Image Classification

Chetan Arora

Chetan Arora
Department of Computer Science and Engineering, IIT Delhi

Visual Recognition Tasks

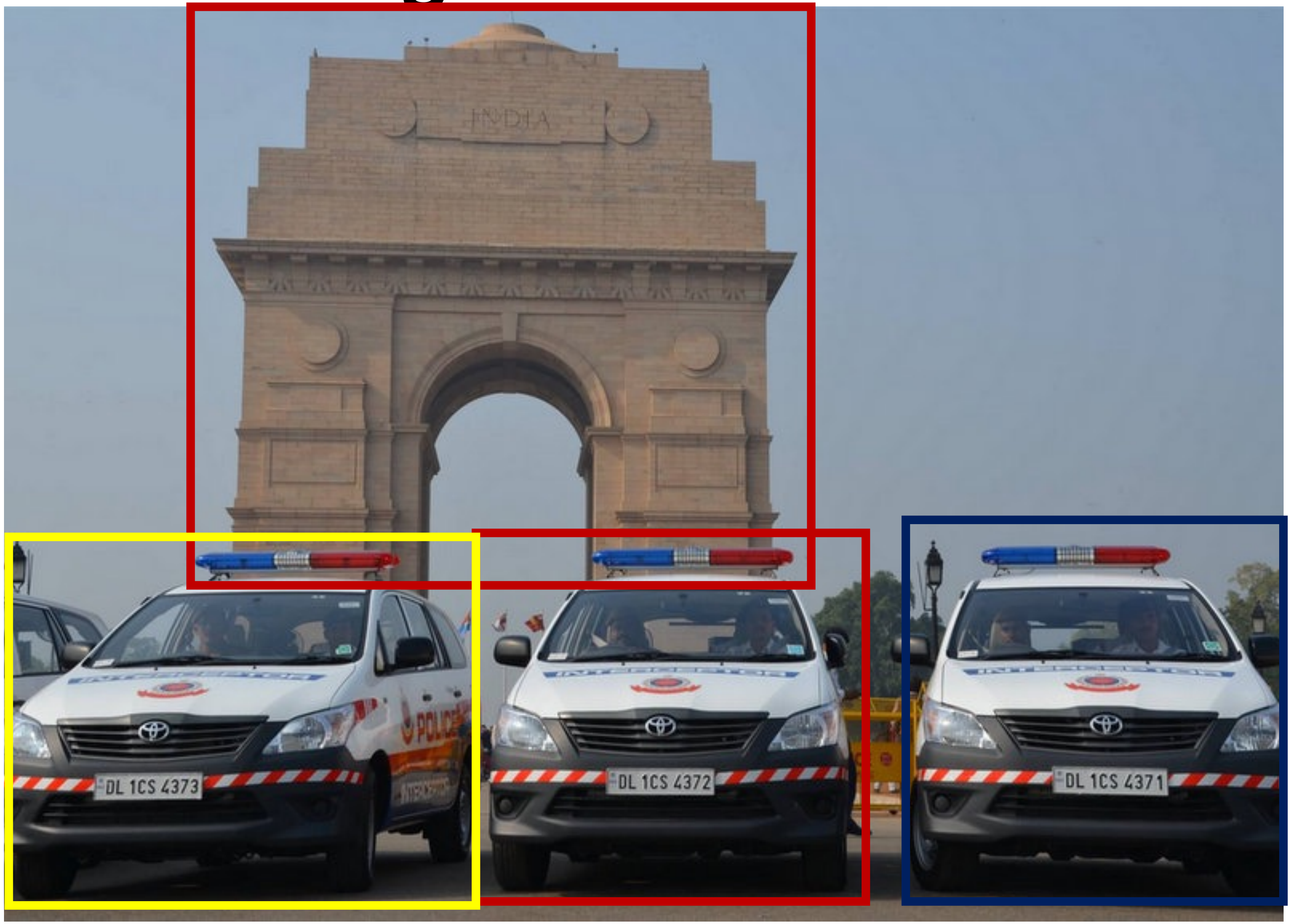
Image Classification

Is it a natural or man made scene

Is it a forest or a beach?

Does this image contains a building?

Visual Recognition Tasks



Object Detection

Does this image contain a building? [where]

Which objects does this image contains?

Chetan Arora
Department of Computer Science and Engineering, IIT Delhi

Visual Recognition Tasks

Instance Segmentation [pixel wise localization]

Which pixels are building?

Chetan Arora
Department of Computer Science and Engineering, IIT Delhi

Applications: Computational Photography

Face Detection

Dynamic Range Enhancement

Applications: Object Attributes

Building:
42 m height
100 m away

Car:
Police Car
Frontal View

Autonomous and Assistive Driving

Chetan Arora
Department of Computer Science and Engineering, IIT Delhi

Applications: Instance Recognition

Does this image contains “India Gate”?

Recognizing landmarks in images

Recognizing products in super market

Chetan Arora
Department of Computer Science and Engineering, IIT Delhi

Applications: Assistive Vision

Applications: Security and Surveillance

Chetan Arora
Department of Computer Science and Engineering, IIT Delhi

Applications: Activity Recognition

What are these guys doing?

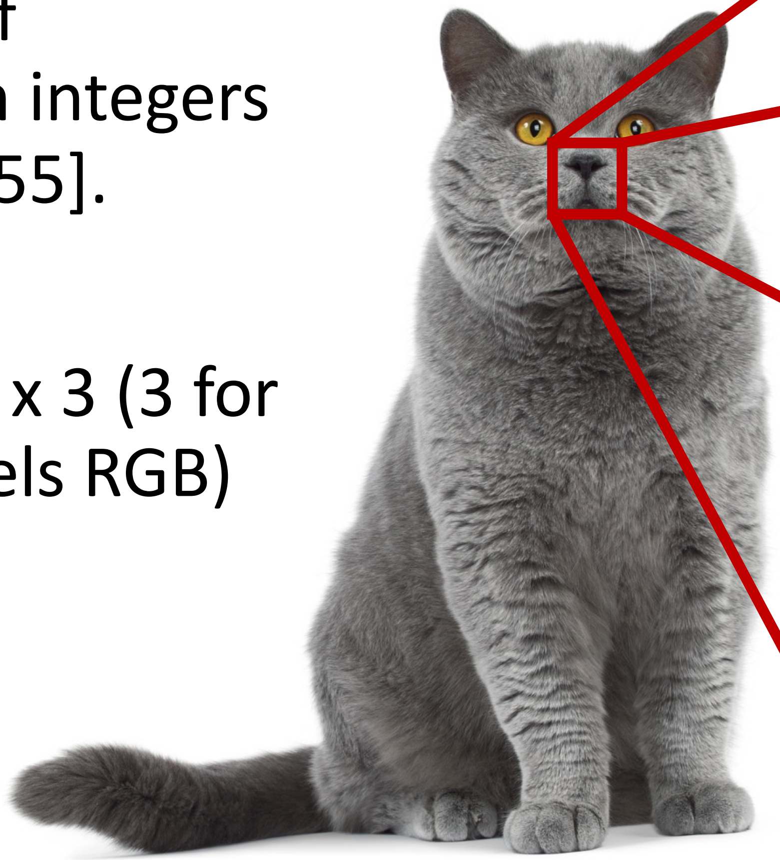
Autonomous Systems

http://fortune.com/2015/10/16/how-tesla-autopilot-learns/

Chetan Arora
Department of Computer Science and Engineering, IIT Delhi

Image Classification Challenge: Semantic Gap

- Images are represented as 3D arrays of numbers, with integers between [0, 255].
- E.g. 300 x 100 x 3 (3 for 3 color channels RGB)



20	56	12	207	12	56	12	20	207	56	207	23	123	12	12
30	78	43	255	43	78	43	30	255	78	255	34	54	43	34
26	96	0	125	27	96	0	26	125	96	125	74	24	0	26
89	78	87	168	49	78	87	89	168	78	168	24	15	87	31
54	56	65	198	63	56	65	54	198	56	198	75	125	65	156
128	45	45	187	82	45	45	128	187	45	187	25	25	45	167
45	98	98	165	63	98	98	45	165	98	165	27	156	98	145
134	67	67	193	82	67	67	134	193	67	193	28	56	67	146
235	45	23	88	76	45	23	235	88	45	88	83	32	23	158
23	145	45	22	126	145	45	23	22	145	22	5	63	45	234
24	234	244	62	139	234	244	24	62	234	62	27	43	244	43
45	65	213	104	176	65	213	45	104	65	104	42	53	213	25
23	213	154	176	174	213	154	23	176	213	176	63	63	154	25
45	54	167	187	27	54	167	45	187	54	187	72	135	167	53
67	76	195	193	26	76	195	67	193	76	193	24	246	195	63

