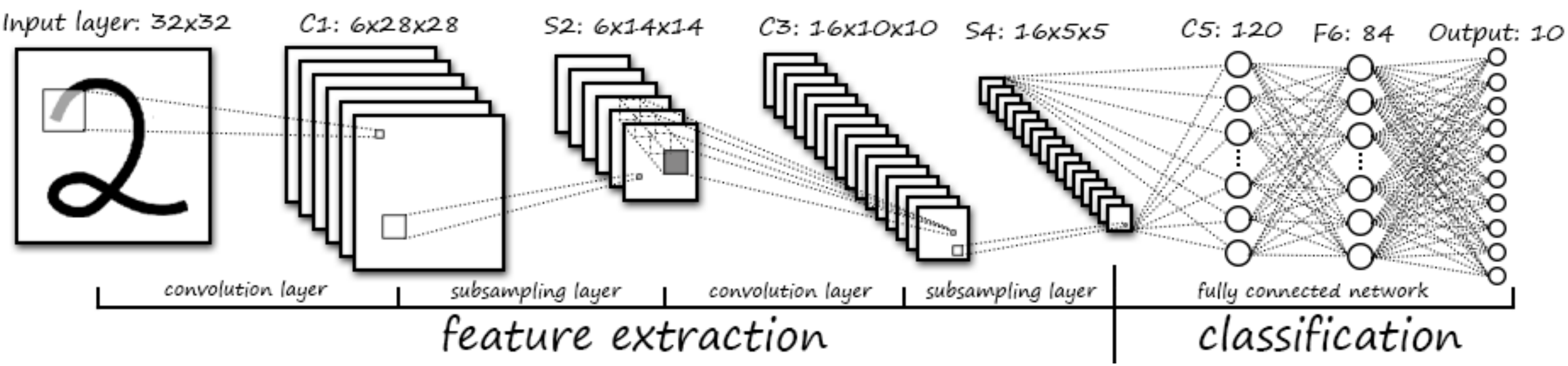
0.00	0.00	0.00	0.00	0.00
0.00	0.00	-2.00	0.00	0.00
0.00	-2.00	8.00	-2.00	0.00
0.00	0.00	-2.00	0.00	0.00
0.00	0.00	0.00	0.00	0.00



Original

0.04	0.04	0.04	0.04	0.04
0.04	0.04	0.04	0.04	0.04
0.04	0.04	0.04	0.04	0.04
0.04	0.04	0.04	0.04	0.04
0.04	0.04	0.04	0.04	0.04

LeNet 5, LeCun 1998

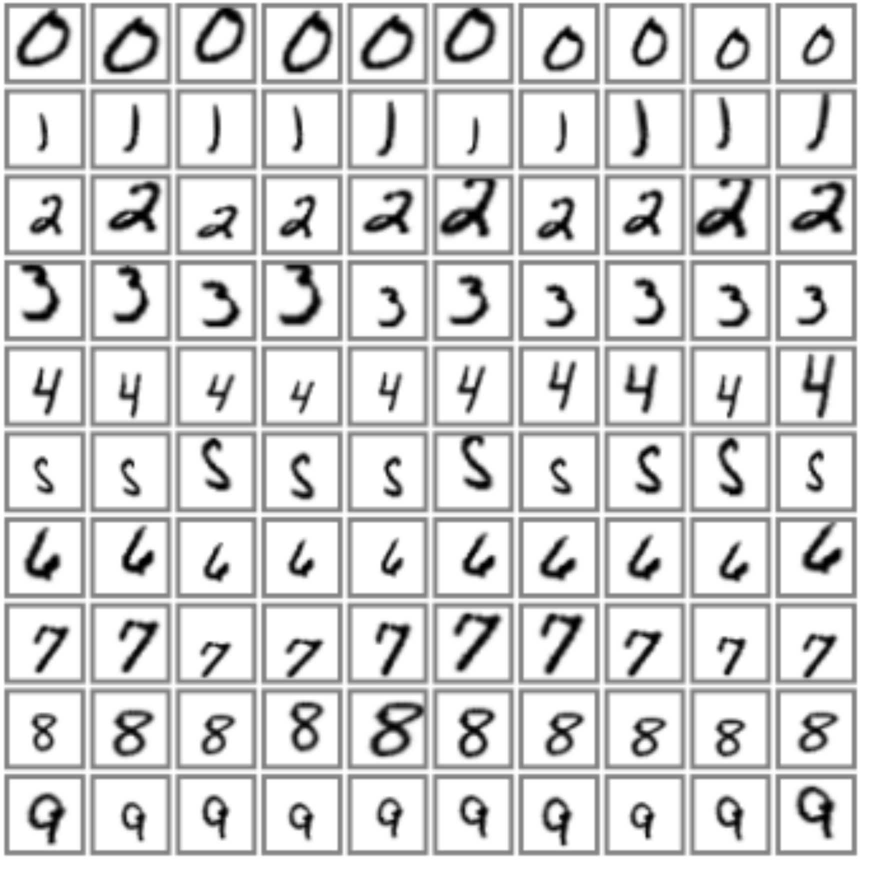


- **Input:** 32x32 pixel image. Largest character is 20x20
  (All important info should be in the center of the receptive fields of the highest level feature detectors)
- **Cx:** Convolutional layer (C1, C3, C5) tanh nonlinear units
- **Sx:** Subsample layer (S2, S4)
- **Fx:** Fully connected layer (F6) logistic/sigmoid units
- Black and White pixel values are normalized:
  E.g. White = -0.1, Black =1.175 (Mean of pixels = 0, Std of pixels =1)

