Deep Learning
Part 3

Yin Li

yin.li@wisc.edu

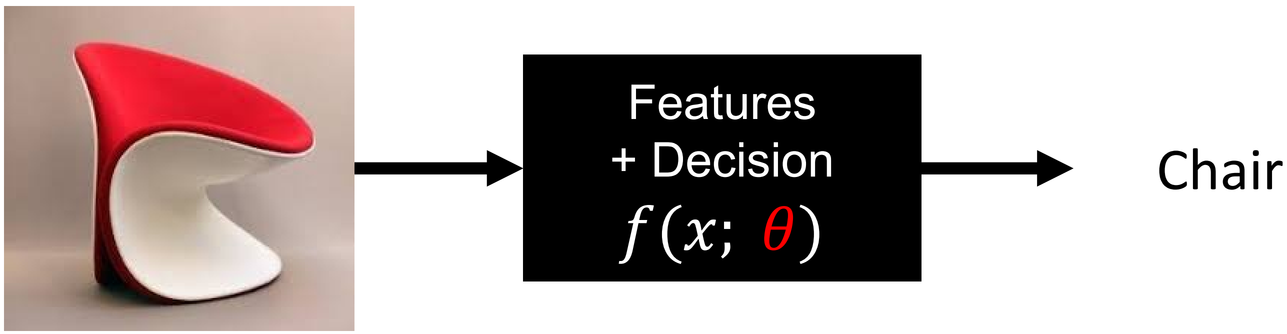
University of Wisconsin, Madison

[Some of the slides from Lana Lazebnik, Kaiming He and others]

Deep Learning = Deep Neural Networks

- Deep Learning: Composing a set of (nonlinear) functions g

$$f(\boldsymbol{x};\boldsymbol{\theta}) = g_1(\dots g_{n-1}(g_n(\boldsymbol{x};\boldsymbol{\theta_n}),\boldsymbol{\theta_{n-1}})\dots,\boldsymbol{\theta_1})$$



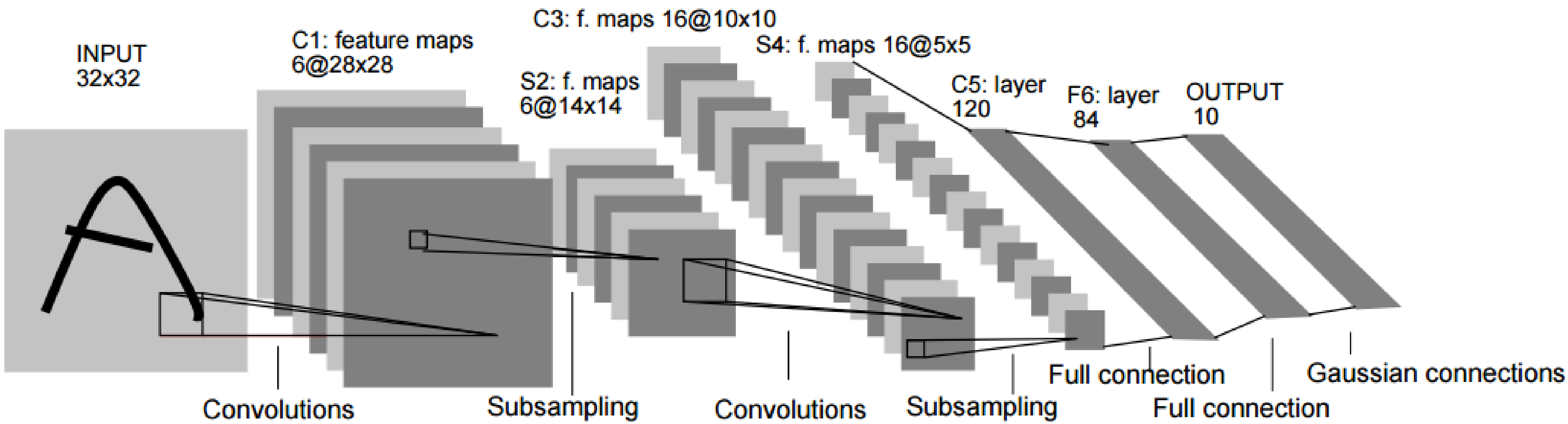
- Each of the function g is represented using a layer of a neural network

- Key element: Linear operations + Nonlinear activations, e.g., $\mathbf{a} = \sigma\left(\boldsymbol{W}^{\boldsymbol{T}}\boldsymbol{x} + \boldsymbol{b}\right)$

What we have talked about so far

- The linear functions
  - Fully connected layer: dense $\boldsymbol{W}$
  - Convolutional layer: sparse and structured $\boldsymbol{W}$
- The nonlinear activations
  - Sigmoid
  - Rectified linear unit (ReLU)
  - Many others …
- Pooling
- Output normalization
- Loss functions

Case study: LeNet-5 (1998)

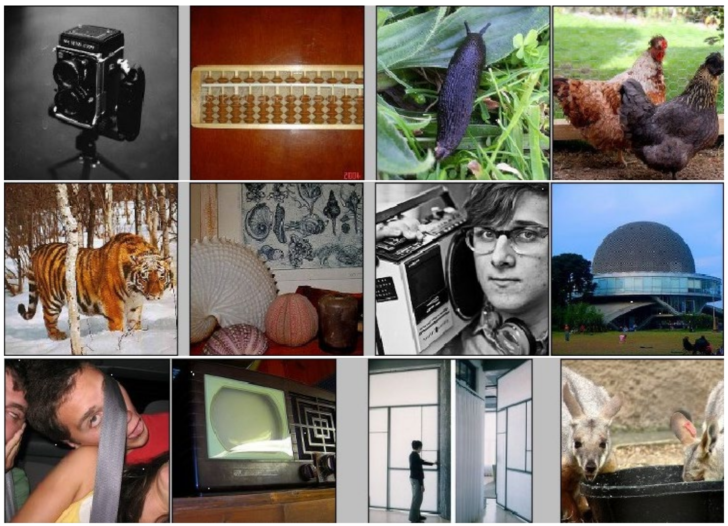


- convolutional layers + fully connected layers
- Sigmoid as the activation function
- [Conv + sigmoid + average pooling] x 2 + fully connected x 3

Figure from *Gradient-based learning applied to document recognition, by Y. LeCun, L. Bottou, Y. Bengio and P. Haffner*

ImageNet challenge (2010-2017)

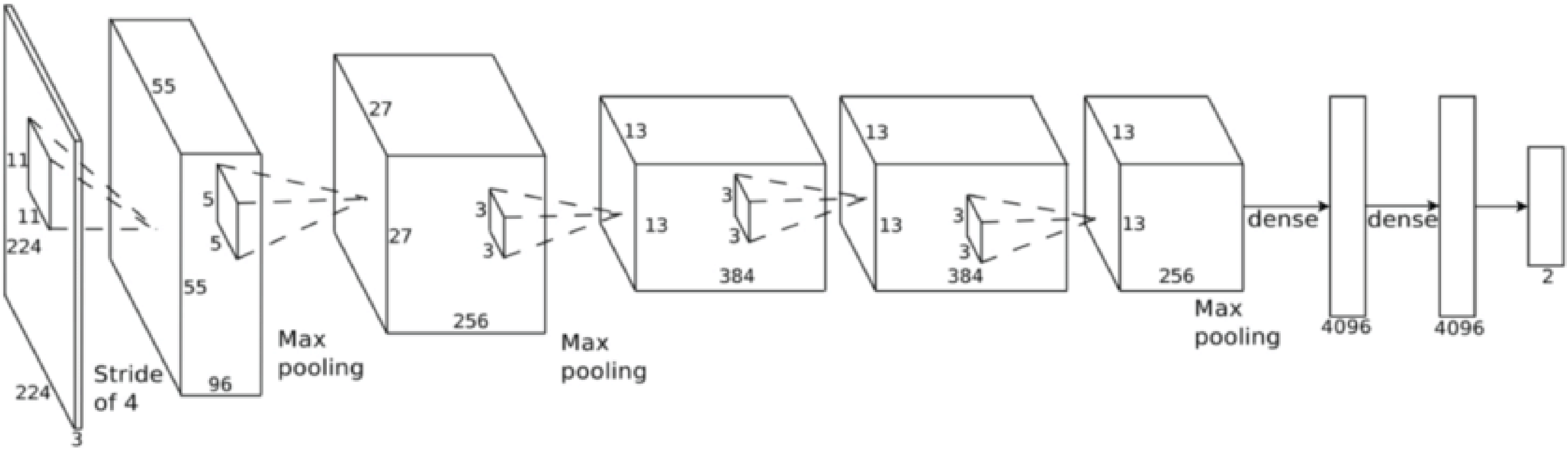
IMAGENET



- ~14 million labeled images, 20K classes
- Images gathered from Internet
- Labels provided by humans
- ImageNet Large-Scale Visual Recognition Challenge (ILSVRC):
  - 1.2 million training images, 1000 classes

www.image-net.org/challenges/LSVRC/

Case study: AlexNet (2012)



- A modern deep neural network
- Winner of ImageNet challenge 2012
- +10% better than everything else in 2012!

Figure from *ImageNet Classification with Deep Convolutional Neural Networks*, *by A. Krizhevsky, I. Sutskever, and G. Hinton*

Case study: AlexNet (2012)

Input: 224 x 224 x 3
(color images)

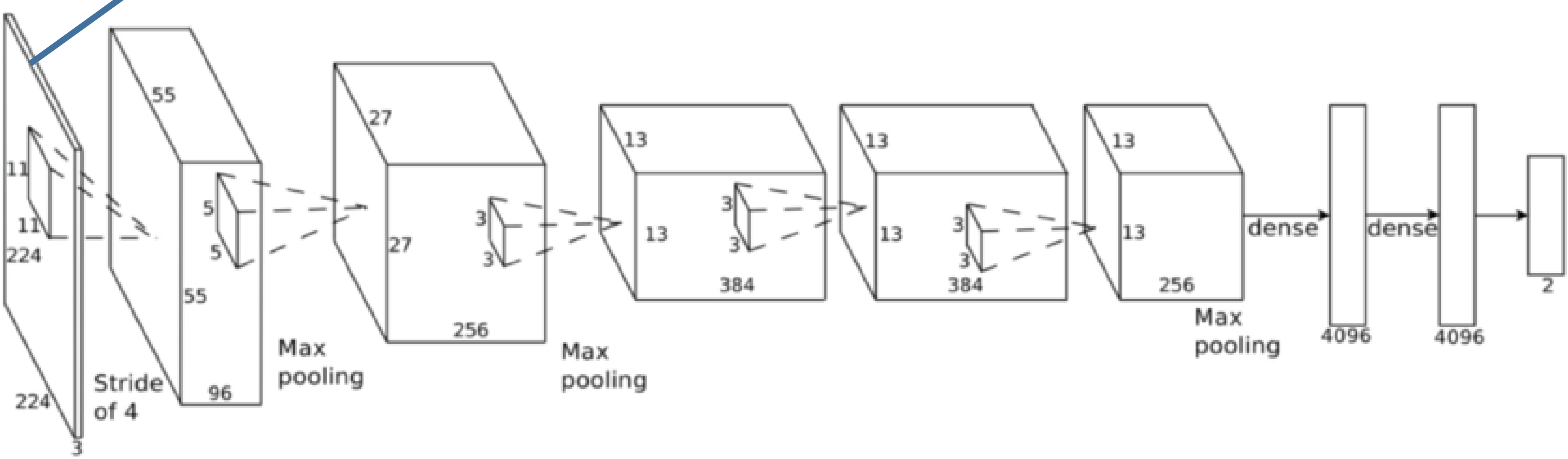


Figure from *ImageNet Classification with Deep Convolutional Neural Networks*,
by A. Krizhevsky, I. Sutskever, and G. Hinton

Case study: AlexNet (2012)

Filter: 11x11x3, stride: 4x4,
#filters: 96
Activation: ReLU

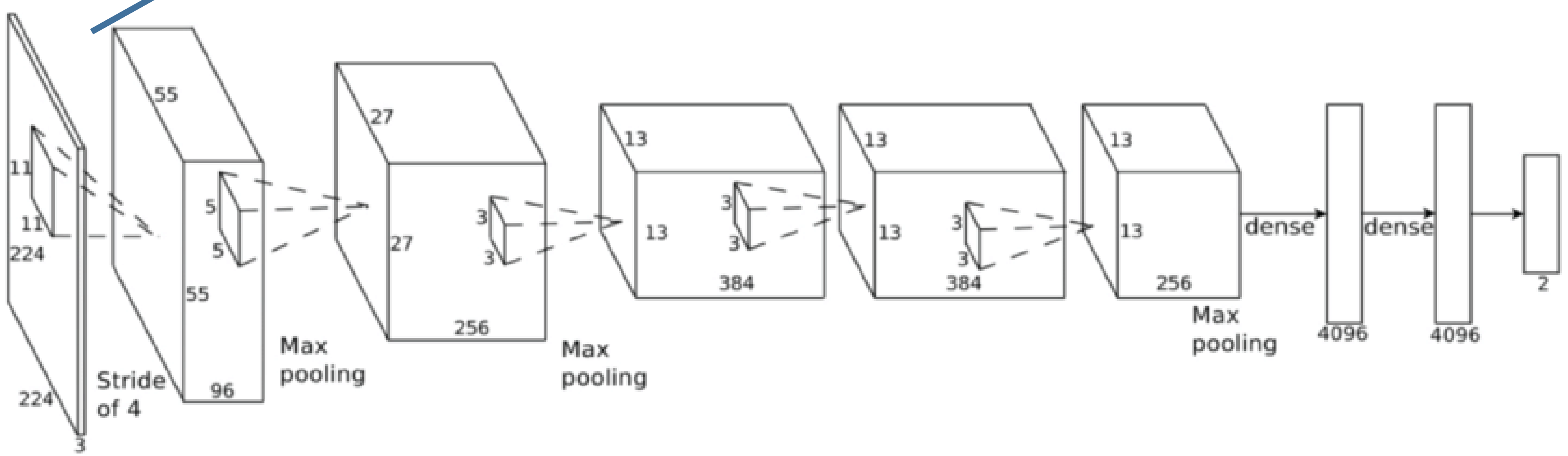


Figure from *ImageNet Classification with Deep Convolutional Neural Networks*,
by A. Krizhevsky, I. Sutskever, and G. Hinton