CS231N Stanford

Chetan Arora
Department of Computer Science and Engineering, IIT Delhi

Image Classification Challenge


Intra class variation


Background Clutter


Deformation


Illumination

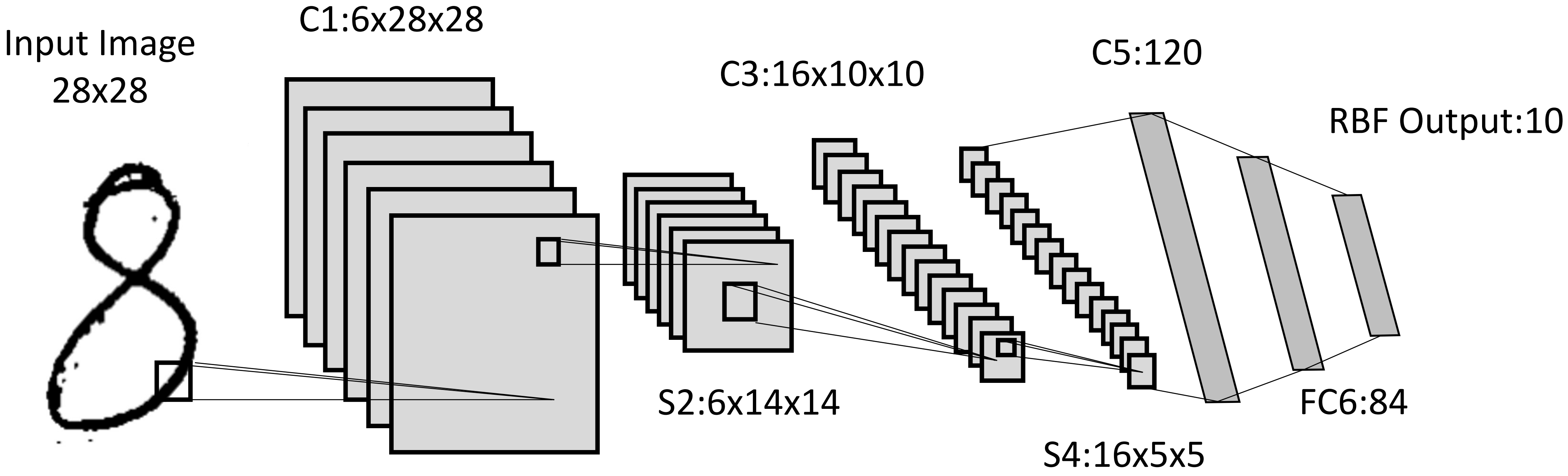

Occlusion


Size

CS231N Stanford

LeNet5

- C1,C3,C5 : Convolutional layers. 5×5 Convolution matrix.
- S2 , S4 : Subsampling layer. Subsampling by factor 2.
- F6 : Fully connected layer.
- All the units of the layers up to FC6 have a sigmoidal activation function

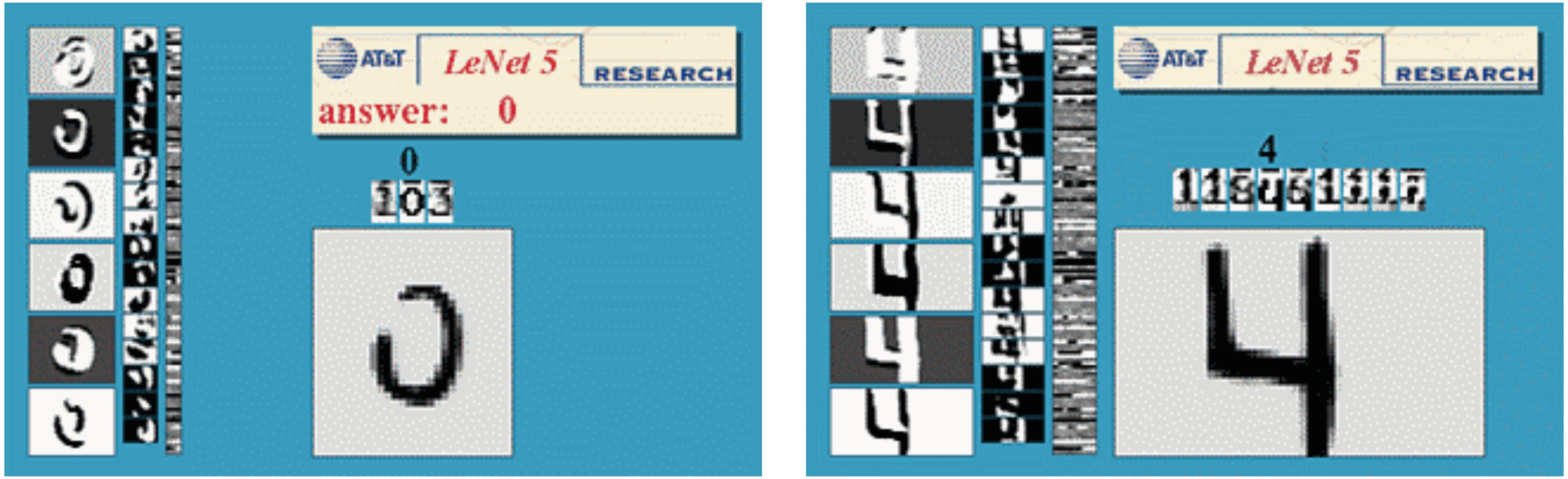


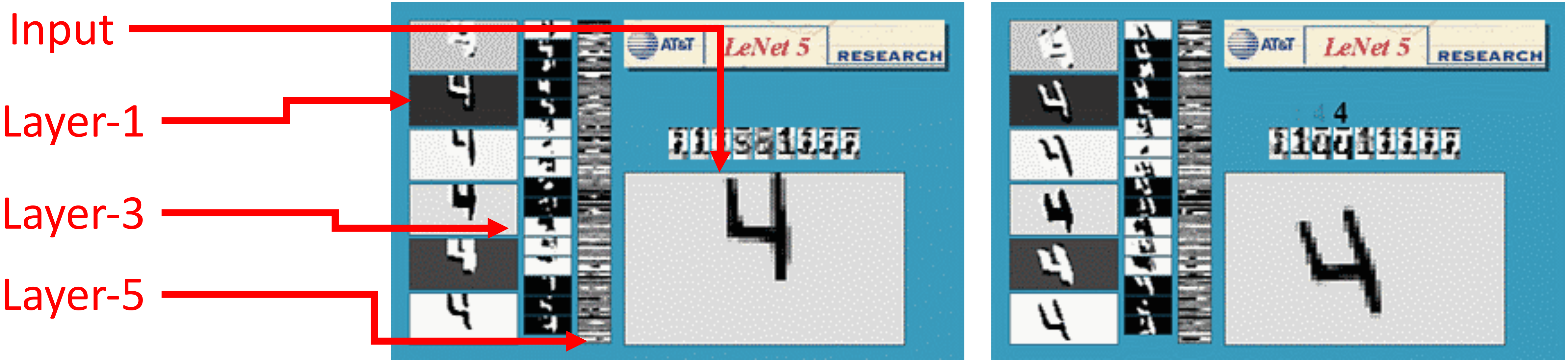
Y. LeCun, P. Haffner, L. Bottou and Y. Bengio: Object Recognition with Gradient-Based Learning. 1999

Chetan Arora
Department of Computer Science and Engineering, IIT Delhi

LeNet5

- About 187,000 connection.
- About 14,000 trainable weight





Chetan Arora
Department of Computer Science and Engineering, IIT Delhi

LeNet5

- Uses knowledge about the invariances to design:
  - the local connectivity,
  - the weight-sharing, and
  - the pooling.

- Achieves about 80 errors:
  - This can be reduced to about 40 errors by using many different transformations of the input and other tricks (Ranzato 2008)

ImageNet Large-Scale Visual Recognition Challenge (ILSVRC)

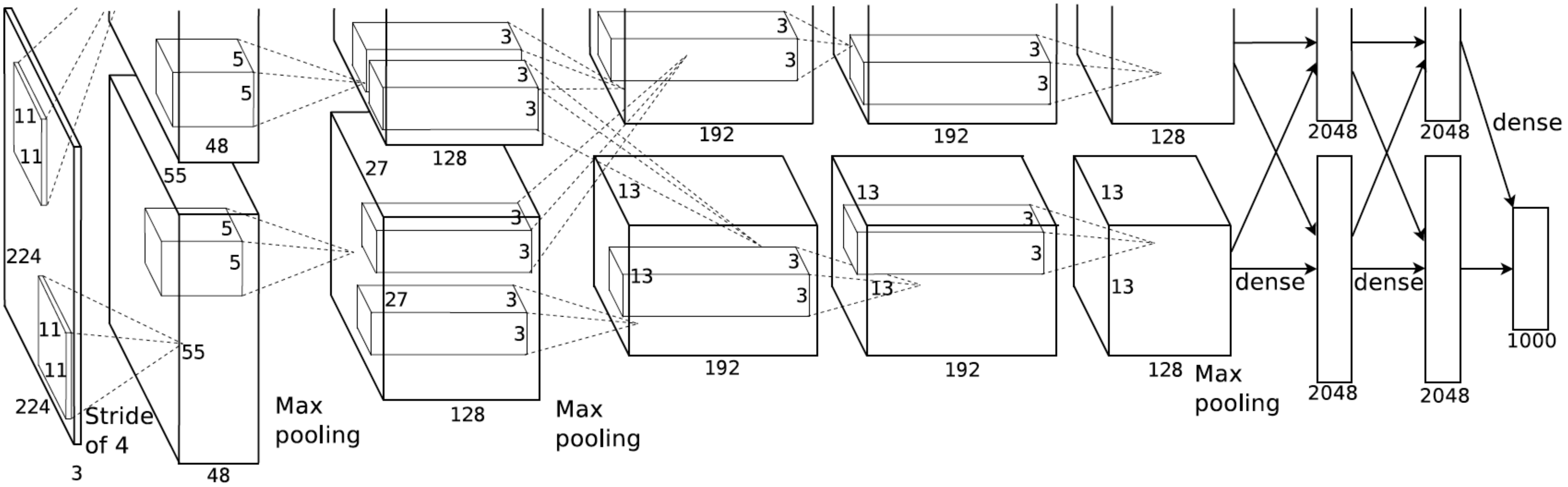
- 10,000,000 labelled images depicting 10,000+ object categories collected from flickr and other search engines.
- ILSVRC 2012
  - Validation and test data of 150,000 photographs, hand labelled with 1000 object categories.
  - A random subset of 50,000 of the images with labels released as validation data
  - The training data, containing the 1000 categories, and 1.2 million images,
- Evaluation
  - Output a list of 5 object categories in descending order of confidence
  - Two error rates: top-1 and top-5

http://image-net.org/challenges/LSVRC/2012/

Chetan Arora
Department of Computer Science and Engineering, IIT Delhi

AlexNet (2012)