MNIST Dataset

60,000 original dataset

Test error: 0.95%



540,000 artificial distortions

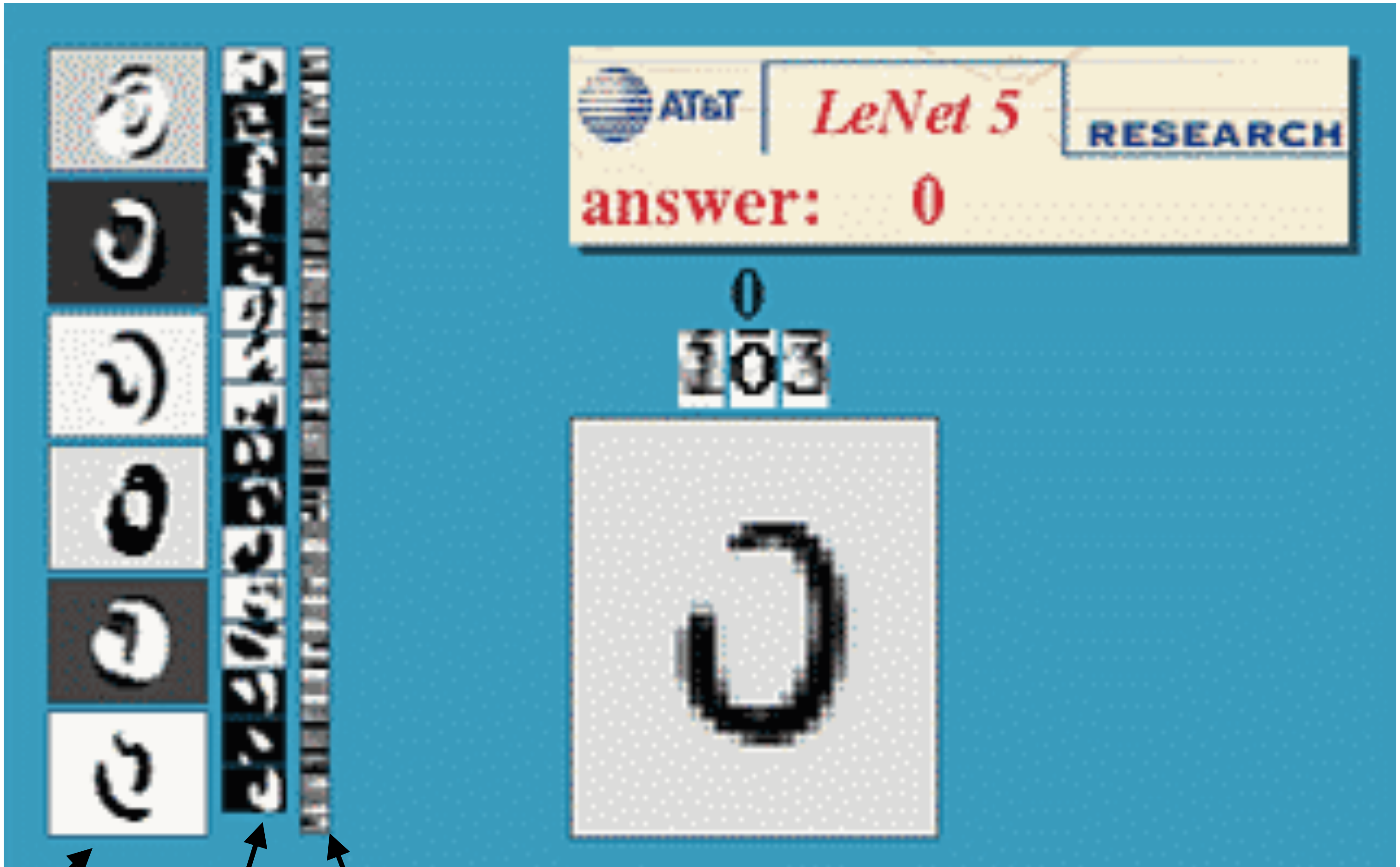
+ 60,000 original

Test error: 0.8%

Misclassified examples

True label -> Predicted label