Case study: AlexNet (2012)

Max pooling
Filter: 3x3, stride 2x2

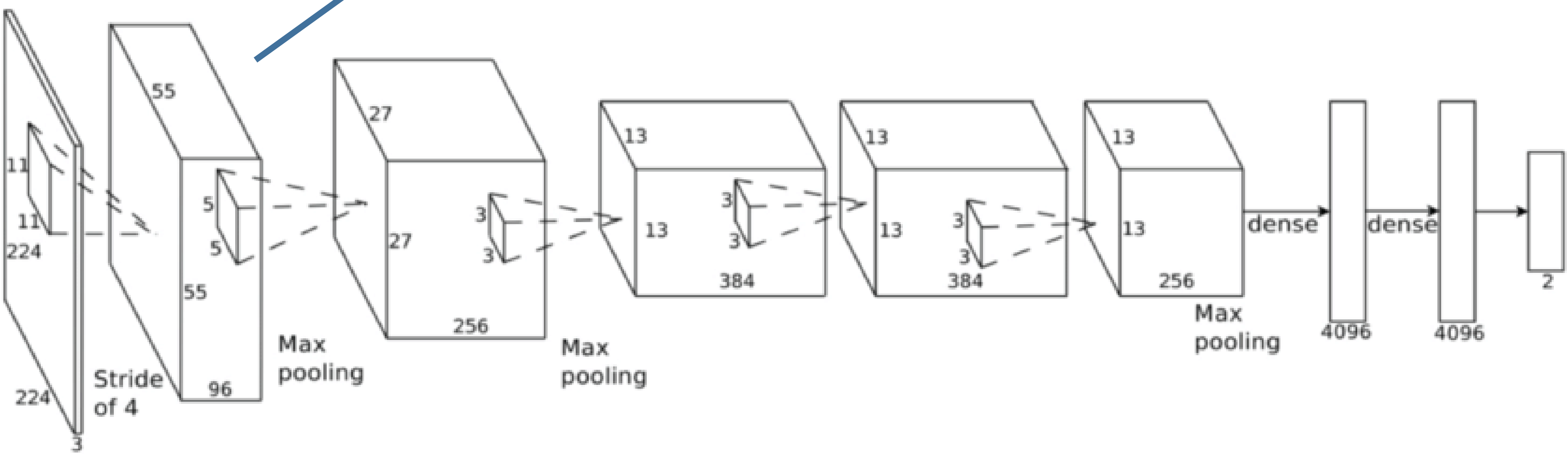


Figure from *ImageNet Classification with Deep Convolutional Neural Networks*, *by A. Krizhevsky, I. Sutskever, and G. Hinton*

Case study: AlexNet (2012)

Filter: 5x5x96, stride: 1x1,
#filters: 256
Activation: ReLU

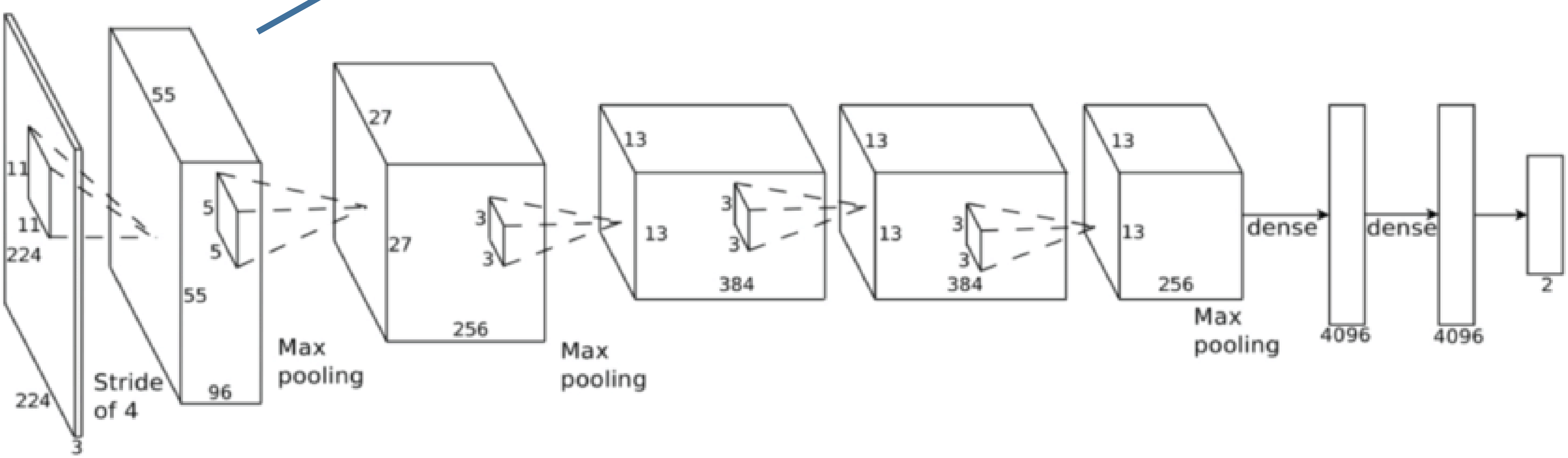


Figure from *ImageNet Classification with Deep Convolutional Neural Networks*, *by A. Krizhevsky, I. Sutskever, and G. Hinton*

Case study: AlexNet (2012)

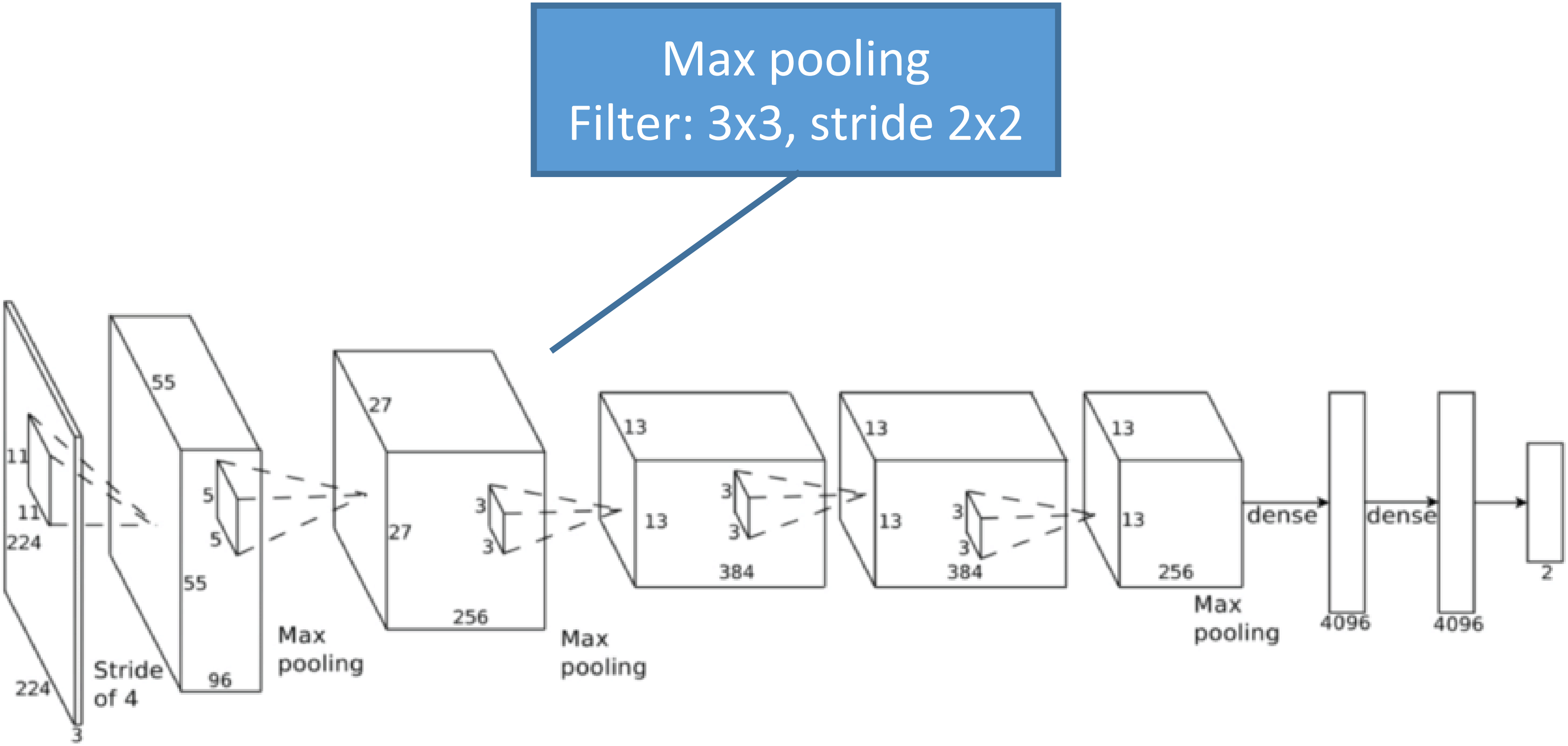


Figure from *ImageNet Classification with Deep Convolutional Neural Networks*, *by A. Krizhevsky, I. Sutskever, and G. Hinton*

Case study: AlexNet (2012)

Filter: 3x3x256, stride: 1x1,
#filters: 384
Activation: ReLU

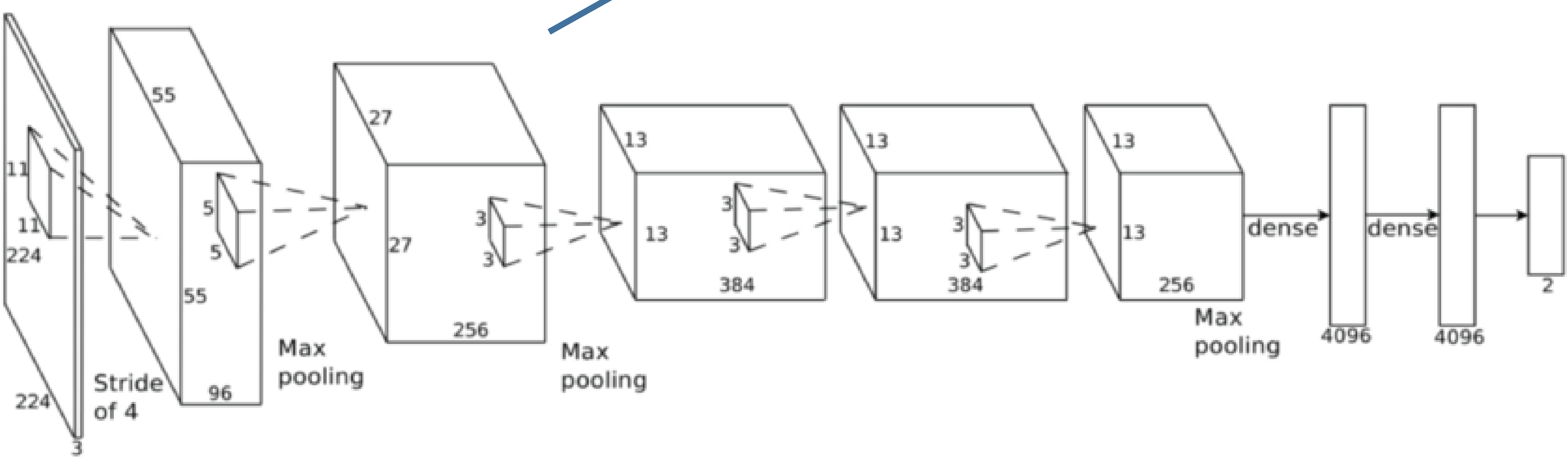


Figure from *ImageNet Classification with Deep Convolutional Neural Networks*, *by A. Krizhevsky, I. Sutskever, and G. Hinton*

Case study: AlexNet (2012)

Filter: 3x3x384, stride: 1x1,
#filters: 384
Activation: ReLU

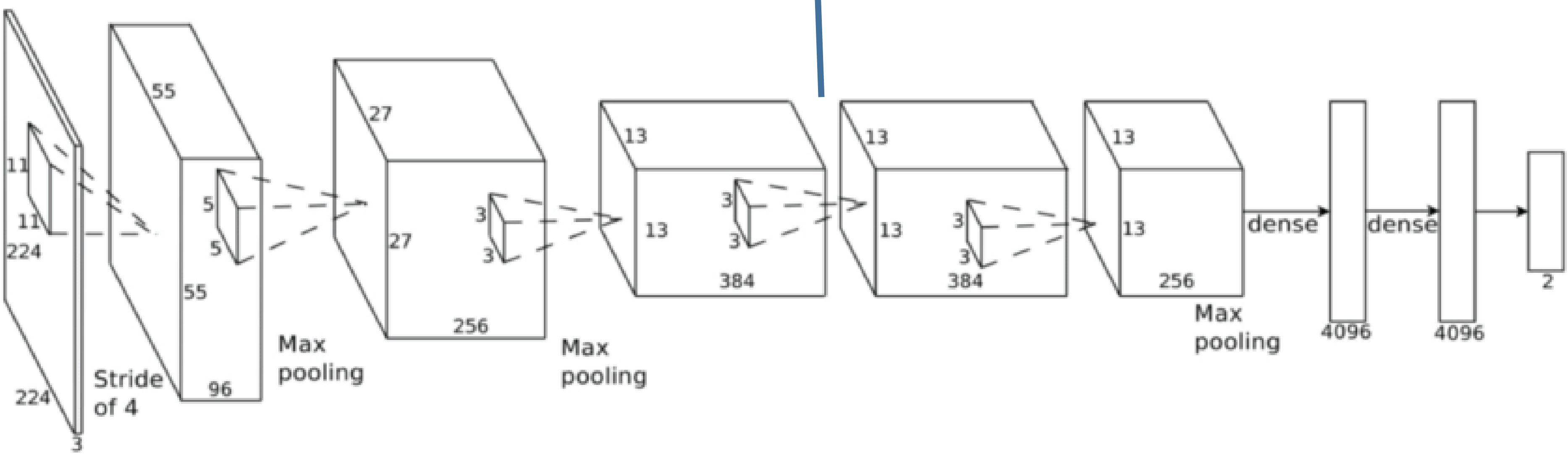


Figure from *ImageNet Classification with Deep Convolutional Neural Networks*, *by A. Krizhevsky, I. Sutskever, and G. Hinton*