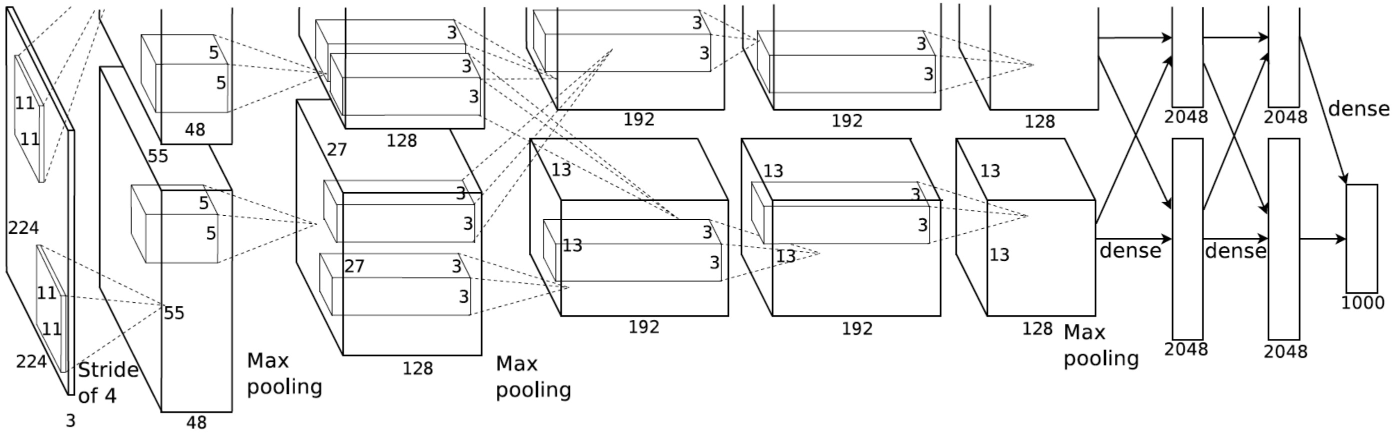
- 5 convolutional layers
- 3 fully connected layers
- 1000-way softmax output layer



Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton: ImageNet Classification with Deep Convolutional Neural Networks

Chetan Arora
Department of Computer Science and Engineering, IIT Delhi

AlexNet

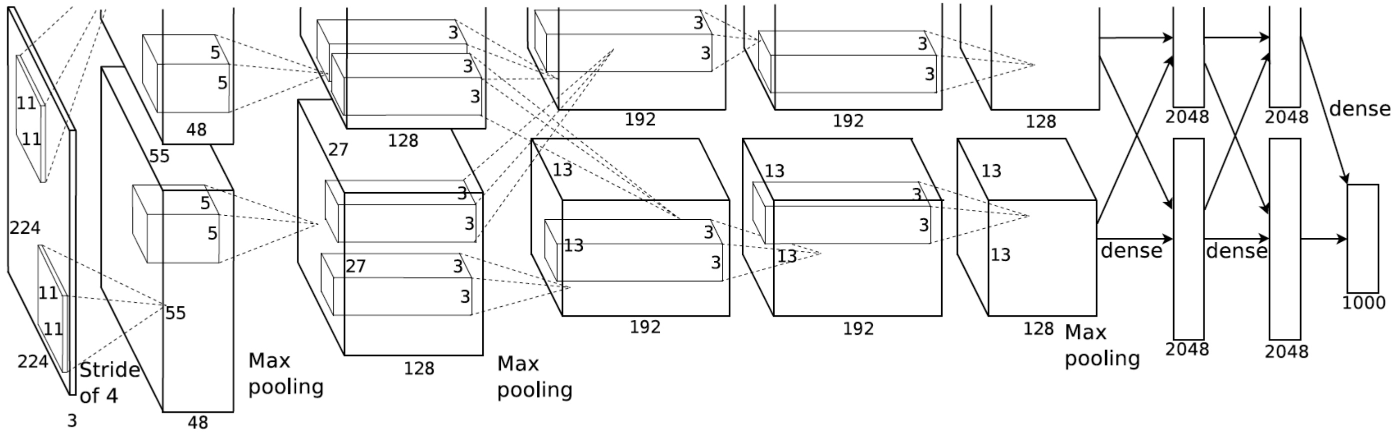


- Input: 227x227x3

- First layer (CONV1): 96 11x11 filters applied at stride 4

Q: What is the output volume size?

AlexNet



- Input: 227x227x3

- First layer (CONV1): 96 11x11 filters applied at stride 4

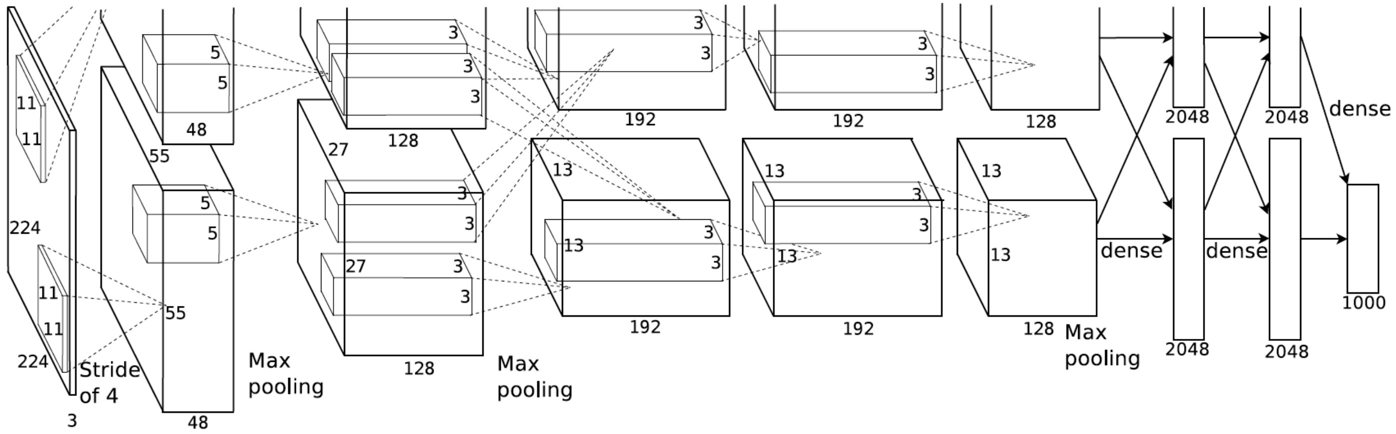
Q: What is the output volume size?

A: (227-11)/4+1 = 55. Output volume [55x55x96]

Q: What is the total number of parameters in this layer?

Chetan Arora
Department of Computer Science and Engineering, IIT Delhi

AlexNet



- Input: 227x227x3

- First layer (CONV1): 96 11x11 filters applied at stride 4

Q: What is the output volume size?

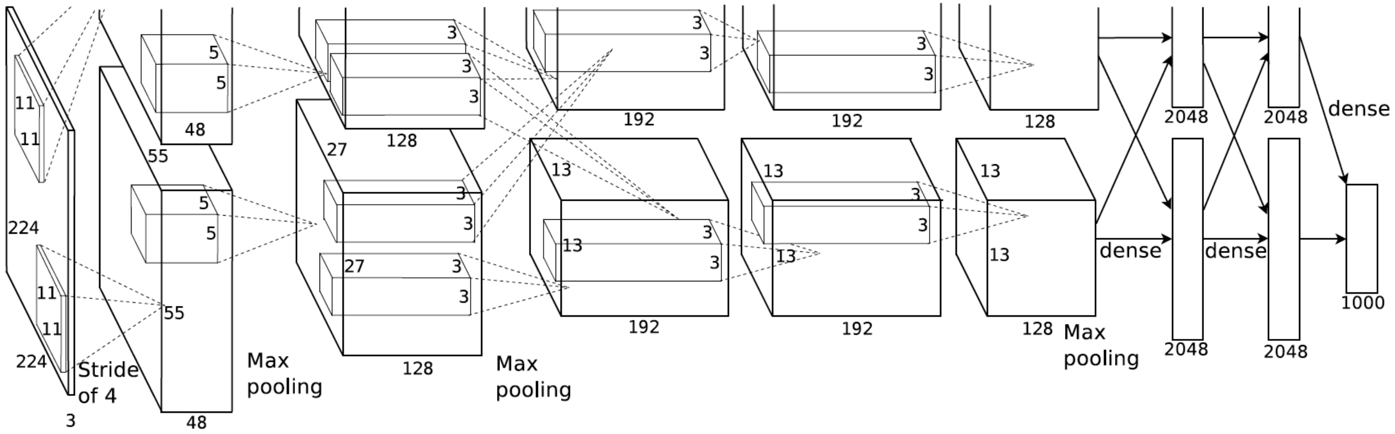
A: (227-11)/4+1 = 55. Output volume [55x55x96]

Q: What is the total number of parameters in this layer?

A: Parameters: (11*11*3)*96 = 35K

Chetan Arora
Department of Computer Science and Engineering, IIT Delhi

AlexNet



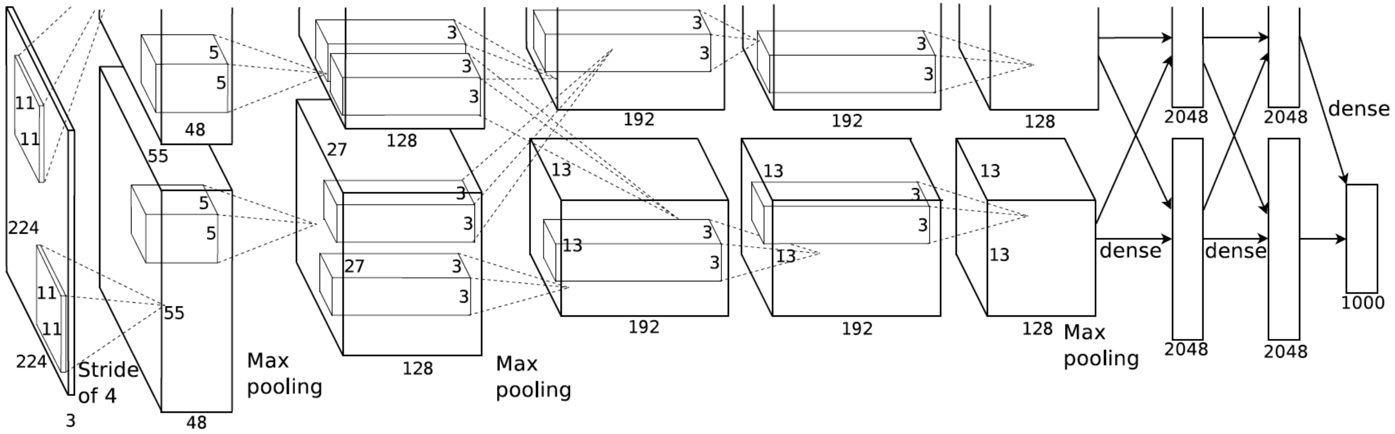
- Input: 227x227x3
- After CONV1: 55x55x96

- Second layer (POOL1): 3x3 filters applied at stride 2

Q: What is the output volume size?