LeNet 5 in Action

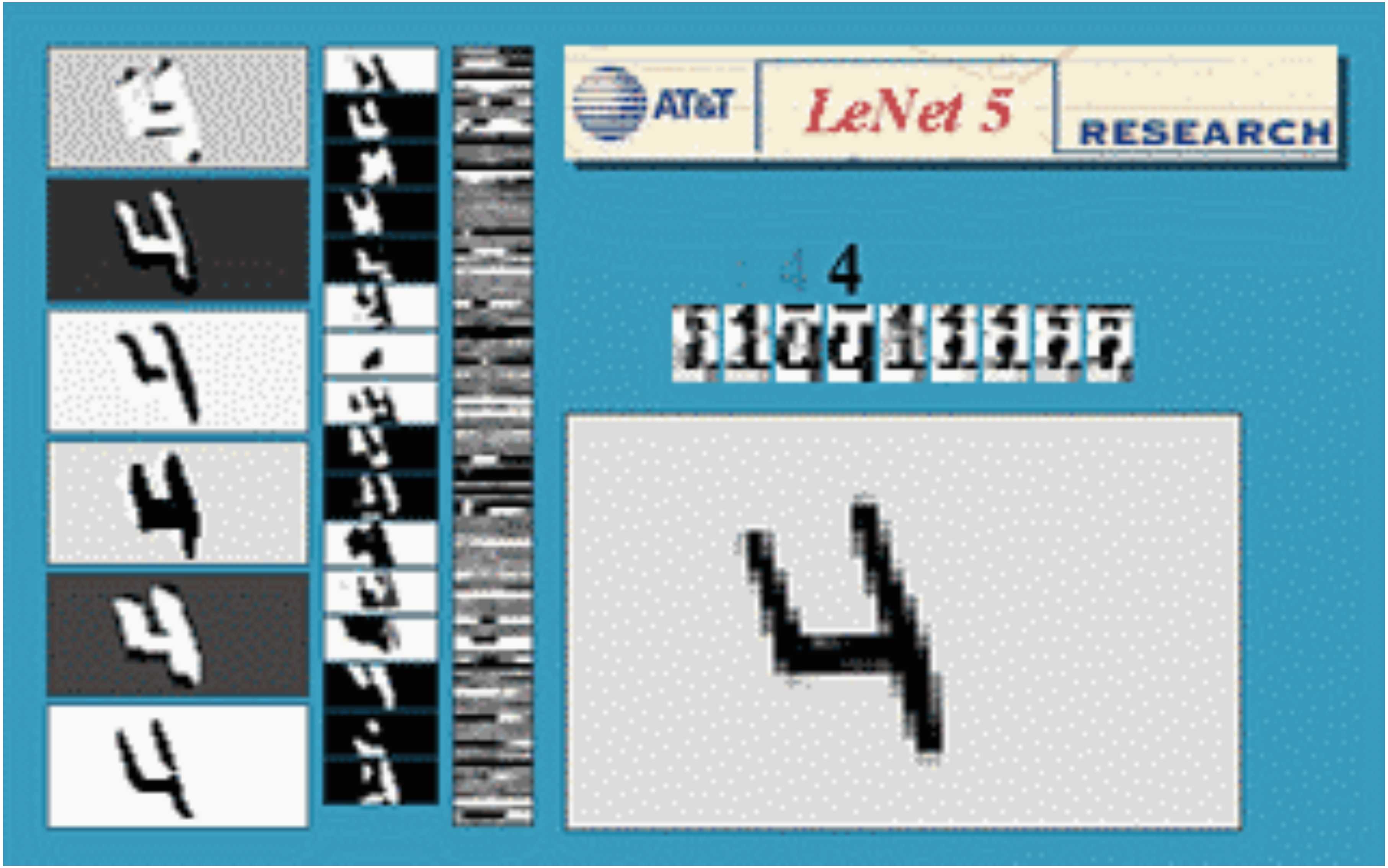


C1 C3 S4

Input

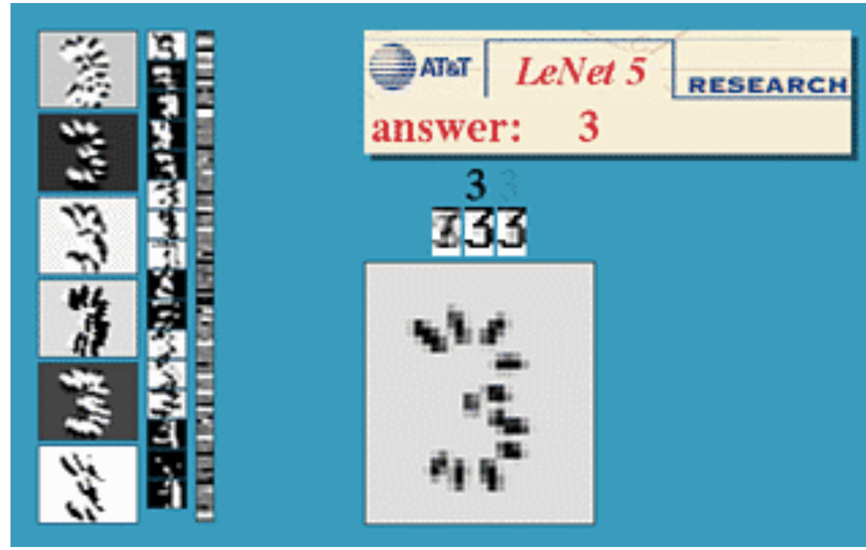
LeNet 5, Shift invariance

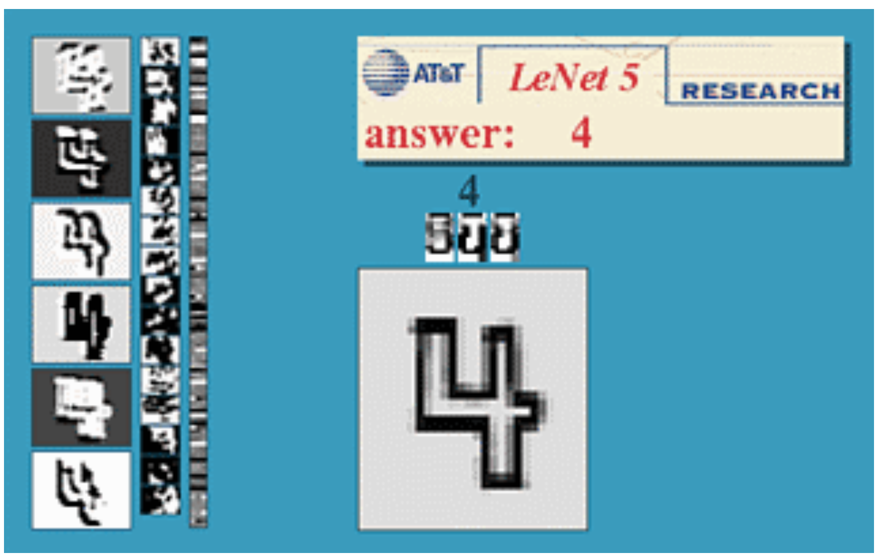
LeNet 5, Rotation invariance

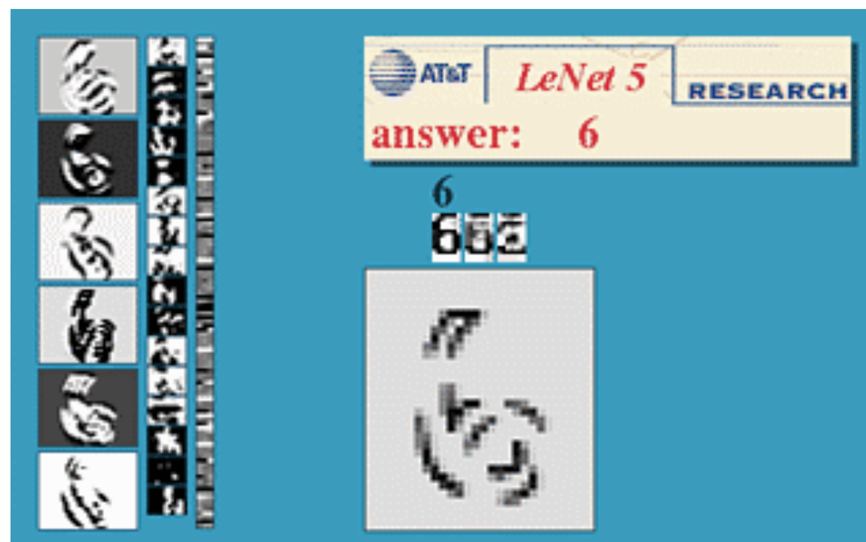


