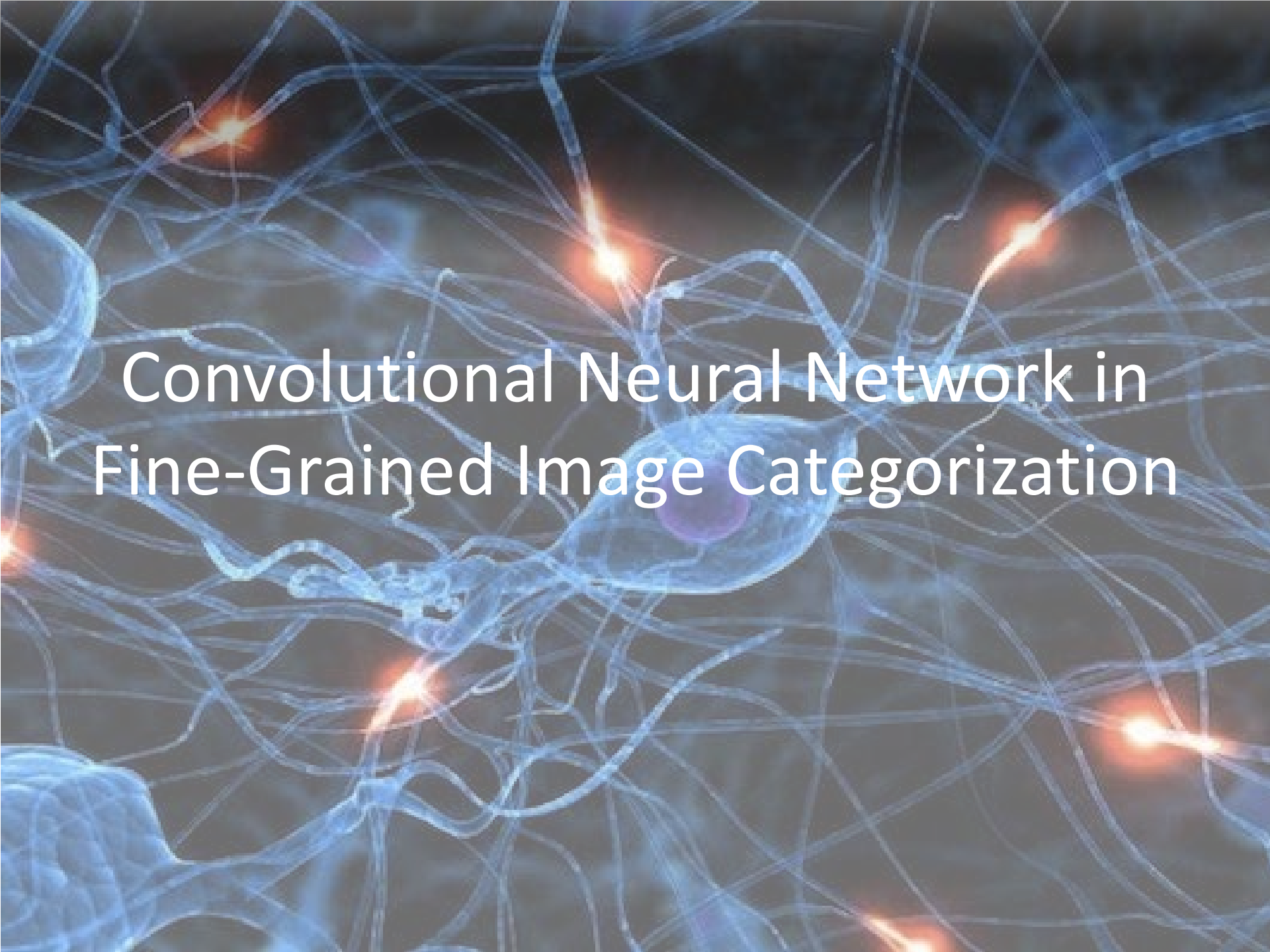
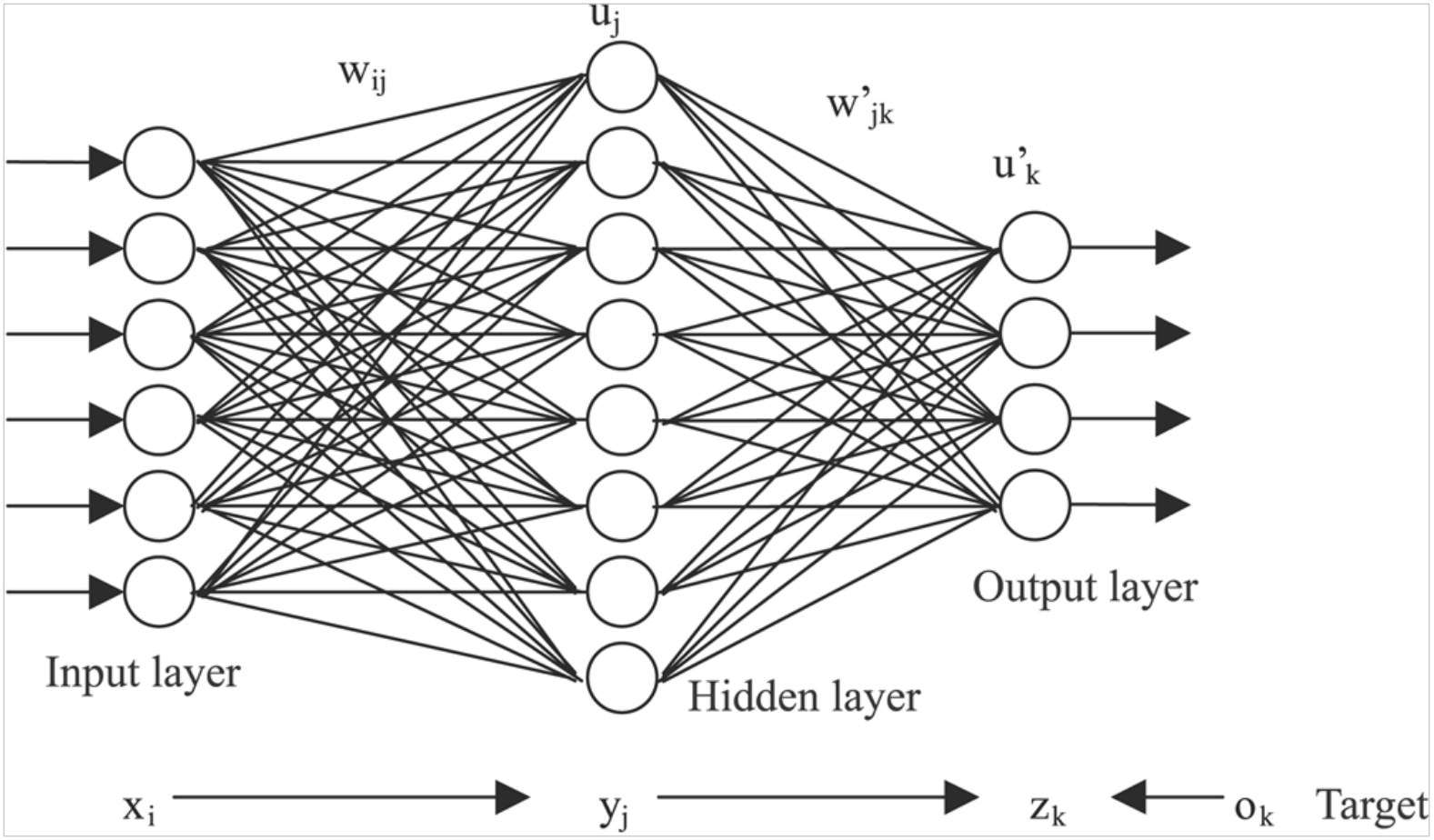
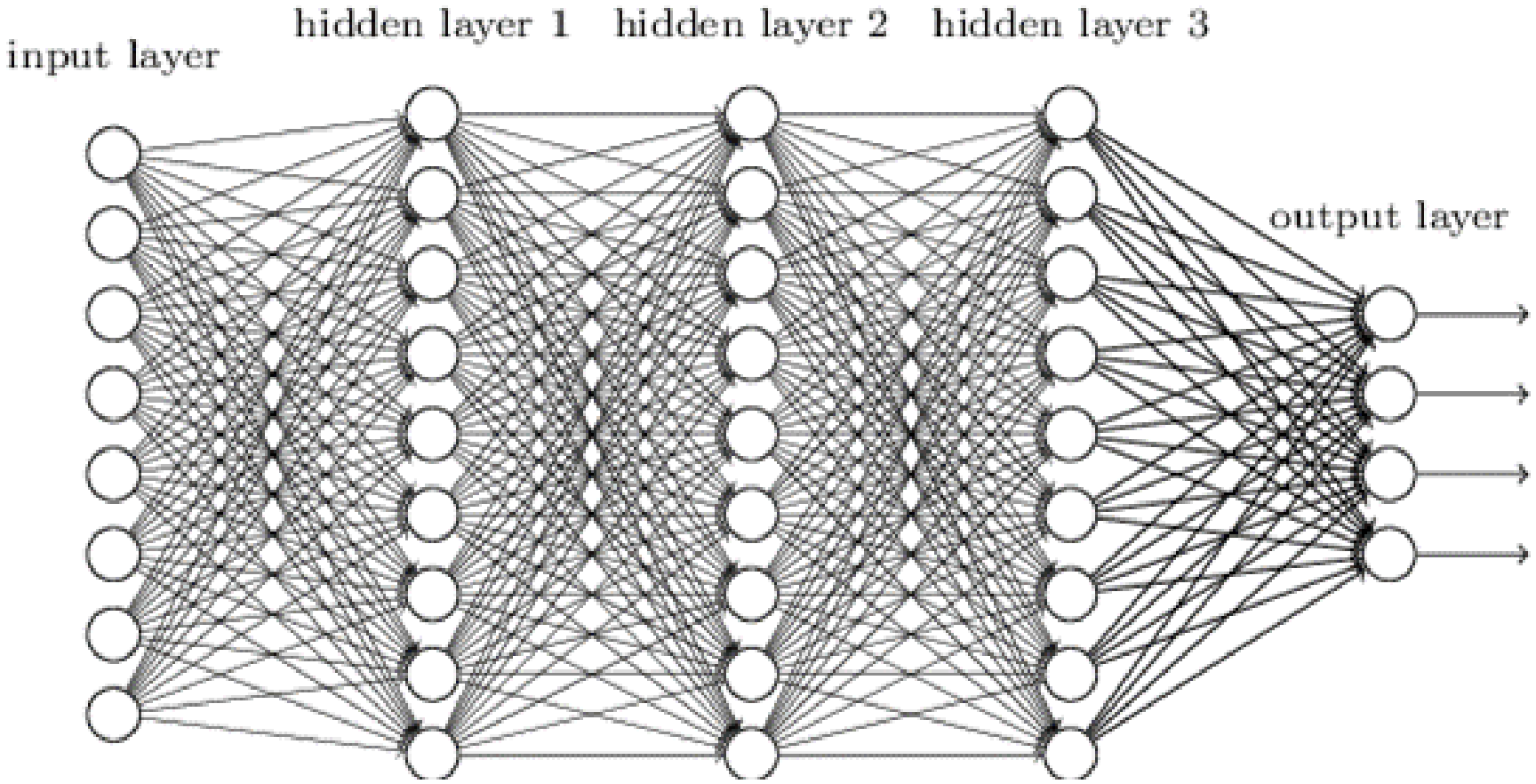