Chetan Arora
Department of Computer Science and Engineering, IIT Delhi

AlexNet



- Input: 227x227x3
- After CONV1: 55x55x96

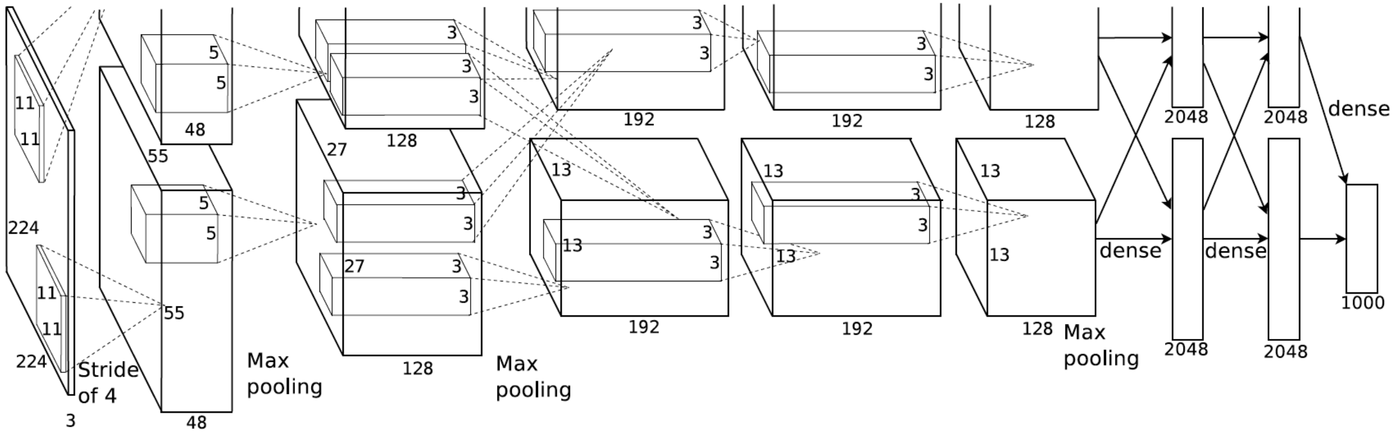
- Second layer (POOL1): 3x3 filters applied at stride 2

Q: What is the output volume size?

A: (55-3)/2+1 = 27. Output volume: 27x27x96

Q: What is the number of parameters in this layer?

AlexNet



- Input: 227x227x3
- After CONV1: 55x55x96

- Second layer (POOL1): 3x3 filters applied at stride 2

Q: What is the output volume size?

A: (55-3)/2+1 = 27. Output volume: 27x27x96

Q: What is the number of parameters in this layer?

A: Parameters: 0!

AlexNet (2012): Key Ideas

- Downsampled images
  - shorter dimension 256 pixels, longer dimension cropped about center to 256 pixels
  - R, G, B channels
- Mean subtraction from inputs

AlexNet (2012): Key Ideas

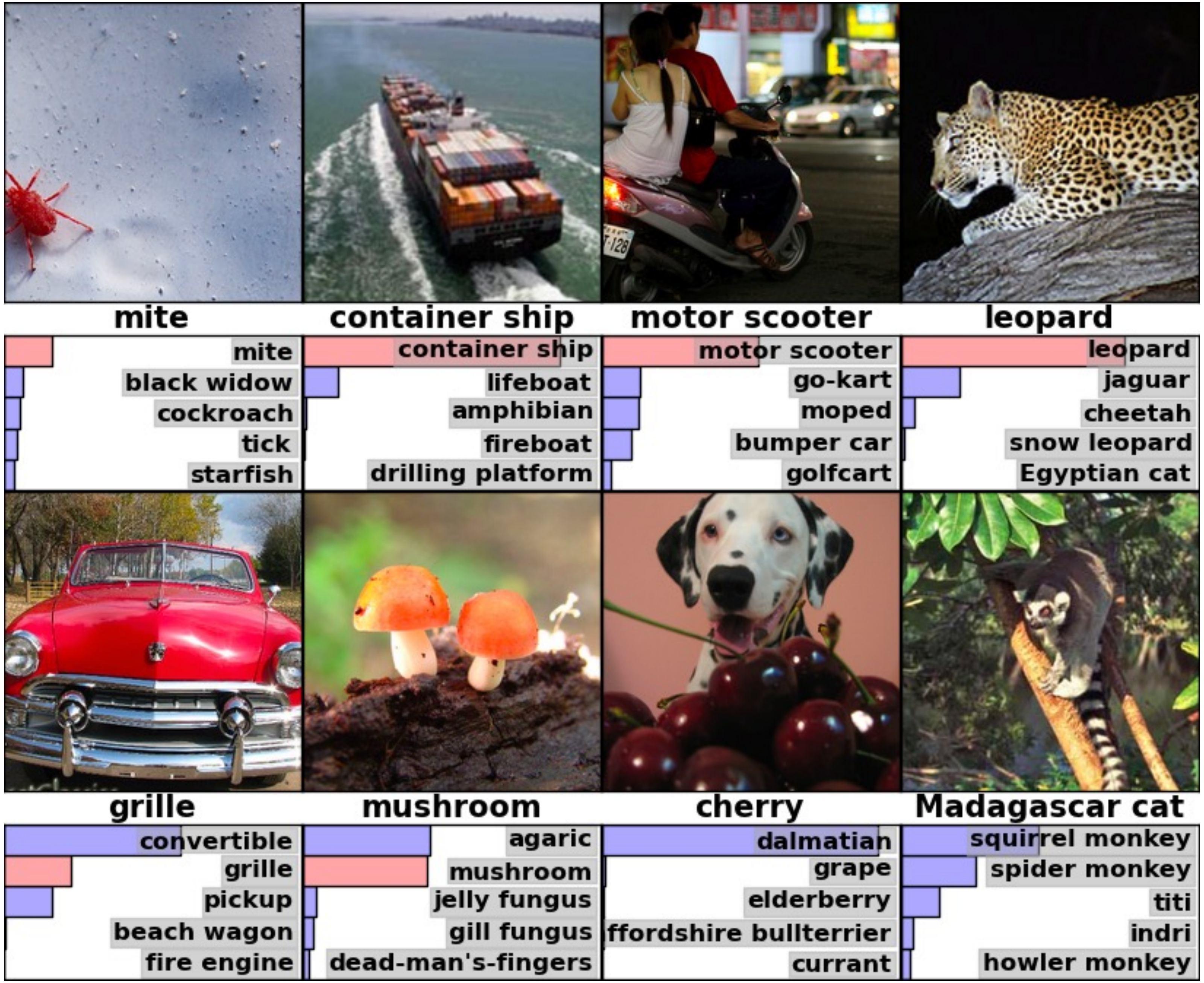
- Data set augmentation
  - Generate image translations by selecting random 224 x 224 sub-images
  - Horizontal reflections (standard trick in computer vision)
  - When testing, extract 10 distinct 224x224 sub-images and average predictions
- More data set augmentation
  - Vary intensity and color of the illumination from epoch to epoch

Chetan Arora
Department of Computer Science and Engineering, IIT Delhi

AlexNet (2012): Key Ideas

- ReLU instead of logistic or tanh units
- DropOut

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

Model	Top-1	Top-5
Sparse coding [2]	*47.1%*	*28.2%*
SIFT + FVs [24]	*45.7%*	*25.7%*
CNN	**37.5%**	**17.0%**

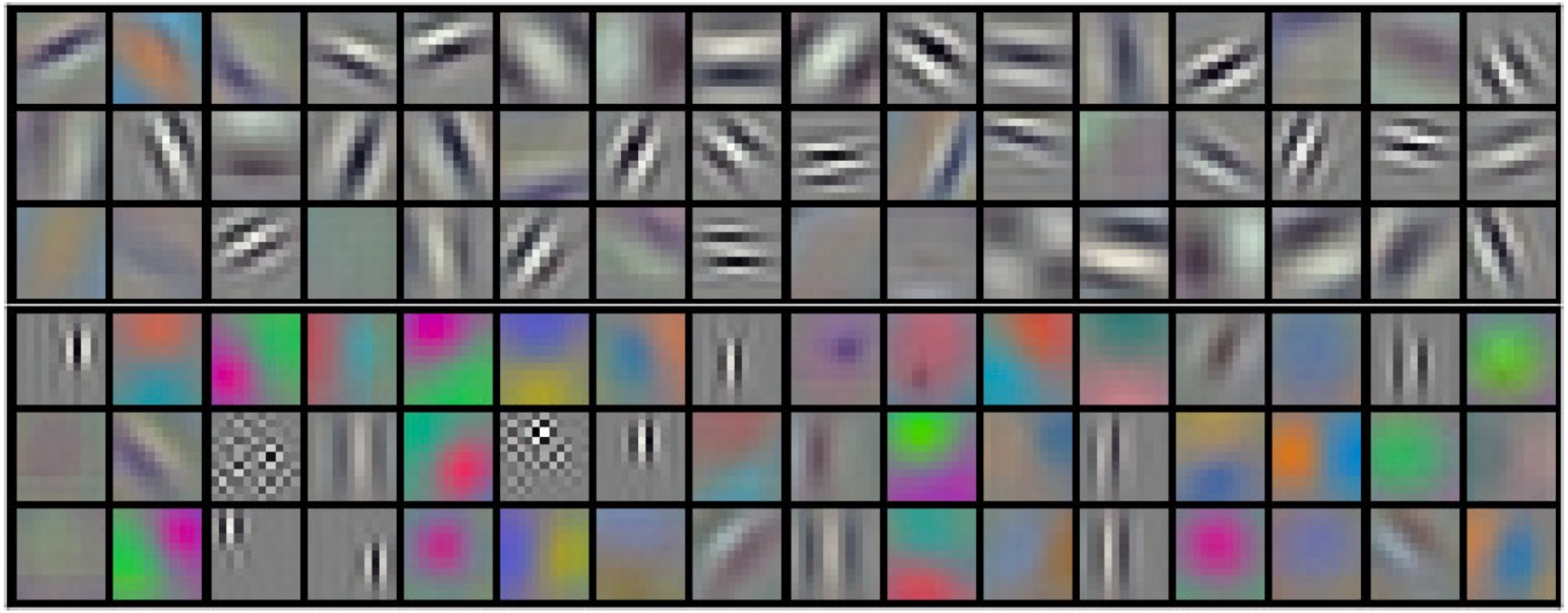
Table 1: Comparison of results on ILSVRC-2010 test set. In *italics* are best results achieved by others.