LeNet 5, Nosie resistance

LeNet 5, Unusual Patterns







ImageNet Classification with Deep Convolutional Neural Networks

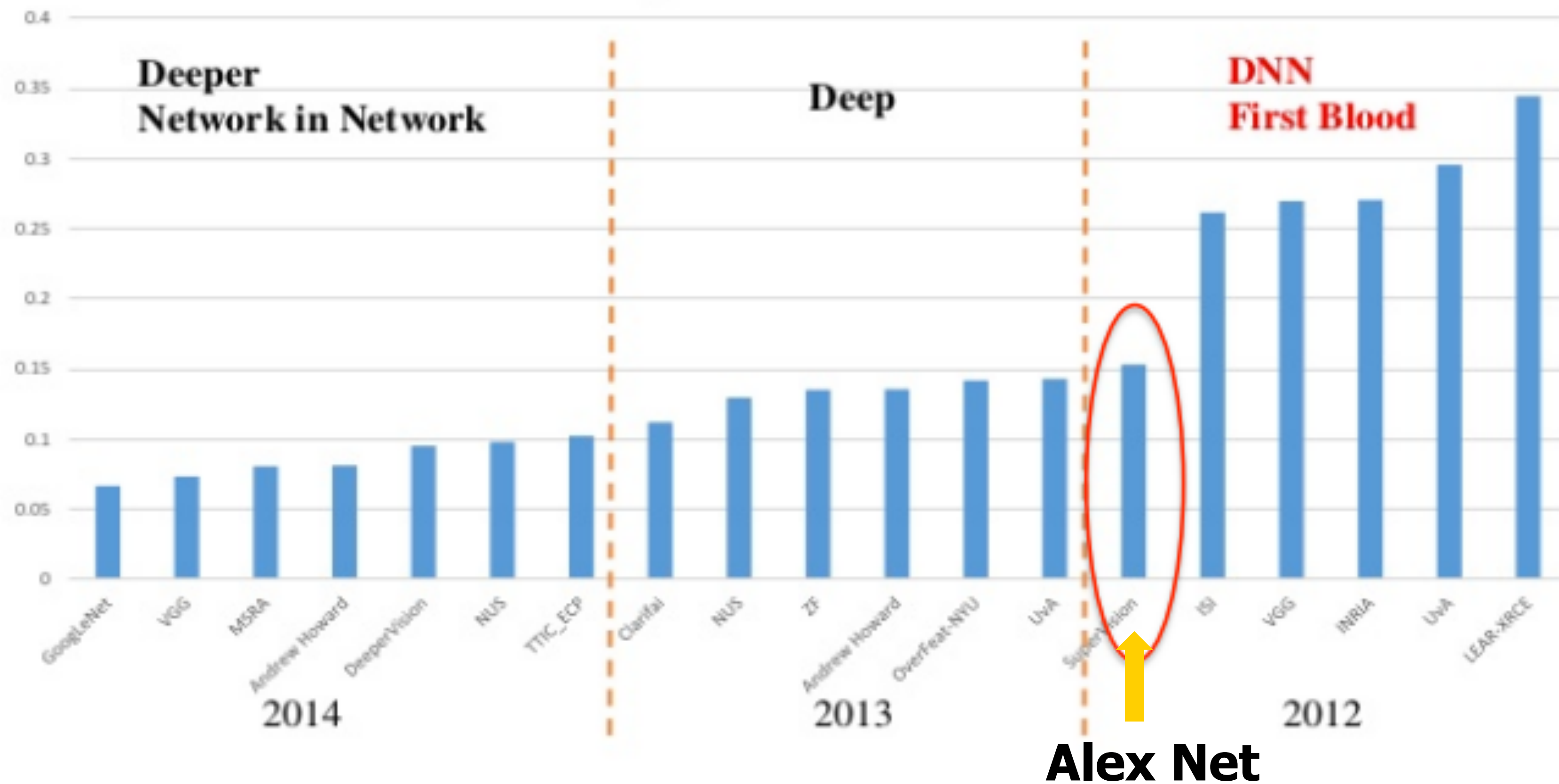
Alex Krizhevsky, Ilya Sutskever, Geoffrey Hinton,