Convolutional Neural Network in
Fine-Grained Image Categorization

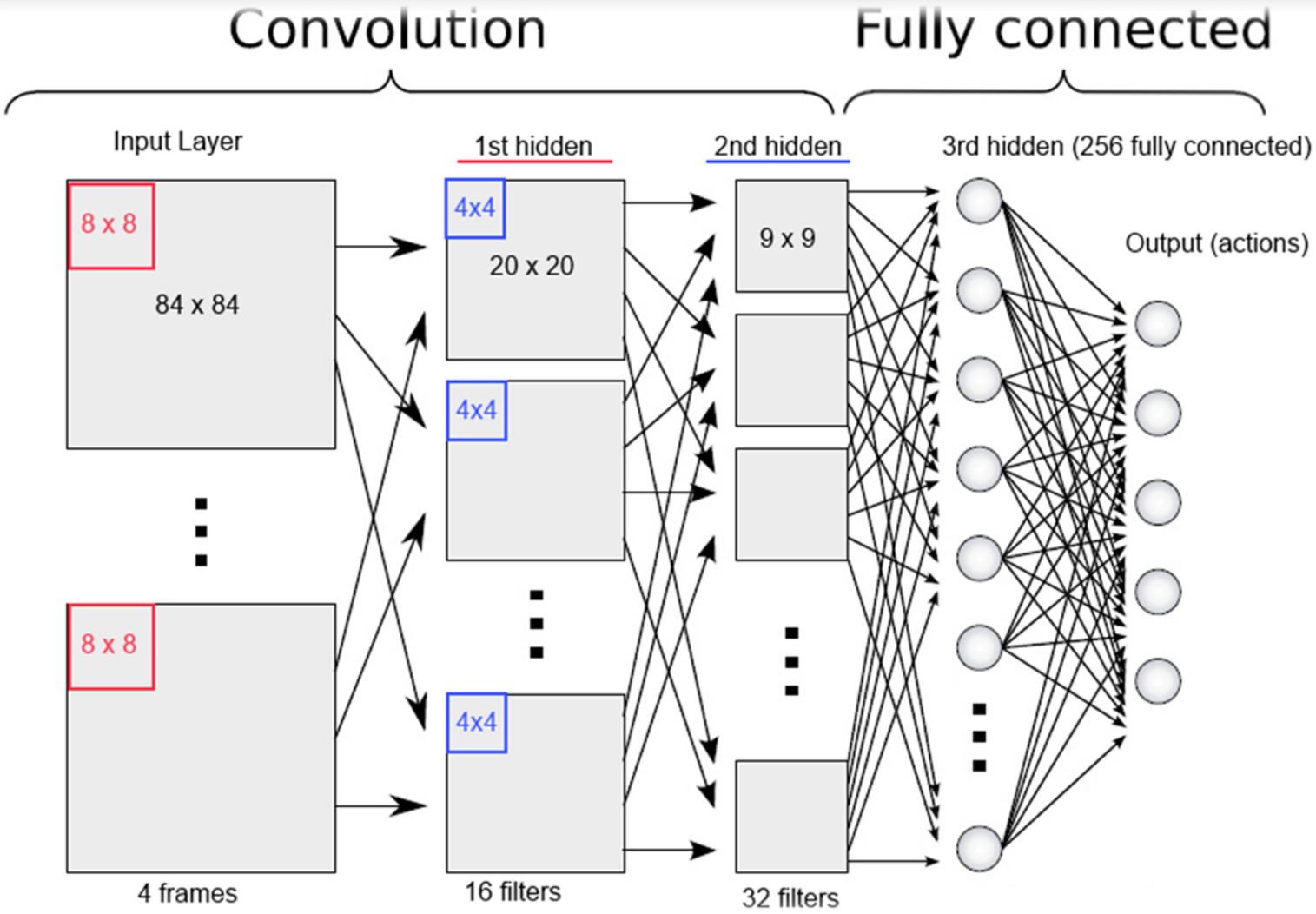
Neural Network



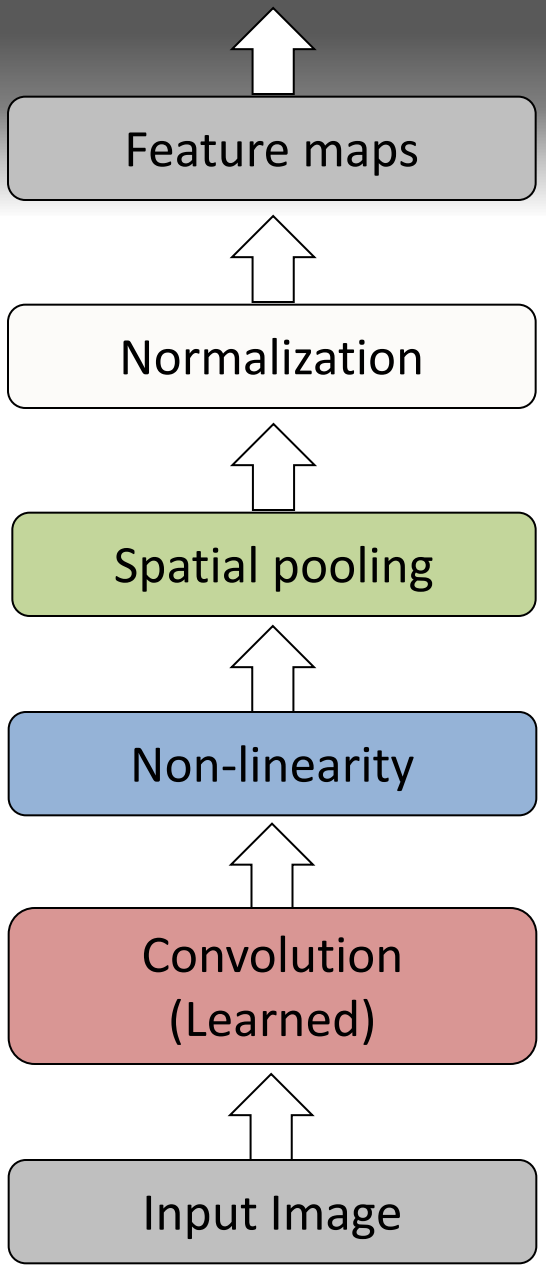
Deep Neural Network



Convolutional Neural Network

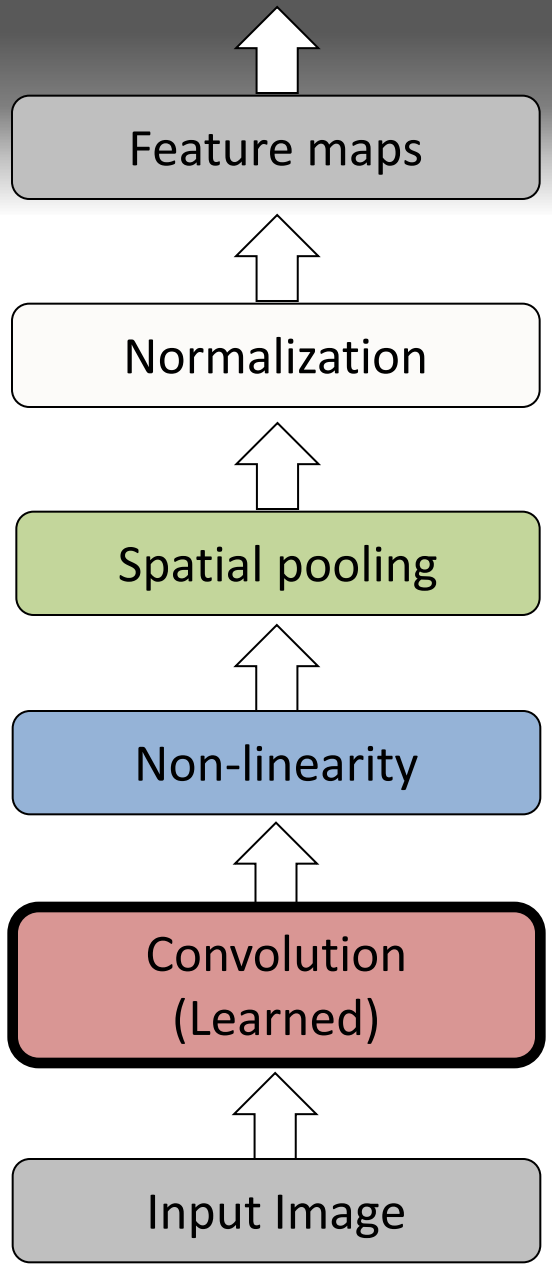


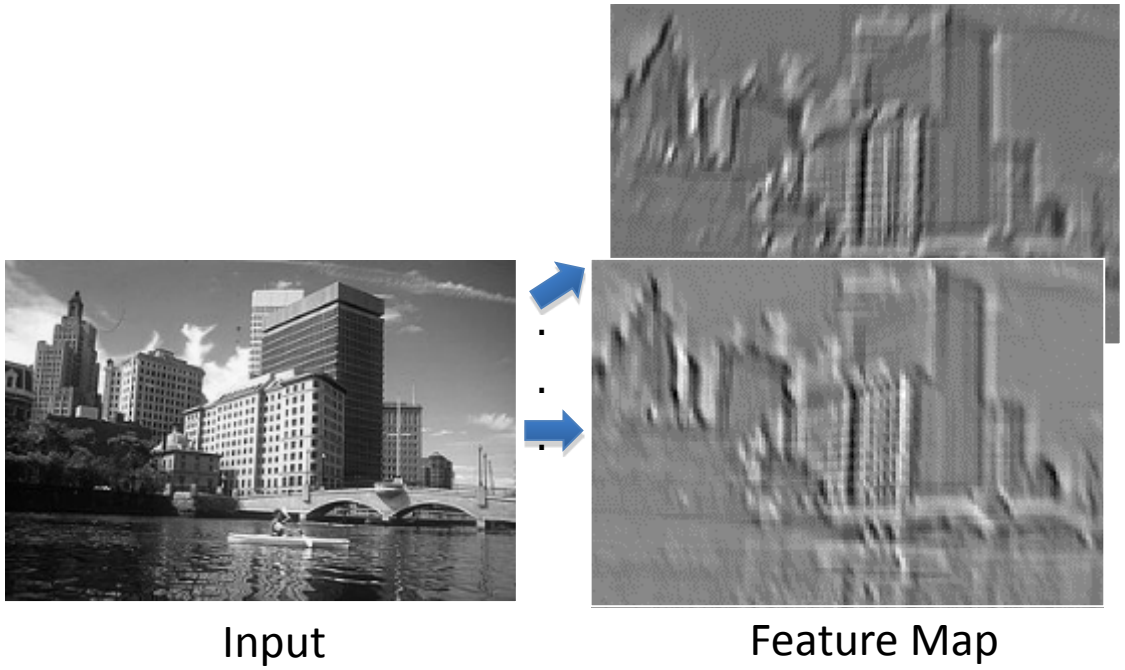
Convolutional Neural Networks