Model	Top-1 (val)	Top-5 (val)	Top-5 (test)
SIFT + FVs [7]	—	—	*26.2%*
1 CNN	40.7%	18.2%	—
5 CNNs	38.1%	16.4%	**16.4%**
1 CNN*	39.0%	16.6%	—
7 CNNs*	36.7%	15.4%	**15.3%**

Table 2: Comparison of error rates on ILSVRC-2012 validation and test sets. In *italics* are best results achieved by others. Models with an asterisk* were "pre-trained" to classify the entire ImageNet 2011 Fall release. See Section 6 for details.

Visualization

- 96 convolutional kernels of size 11X11X3 learned by the first convolutional layer on the 224X224X3 input images.
- The top 48 kernels were learned on GPU 1 while the bottom 48 kernels were learned on GPU 2



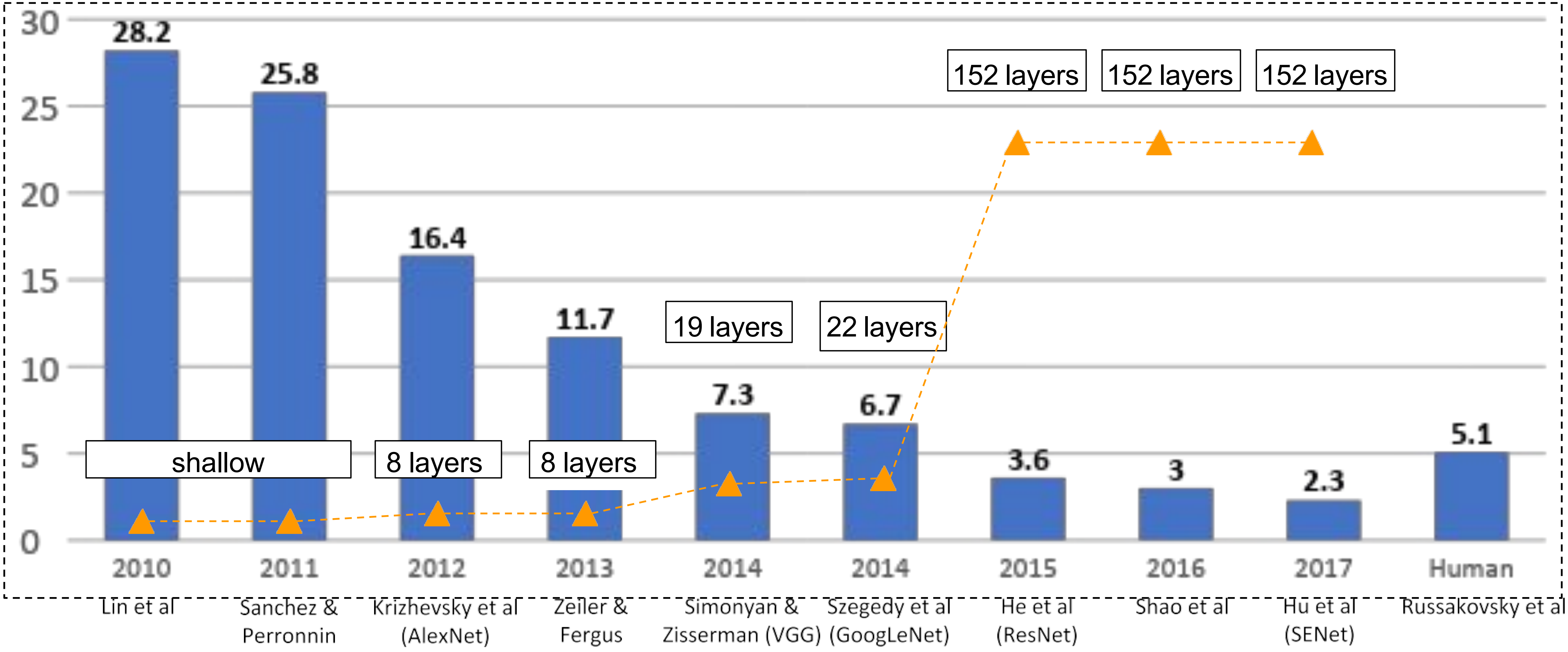
Generic Feature Vectors?

- Five ILSVRC-2010 test images in the first column.

- The remaining columns show the 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.