Advances in Neural Information Processing Systems 2012

Alex Net

ILSVRC

ImageNet Classification error throughout years and groups



Li Fei-Fei: ImageNet Large Scale Visual Recognition Challenge, 2014 http://image-net.org/

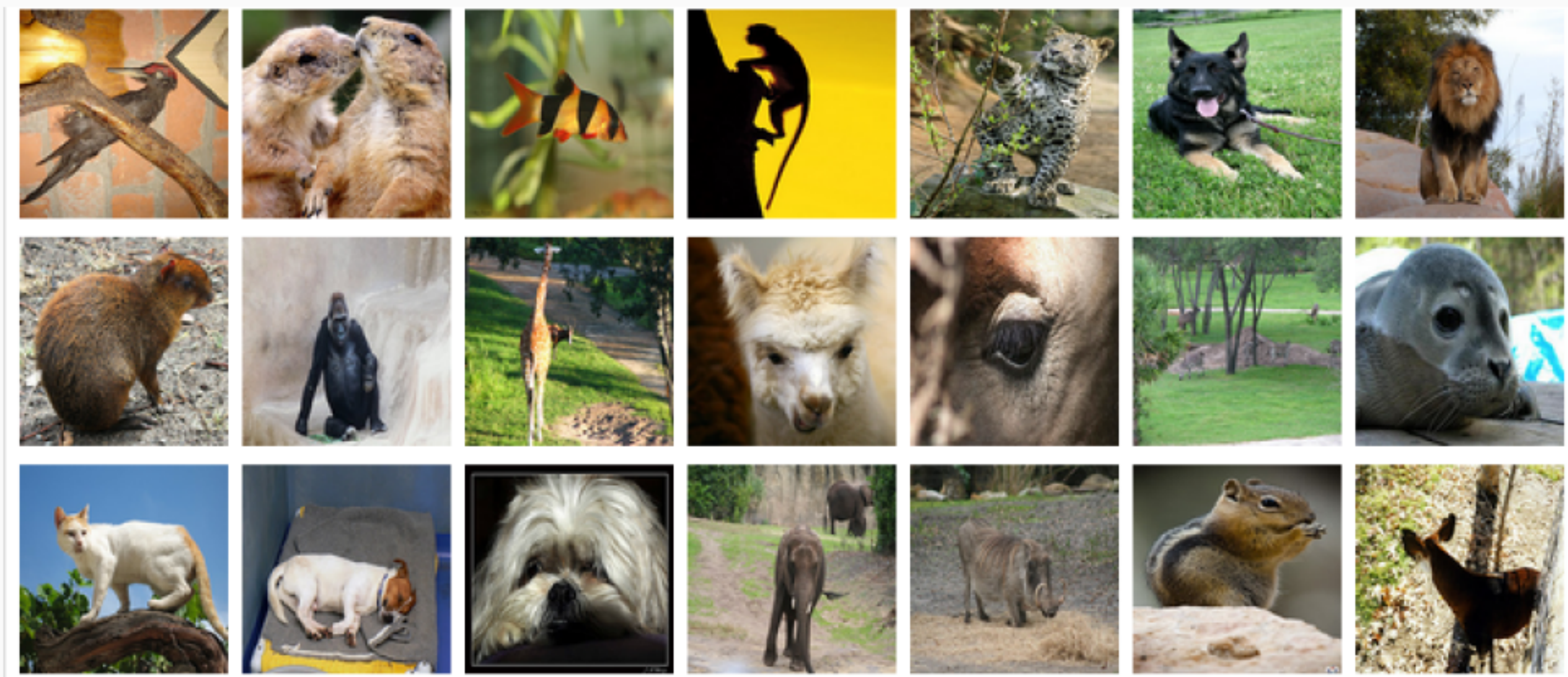
ImageNet

- 15M images
- 22K categories
- Images collected from Web
- Human labelers (Amazon's Mechanical Turk crowd-sourcing)
- ImageNet Large Scale Visual Recognition Challenge (ILSVRC-2010)
    - 1K categories
    - 1.2M training images (~1000 per category)
    - 50,000 validation images
    - 150,000 testing images
- RGB images
- Variable-resolution, but this architecture scales them to 256x256 size

ImageNet

Classification goals:

- ❑ Make 1 guess about the label (Top-1 error)
- ❑ make 5 guesses about the label (Top-5 error)