slide credit: S. Lazebnik

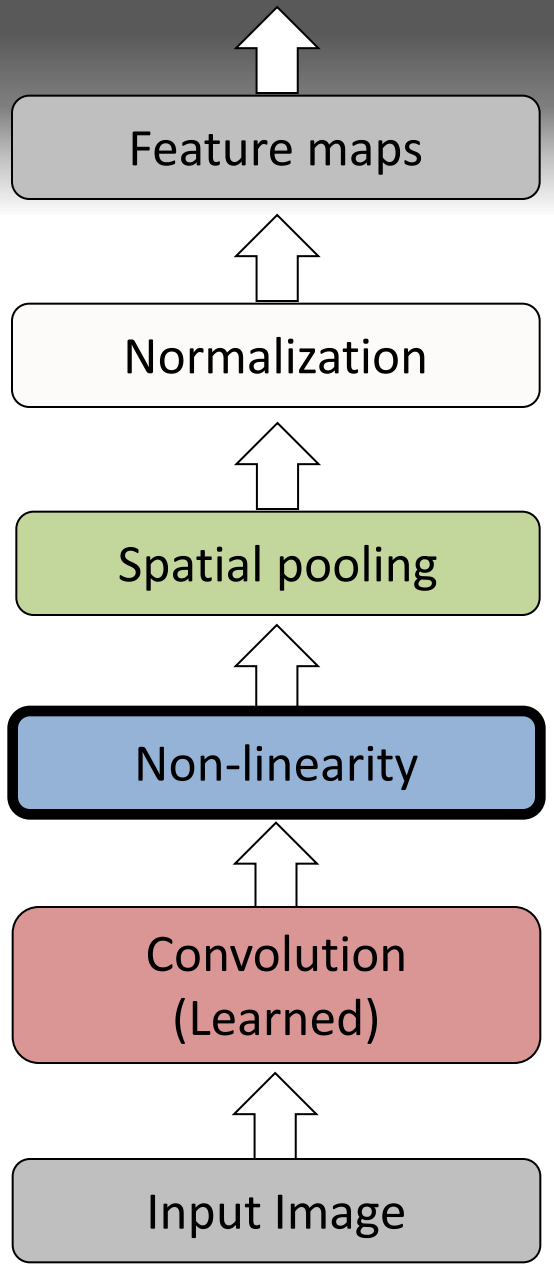
Convolutional Neural Networks



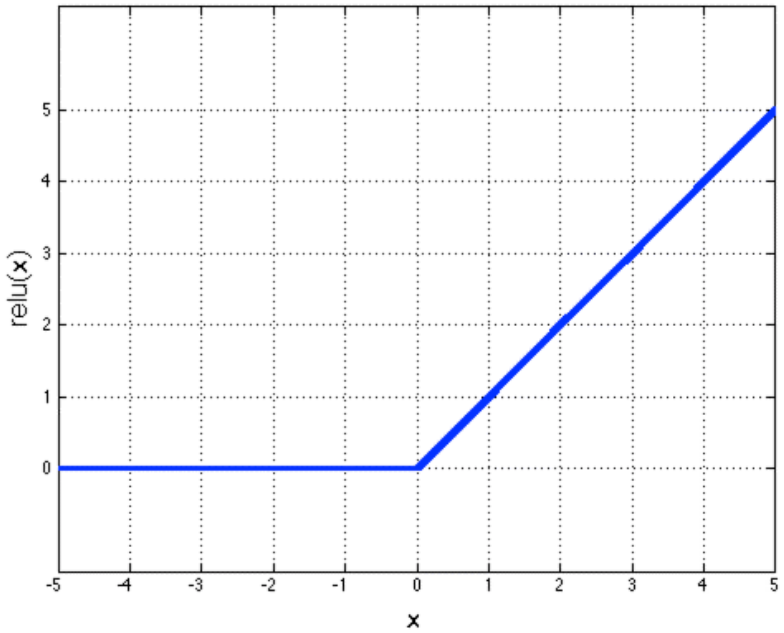


slide credit: S. Lazebnik

Convolutional Neural Networks

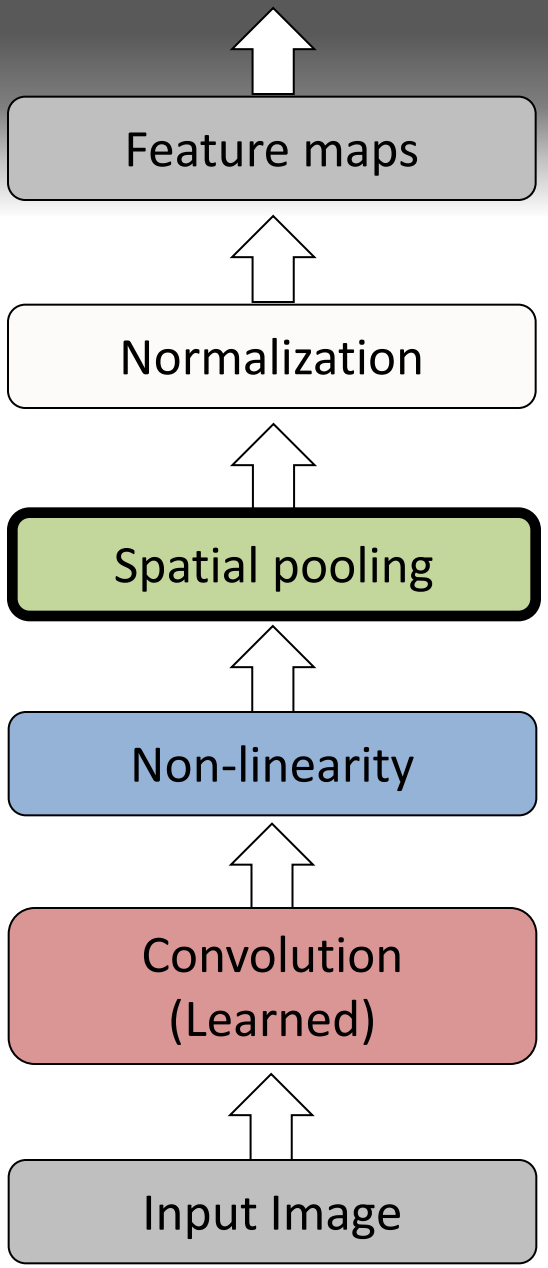


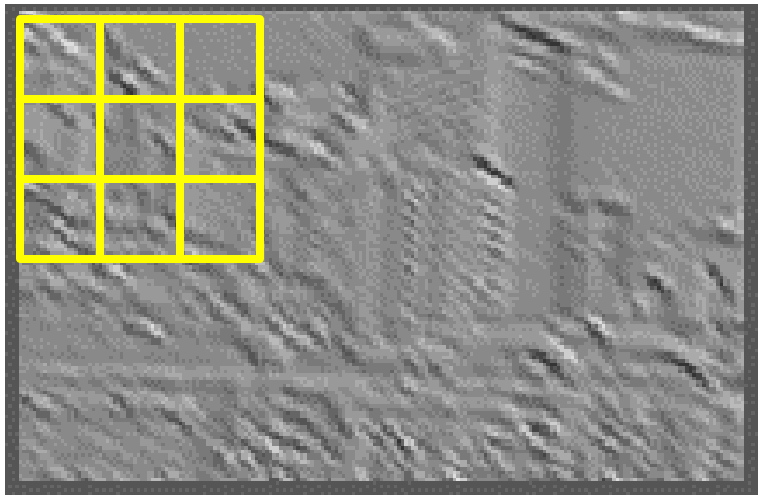
Rectified Linear Unit (ReLU)



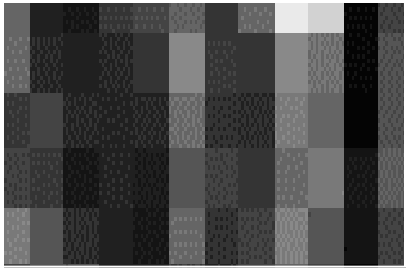
slide credit: S. Lazebnik

Convolutional Neural Networks



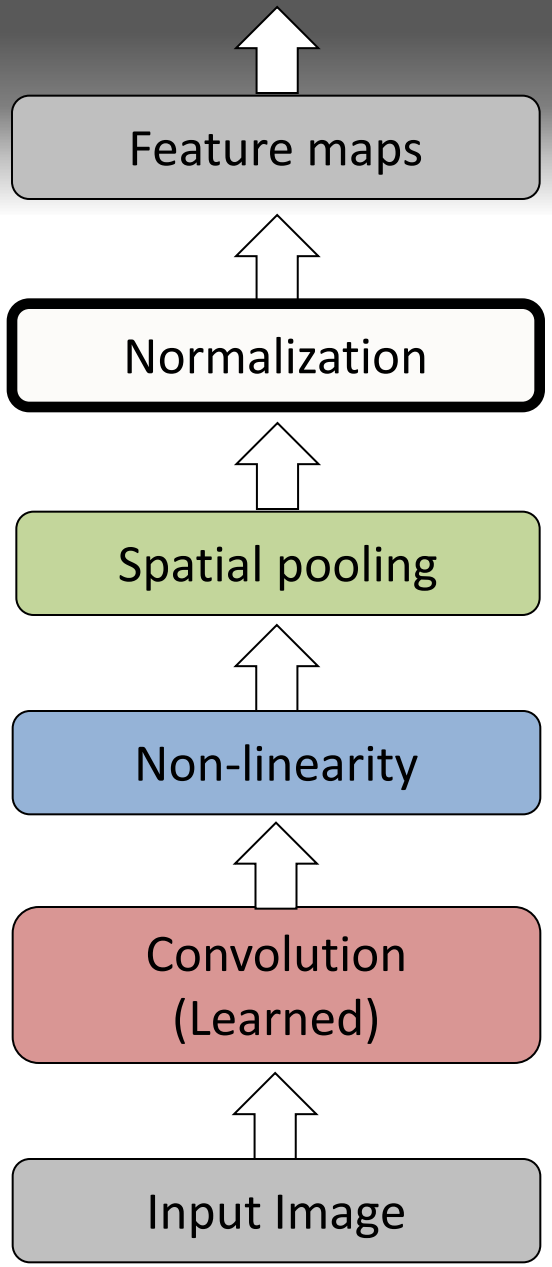


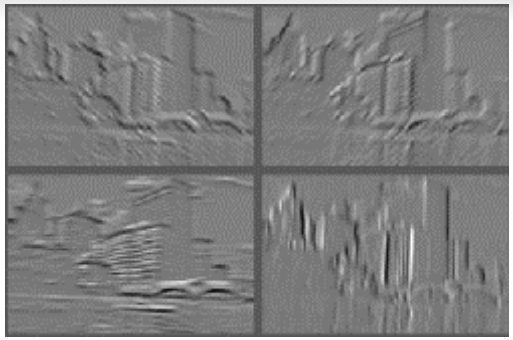
Max pooling



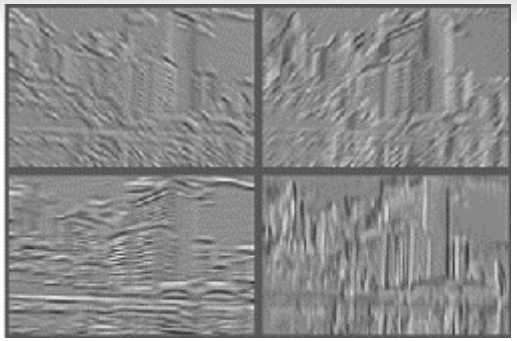
slide credit: S. Lazebnik

Convolutional Neural Networks





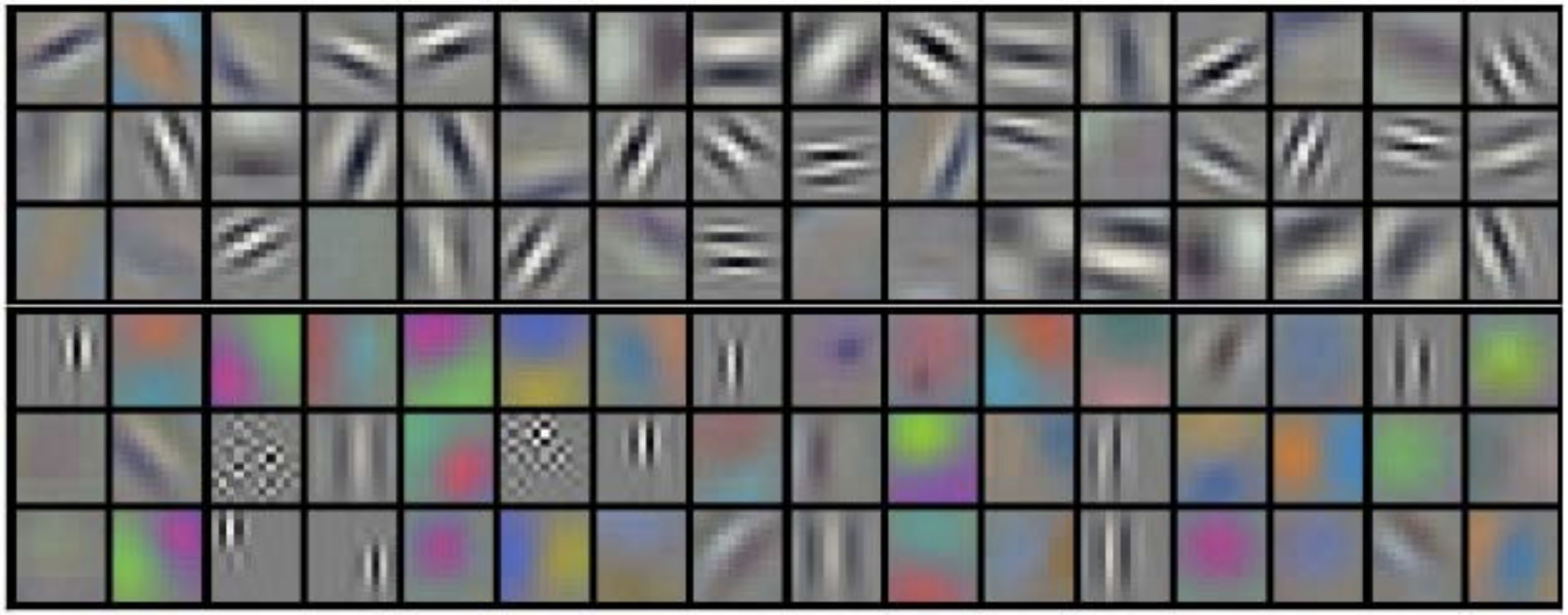
Feature Maps



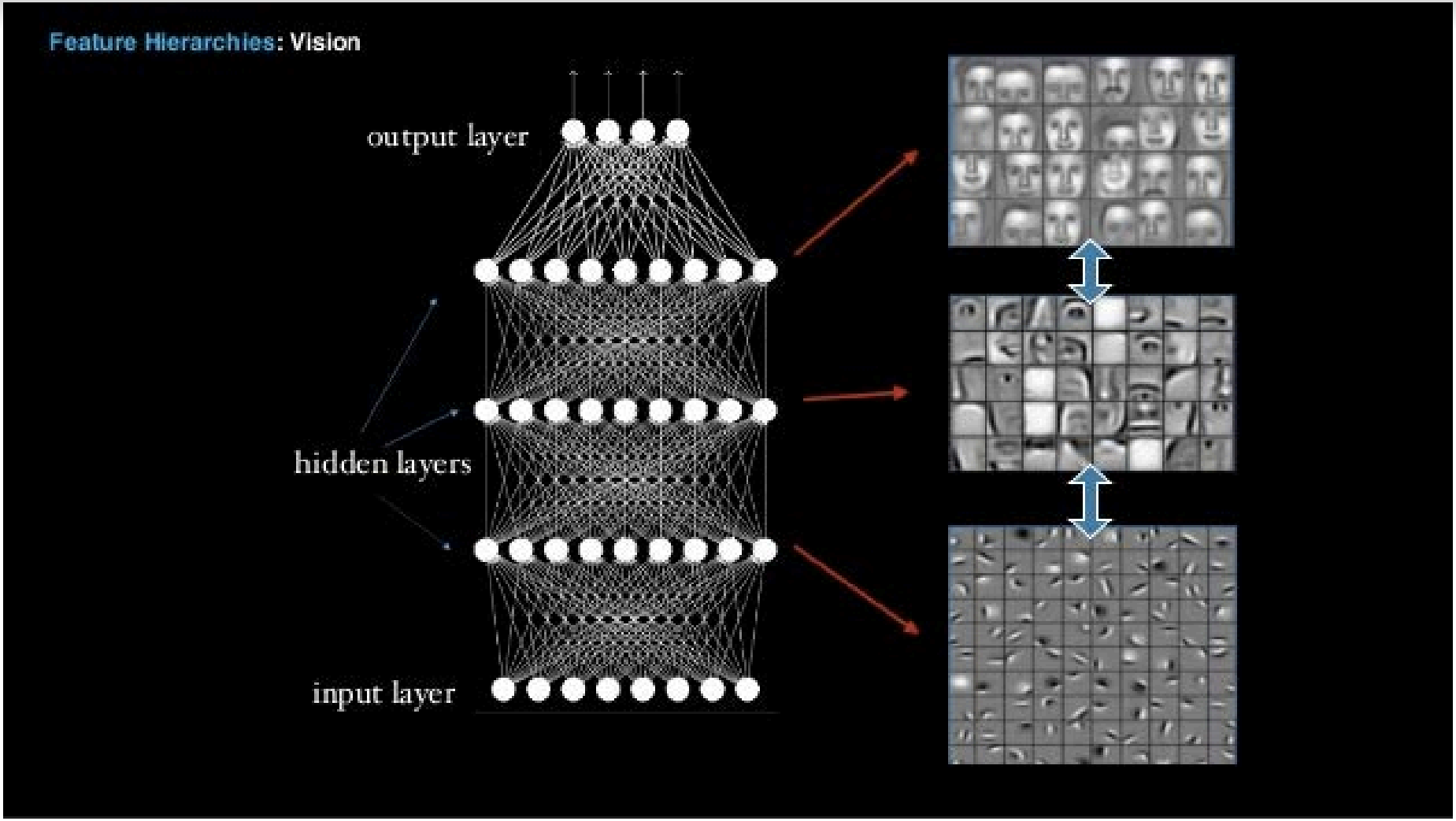
Feature Maps
After Contrast
Normalization

slide credit: S. Lazebnik

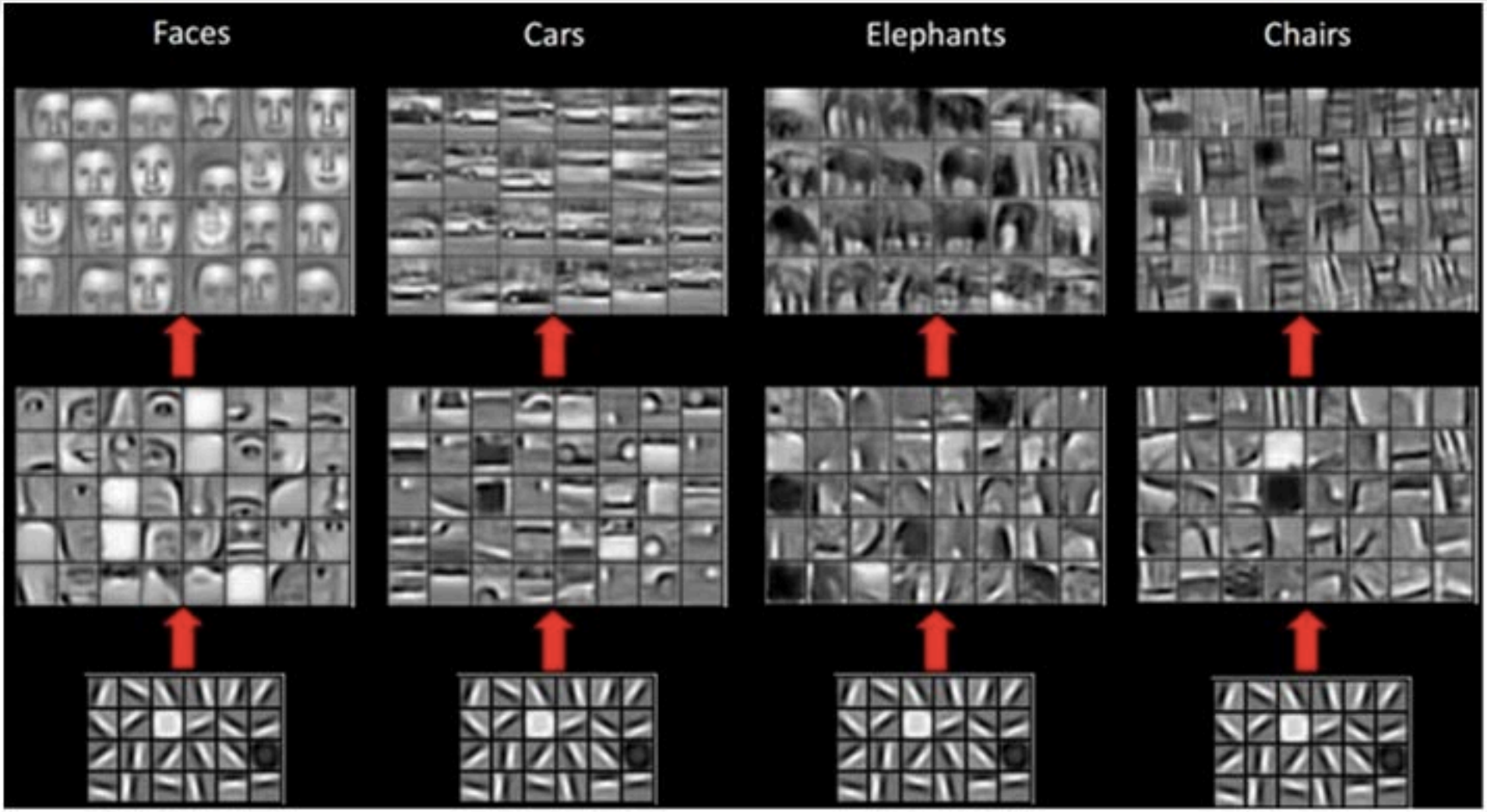
Filters in First Conv Layer



Filters in Different Layers

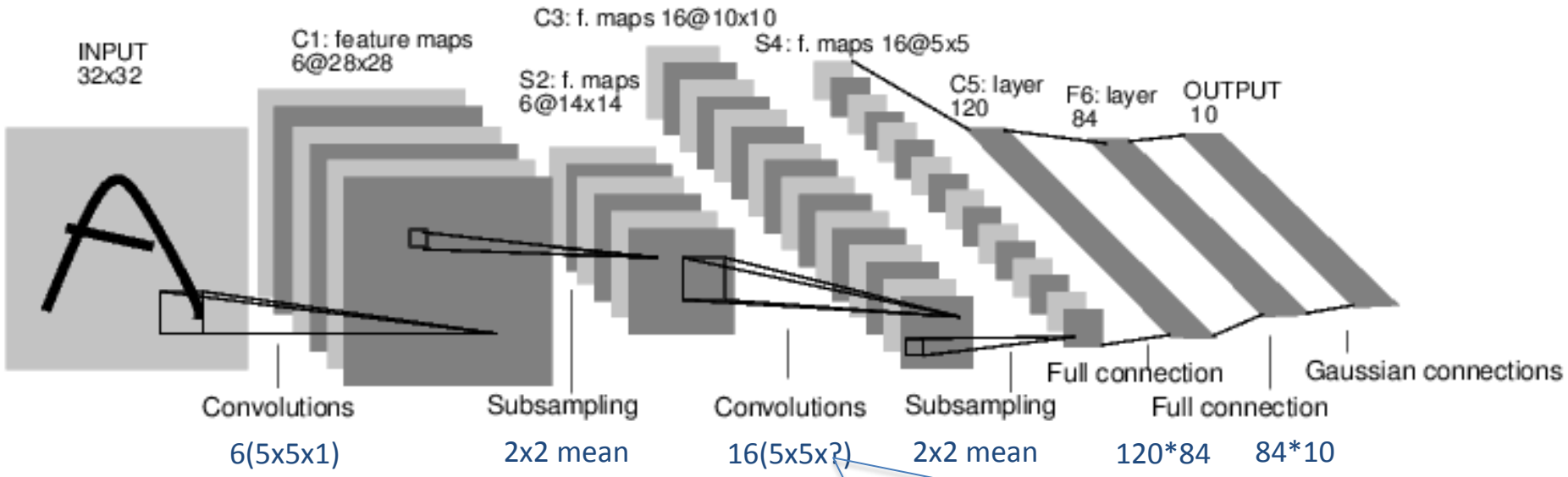


Filters for Different Categories



GO!

LeNet



	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
0	X				X	X	X			X	X	X	X		X	X
1	X	X				X	X	X			X	X	X	X		X
2	X	X	X				X	X	X			X		X	X	X
3		X	X	X			X	X	X	X			X		X	X
4			X	X	X			X	X	X	X		X	X		X
5				X	X	X			X	X	X	X		X	X	X

TABLE I

EACH COLUMN INDICATES WHICH FEATURE MAP IN S2 ARE COMBINED BY THE UNITS IN A PARTICULAR FEATURE MAP OF C3.

Y. LeCun, et al., Proc. of IEEE, 1998

LeNet

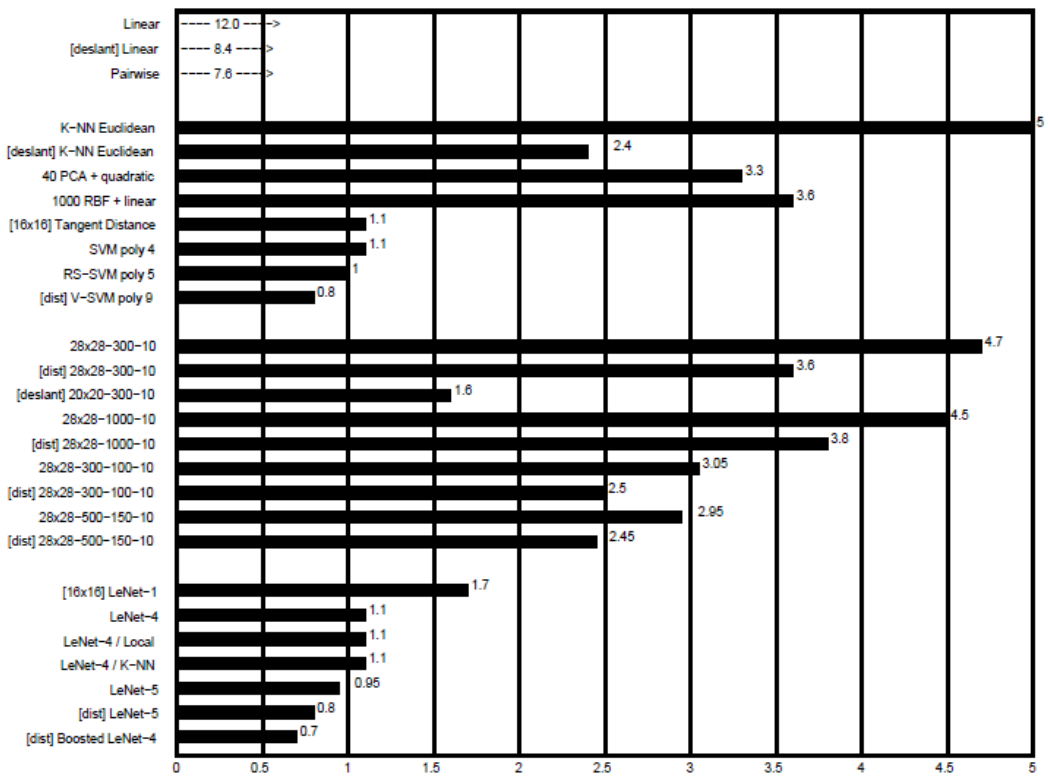
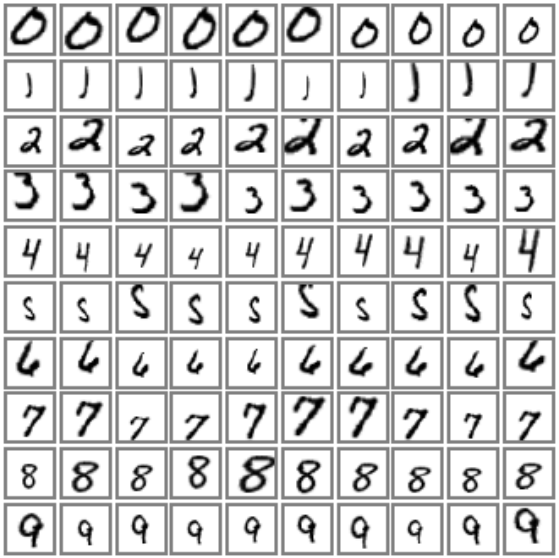
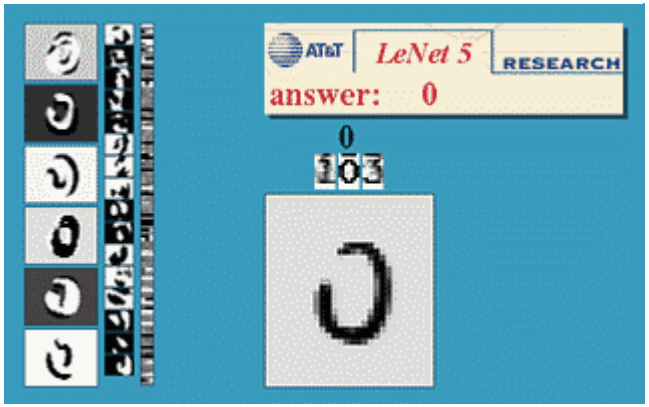


Fig. 9. Error rate on the test set (%) for various classification methods. [deslant] indicates that the classifier was trained and tested on the deslanted version of the database. [dist] indicates that the training set was augmented with artificially distorted examples. [16x16] indicates that the system used the 16x16 pixel images. The uncertainty in the quoted error rates is about 0.1%.





Y. LeCun, et al., Proc. of IEEE, 1998

ImageNet Competition

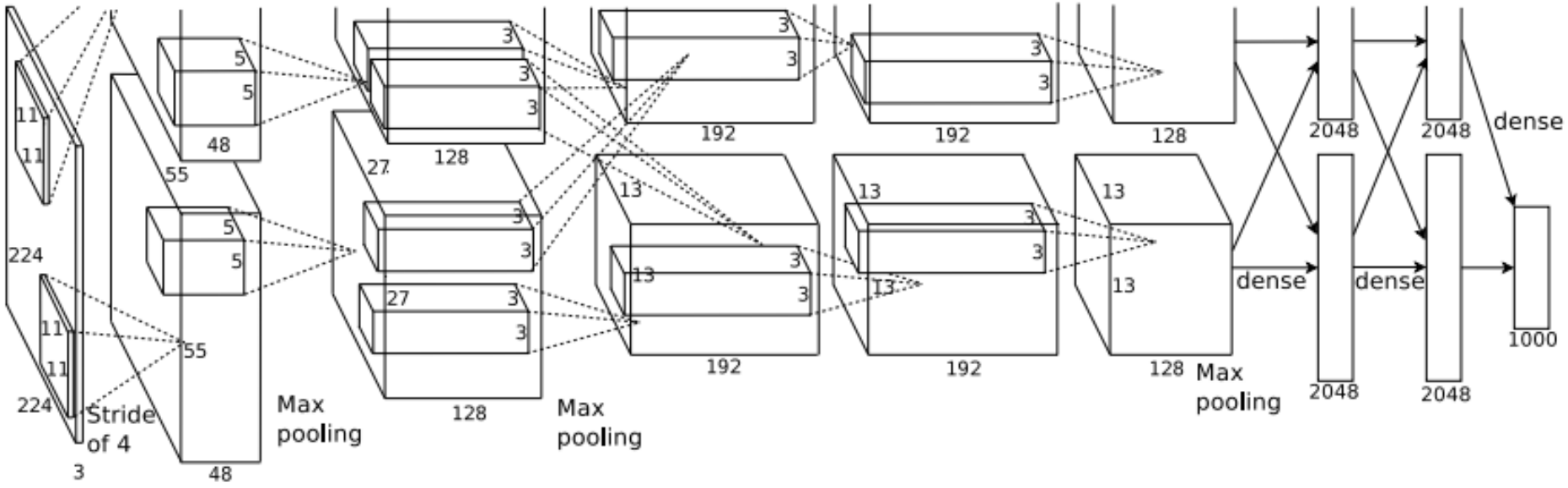
IMAGENET

- Large Scale Visual Recognition Challenge
- Using a subset of the large hand-labeled **ImageNet** dataset (1.2 million images from 1000 object categories) .
- For each image, algorithms will produce a list of at most **5 object categories** in the descending order of confidence. The quality of a labeling will be evaluated based on the label that best matches the ground truth label for the image.

GO!

Deep

AlexNet

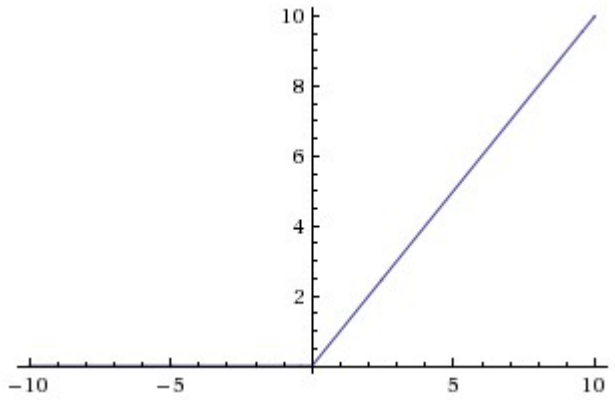


96(11x11x3) 256(5x5x48) 384(3x3x256) 384(3x3x192) 256(3x3x192) 4096x4096 4096x1000

A. Krizhevsky, et al., NIPS 2012

AlexNet

- Data Augmentation
- Overlapping Pooling
  - Prevent overfitting
- ReLU Nonlinearity
  - Faster Learning



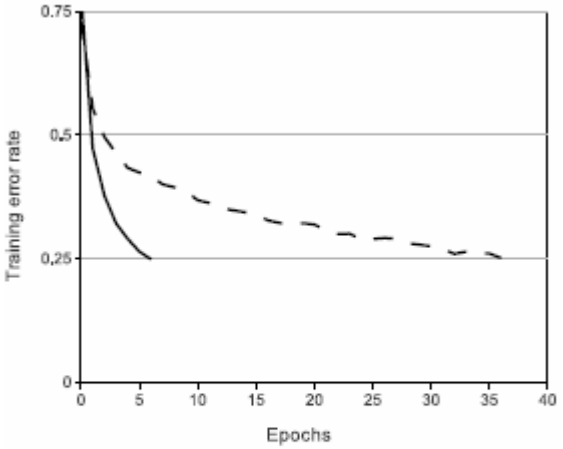


Figure 1: A four-layer convolutional neural network with ReLUs (**solid line**) reaches a 25% training error rate on CIFAR-10 six times faster than an equivalent netv
(**dashed line**). The learning rates for each net-

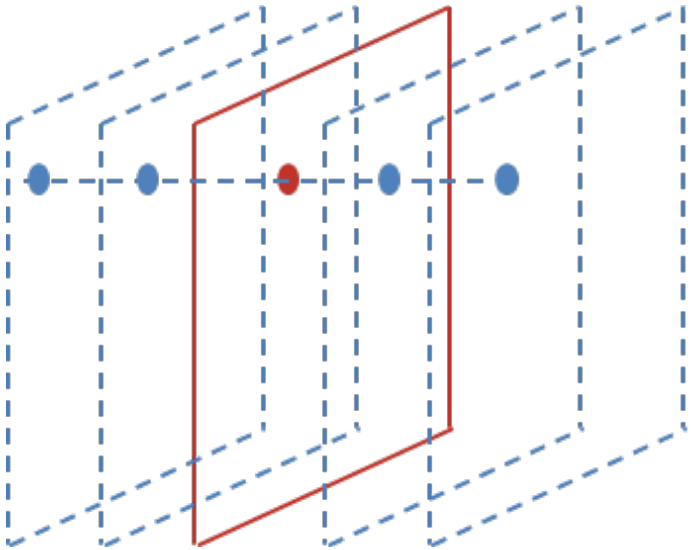
A. Krizhevsky, et al., NIPS 2012

AlexNet

- Local Response Normalization
  - aids generalization

$$b^i_{x,y} = a^i_{x,y} / \left(k + \alpha \sum_{j=\max(0,i-n/2)}^{\min(N-1,i+n/2)} (a^j_{x,y})^2 \right)^{\beta}$$

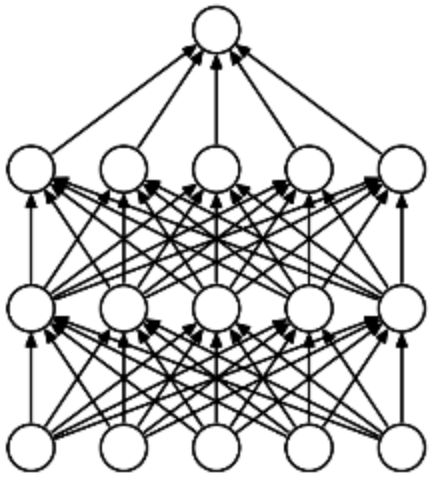
k = 2, n = 5, α = 10−4, and β = 0.75



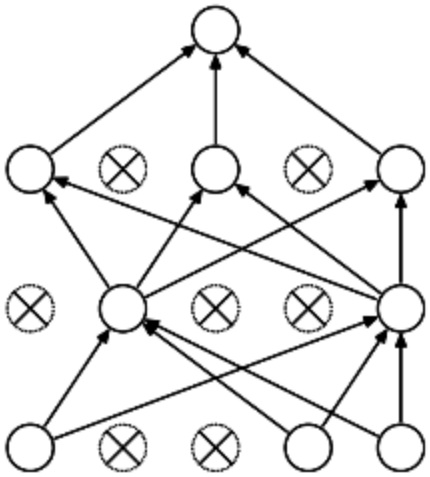
A. Krizhevsky, et al., NIPS 2012

AlexNet

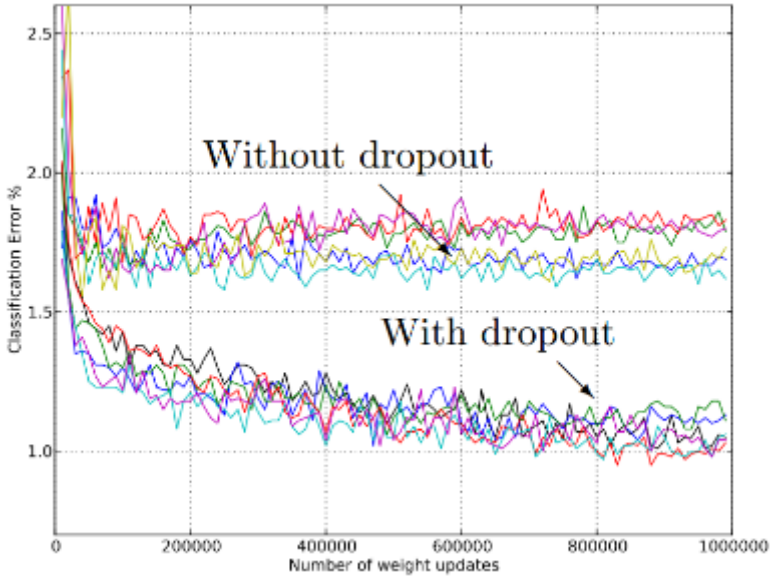
- Dropout
  - Consists of setting to zero the output of each hidden neuron with probability 0.5.



(a) Standard Neural Net



(b) After applying dropout.



Dropout: A simple way to prevent neural networks from overfitting [Srivastava JMLR 2014]

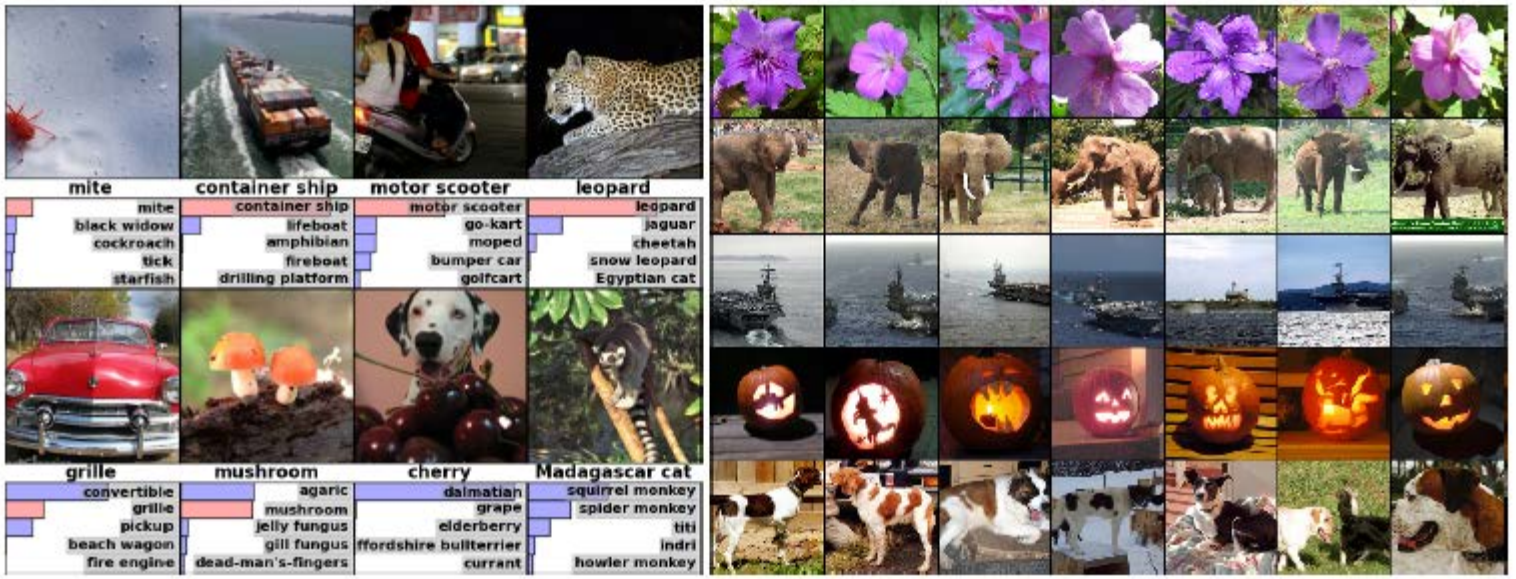
AlexNet

Winner of ILSVRC 2010, NEC-UIUC Group

Winner of ILSVRC 2012, Univ. of Toronto

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

Table 1: Comparison of results on ILSVRC-2010 test set. In *italics* are best results achieved by others.



GO!

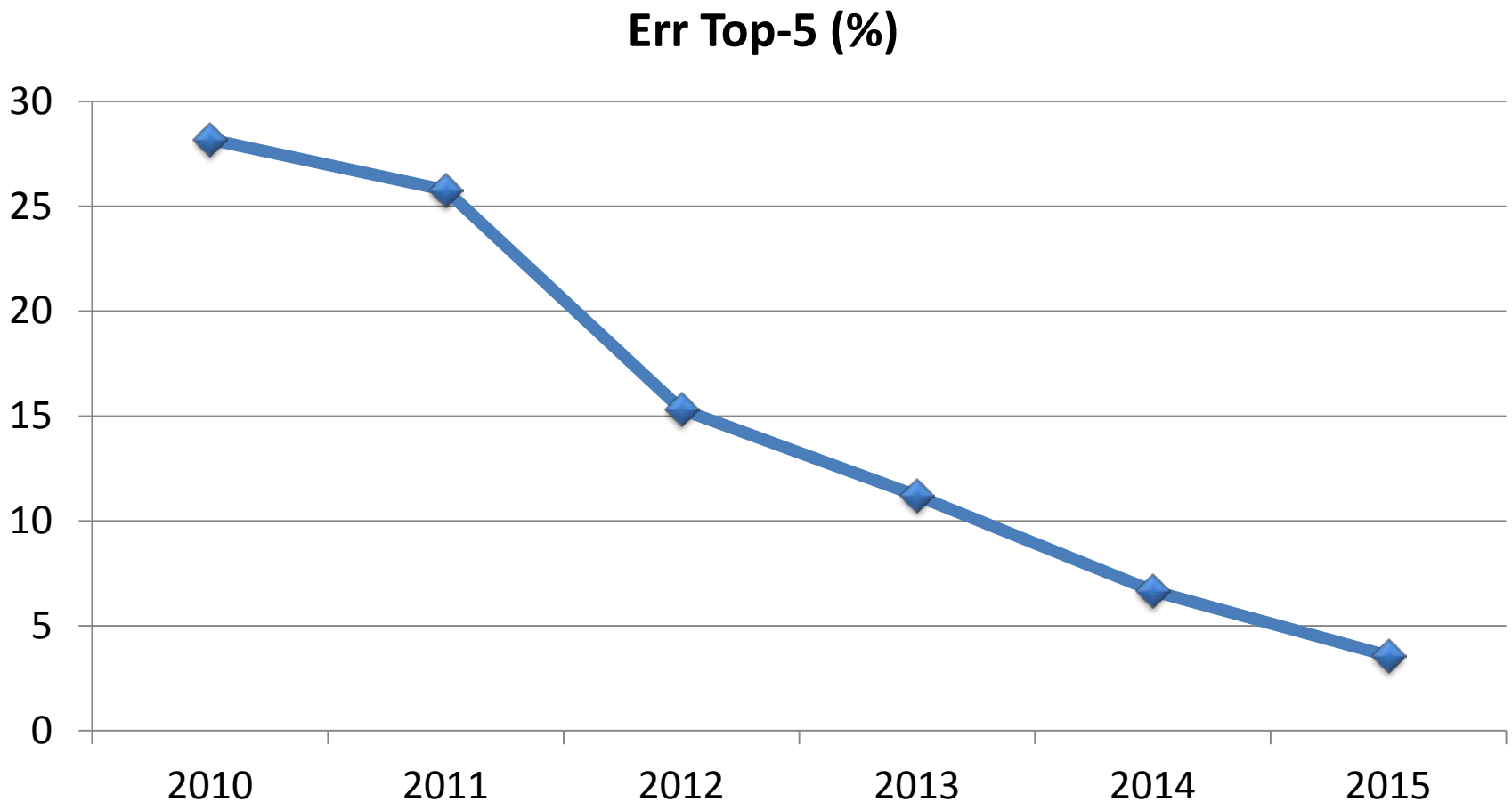
Deep

Deeper

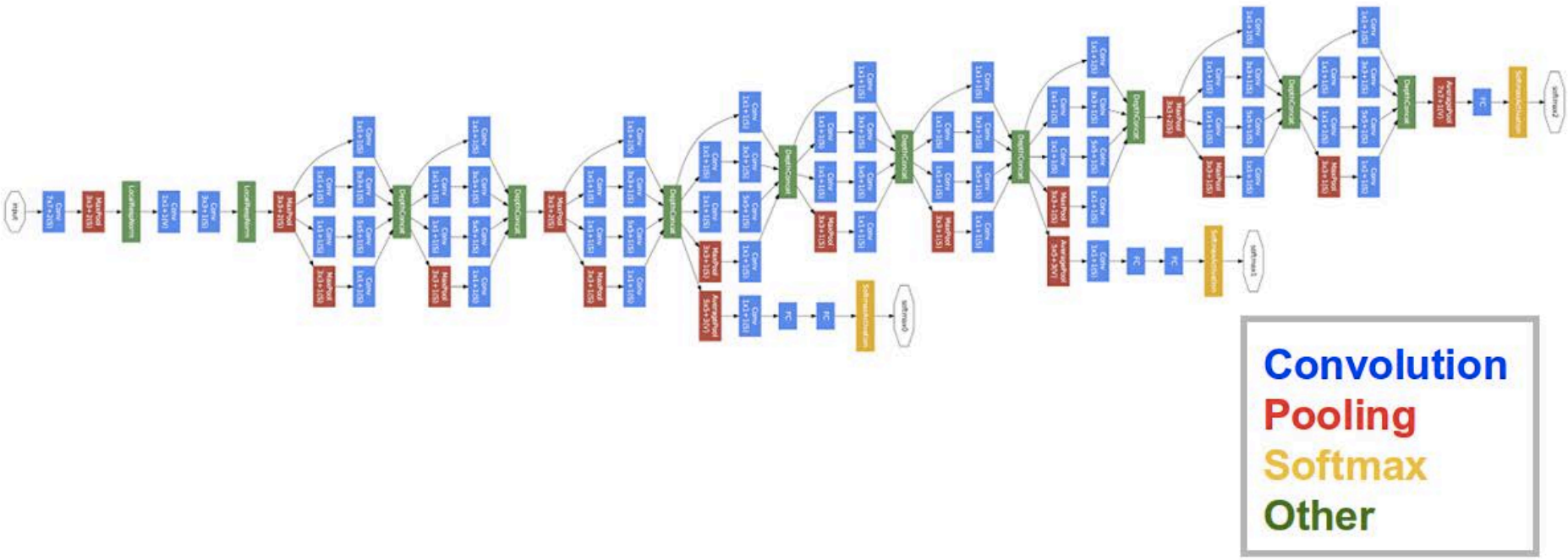
1) Overfitting

2) Hard to optimize

3) Huge computing recourse

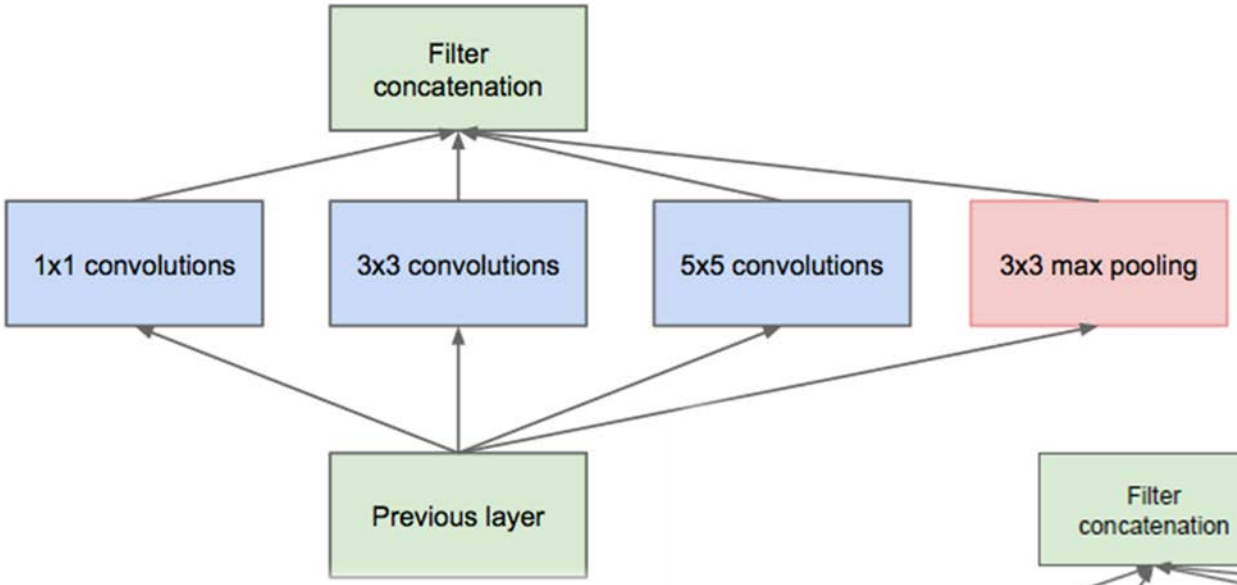

Err Top-5 (%)
30
25
20
15
10
5
0
2010
2011
2012
2013
2014
2015

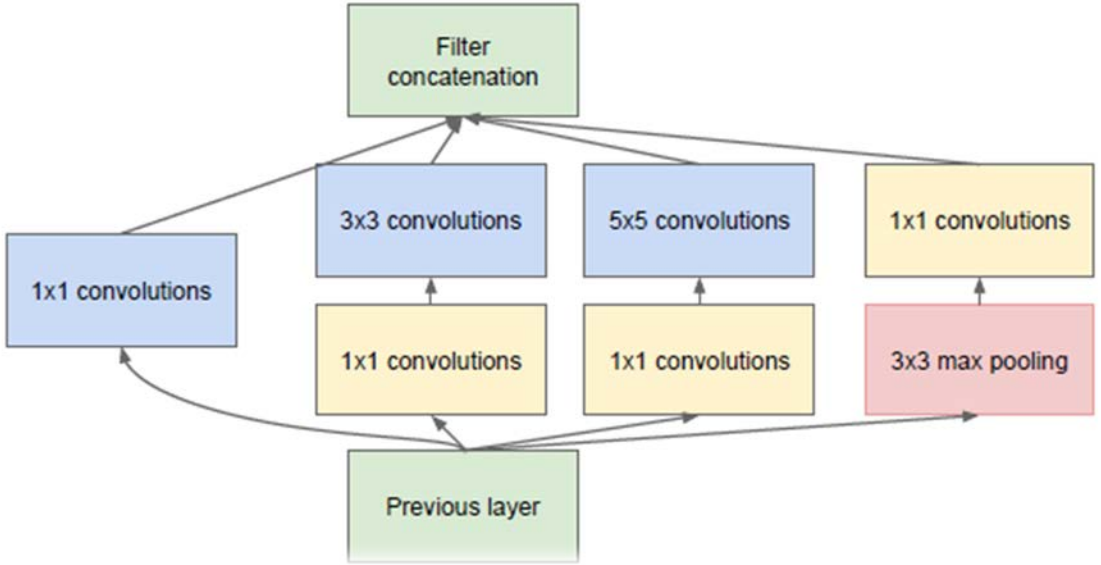
GoogLeNet



C. Szegedy, et al., CVPR 2015

GoogLeNet





C. Szegedy, et al., CVPR 2015

GoogLeNet

Team	Year	Place	Error (top-5)	Uses external data
SuperVision	2012	1st	16.4%	no
SuperVision	2012	1st	15.3%	Imagenet 22k
Clarifai	2013	1st	11.7%	no
Clarifai	2013	1st	11.2%	Imagenet 22k
MSRA	2014	3rd	7.35%	no
VGG	2014	2nd	7.32%	no
GoogLeNet	2014	1st	6.67%	no

Winner of ILSVRC 2014, Google

Table 2: Classification performance.

C. Szegedy, et al., CVPR 2015

GO!

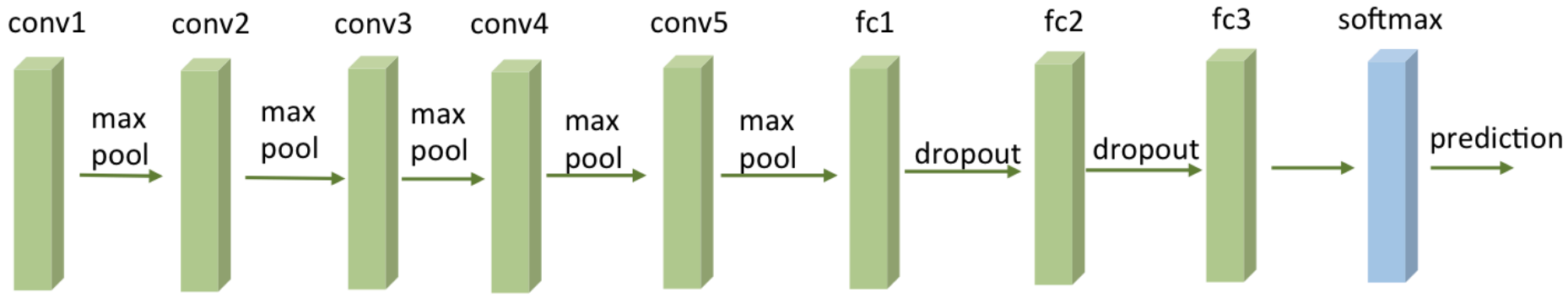
Deep

Deeper

Deeper

VGG

each conv includes 3 convolutional layers



K Simonyan, A Zisserman, ICLR 2015

VGG

Table 7: **Comparison with the state of the art in ILSVRC classification**. Our method is denoted as "VGG". Only the results obtained without outside training data are reported.

Method	top-1 val. error (%)	top-5 val. error (%)	top-5 test error (%)
VGG (2 nets, multi-crop & dense eval.)	**23.7**	**6.8**	**6.8**
VGG (1 net, multi-crop & dense eval.)	24.4	7.1	7.0
VGG (ILSVRC submission, 7 nets, dense eval.)	24.7	7.5	7.3
GoogLeNet (Szegedy et al., 2014) (1 net)	-	7.9	
GoogLeNet (Szegedy et al., 2014) (7 nets)	-	**6.7**	
MSRA (He et al., 2014) (11 nets)	-	-	8.1
MSRA (He et al., 2014) (1 net)	27.9	9.1	9.1
Clarifai (Russakovsky et al., 2014) (multiple nets)	-	-	11.7
Clarifai (Russakovsky et al., 2014) (1 net)	-	-	12.5
Zeiler & Fergus (Zeiler & Fergus, 2013) (6 nets)	36.0	14.7	14.8
Zeiler & Fergus (Zeiler & Fergus, 2013) (1 net)	37.5	16.0	16.1
OverFeat (Sermanet et al., 2014) (7 nets)	34.0	13.2	13.6
OverFeat (Sermanet et al., 2014) (1 net)	35.7	14.2	-
Krizhevsky et al. (Krizhevsky et al., 2012) (5 nets)	38.1	16.4	16.4
Krizhevsky et al. (Krizhevsky et al., 2012) (1 net)	40.7	18.2	-

K Simonyan, A Zisserman, ICLR 2015

GO!

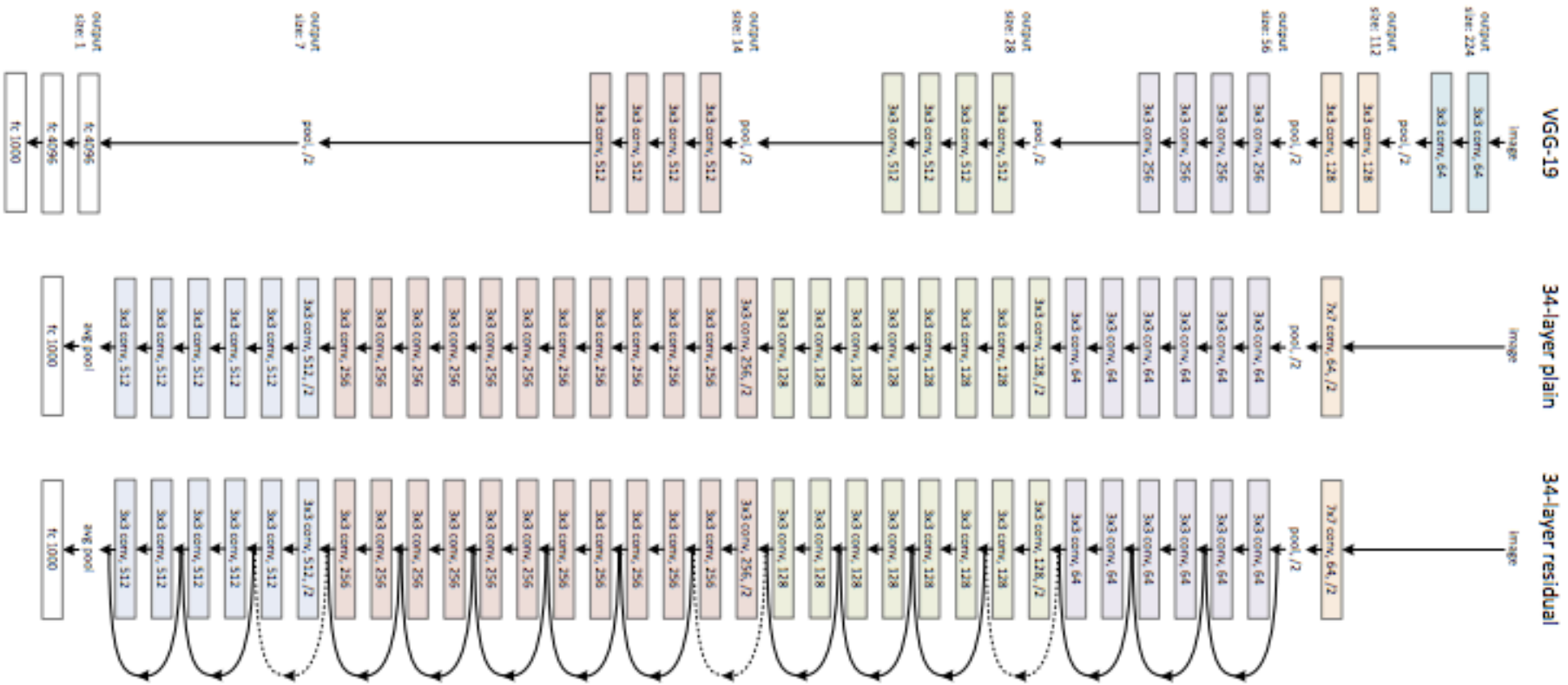
Deep

Deeper

Deeper

Deeper

Deep Residual Learning



K. He, et al., (MSRA)

Deep Residual Learning

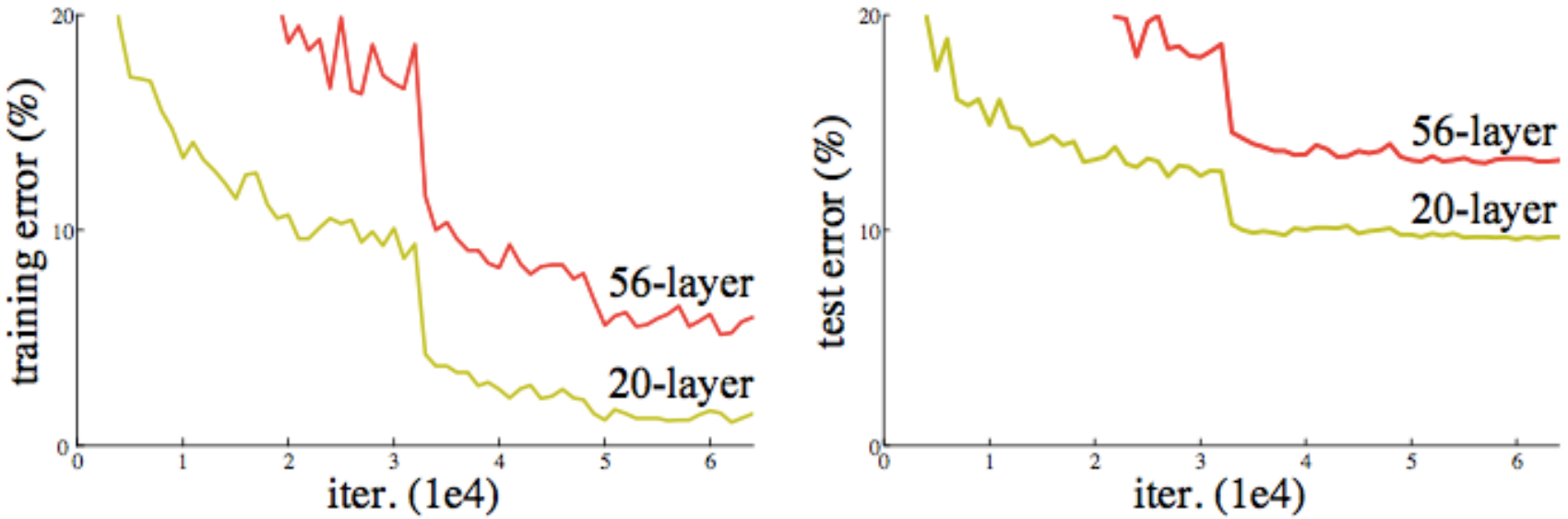


Figure 1. Training error (left) and test error (right) on CIFAR-10 with 20-layer and 56-layer "plain" networks. The deeper network has higher training error, and thus test error. Similar phenomena on ImageNet is presented in Fig. 4.

K. He, et al., (MSRA)

Deep Residual Learning

Desired underlying mapping:

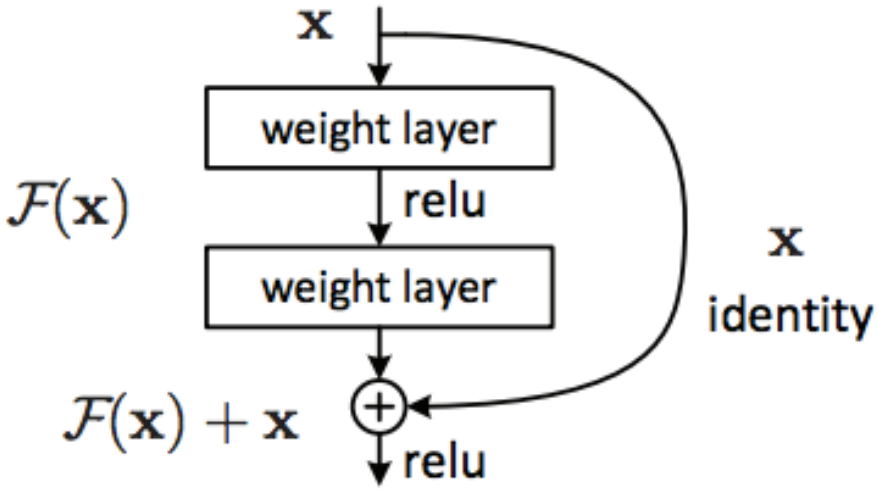
$$\mathcal{H}(\mathbf{x})$$

Residual function:

$$\mathcal{F}(\mathbf{x}) := \mathcal{H}(\mathbf{x}) - \mathbf{x}.$$

Original function becomes:

$$\mathcal{F}(\mathbf{x}) + \mathbf{x}.$$



K. He, et al., (MSRA)

Deep Residual Learning

method	top-1 err.	top-5 err.
VGG [41] (ILSVRC'14)	-	8.43†
GoogLeNet [44] (ILSVRC'14)	-	7.89
VGG [41] (v5)	24.4	7.1
PReLU-net [13]	21.59	5.71
BN-inception [16]	21.99	5.81
ResNet-34 B	21.84	5.71
ResNet-34 C	21.53	5.60
ResNet-50	20.74	5.25
ResNet-101	19.87	4.60
ResNet-152	**19.38**	**4.49**

Table 4. Error rates (%) of **single-model** results on the ImageNet validation set (except † reported on the test set).

Deep Residual Learning

method	top-5 err. (**test**)
VGG [41] (ILSVRC'14)	7.32
GoogLeNet [44] (ILSVRC'14)	6.66
VGG [41] (v5)	6.8
PReLU-net [13]	4.94
BN-inception [16]	4.82
ResNet (ILSVRC'15)	**3.57**

Winner of ILSVRC 2015, MSRA

Table 5. Error rates (%) of **ensembles**. The top-5 error is on the test set of ImageNet and reported by the test server.