Chetan Arora
Department of Computer Science and Engineering, IIT Delhi

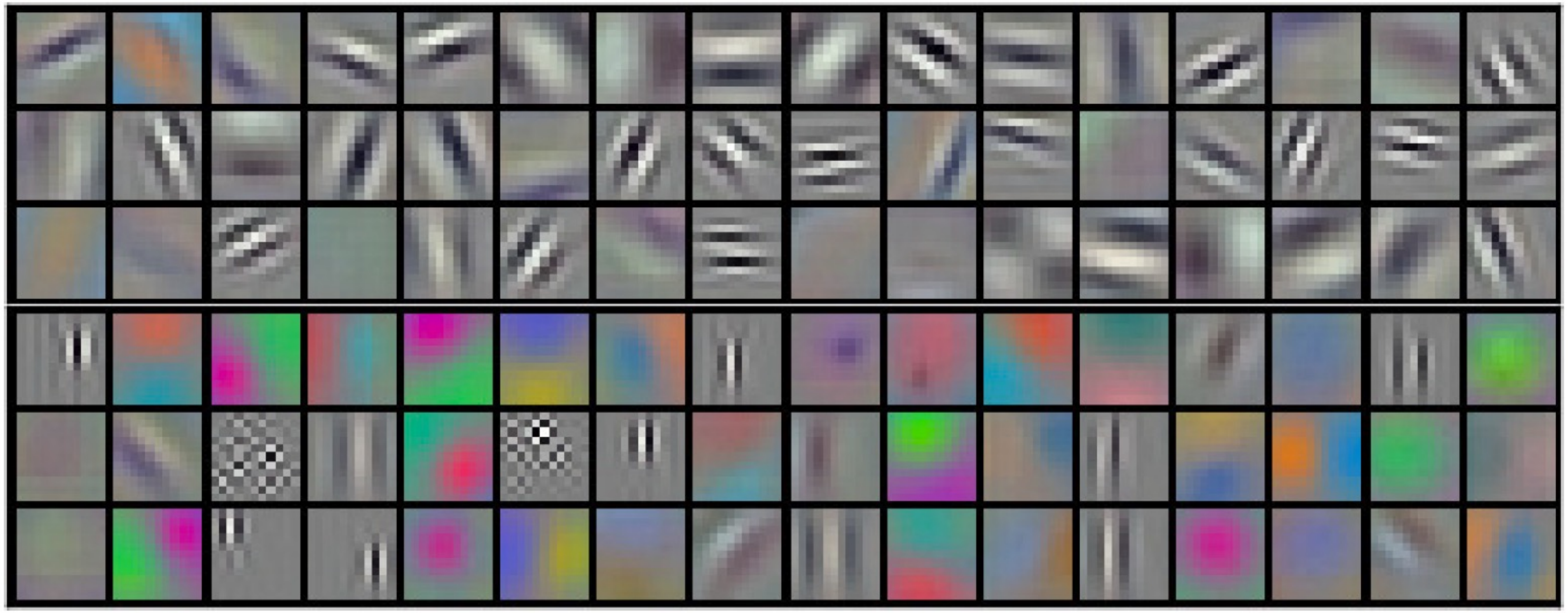
ILSVRC Winners



CS231N Stanford

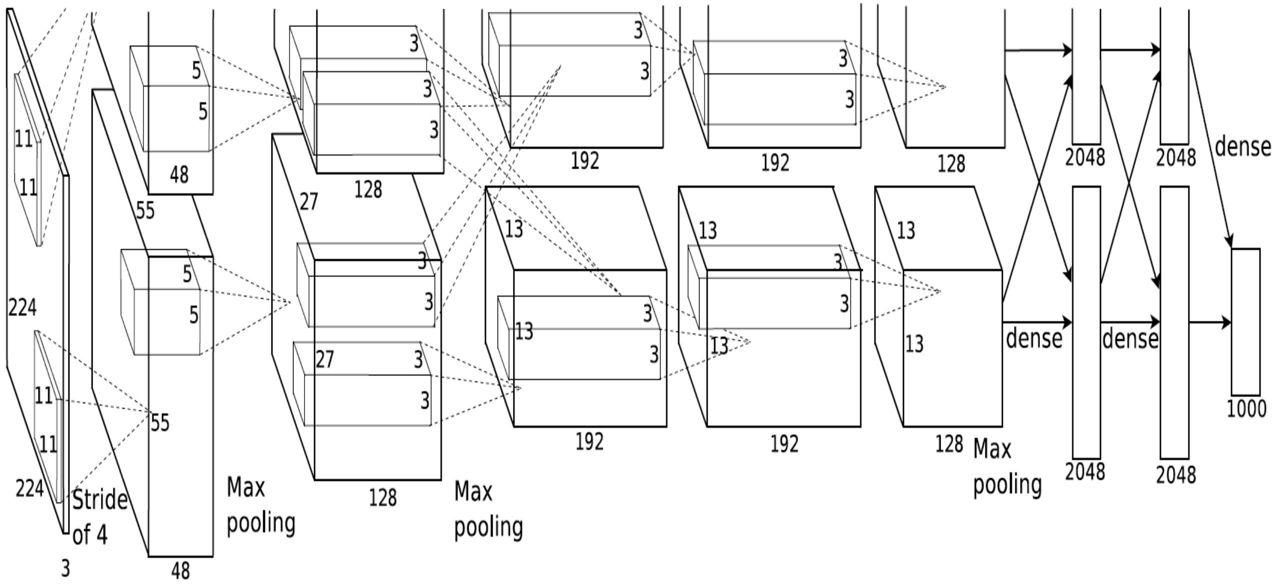
AlexNet: Some Obvious Problems

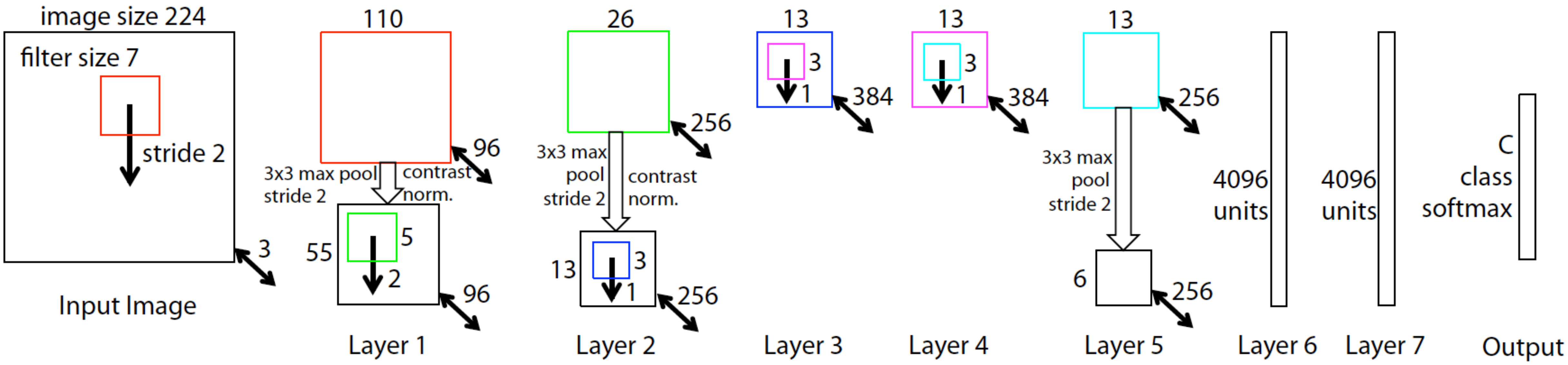
- The first layer filters are a mix of extremely high and low frequency information, with little coverage of the mid frequencies.
- 2nd layer visualization shows aliasing artefacts caused by the large stride(4) used in the 1st layer convolutions.



ZFNet

- 1st layer filter size: 11x11 to 7x7
- Stride of the convolution: 4 to 2
- Sparse connections to Dense (used in AlexNet due to the model being split across 2 GPUs)





Matthew D. Zeiler and Rob Fergus: Visualizing and Understanding Convolutional Networks

VGGNet: Key Ideas

- Rather than using relatively large receptive fields in the first conv. layers (11X11 with stride 4 in AlexNet, 7X7 with stride 2 in ZFNet), use very small 3X3 receptive fields.

Advantage:

- A stack of three 3X3 conv. layers (without spatial pooling in between) has an effective receptive field of 7X7.
- Number of parameters in a conv. layer with C channel input and output: $(k \times k \times C) \times C$.
  - In a single 7X7 layer: $49 \times C^2$.
  - In three 3X3 layers: $3 \times 9 \times C^2$.
- Lesser parameters allows deeper networks.

Karen Simonyan & Andrew Zisserman: VERY DEEP CONVOLUTIONAL NETWORKS FOR LARGE-SCALE IMAGE RECOGNITION

VGGNet: Key Ideas

- The incorporation of 1×1 conv. layers

Advantage:

- Increase the nonlinearity
- No affect in the receptive fields of the conv. layers.

Chetan Arora
Department of Computer Science and Engineering, IIT Delhi

VGGNet

Network	A,A-LRN	B	C	D	E
Number of parameters	133	133	134	138	144

*LRN = Local Response Normalization (as in AlexNet)

ConvNet Configuration					
A	A-LRN	B	C	D	E
11 weight layers	11 weight layers	13 weight layers	16 weight layers	16 weight layers	19 weight layers
input (224×224 RGB image)					
conv3-64	conv3-64 **LRN**	conv3-64 **conv3-64**	conv3-64 conv3-64	conv3-64 conv3-64	conv3-64 conv3-64
maxpool					
conv3-128	conv3-128	conv3-128 **conv3-128**	conv3-128 conv3-128	conv3-128 conv3-128	conv3-128 conv3-128
maxpool					
conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256 **conv1-256**	conv3-256 conv3-256 **conv3-256**	conv3-256 conv3-256 conv3-256 **conv3-256**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
FC-4096					
FC-4096					
FC-1000					
soft-max					

Network in Network (NIN)

- The convolutional layers generate feature maps by linear convolutional filters followed by nonlinear activation functions.
- Linear convolution is sufficient for abstraction when the instances of the latent concepts are linearly separable. Achievable level of abstraction is low.

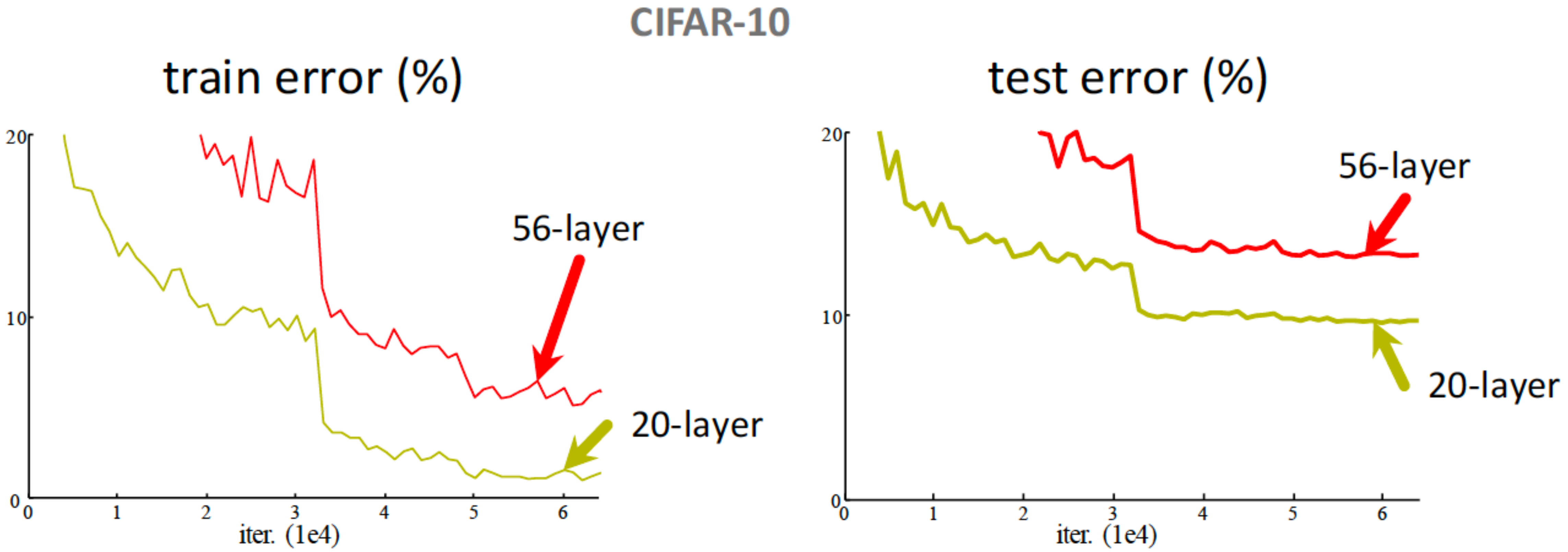
Key Idea:

- It is beneficial to do a better abstraction on each local patch, before combining them into higher level concepts.
- Replaced a convolution with a “micro network” structure which is a better/more general nonlinear function approximator.

Min Lin, Qiang Chen, and Shuicheng Yan: Network In Network

Deeper Networks

- Stacking 3x3 conv layers.
- 56-layer net has higher training error and test error than 20-layer net



Kaiming He, Xiangyu Zhang, Shaoqing Ren, & Jian Sun. "Deep Residual Learning for Image Recognition". CVPR 2016

Deeper Networks

- The degradation is not caused by overfitting (training error is also high)
- The degradation indicates that not all systems are similarly easy to optimize.

Shallow → Deep

- Consider a shallower architecture and its deeper counterpart that adds more layers onto it.
- There exists a solution by construction to the deeper model: the added layers are identity mapping. Other layers are copied from the learned shallower model.