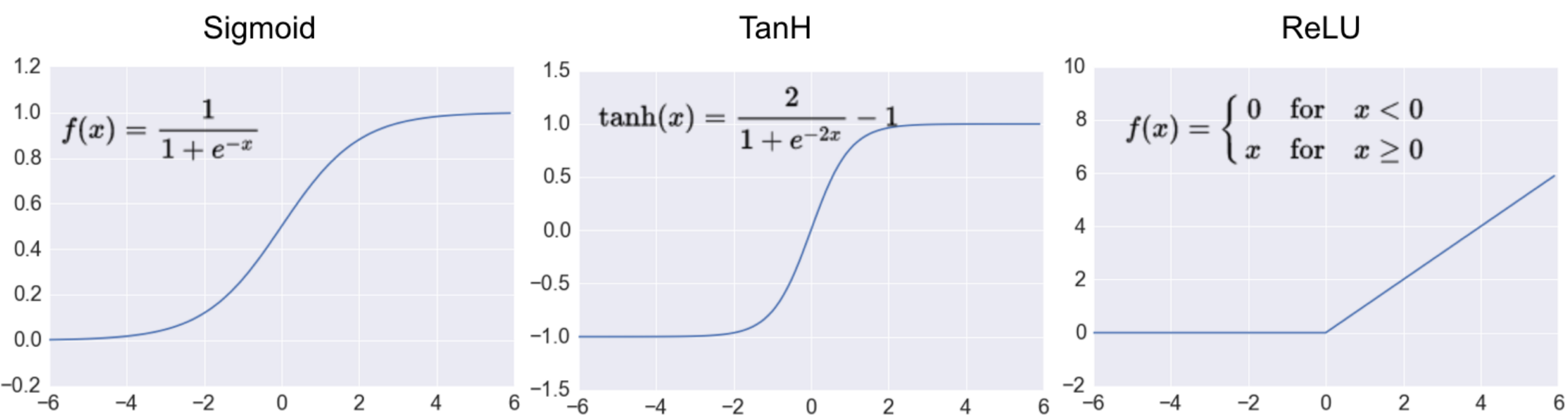
The Architecture

Typical nonlinearities: $f(x) = \tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$

$f(x) = (1 + e^{-x})^{-1}$ (logistic function)

Here, however, Rectified Linear Units (ReLU) are used: $f(x) = \max(0, x)$

Non-saturating/Gradients don't vanish – faster training



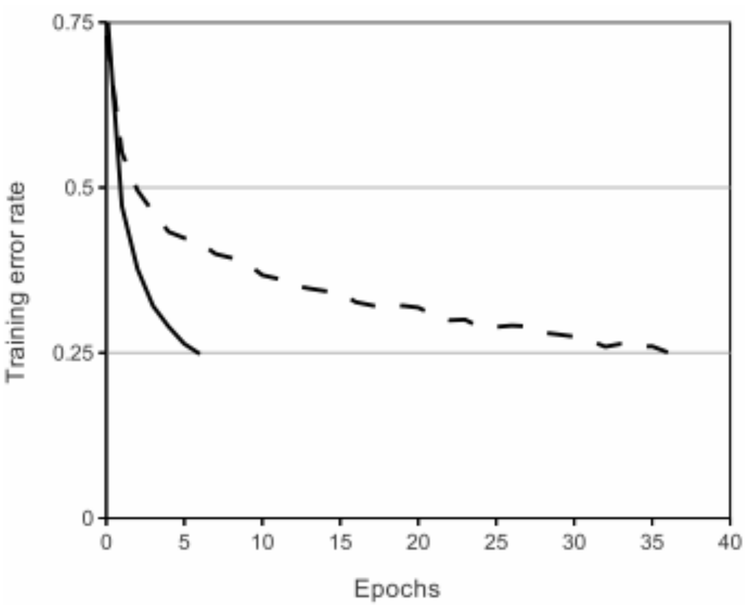
The Architecture

Typical nonlinearities: $f(x) = \tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$

$f(x) = (1 + e^{-x})^{-1}$ (logistic function)

Here, however, Rectified Linear Units (ReLU) are used: $f(x) = \max(0, x)$

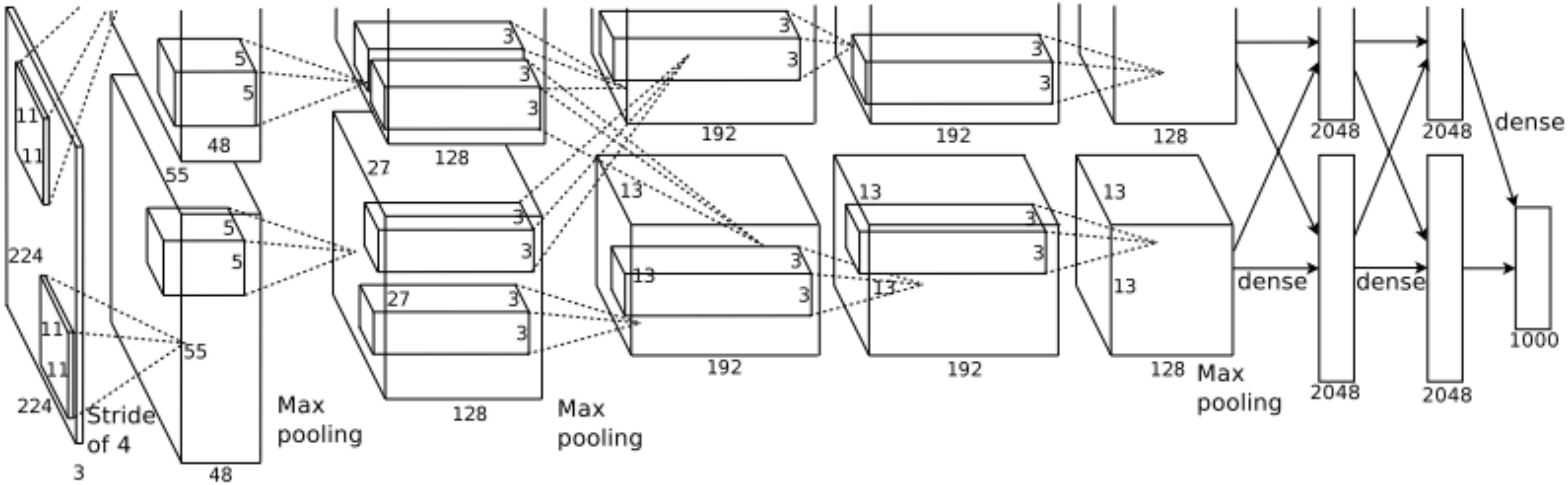
Non-saturating/Gradients don't vanish – faster training



A four-layer convolutional neural network with ReLUs (solid line) reaches a 25% training error rate on CIFAR-10 six times faster than an equivalent network with tanh neurons

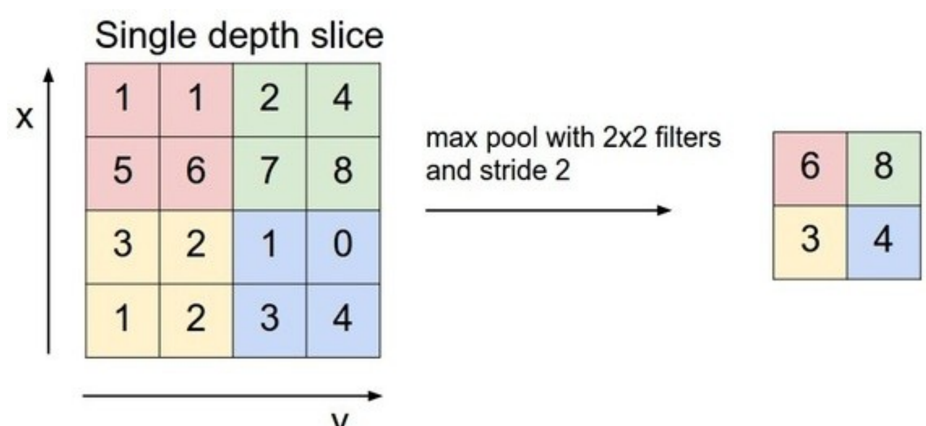
(dashed line)

The Architecture

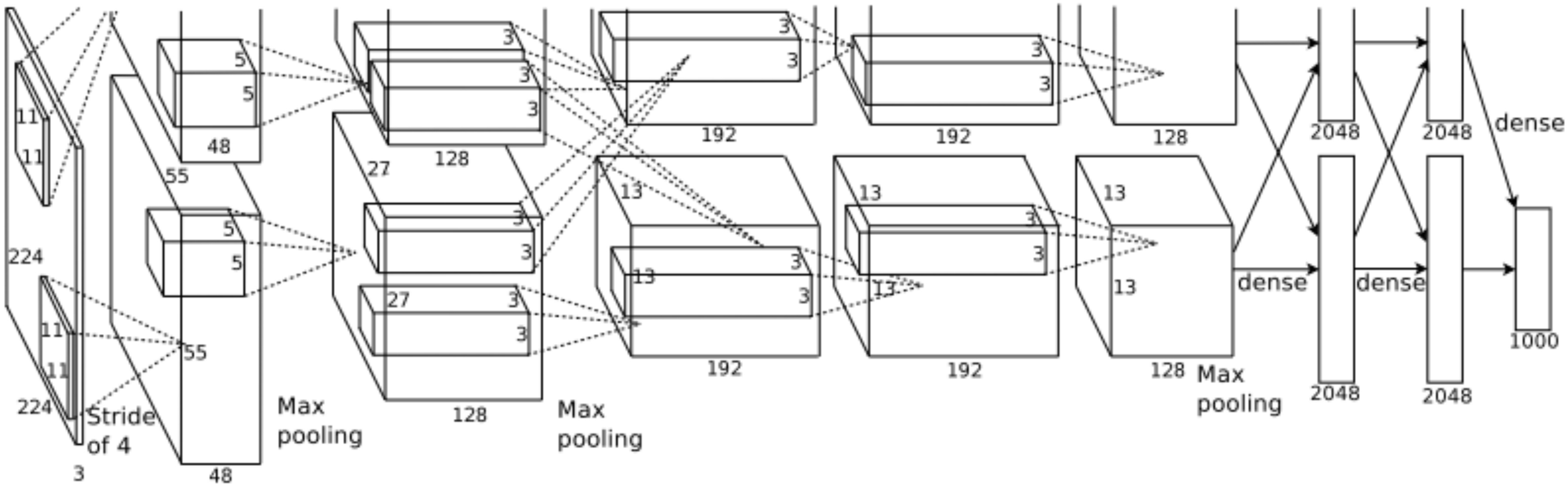


5 convolution layers (ReLU)

3 overlapping max pooling – nonlinear downsampling (max value of regions)



The Architecture



5 convolution layers (ReLU)

3 overlapping max pooling – nonlinear downsampling (max value of regions)

2 fully connected layers

output softmax

The Architecture

- Trained with stochastic gradient descent
- on two NVIDIA GTX 580 3GB GPUs
- for about a week

- 650,000 neurons
- 60,000,000 parameters
- 630,000,000 connections
- 5 convolutional layer with Rectified Linear Units (ReLUs), 3 overlapping max pooling, 2 fully connected layer
- Final feature layer: 4096-dimensional

- Prevent overfitting – data augmentation, dropout trick
- Randomly extracted 224x224 patches for more data

Preventing overfitting

1) The easiest and most common method to **reduce overfitting** on image data is to artificially **enlarge the dataset** using label-preserving transformations.

We employ two distinct forms of **data augmentation**:

- image translation
- horizontal reflections
- changing RGB intensities

2) **Dropout**: set the output of each hidden neuron to zero w.p. 0.5.

- So every time an input is presented, the neural network samples a different architecture, but all these architectures share weights.
- This technique reduces complex co-adaptations of neurons, since a neuron cannot rely on the presence of particular other neurons.
- forced to learn more robust features that are useful in conjunction with many different random subsets of the other neurons.

Results

Results on the test data:

top-1 error rate: 37.5%

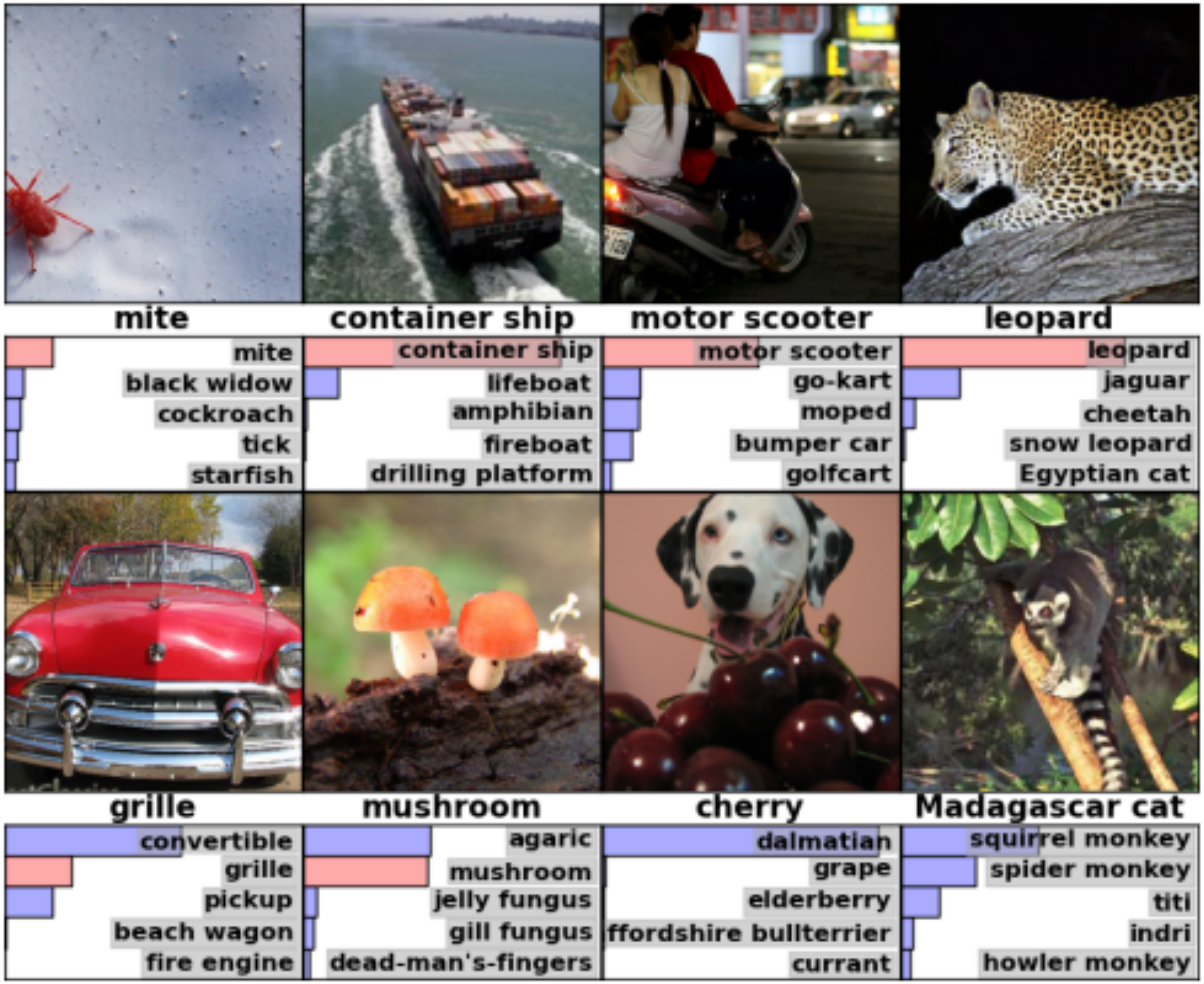
top-5 error rate: 17.0%

ILSVRC-2012 competition:

15.3% classification error

2nd best team: 26.2% classification error

Results


mite
mite
black widow
cockroach
tick
starfish
container ship
container ship
lifeboat
amphibian
fireboat
drilling platform
motor scooter
motor scooter
go-kart
moped
bumper car
golfcart
leopard
leopard
jaguar
cheetah
snow leopard
Egyptian cat
grille
convertible
grille
pickup
beach wagon
fire engine
mushroom
agaric
mushroom
jelly fungus
gill fungus
dead-man's-fingers
cherry
dalmatian
grape
elderberry
ffordshire bullterrier
currant
Madagascar cat
squirrel monkey
spider monkey
titi
indri
howler monkey

Results: Image similarity

Test column

six training images that produce feature vectors in the last hidden layer with the smallest Euclidean distance from the feature vector for the test image.