Case study: AlexNet (2012)

Filter: 3x3x384, stride: 1x1,
#filters: 256
Activation: ReLU

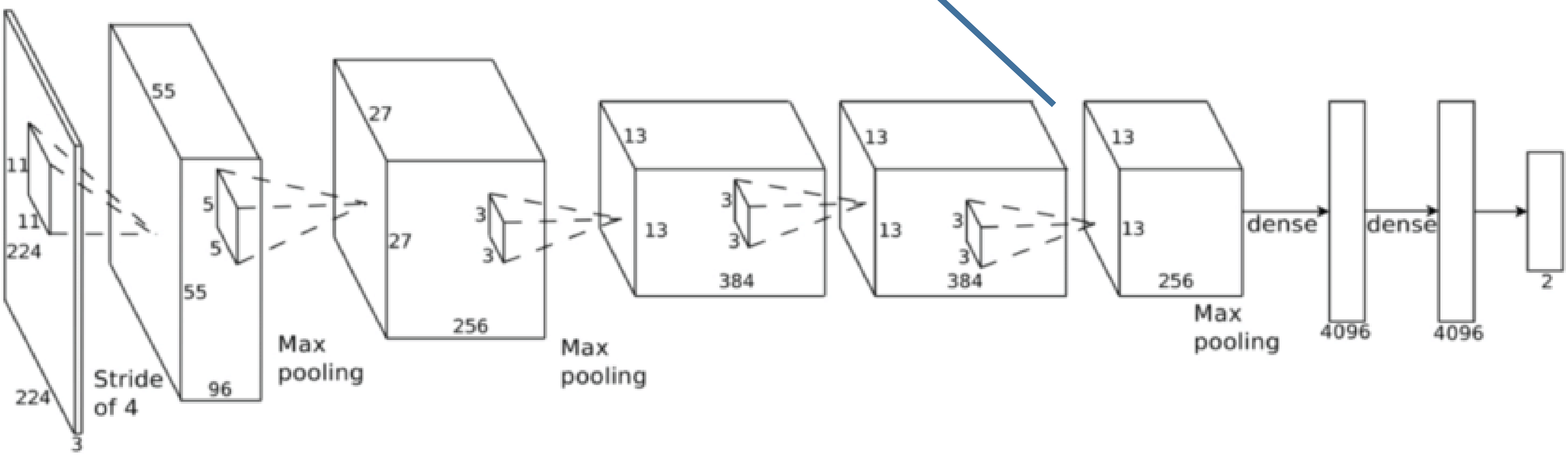


Figure from *ImageNet Classification with Deep Convolutional Neural Networks*, *by A. Krizhevsky, I. Sutskever, and G. Hinton*

Case study: AlexNet (2012)

Max pooling
Filter: 3x3, stride 2x2

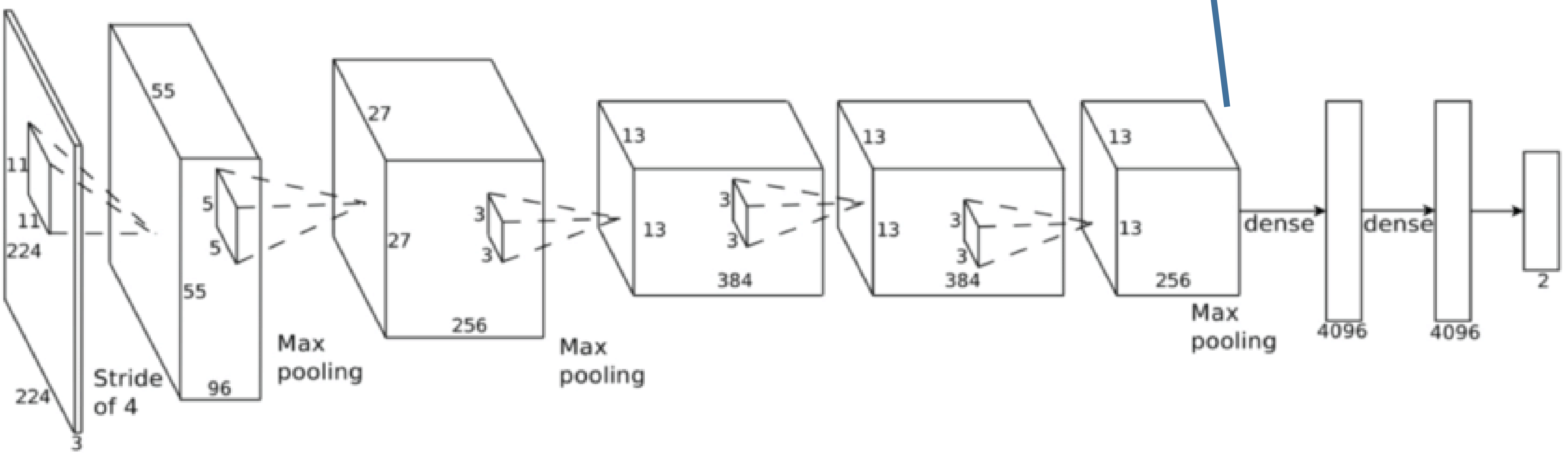


Figure from *ImageNet Classification with Deep Convolutional Neural Networks*, *by A. Krizhevsky, I. Sutskever, and G. Hinton*

Case study: AlexNet (2012)

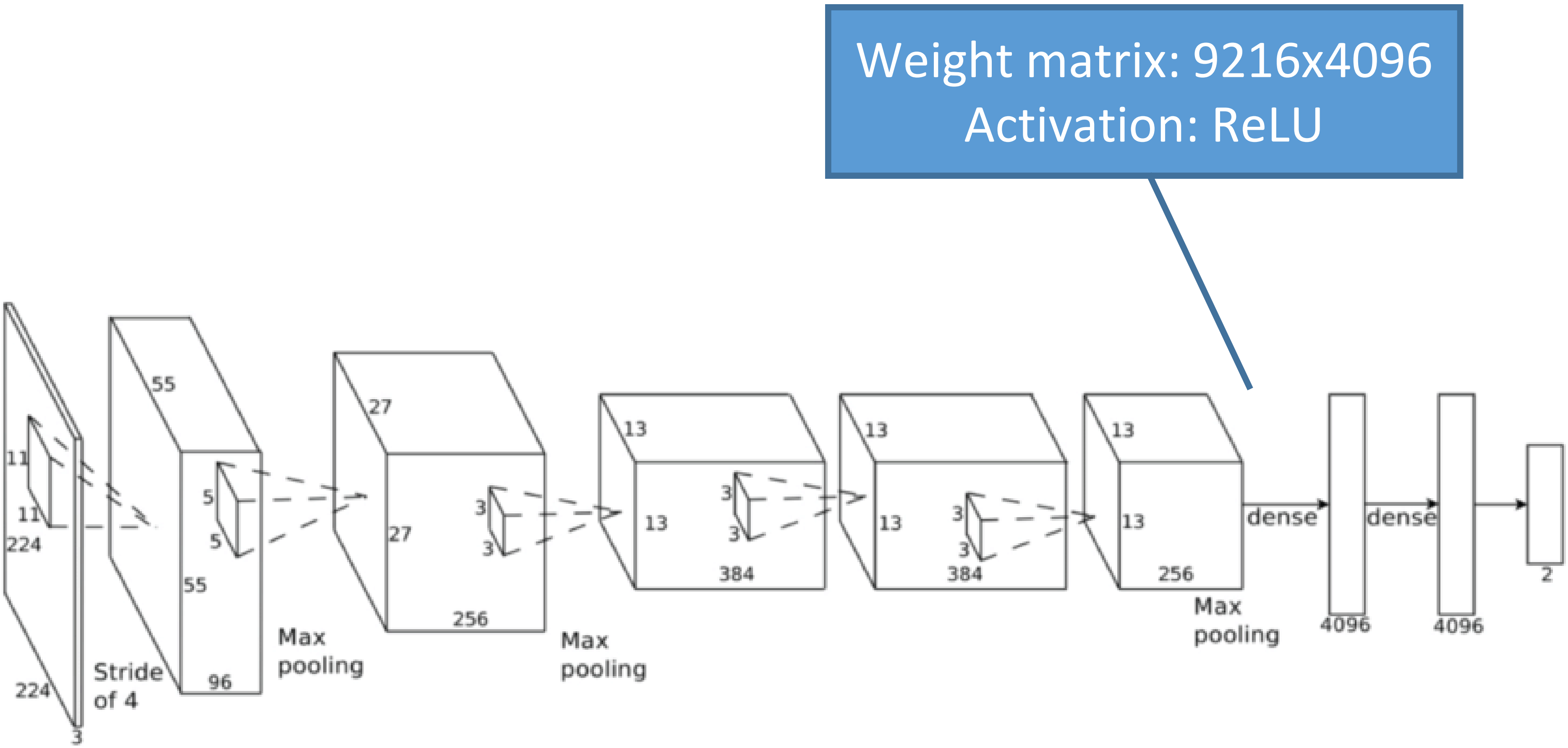


Figure from *ImageNet Classification with Deep Convolutional Neural Networks*, *by A. Krizhevsky, I. Sutskever, and G. Hinton*

Case study: AlexNet (2012)

Weight matrix: 4096x4096
Activation: ReLU

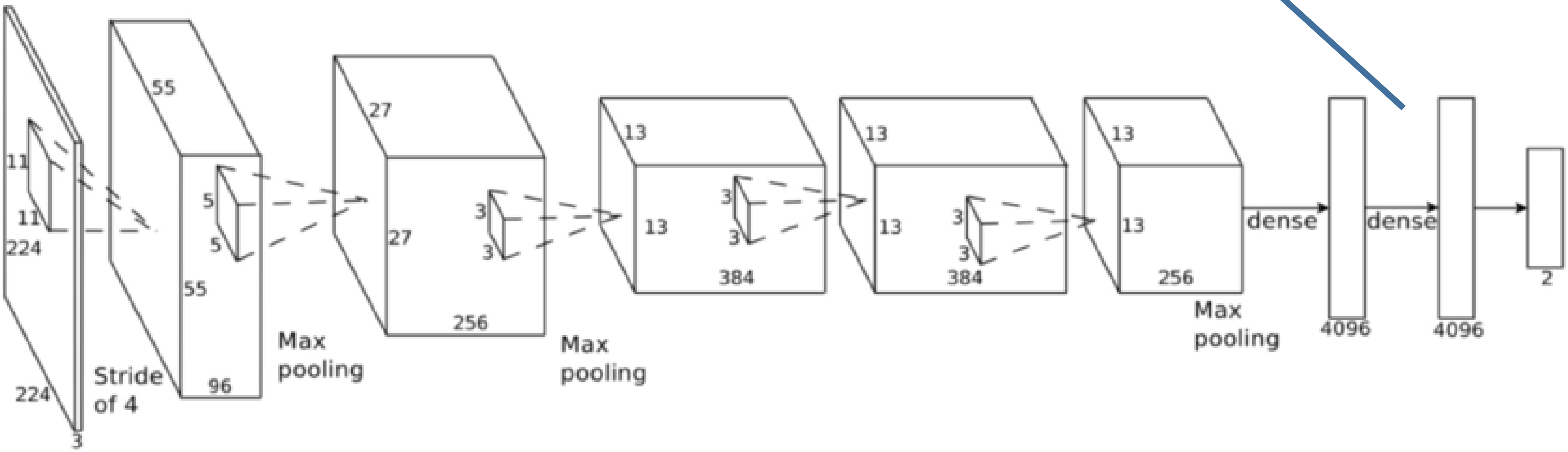


Figure from *ImageNet Classification with Deep Convolutional Neural Networks*, *by A. Krizhevsky, I. Sutskever, and G. Hinton*

Case study: AlexNet (2012)

Weight matrix: 4096x1000
Output normalization: Softmax

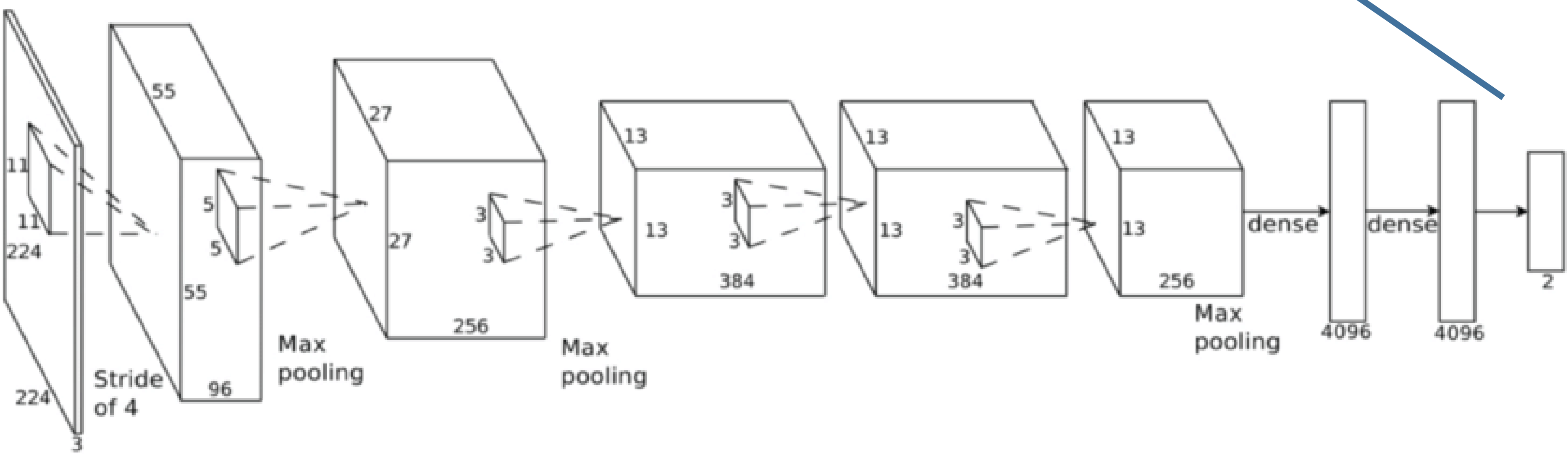
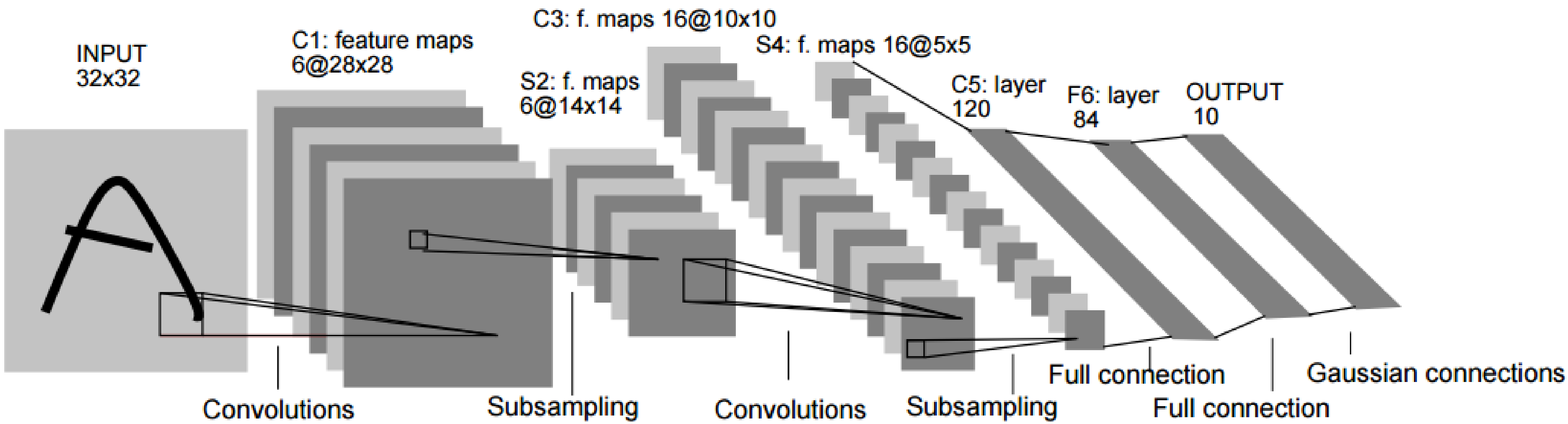


Figure from *ImageNet Classification with Deep Convolutional Neural Networks*, *by A. Krizhevsky, I. Sutskever, and G. Hinton*

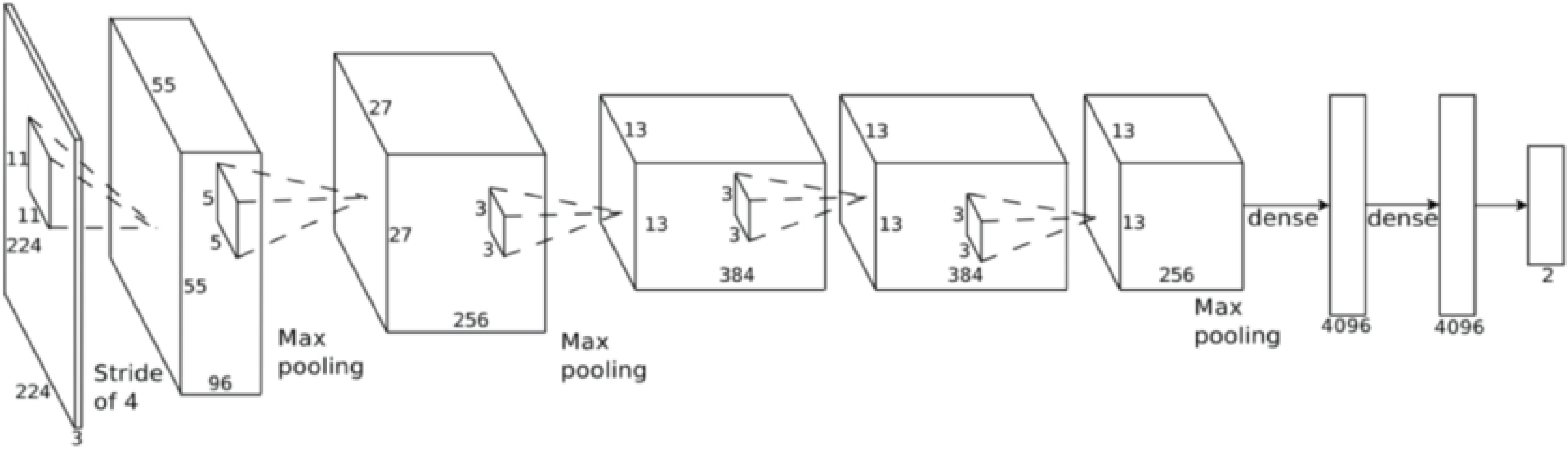
Case study: LeNet-5 (1998)



- convolutional layers + fully connected layers
- Sigmoid as the activation function
- [Conv + sigmoid + average pooling] x 2 + fully connected x 3
- Trained on MNIST with 60K training samples

Figure from *Gradient-based learning applied to document recognition, by Y. LeCun, L. Bottou, Y. Bengio and P. Haffner*

Case study: AlexNet (2012)



- convolutional layers + fully connected layers
- ReLU as the activation function
- [Conv + ReLU + max pooling] x 5 + fully connected x 3
- Most importantly: 1.2 millions of training images!

Figure from *ImageNet Classification with Deep Convolutional Neural Networks*, *by A. Krizhevsky, I. Sutskever, and G. Hinton*

Deep convolutional networks: basic design

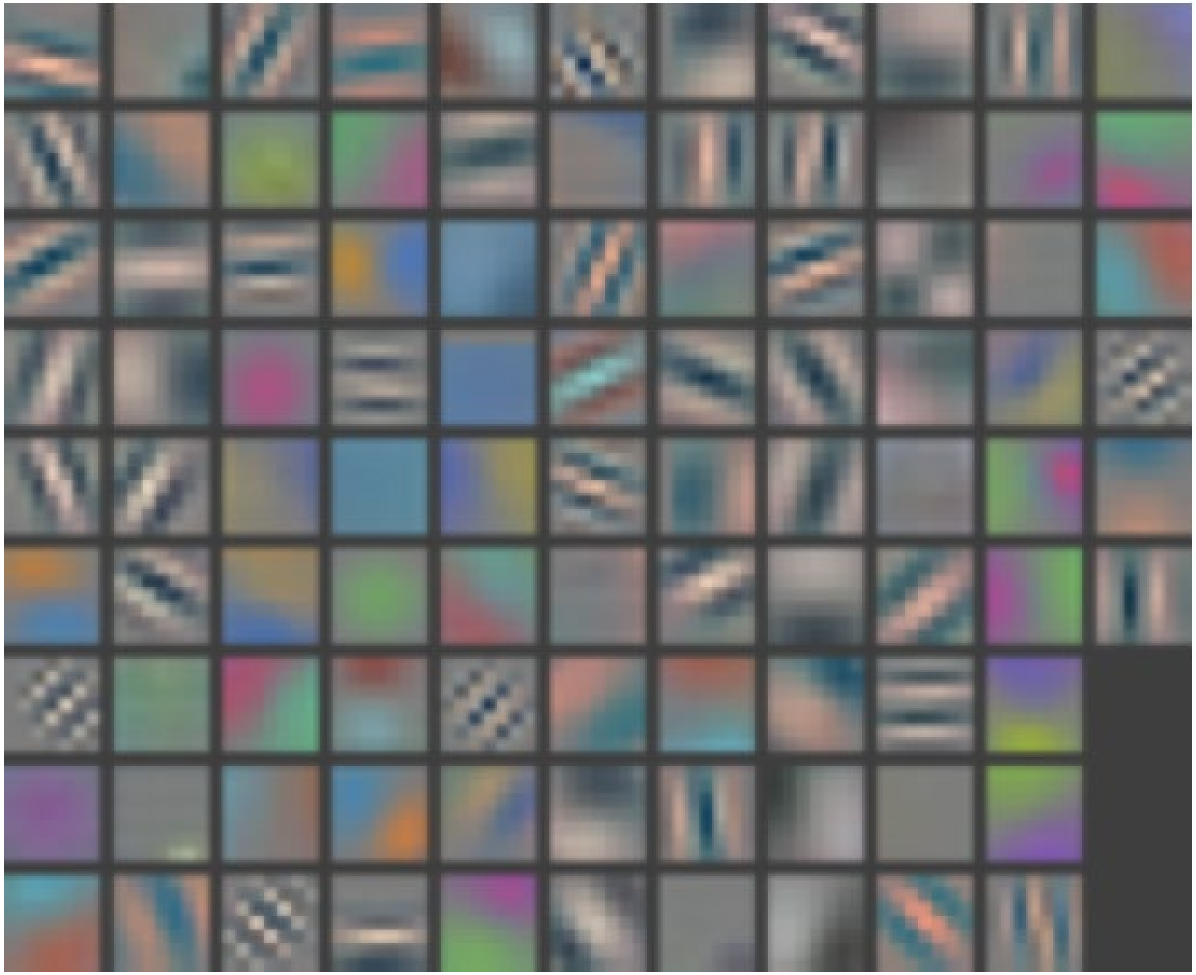
- [Conv + ReLU] + Pooling
- A few fully connected (FC) layers at the end
- Output normalization + Loss function
- Training: mini-batch stochastic gradient descent
- Inference: use the (normalized) outputs

Case study: AlexNet (2012)

- What makes deep learning work?
  - A modern design of neural network architectures
  - Large scale training dataset
  - A lot of computing power (GPUs)

- Why does deep learning work so well?
  - Still a mystery to a large extent
  - Intuitively, a deep neural network builds a hierarchical representation of data

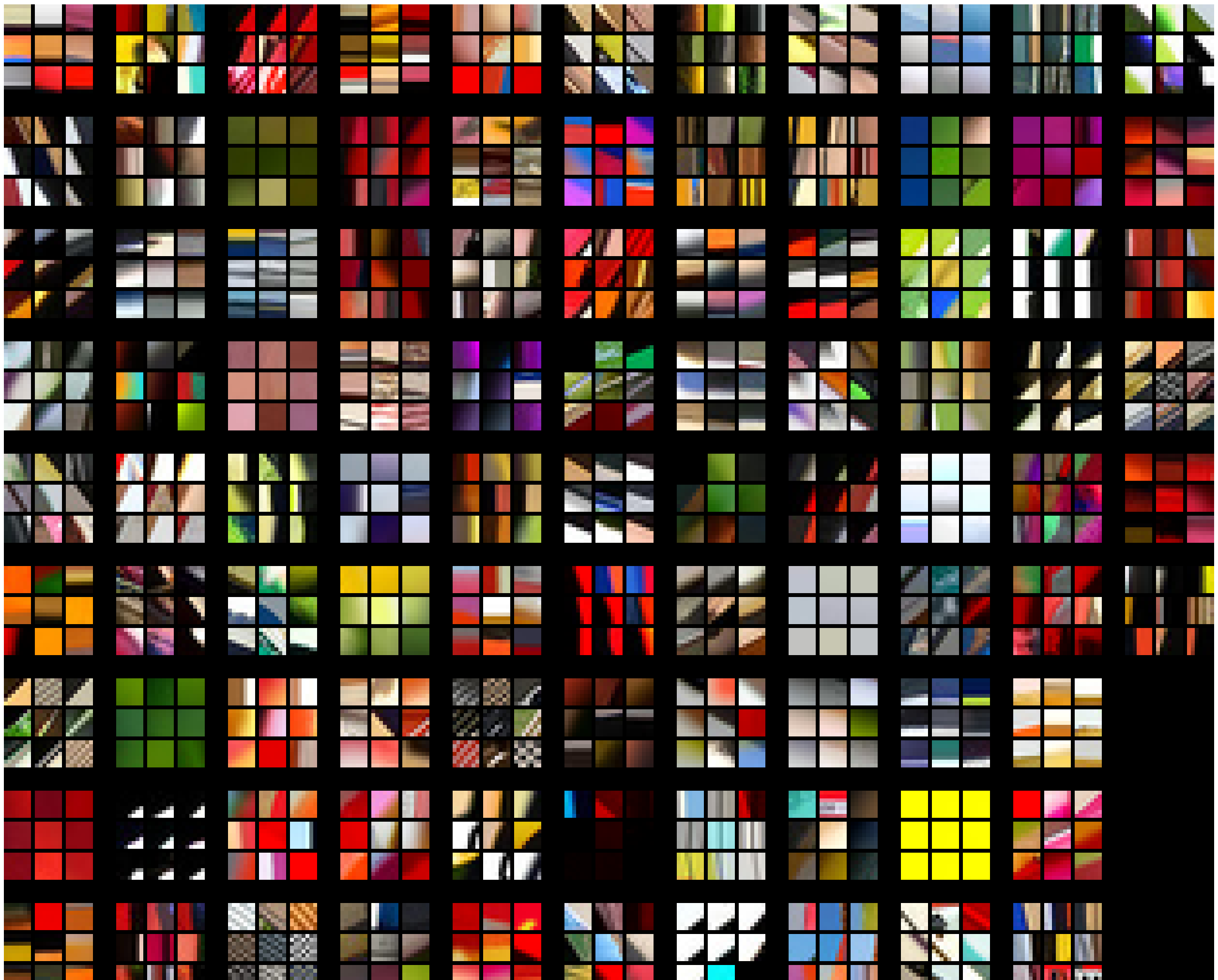
Layer 1 Filters

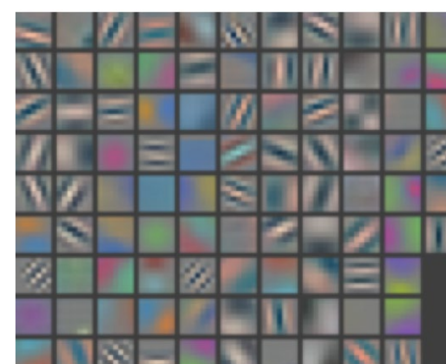


Figures from Visualizing and Understanding Convolutional Networks
by *M. Zeiler and R. Fergus*

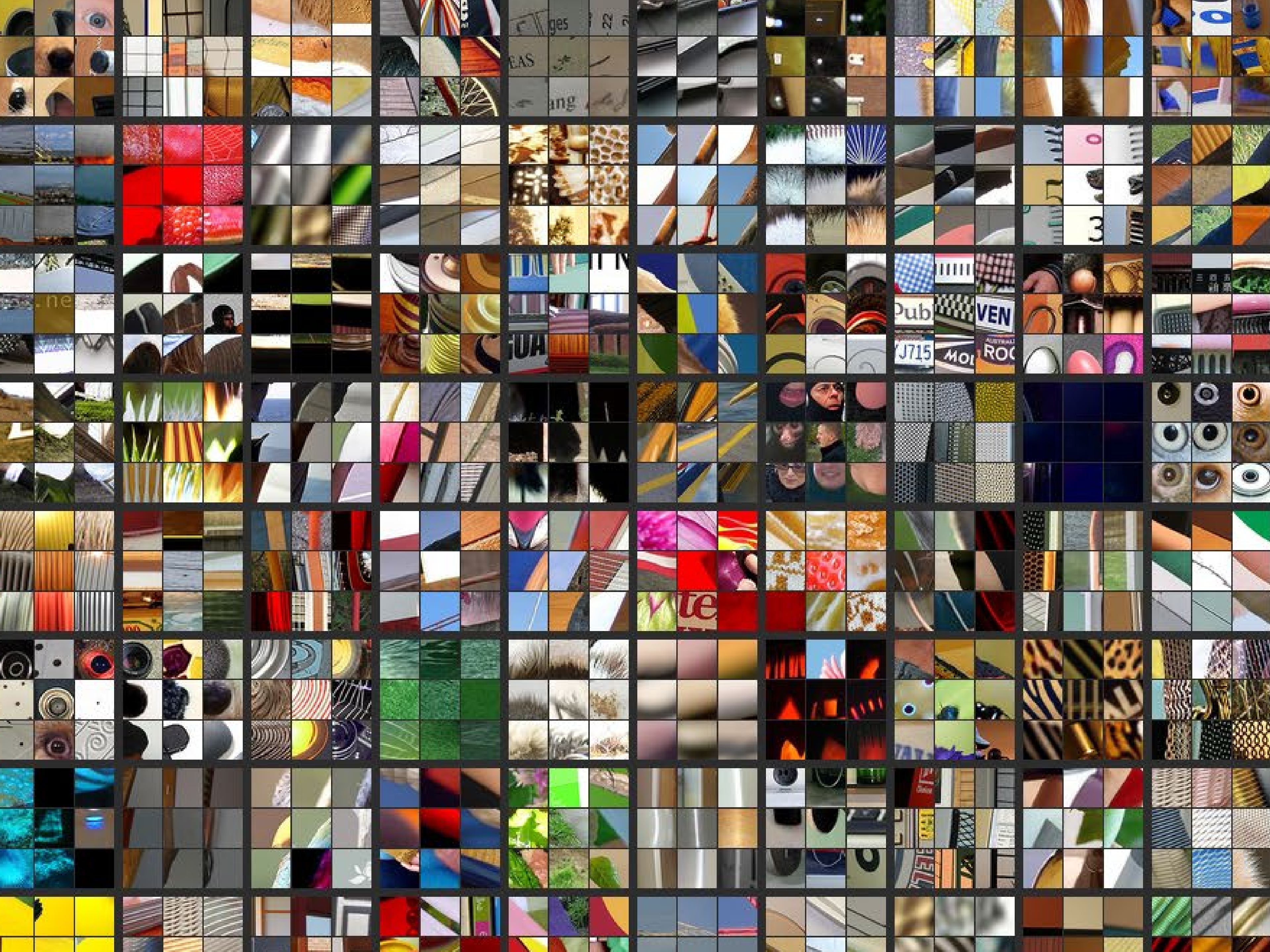
Layer 1: Top-9 Patches

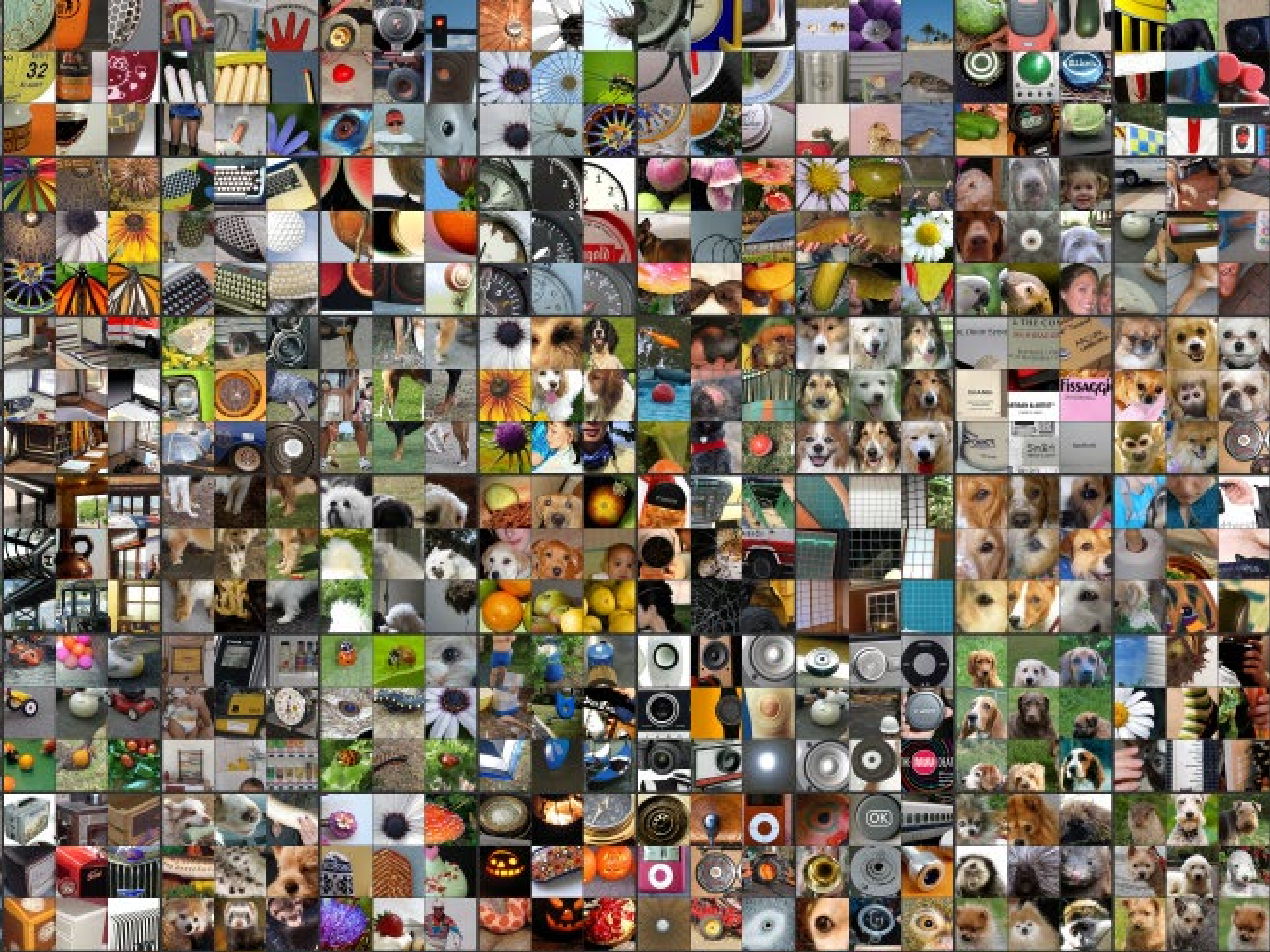
- Select patches on the validation set with maximum activation of a given convolutional filter / kernel

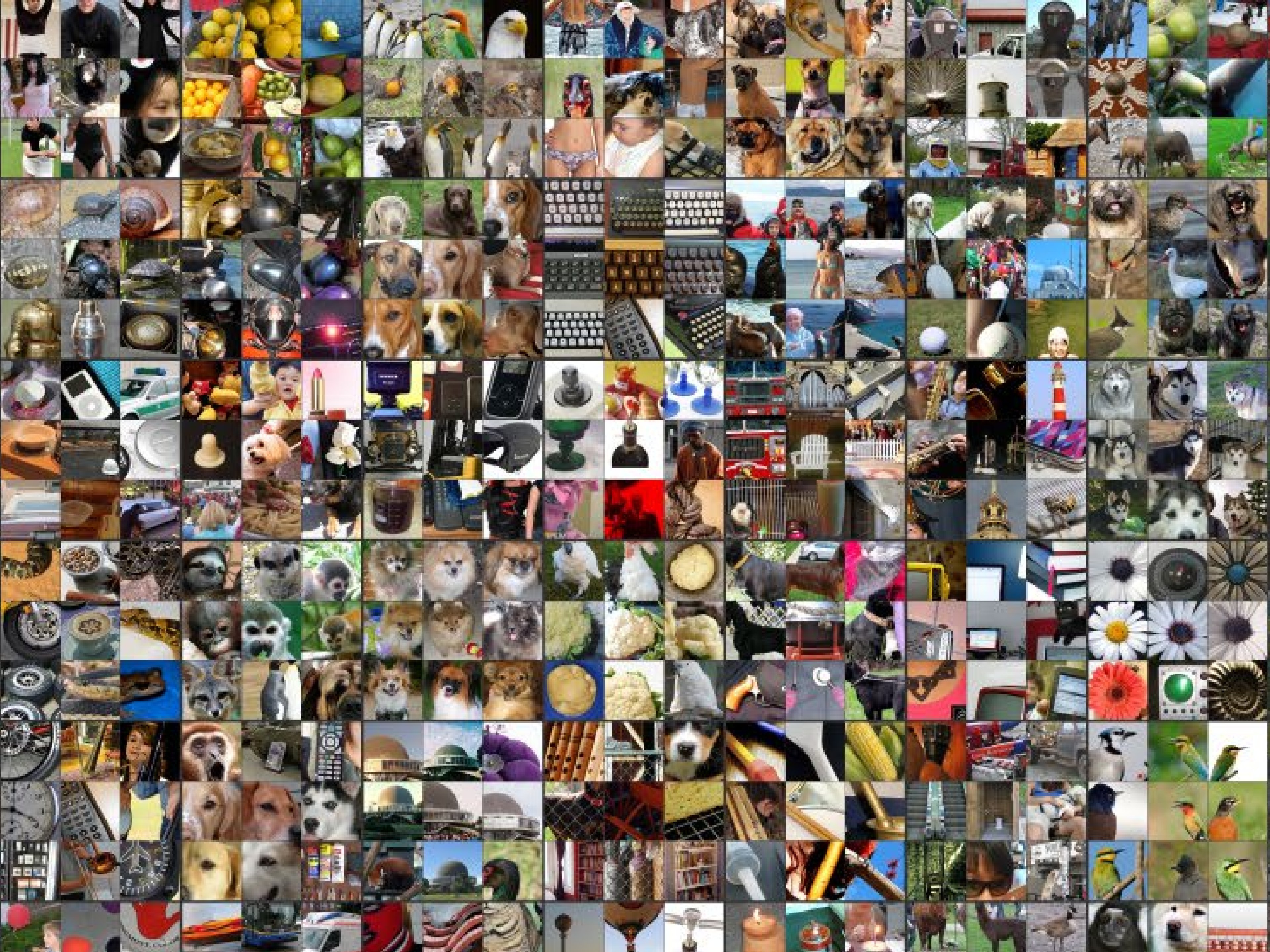




Layer 2 - 5: Top-9 Patches




32
fissaggi
OK



Case study: VGGNet (2014)

VGG16 VGG19

Image source

- 2nd place in ImageNet challenge 2014
- Make the network deeper
  - AlexNet (5 conv + 3 FC)
  - VGGNet (12/14 conv + 3 FC)
- Use smaller filter / kernel size
  - AlexNet (11x11, 5x5, 3x3)
  - VGGNet (3x3)
- Large receptive fields replaced by successive layers of 3x3 convolutions (with ReLU in between)

Case study: GoogLeNet (2014)

- The Inception Module
  - Parallel paths with different receptive field sizes and operations

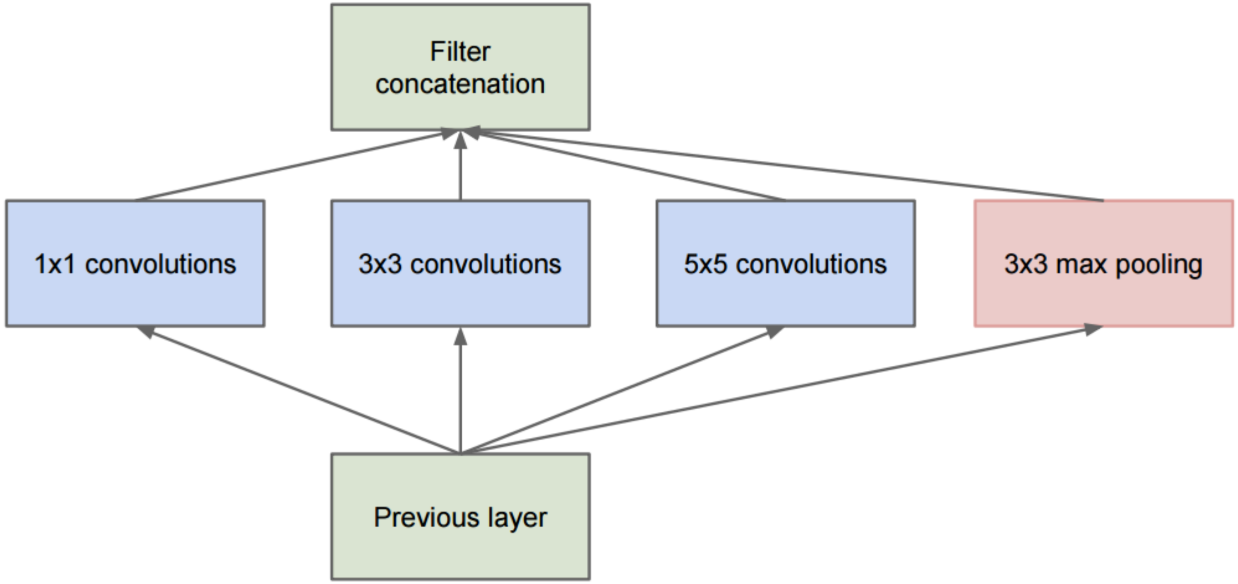


Figure from Going deeper with convolutions by *C. Szegedy et al.*

Case study: GoogLeNet (2014)

- The Inception Module
  - Parallel paths with different receptive field sizes and operations
  - Use 1x1 convolutions for dimensionality reduction before expensive convolutions

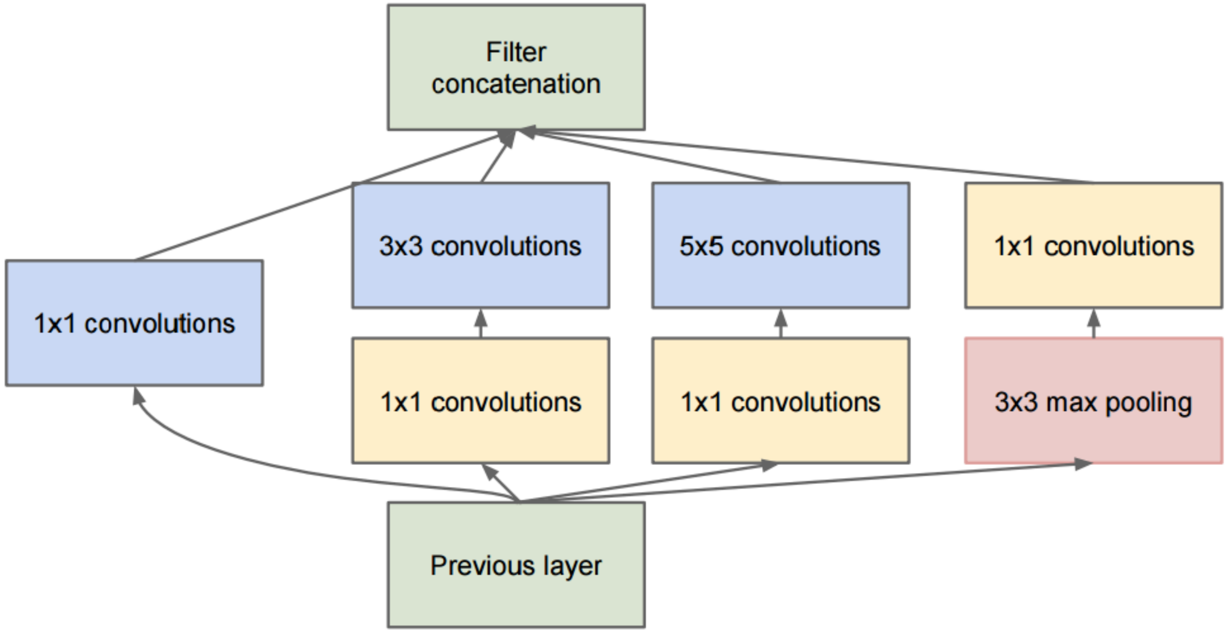
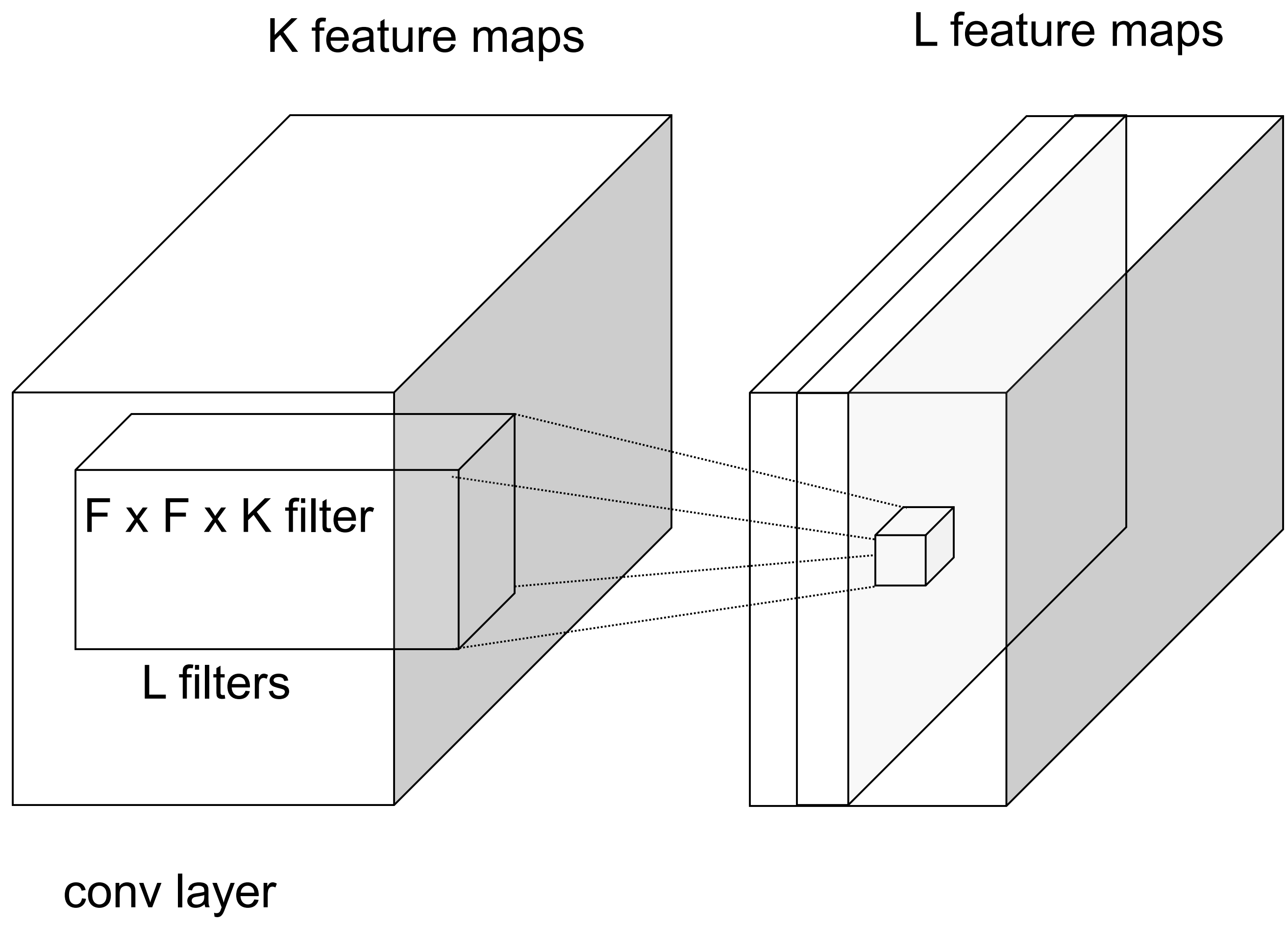
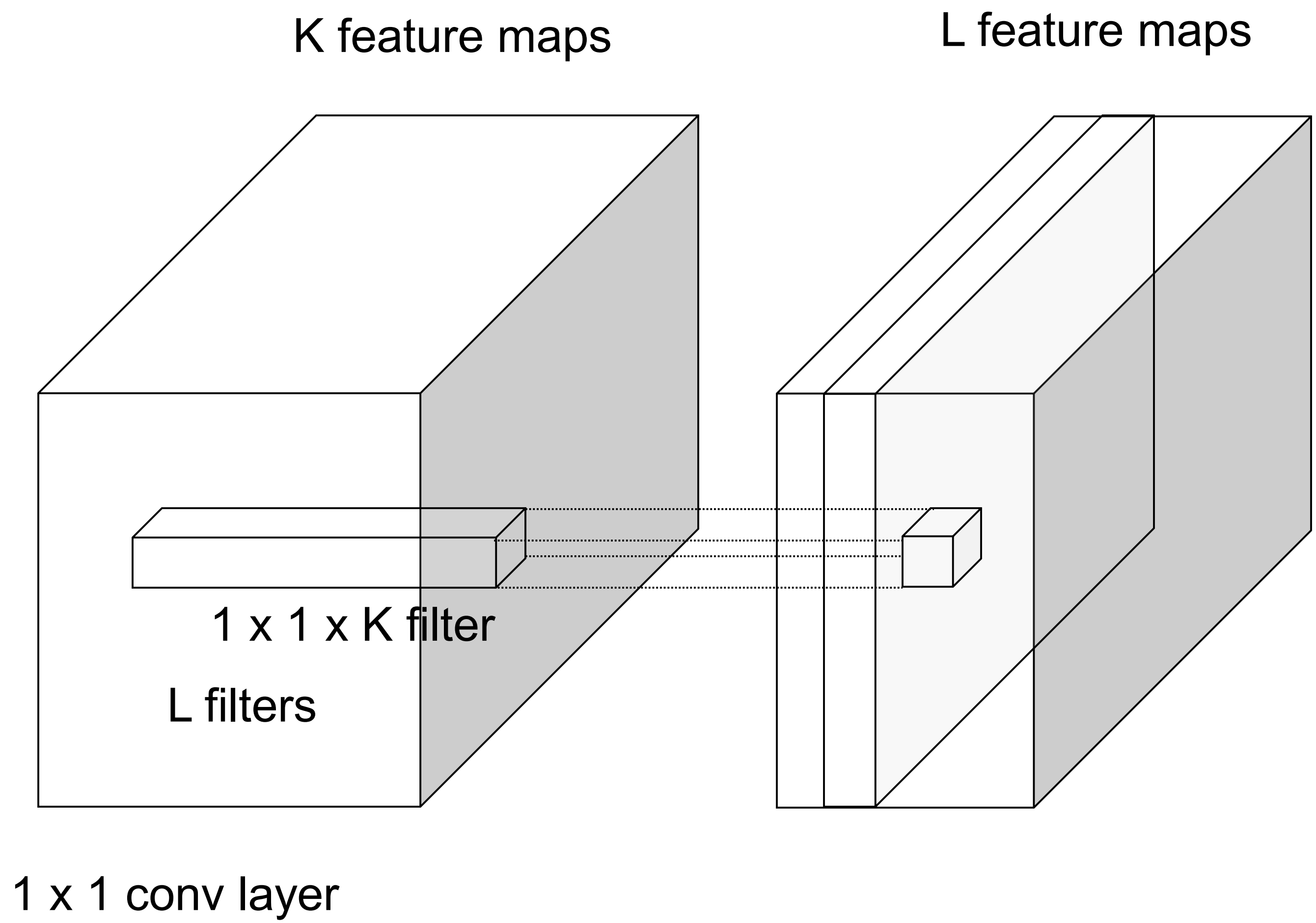


Figure from Going deeper with convolutions by *C. Szegedy et al.*

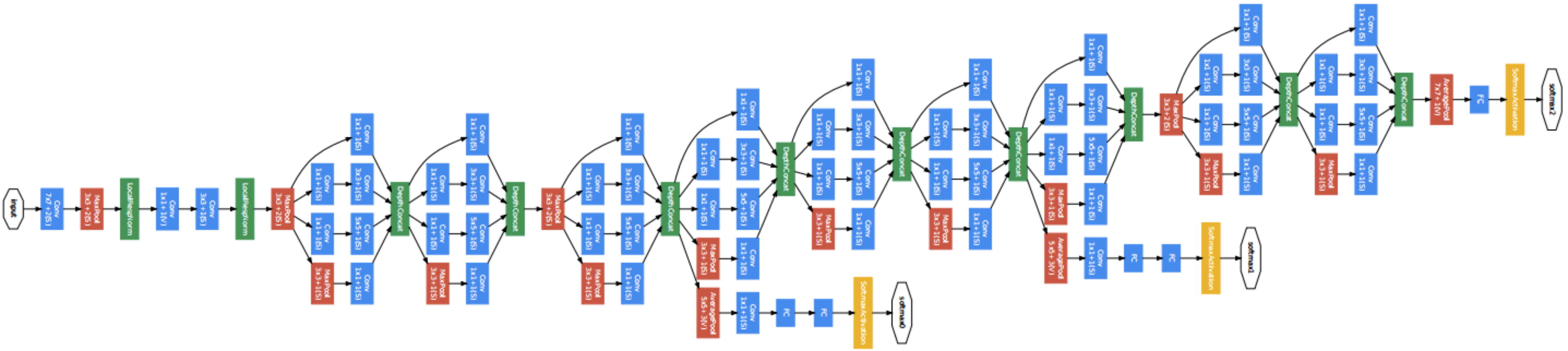
FxF convolutions



1x1 convolutions



Case study: GoogLeNet (2014)



Inception module

Figure from Going deeper with convolutions by *C. Szegedy et al.*

Case study: GoogLeNet (2014)

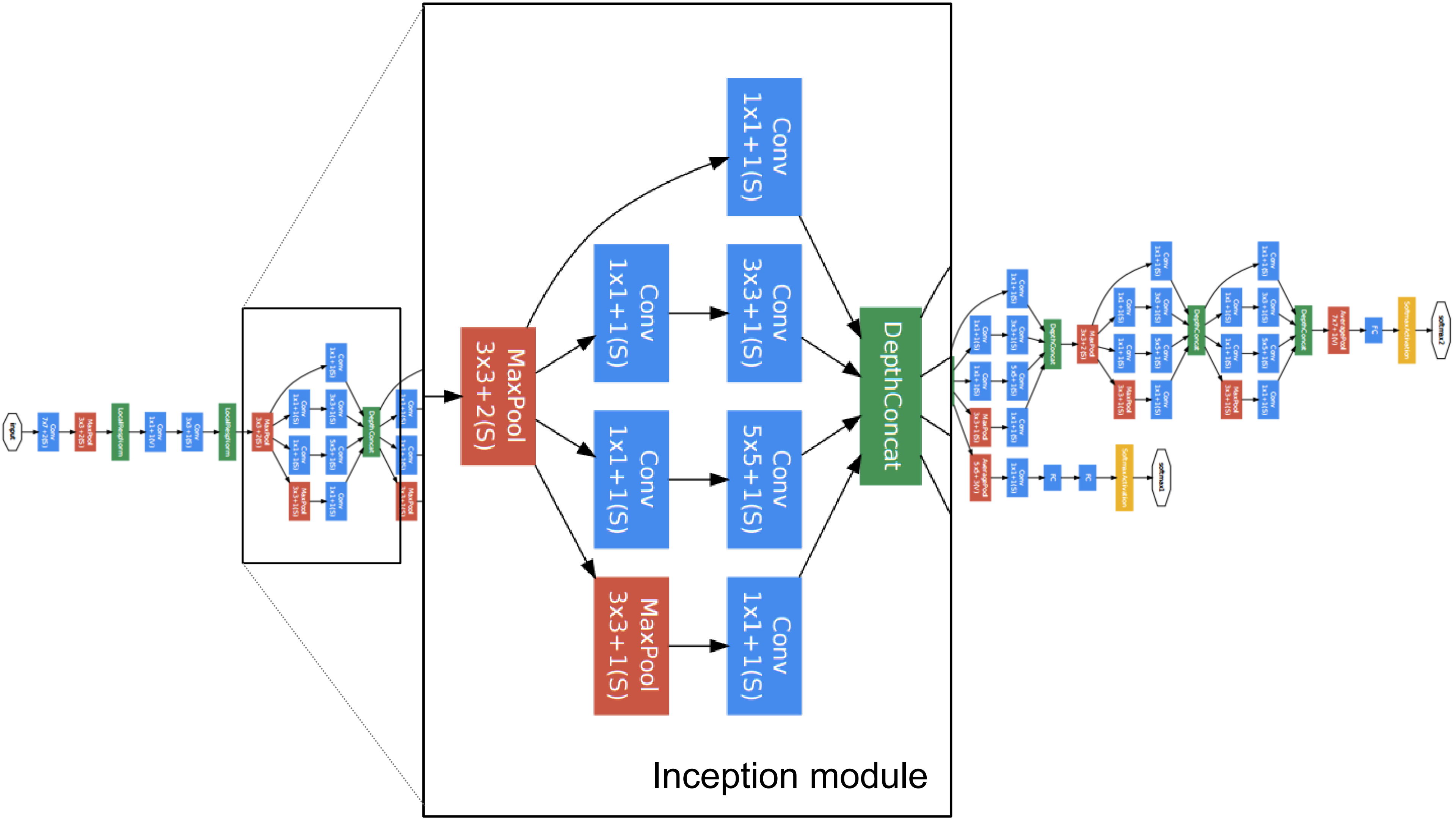


Figure from Going deeper with convolutions by *C. Szegedy et al.*

Case study: GoogLeNet (2014)

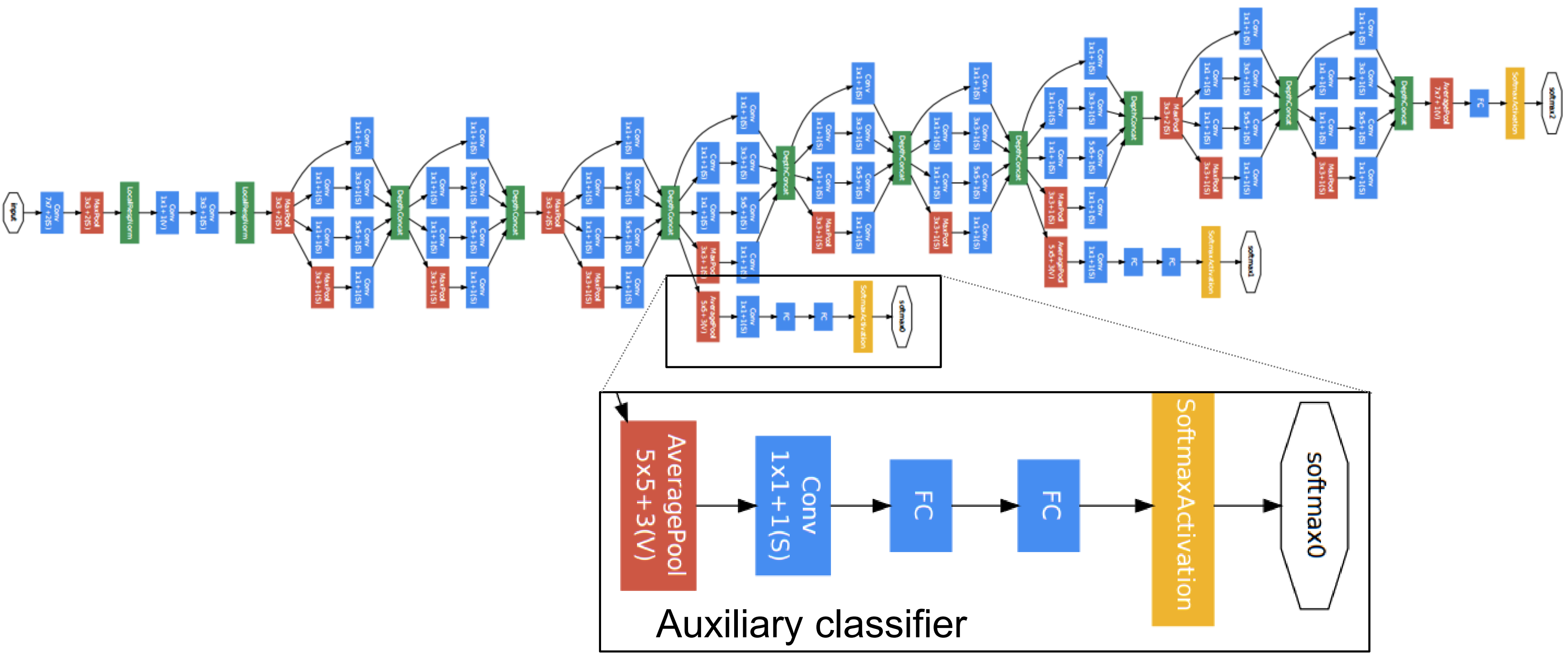


Figure from Going deeper with convolutions by *C. Szegedy et al.*

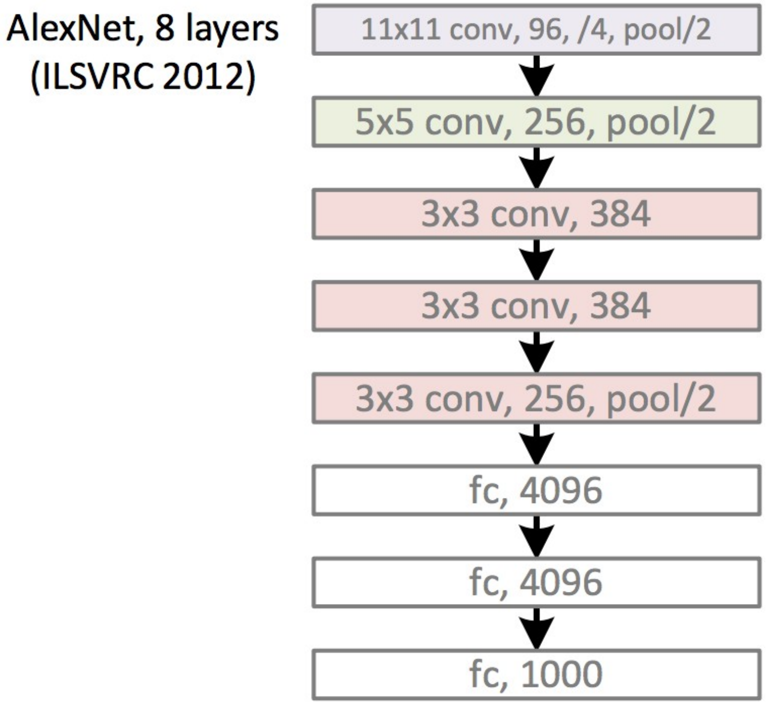
ImageNet challenge 2012-2014

Team	Year	Place	Error (top-5)	External data
SuperVision – Toronto (8 layers)	2012	-	16.4%	no
SuperVision	2012	1st	15.3%	ImageNet 22k
Clarifai – NYU (7 layers)	2013	-	11.7%	no
Clarifai	2013	1st	11.2%	ImageNet 22k
VGG – Oxford (16/19 layers)	2014	2nd	7.32%	no
GoogLeNet (22 layers)	2014	1st	6.67%	no
Human expert*			5.1%	

http://karpathy.github.io/2014/09/02/what-i-learned-from-competing-against-a-convnet-on-imagenet/

Convolutional networks: depth

Revolution of Depth



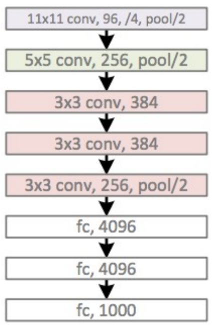
Kaiming He, Xiangyu Zhang, Shaoqing Ren, & Jian Sun. “Deep Residual Learning for Image Recognition”. arXiv 2015.

Slide Credit: Kaiming He

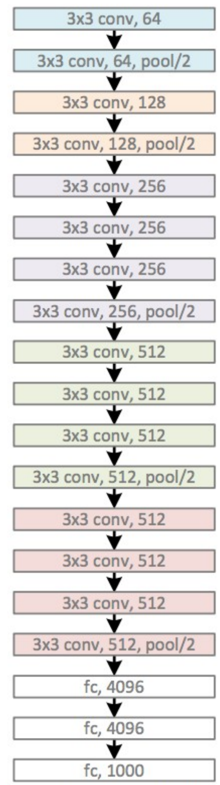
Convolutional networks: depth

Revolution of Depth

AlexNet, 8 layers
(ILSVRC 2012)



VGG, 19 layers
(ILSVRC 2014)



GoogleNet, 22 layers
(ILSVRC 2014)

Kaiming He, Xiangyu Zhang, Shaoqing Ren, & Jian Sun. "Deep Residual Learning for Image Recognition". arXiv 2015.

Slide Credit: Kaiming He

Case study: ResNet (2015)

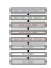
AlexNet, 8 layers
(ILSVRC 2012)

VGG, 19 layers
(ILSVRC 2014)

Slide Credit: Kaiming He

Case study: ResNet (2015)

AlexNet, 8 layers
(ILSVRC 2012)



VGG, 19 layers
(ILSVRC 2014)

ResNet, 152 layers
(ILSVRC 2015)

- Winner of ImageNet challenge 2015

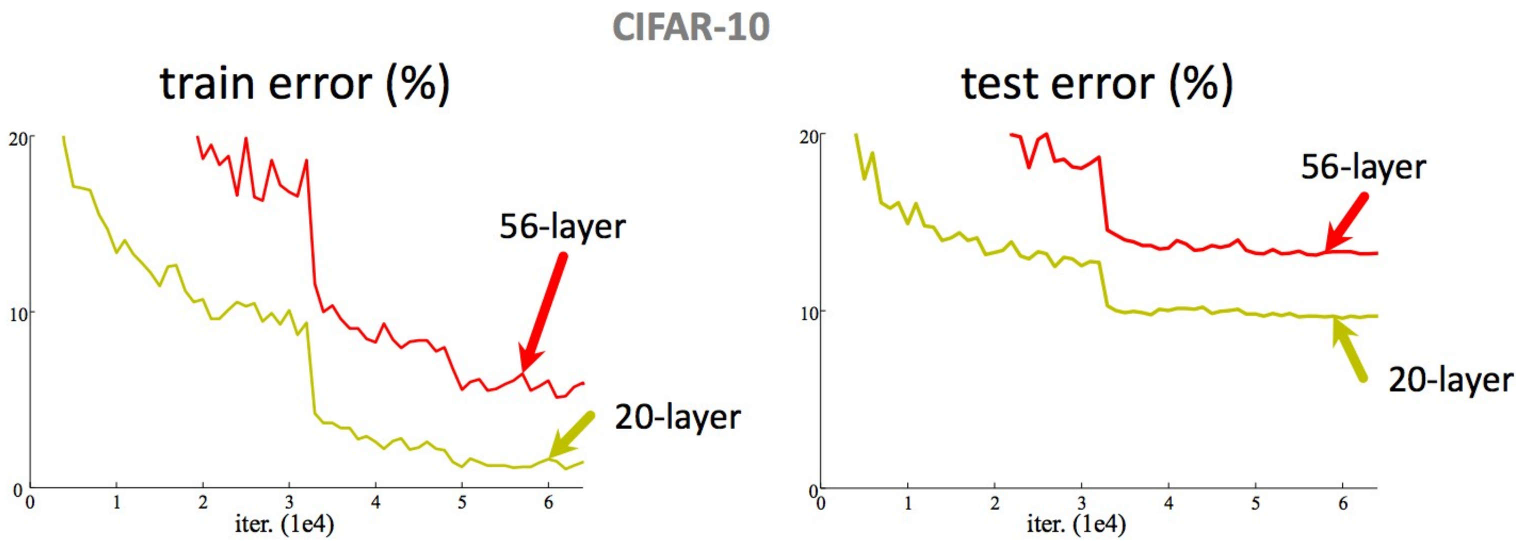
Slide Credit: Kaiming He


I WAS WINNING IMAGENET
UNTIL A DEEPER MODEL CAME ALONG
imgflip.com

Source (?)

Convolutional networks: depth

Simply stacking layers?

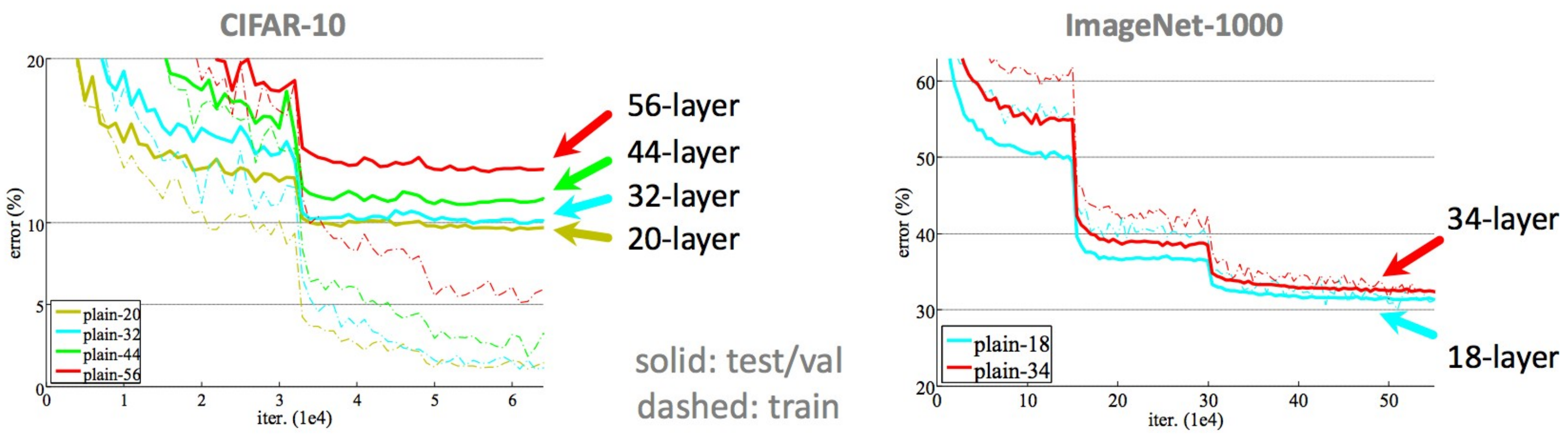


- *Plain* nets: stacking 3x3 conv layers...
- 56-layer net has **higher training error** and test error than 20-layer net

Slide Credit: Kaiming He

Convolutional networks: depth

Simply stacking layers?

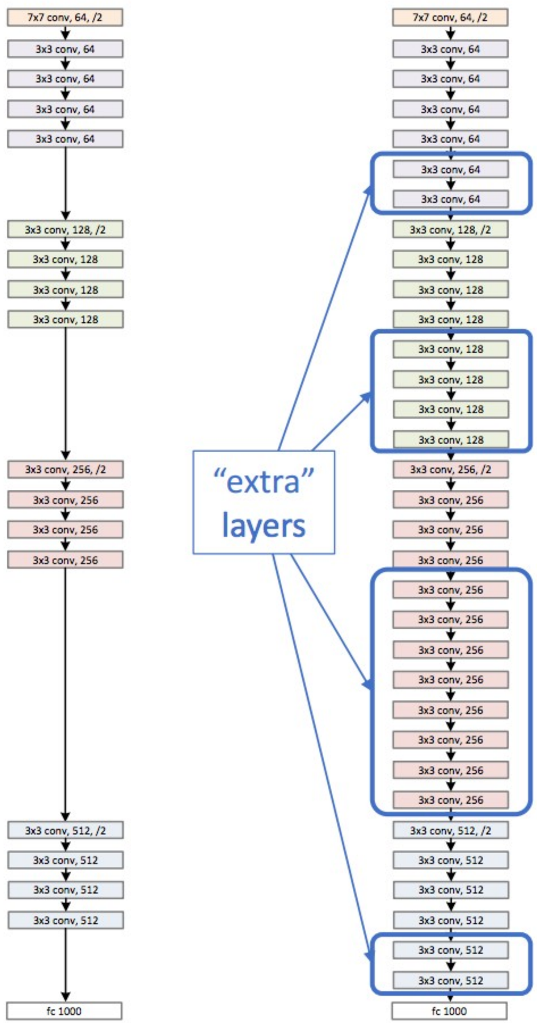


- "Overly deep" plain nets have **higher training error**
- A general phenomenon, observed in many datasets

Slide Credit: Kaiming He

Convolutional networks: depth

a shallower model (18 layers)



a deeper counterpart (34 layers)

- A deeper model should not have **higher training error**
- A solution *by construction*:
  - original layers: copied from a learned shallower model
  - extra layers: set as identity
  - at least the same training error
- Optimization difficulties: solvers cannot find the solution when going deeper… ***e.g., Gradient vanishing?***

Kaiming He, Xiangyu Zhang, Shaoqing Ren, & Jian Sun. "Deep Residual Learning for Image Recognition". arXiv 2015.

Slide Credit: Kaiming He

Case study: ResNet (2015)

- The residual module
  - Introduce *skip* or *shortcut* connections (existing before in various forms in literature)
  - Make it easy for network layers to represent the identity mapping
  - Also produce better gradients during training

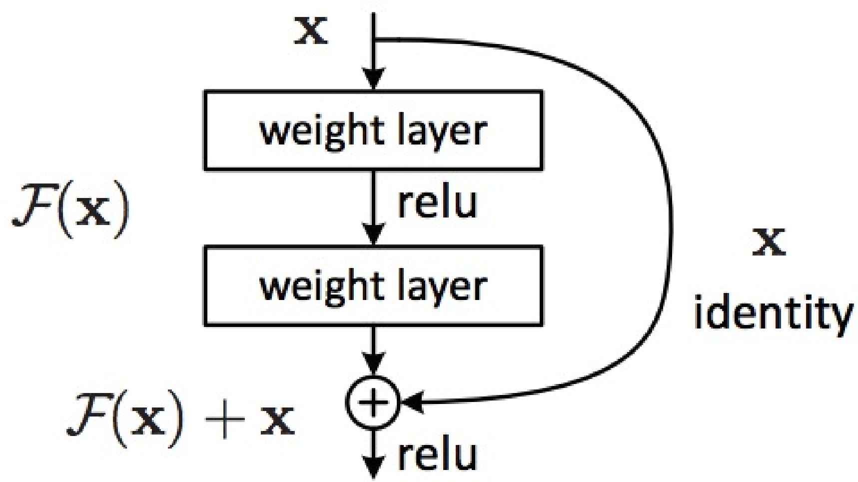


Figure from Deep Residual Learning for Image Recognition
by *K. He, X. Zhang, S. Ren, and J. Sun*

Case study: ResNet (2015)

- Architectures for ImageNet:

layer name	output size	18-layer	34-layer	50-layer	101-layer	152-layer
conv1	112×112	7×7, 64, stride 2				
conv2_x	56×56	3×3 max pool, stride 2				
		$\begin{bmatrix} 3\times3, 64 \\ 3\times3, 64 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 64 \\ 3\times3, 64 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}\times3$
conv3_x	28×28	$\begin{bmatrix} 3\times3, 128 \\ 3\times3, 128 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 128 \\ 3\times3, 128 \end{bmatrix}\times4$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix}\times4$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix}\times4$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix}\times8$
conv4_x	14×14	$\begin{bmatrix} 3\times3, 256 \\ 3\times3, 256 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 256 \\ 3\times3, 256 \end{bmatrix}\times6$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix}\times6$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix}\times23$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix}\times36$
conv5_x	7×7	$\begin{bmatrix} 3\times3, 512 \\ 3\times3, 512 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 512 \\ 3\times3, 512 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix}\times3$
	1×1	average pool, 1000-d fc, softmax				
FLOPs		1.8×10^9	3.6×10^9	3.8×10^9	7.6×10^9	11.3×10^9

Figure from Deep Residual Learning for Image Recognition
by *K. He, X. Zhang, S. Ren, and J. Sun*

ImageNet challenge 2012-2016

Team	Year	Place	Error (top-5)	External data
SuperVision – Toronto (AlexNet, 8 layers)	2012	-	16.4%	no
SuperVision	2012	1st	15.3%	ImageNet 22k
Clarifai – NYU (7 layers)	2013	-	11.7%	no
Clarifai	2013	1st	11.2%	ImageNet 22k
VGG – Oxford (16/19 layers)	2014	2nd	7.32%	no
GoogLeNet (22 layers)	2014	1st	6.67%	no
ResNet (152 layers)	2015	1st	3.57%	
Human expert*			5.1%	

http://karpathy.github.io/2014/09/02/what-i-learned-from-competing-against-a-convnet-on-imagenet/

Summary: Deep Convolutional Networks

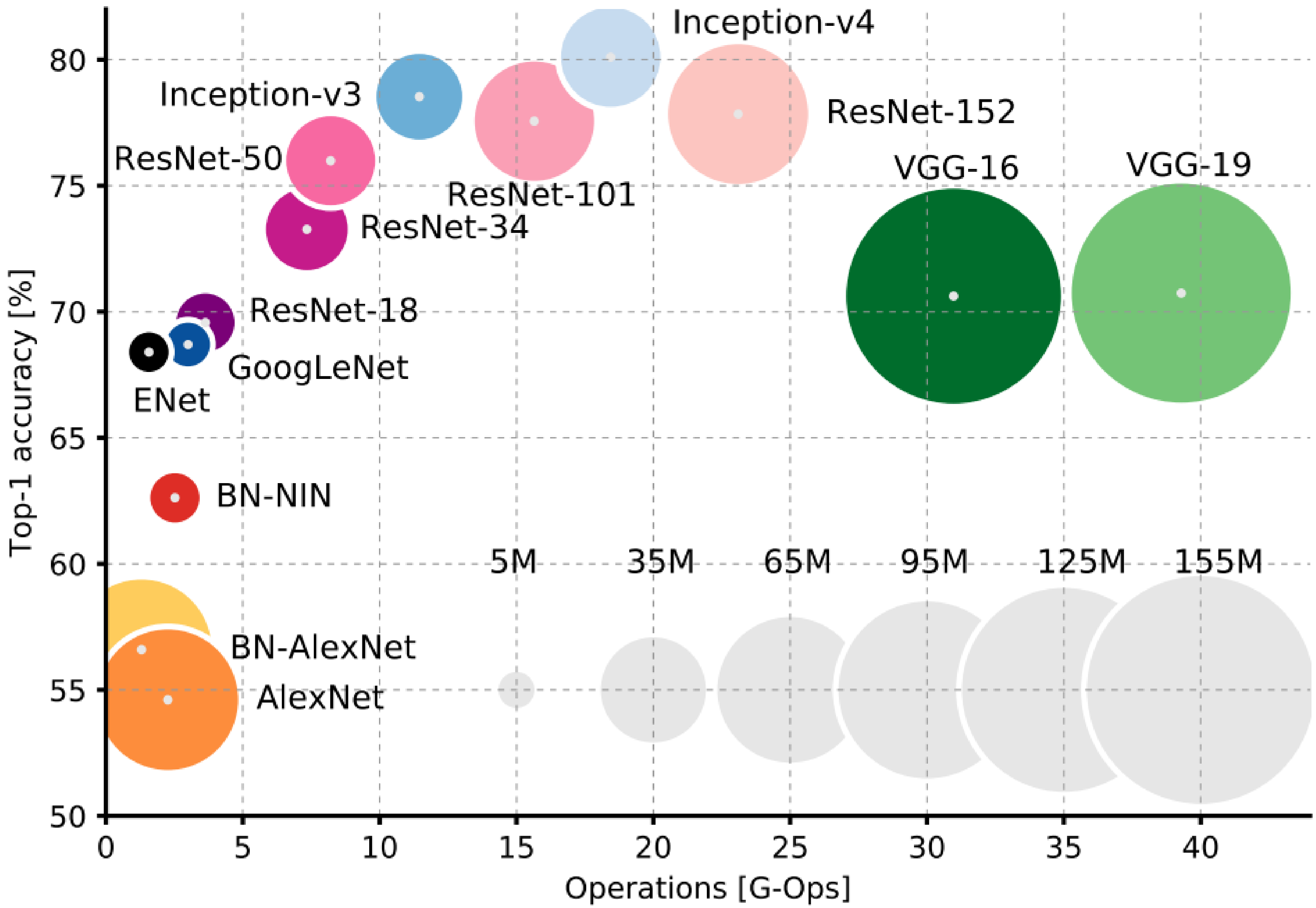


Figure from An Analysis of Deep Neural Network Models for Practical Applications by A. Canziani, A. Paszke, E. Culurciello