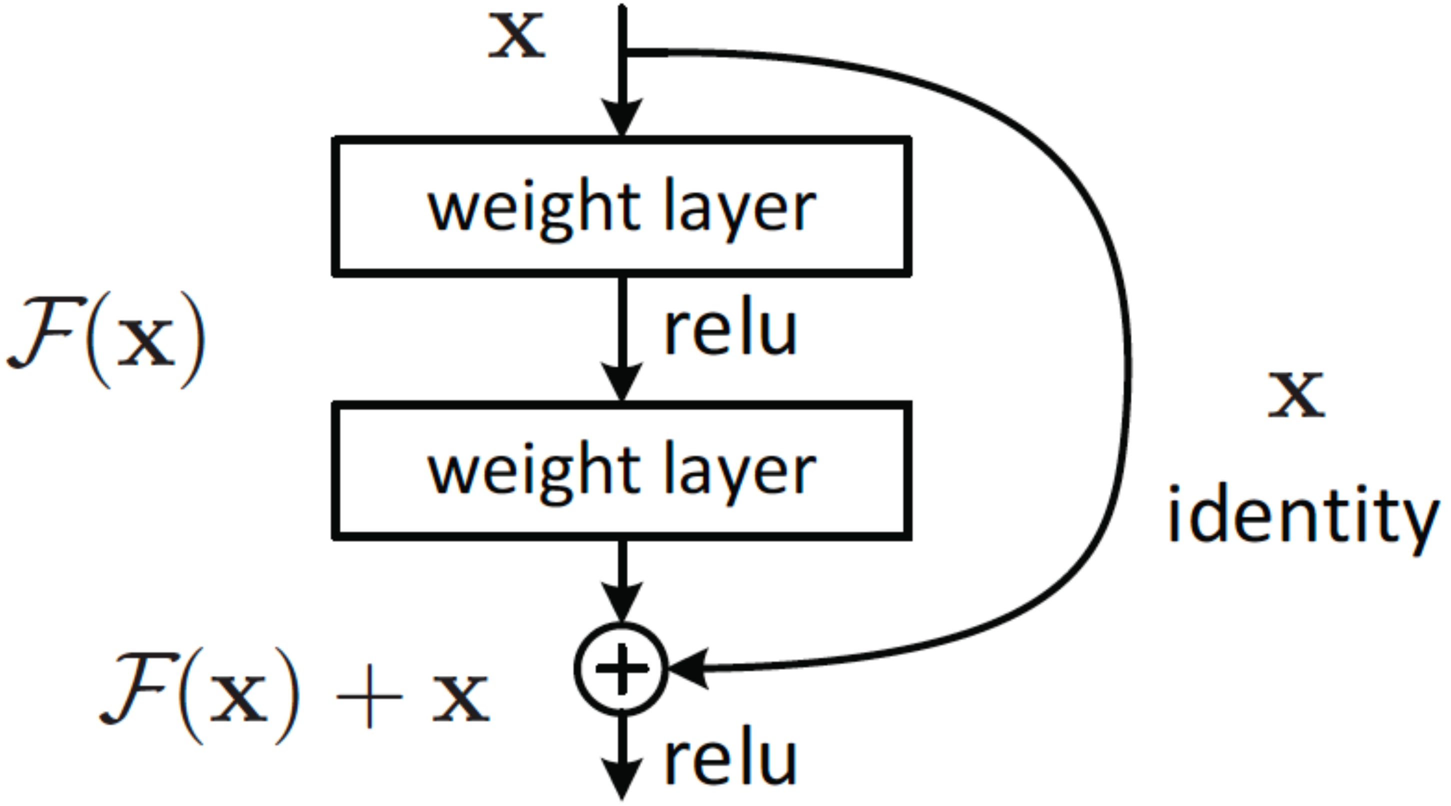
Kaiming He, Xiangyu Zhang, Shaoqing Ren, & Jian Sun. “Deep Residual Learning for Image Recognition”. CVPR 2016

Residual Learning

- Instead of hoping each few stacked layers directly fit a desired underlying mapping, explicitly let these layers fit a residual mapping.
- Denoting the desired underlying mapping as H(x). The stacked nonlinear layers fit another mapping:

$$F(x) = H(x) - x$$

- The original mapping is recast into $F(x) + x$.

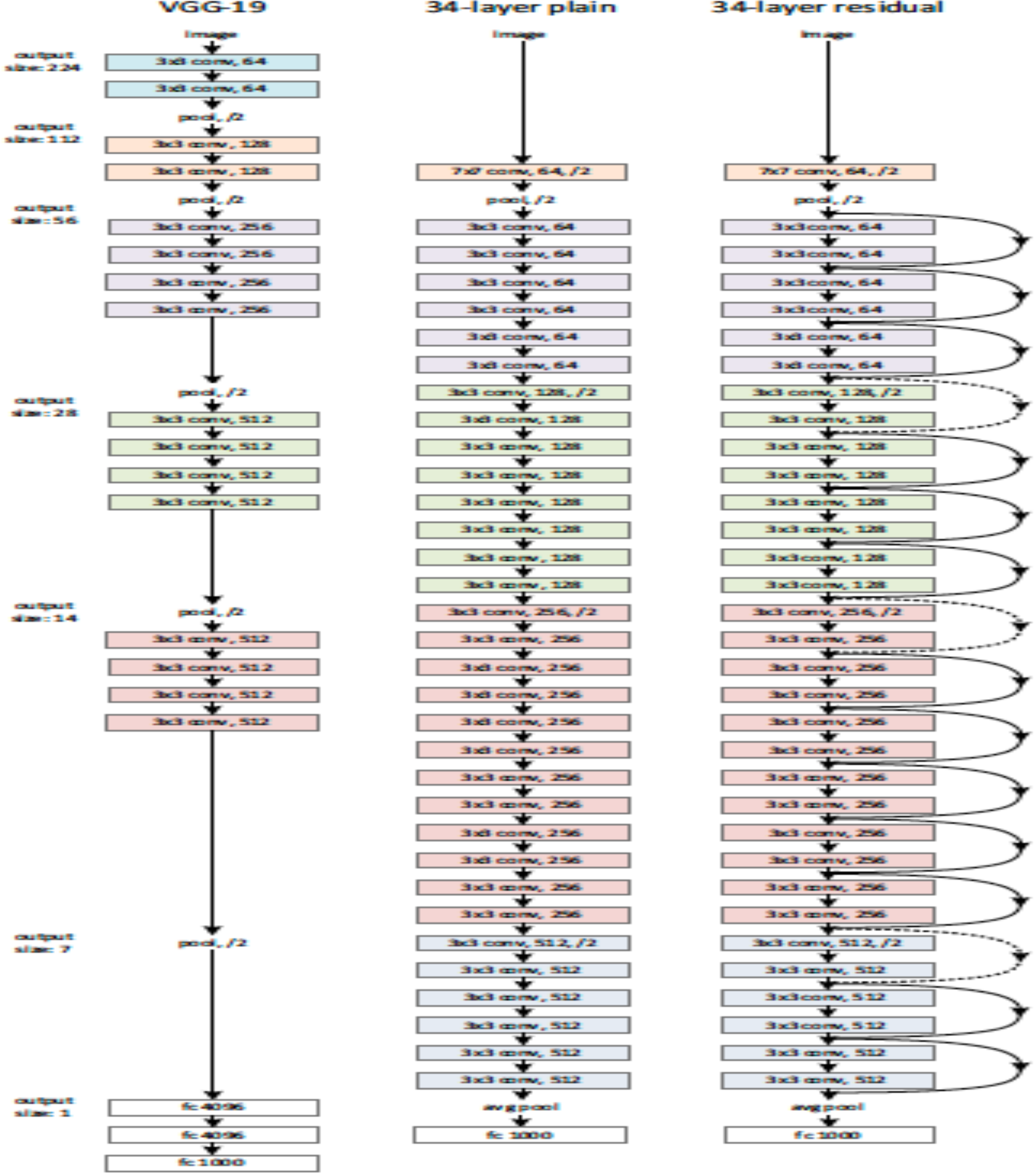


Kaiming He, Xiangyu Zhang, Shaoqing Ren, & Jian Sun. "Deep Residual Learning for Image Recognition". CVPR 2016

ResNet

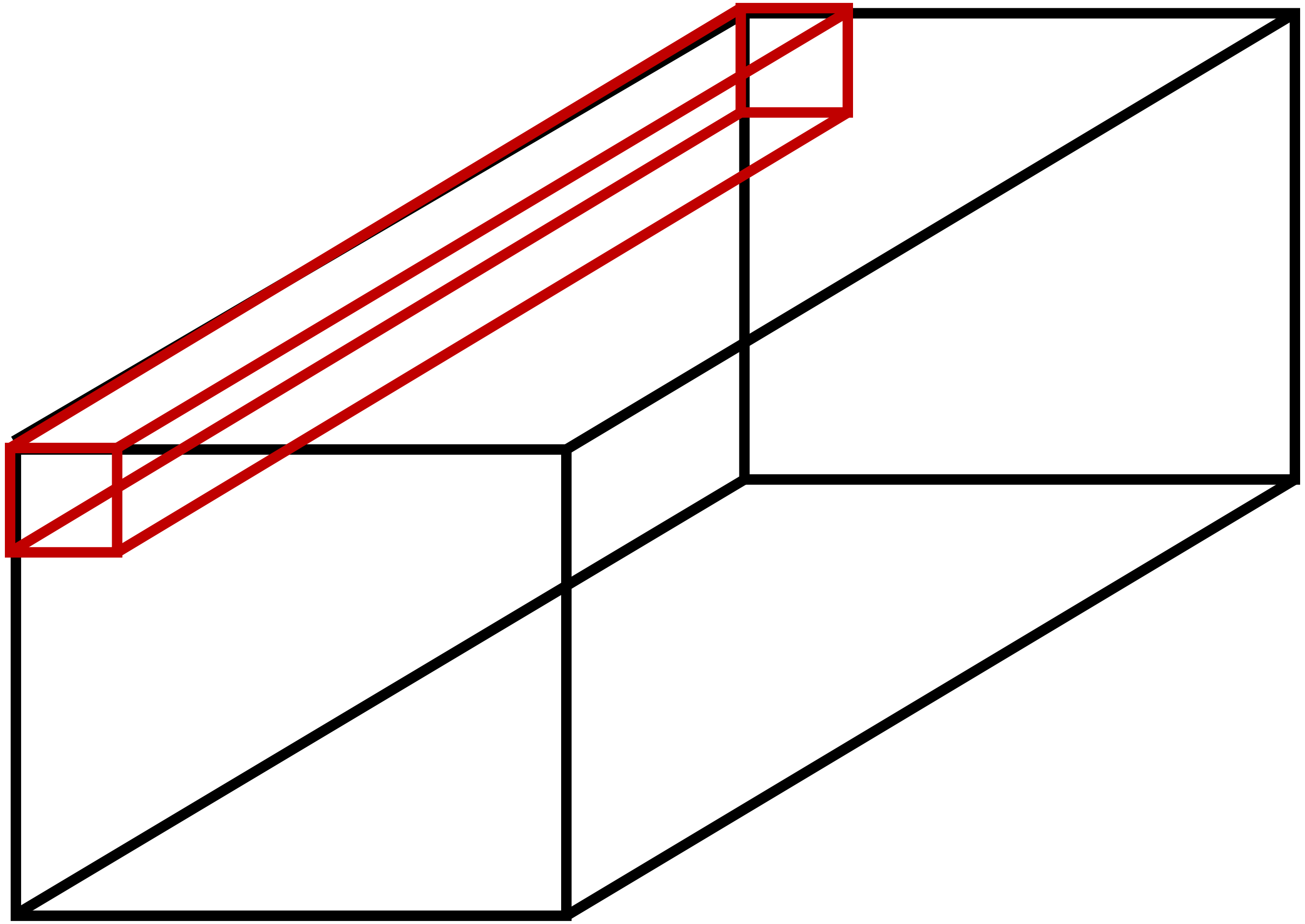
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, & Jian Sun. “Deep Residual Learning for Image Recognition”. CVPR 2016

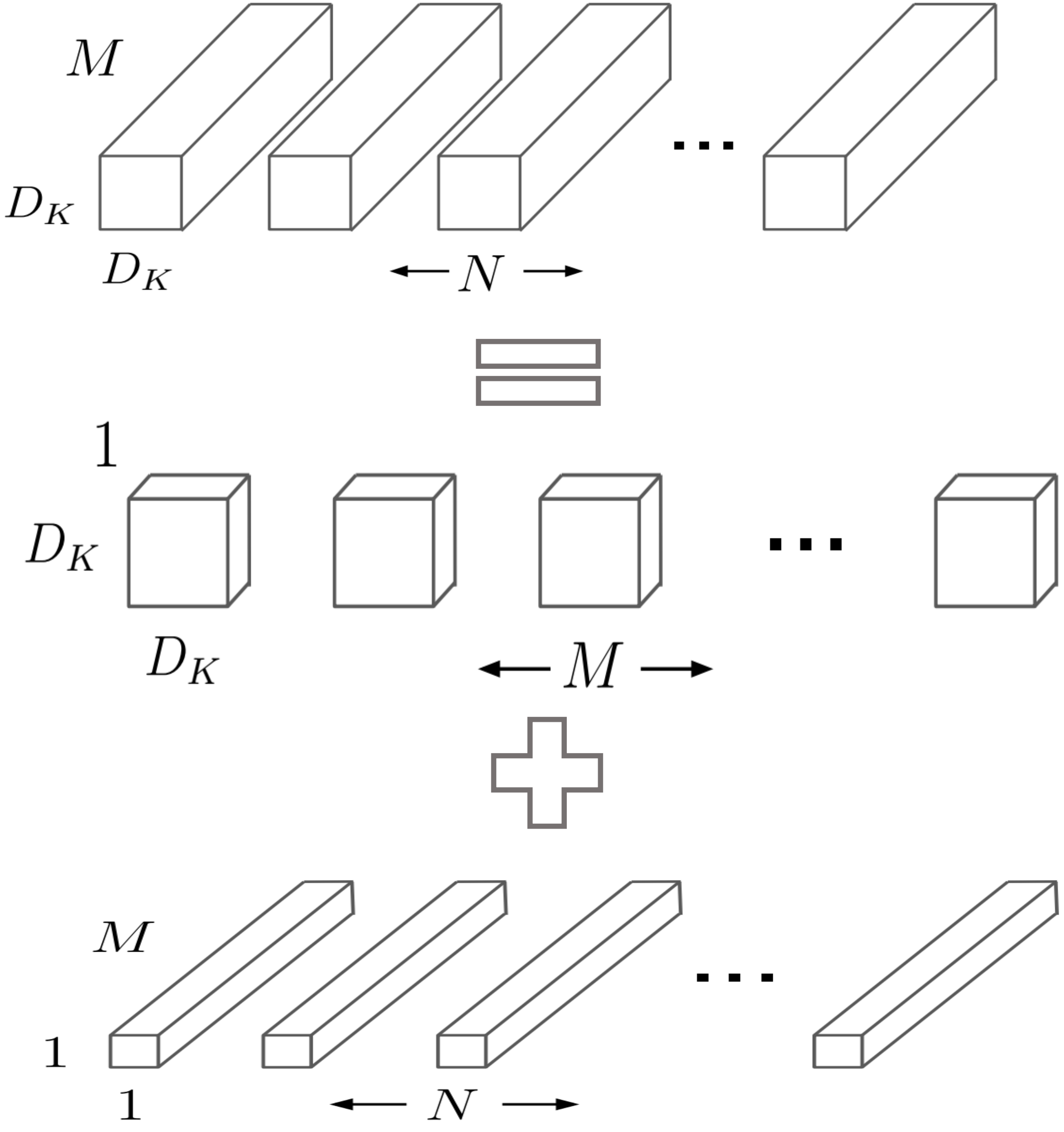
- Stack residual learning modules



Chetan Arora
Department of Computer Science and Engineering, IIT Delhi

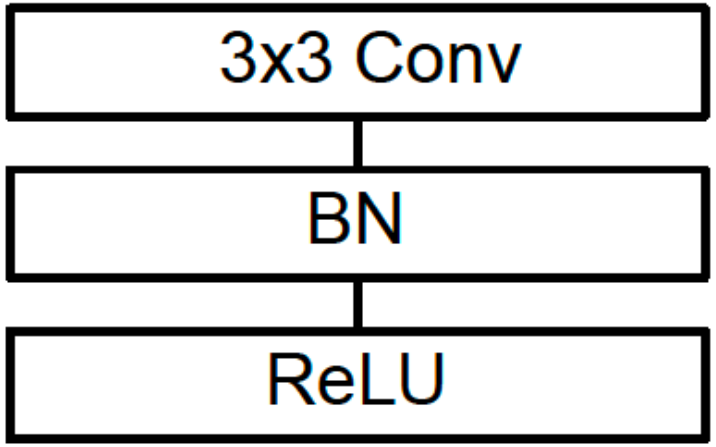
MobileNets



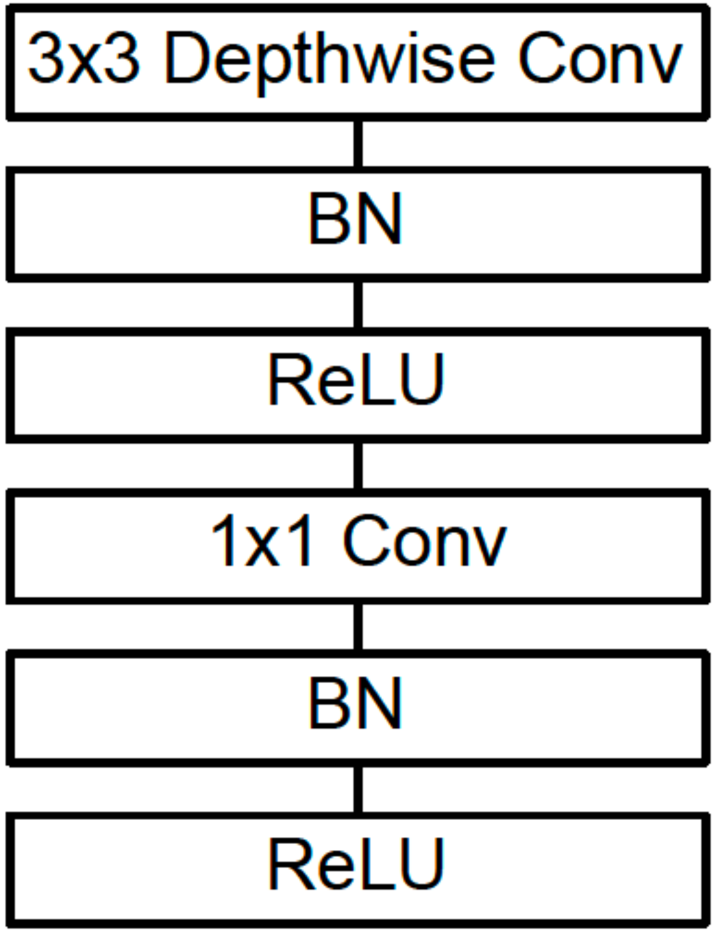


A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, H. Adam
MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. ArXiv 2017

MobileNets



Original Block



MobileNet Block

Table 8. MobileNet Comparison to Popular Models

Model	ImageNet Accuracy	Million Mult-Adds	Million Parameters
1.0 MobileNet-224	70.6%	569	4.2
GoogleNet	69.8%	1550	6.8
VGG 16	71.5%	15300	138

Table 9. Smaller MobileNet Comparison to Popular Models

Model	ImageNet Accuracy	Million Mult-Adds	Million Parameters
0.50 MobileNet-160	60.2%	76	1.32
Squeezenet	57.5%	1700	1.25
AlexNet	57.2%	720	60

Andrew G. Howard Menglong Zhu Bo Chen Dmitry Kalenichenko
MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications