Day 2 Lecture 4

Imagenet

[Xavier Giró-i-Nieto]

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Signal Theory and Communications
Image Processing Group

ImageNet ILSRVC

Li Fei-Fei, [“How we’re teaching computers to understand pictures”](#) TEDTalks 2014.

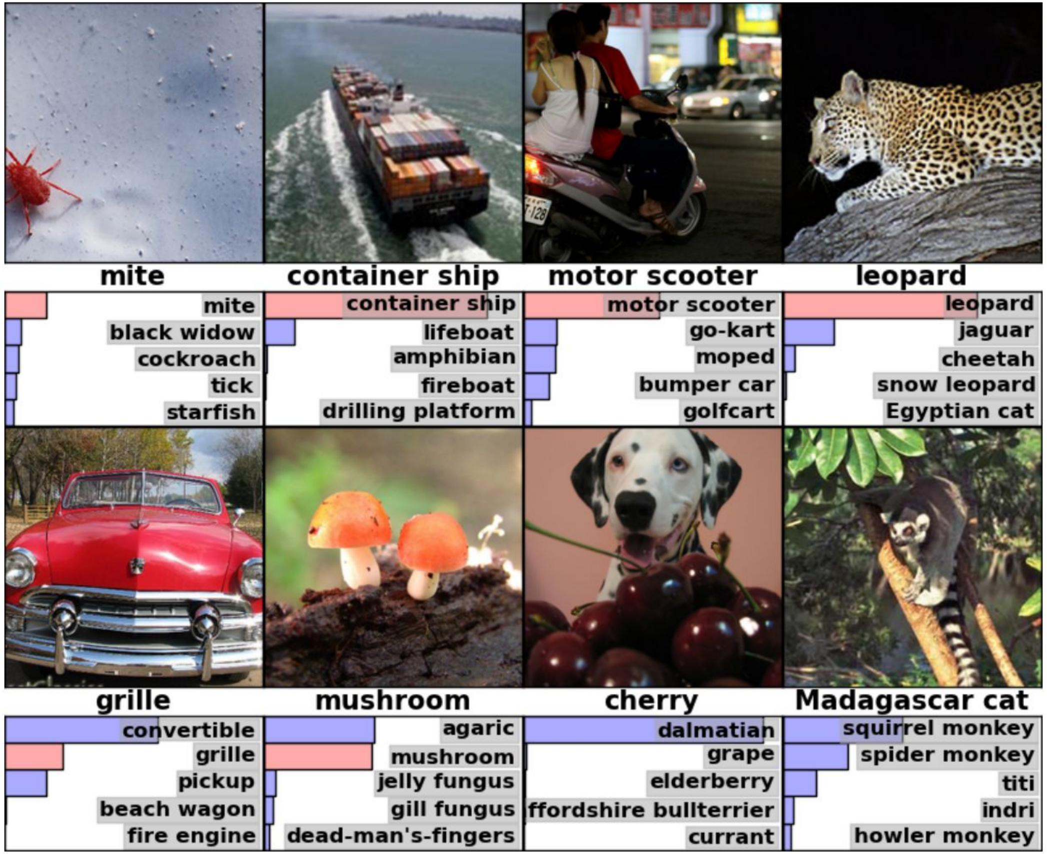
Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., ... & Fei-Fei, L. (2015). [Imagenet large scale visual recognition challenge](#). *arXiv preprint arXiv:1409.0575*. [web]

ImageNet ILSRVC

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., ... & Fei-Fei, L. (2015). Imagenet large scale visual recognition challenge. *arXiv preprint arXiv:1409.0575*. [web]

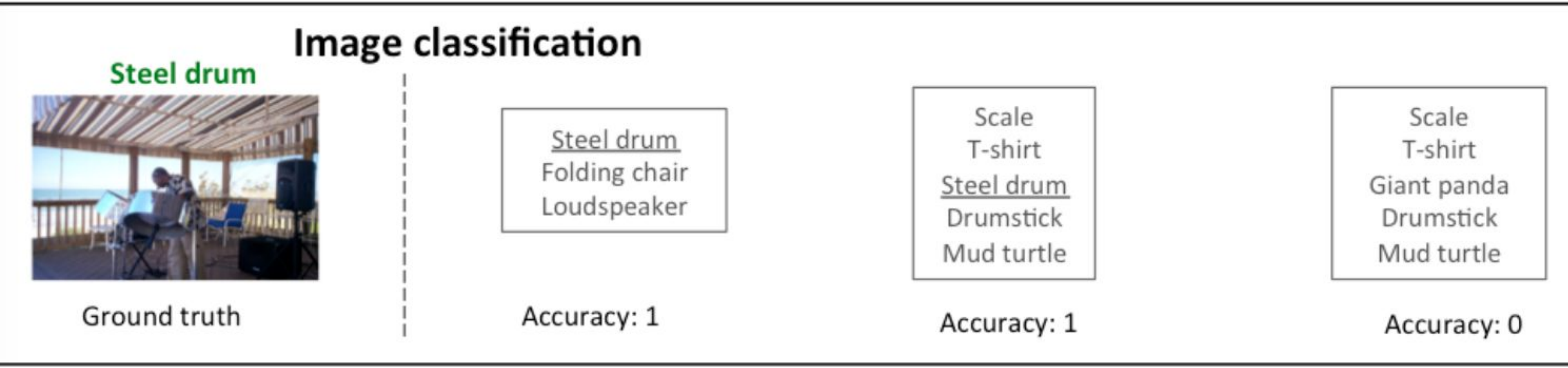
ImageNet ILSRVC

- 1,000 object classes (categories).
- Images:
  - 1.2 M train
  - 100k test.



ImageNet ILSRVC

- Top 5 error rate



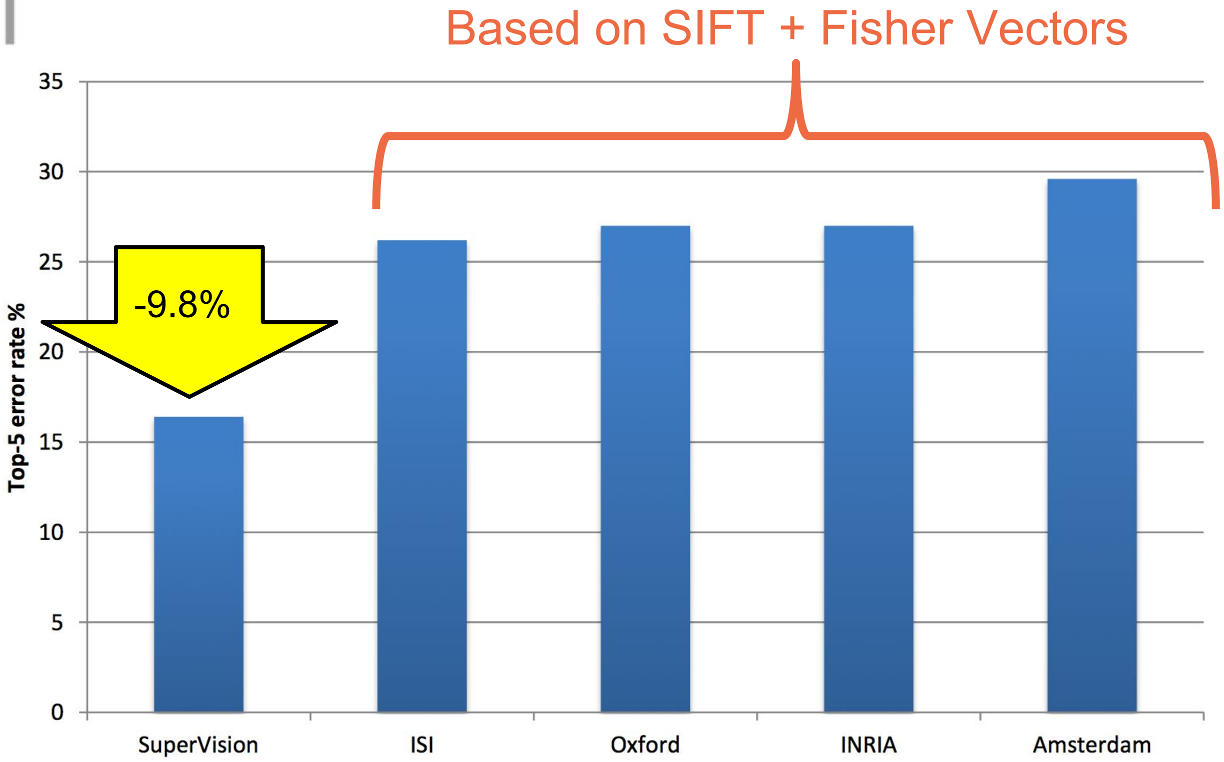
Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., ... & Fei-Fei, L. (2015). [Imagenet large scale visual recognition challenge](). *arXiv preprint arXiv:1409.0575*. [web]

ImageNet ILSRVC

Image Classification 2012

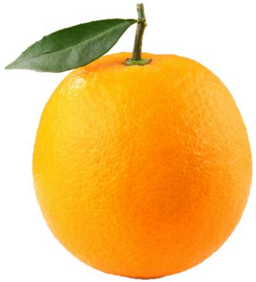
IMAGENET

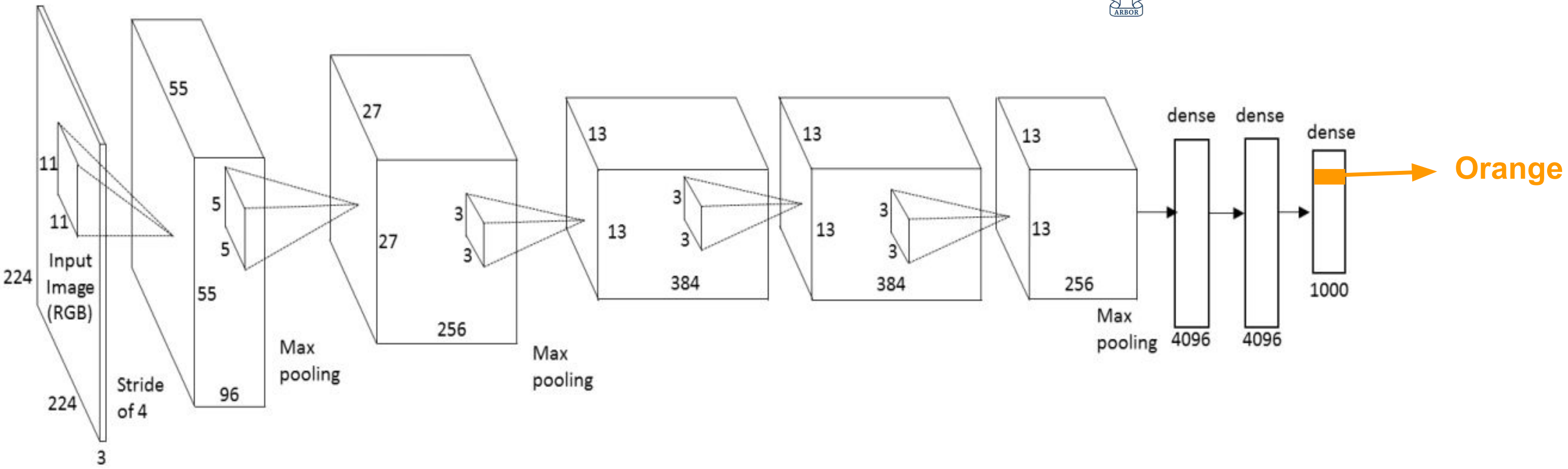
Slide credit: Rob Fergus (NYU)



Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., ... & Fei-Fei, L. (2014). Imagenet large scale visual recognition challenge. *arXiv preprint arXiv:1409.0575*. [web]

AlexNet (Supervision)

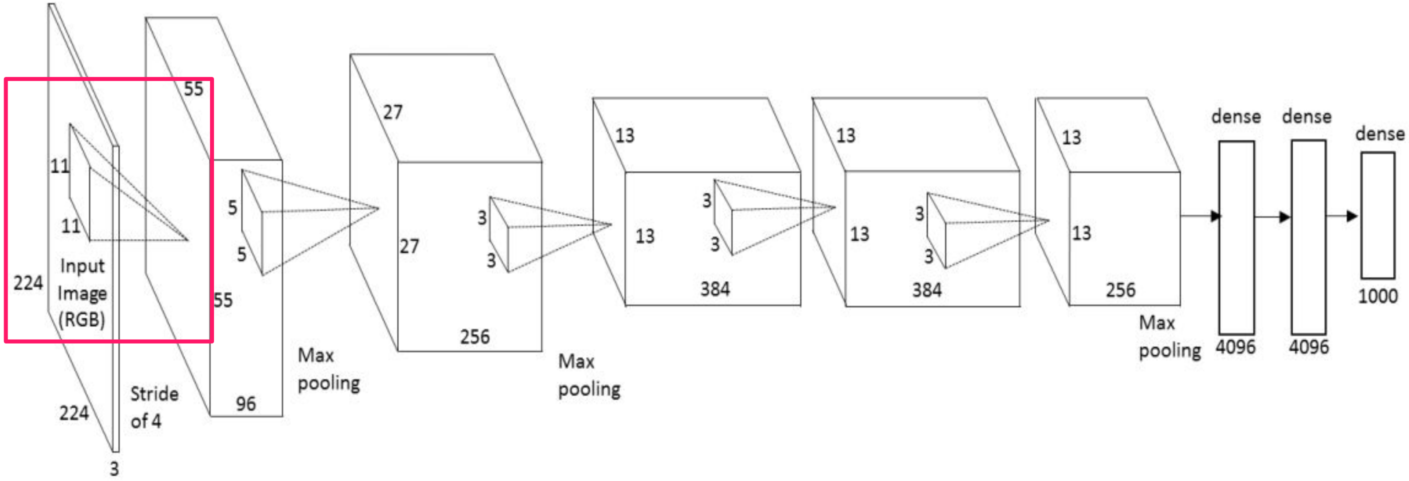




A Krizhevsky, I Sutskever, GE Hinton "[Imagenet classification with deep convolutional neural networks]" Part of: [Advances in Neural Information Processing Systems 25 (NIPS 2012)]

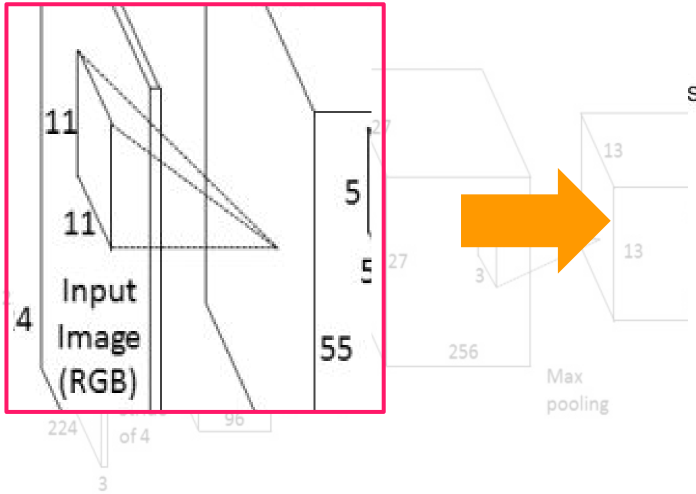
Slide credit: Junting Pan, "Visual Saliency Prediction using Deep Learning Techniques" (ETSETB-UPC 2015)

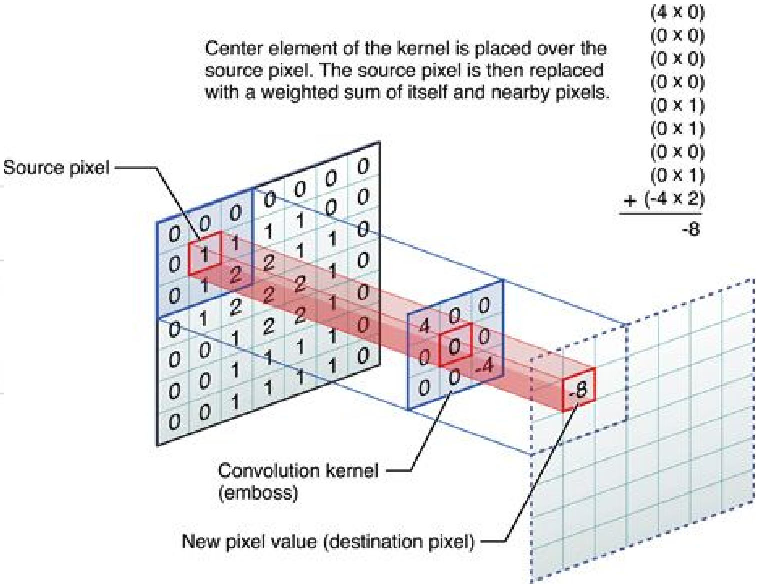
AlexNet (Supervision)



Slide credit: Junting Pan, "Visual Saliency Prediction using Deep Learning Techniques" (ETSETB-UPC 2015)

AlexNet (Supervision)





Slide credit: Junting Pan, "Visual Saliency Prediction using Deep Learning Techniques" (ETSETB-UPC 2015)

AlexNet (Supervision)

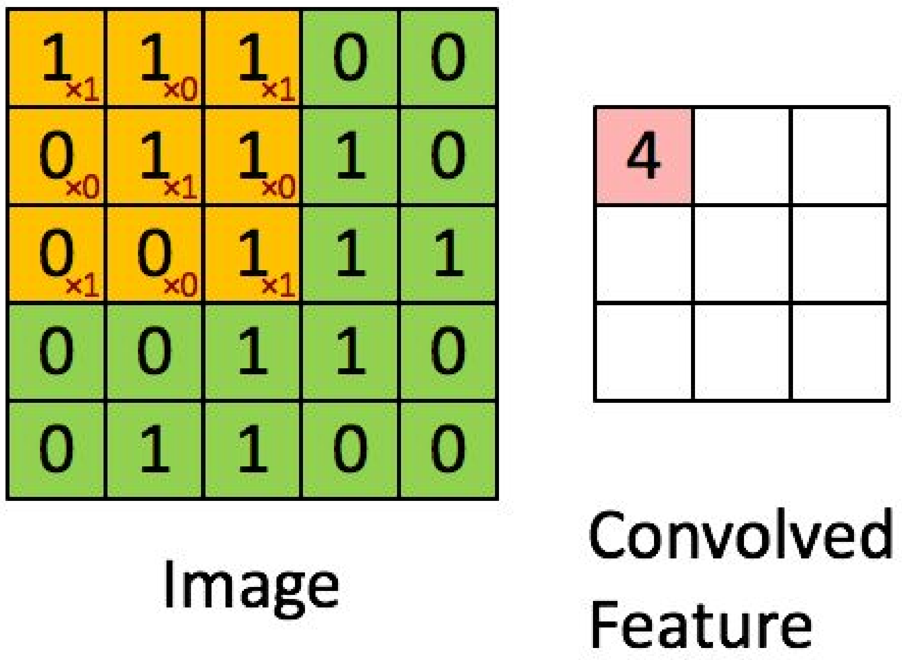


Image credit: Deep learning Tutorial (Stanford University)

AlexNet (Supervision)

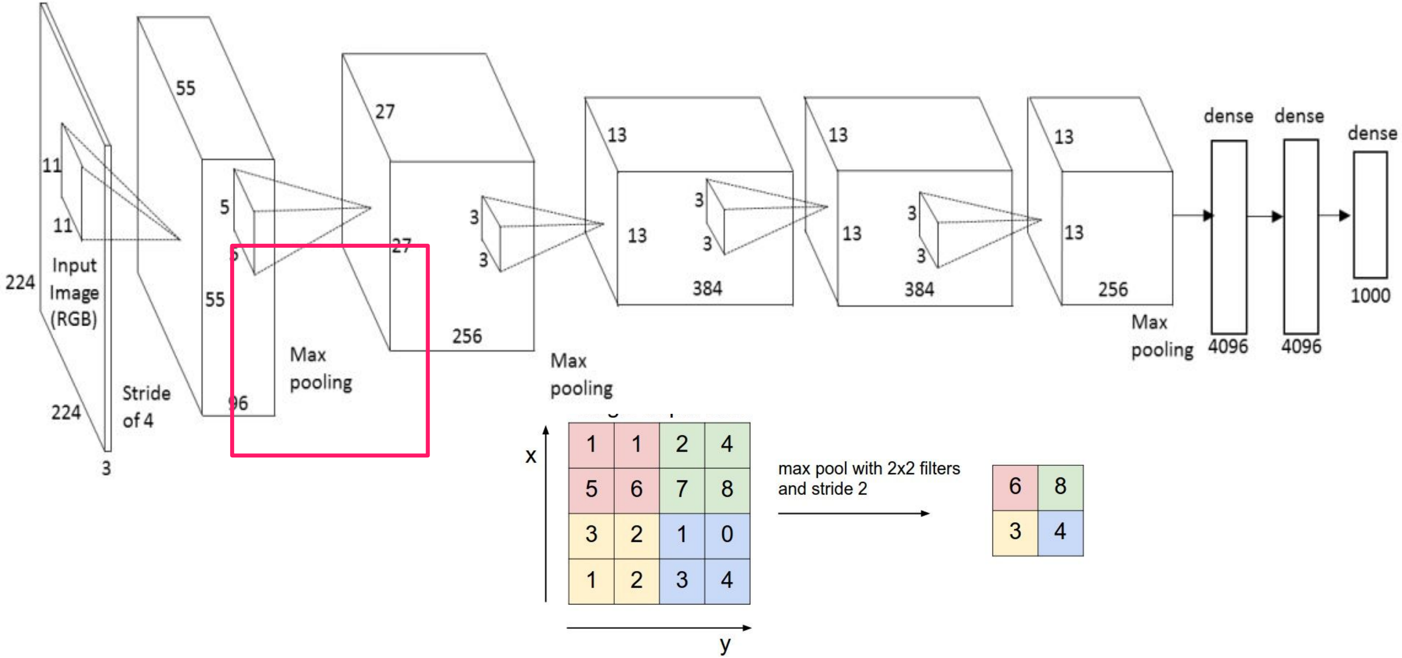


Image credit: Deep learning Tutorial (Stanford University)

AlexNet (Supervision)

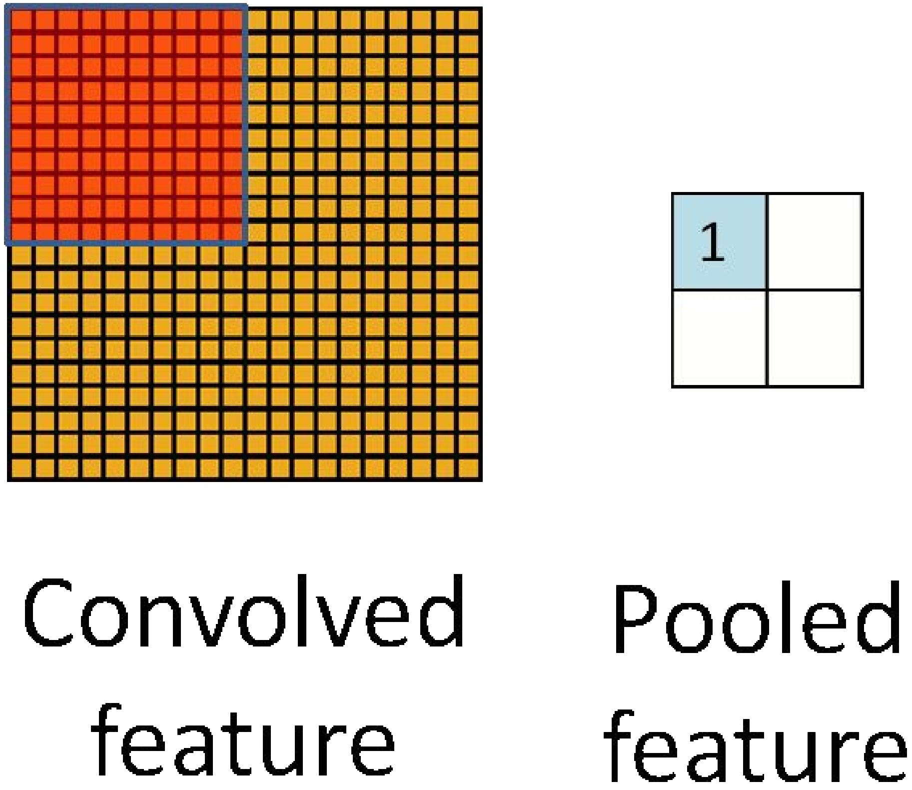
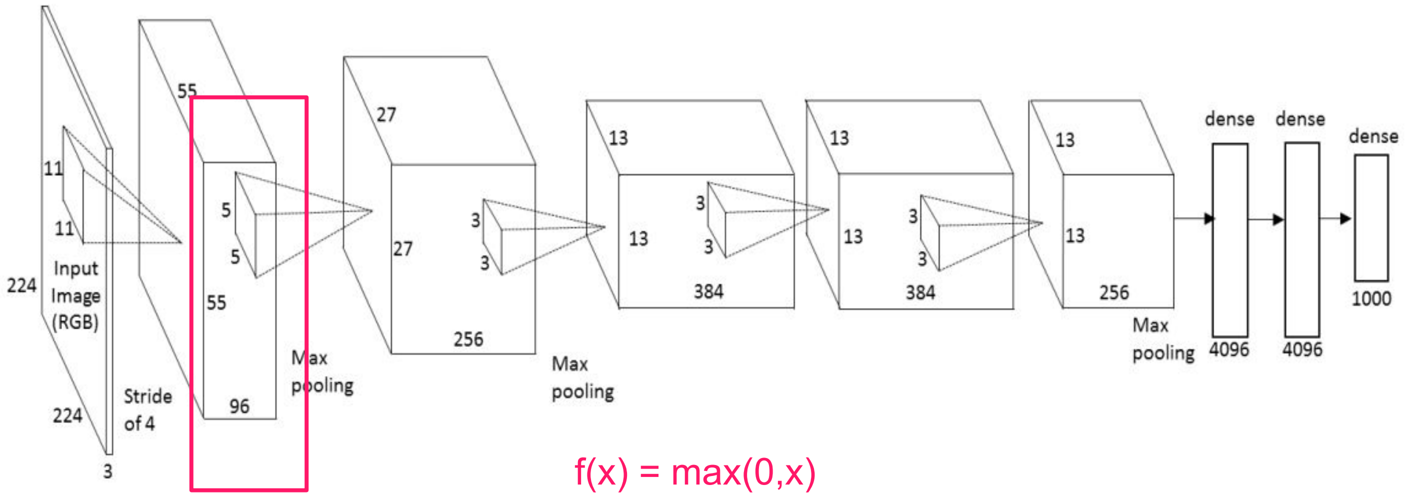


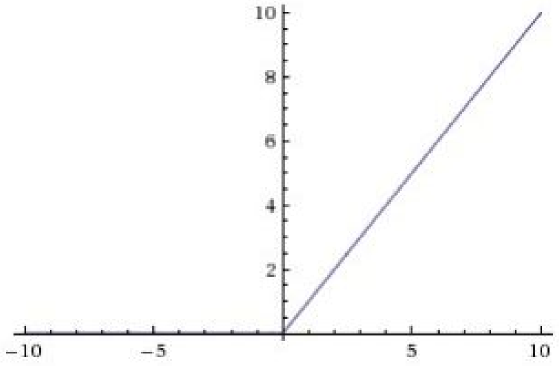
Image credit: Deep learning Tutorial (Stanford University)

AlexNet (Supervision)



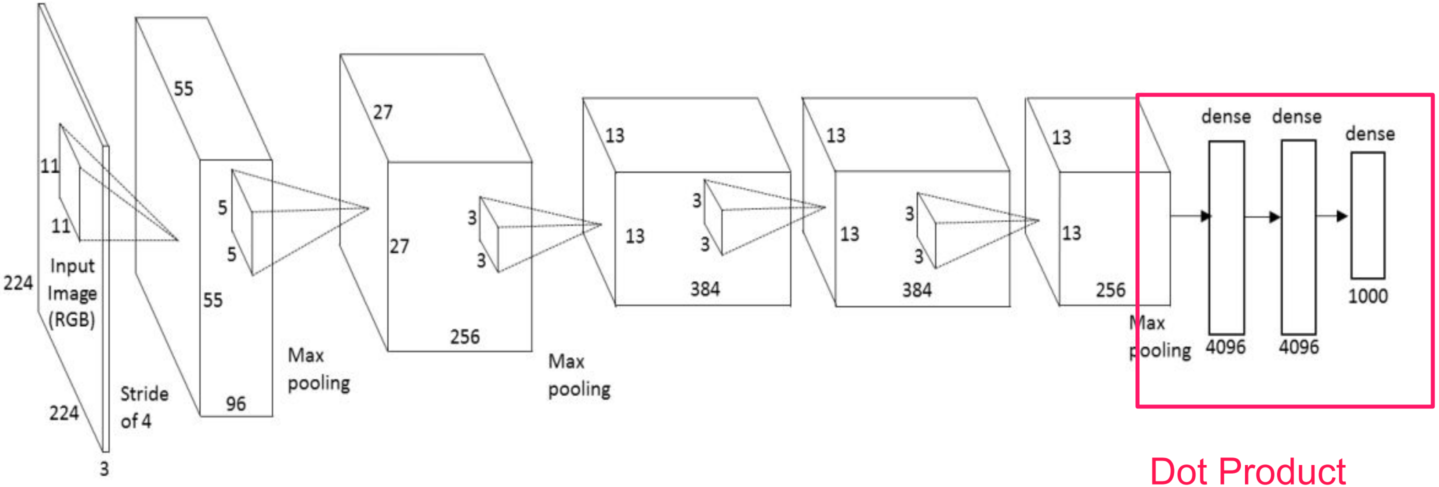
Rectified
Linear
Unit
(non-linearity)

$f(x) = \max(0,x)$



Slide credit: Junting Pan, "Visual Saliency Prediction using Deep Learning Techniques" (ETSETB-UPC 2015)

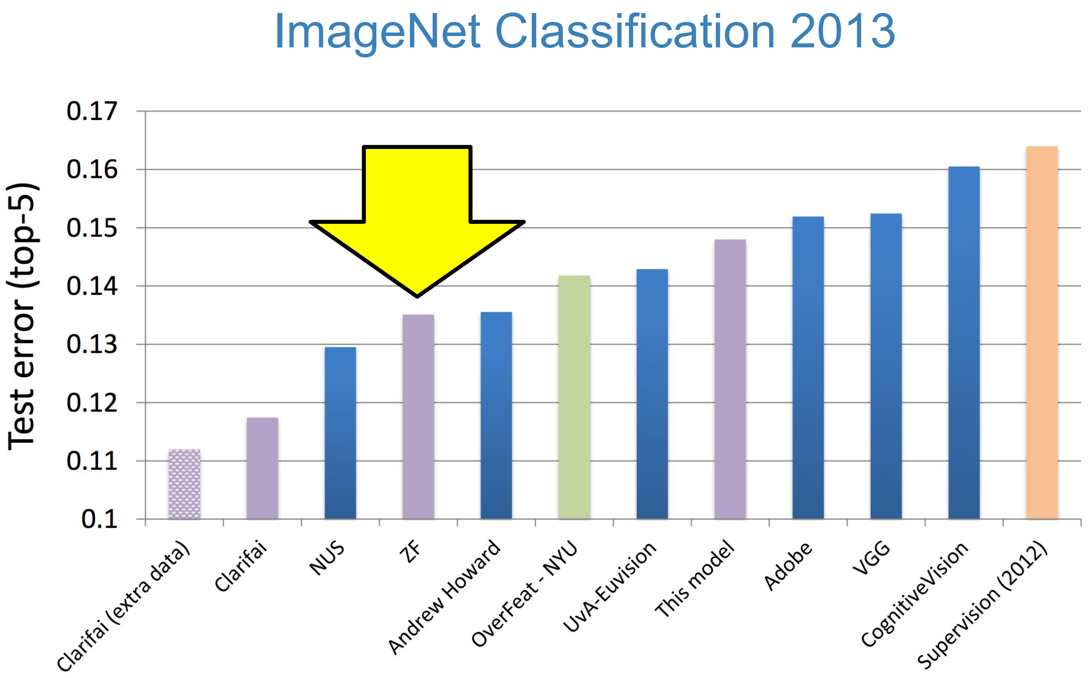
AlexNet (Supervision)



Dot Product

Slide credit: Junting Pan, "Visual Saliency Prediction using Deep Learning Techniques" (ETSETB-UPC 2015)

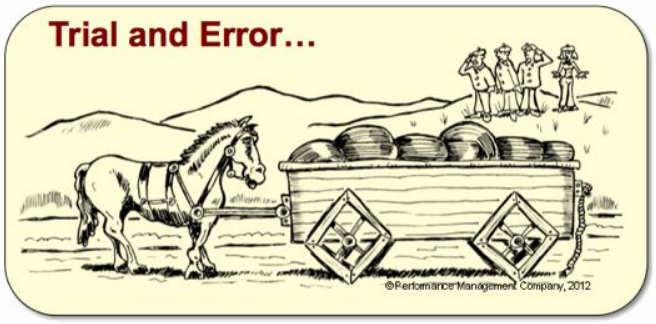
ImageNet ILSRVC



Slide credit:
Rob Fergus (NYU)

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., ... & Fei-Fei, L. (2015). Imagenet large scale visual recognition challenge. *arXiv preprint arXiv:1409.0575*. [web]

Zeiler-Fergus (ZF)



The development of better convnets is reduced to trial-and-error.

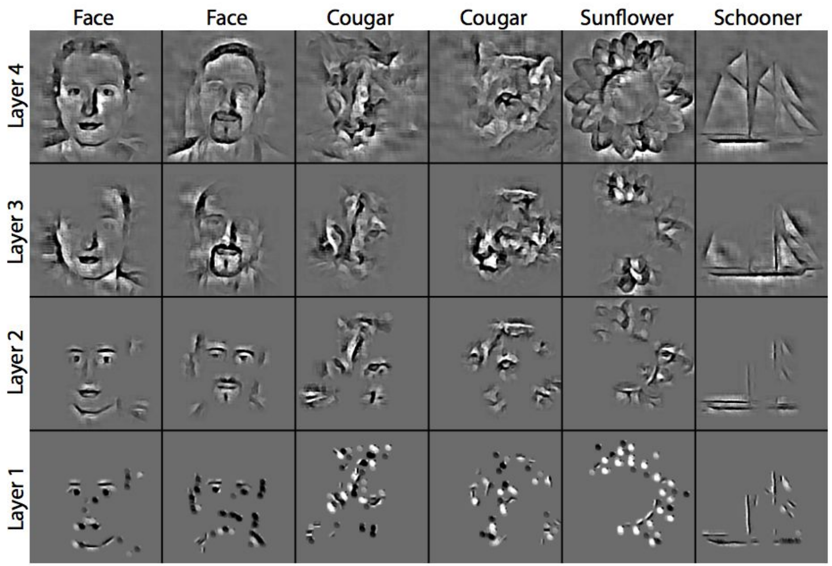
<u>Visualization</u> can help in proposing better architectures.

Zeiler, M. D., & Fergus, R. (2014). [Visualizing and understanding convolutional networks](). In *Computer Vision–ECCV 2014* (pp. 818-833). Springer International Publishing.

Zeiler-Fergus (ZF)

“A convnet model that uses the same components (filtering, pooling) but in reverse, so instead of mapping pixels to features does the opposite.”

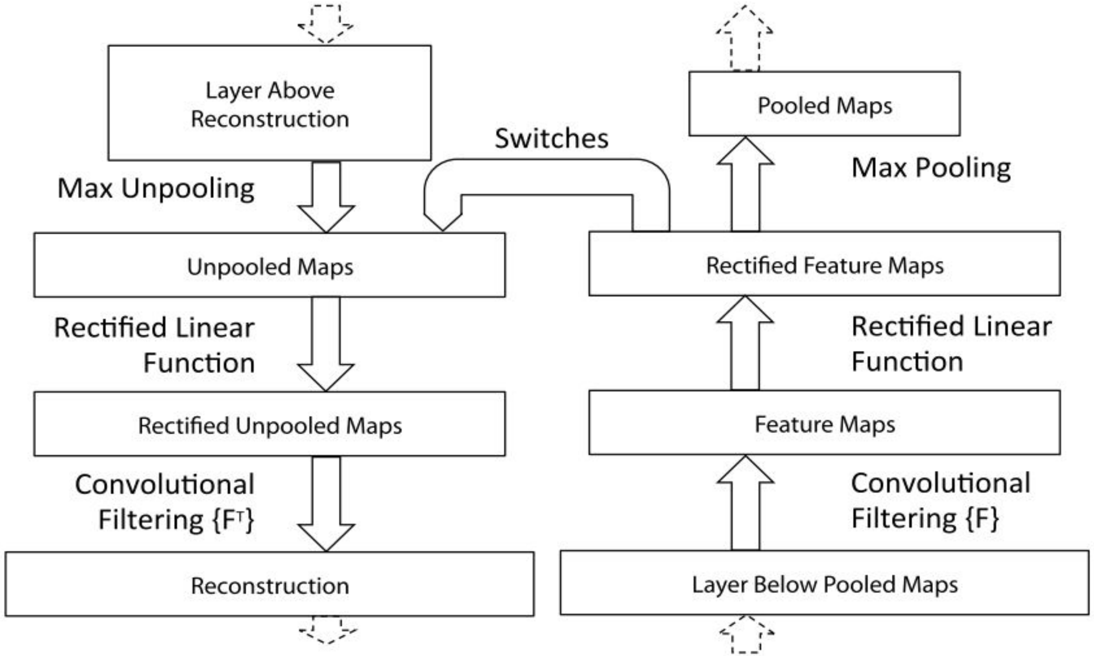
NEW YORK UNIVERSITY



Zeiler, Matthew D., Graham W. Taylor, and Rob Fergus. "Adaptive deconvolutional networks for mid and high level feature learning." *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE, 2011.

Zeiler-Fergus (ZF)

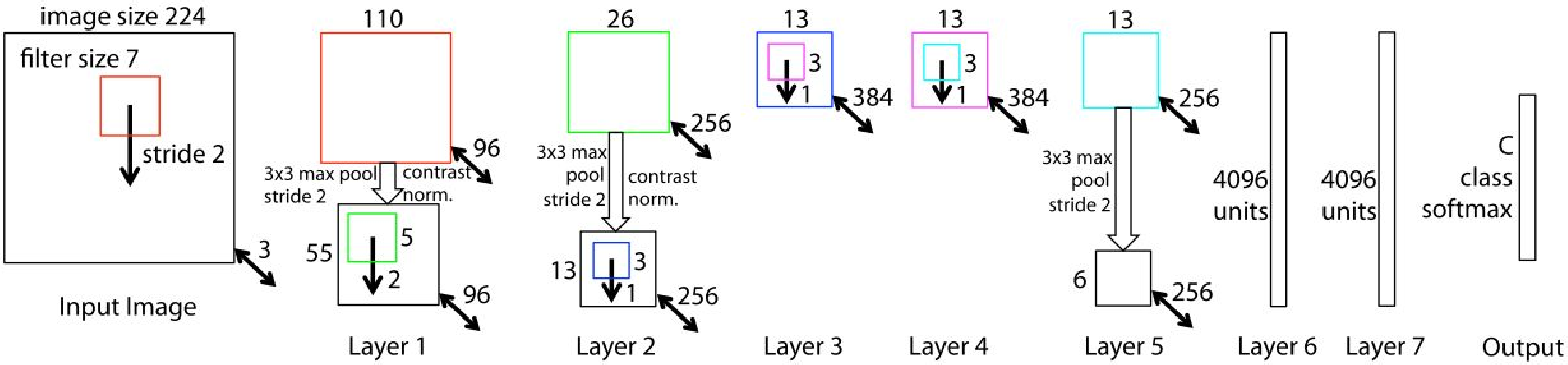
DeconvN



Conv

Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. In *Computer Vision–ECCV 2014* (pp. 818-833). Springer International Publishing.

Zeiler-Fergus (ZF)



Zeiler, M. D., & Fergus, R. (2014). [Visualizing and understanding convolutional networks](). In *Computer Vision–ECCV 2014* (pp. 818-833). Springer International Publishing.

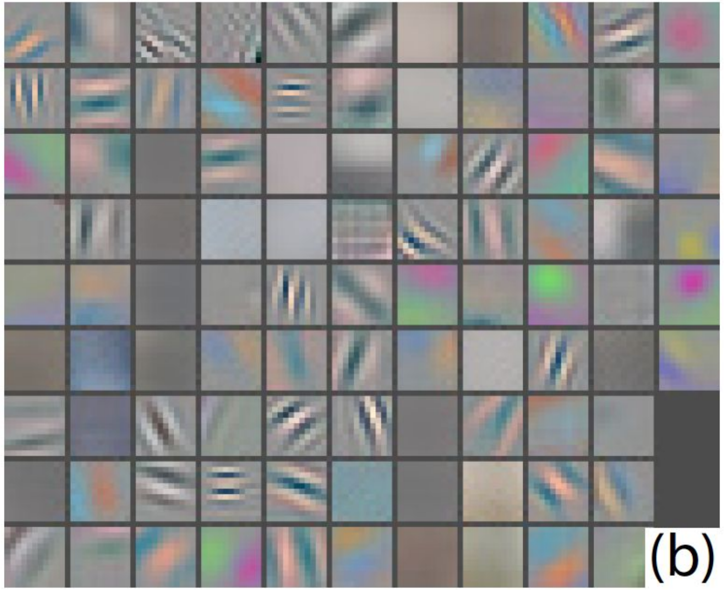
Zeiler-Fergus (ZF)

layer	size-in	size-out	kernel	param	FLPS
conv1	220×220×3	110×110×64	7×7×3, 2	9K	115M
pool1	110×110×64	55×55×64	3×3×64, 2	0	
rnorm1	55×55×64	55×55×64		0	
conv2a	55×55×64	55×55×64	1×1×64, 1	4K	13M
conv2	55×55×64	55×55×192	3×3×64, 1	111K	335M
rnorm2	55×55×192	55×55×192		0	
pool2	55×55×192	28×28×192	3×3×192, 2	0	
conv3a	28×28×192	28×28×192	1×1×192, 1	37K	29M
conv3	28×28×192	28×28×384	3×3×192, 1	664K	521M
pool3	28×28×384	14×14×384	3×3×384, 2	0	
conv4a	14×14×384	14×14×384	1×1×384, 1	148K	29M
conv4	14×14×384	14×14×256	3×3×384, 1	885K	173M
conv5a	14×14×256	14×14×256	1×1×256, 1	66K	13M
conv5	14×14×256	14×14×256	3×3×256, 1	590K	116M
conv6a	14×14×256	14×14×256	1×1×256, 1	66K	13M
conv6	14×14×256	14×14×256	3×3×256, 1	590K	116M
pool4	14×14×256	7×7×256	3×3×256, 2	0	
concat	7×7×256	7×7×256		0	
fc1	7×7×256	1×32×128	maxout p=2	103M	103M
fc2	1×32×128	1×32×128	maxout p=2	34M	34M
fc7128	1×32×128	1×1×128		524K	0.5M
L2	1×1×128	1×1×128		0	
total				140M	1.6B

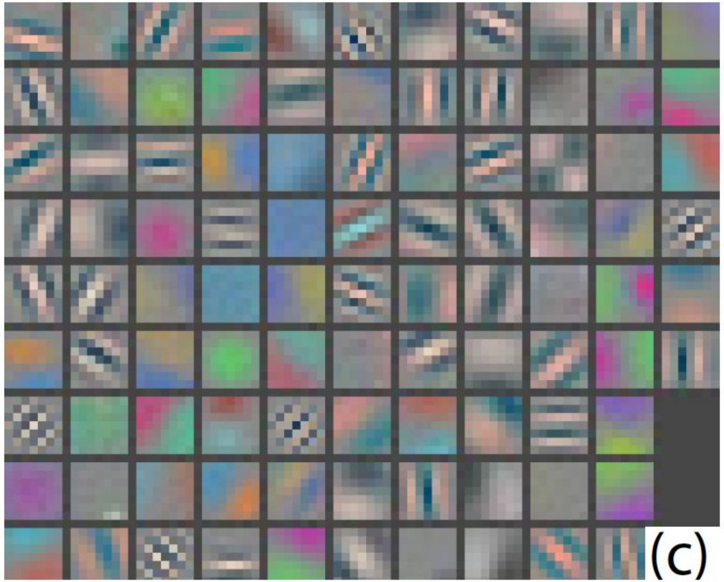
Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. In *Computer Vision–ECCV 2014* (pp. 818-833). Springer International Publishing.

Zeiler-Fergus (ZF): Stride & filter size

The smaller stride (2 vs 4) and filter size (7x7 vs 11x11) results in more distinctive features and fewer “dead" features.



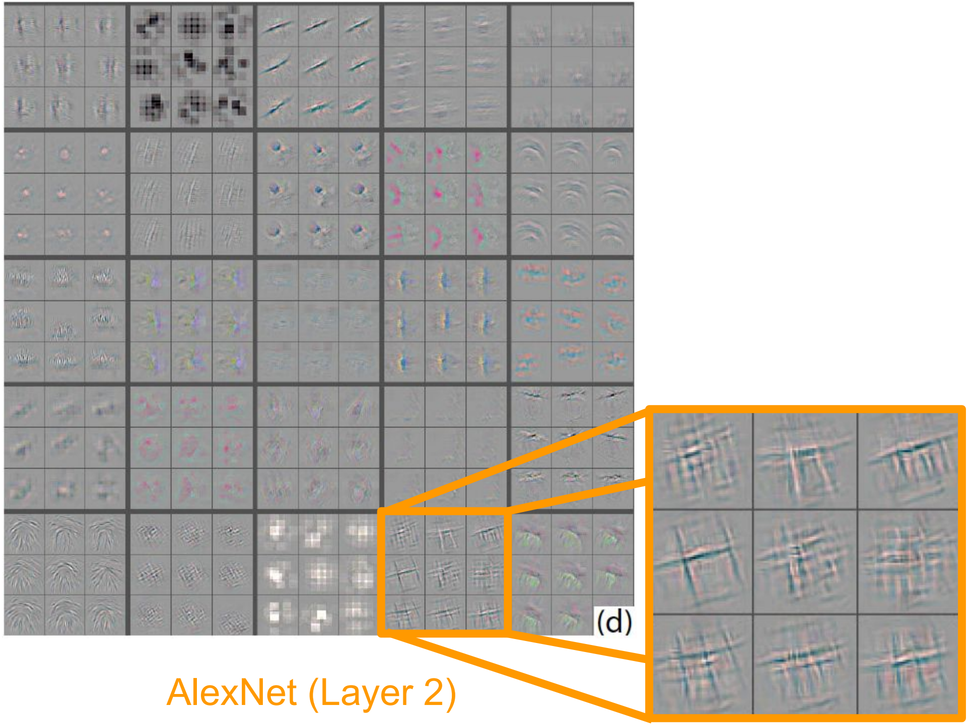
AlexNet (Layer 1)



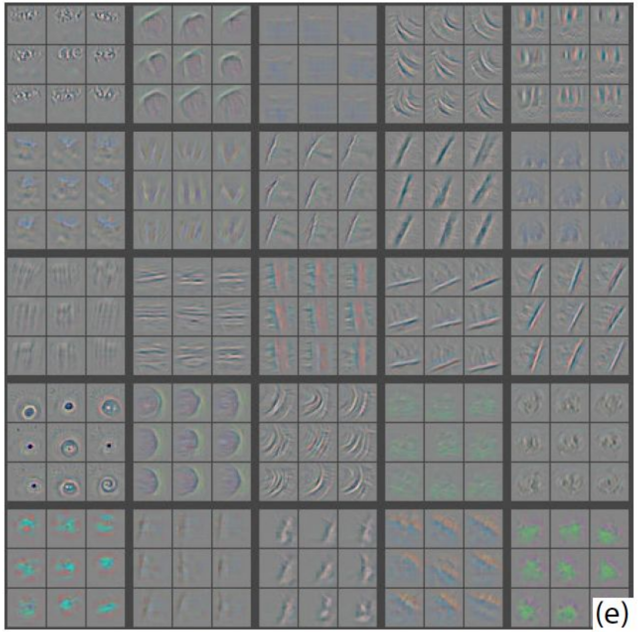
ZF (Layer 1)

Zeiler-Fergus (ZF)

Cleaner features in ZF, without the aliasing artifacts caused by the stride 4 used in AlexNet.



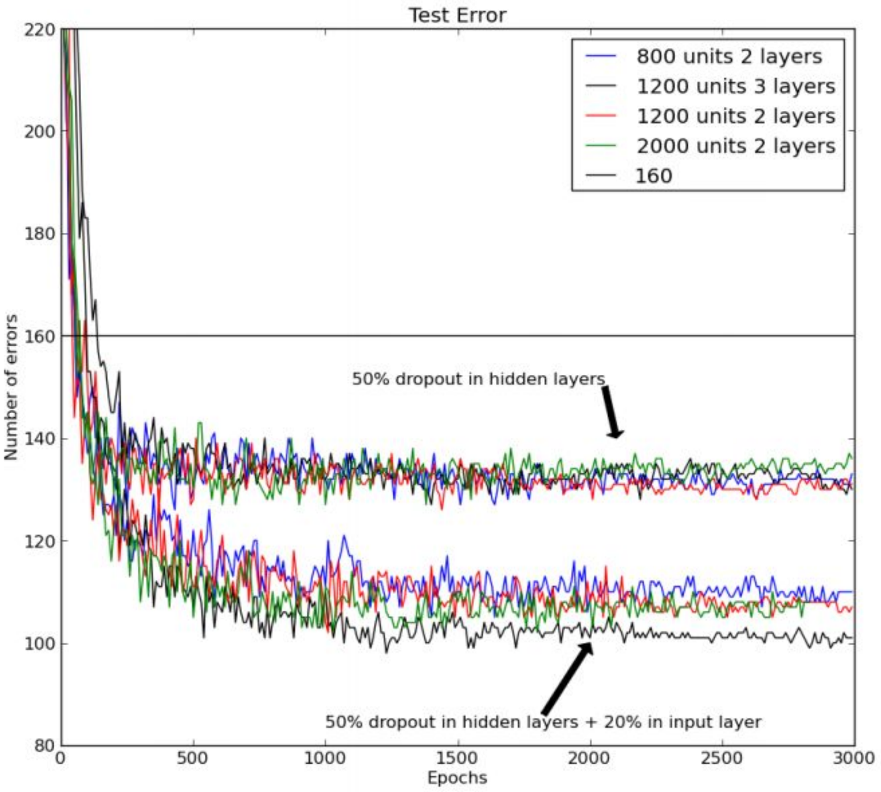
AlexNet (Layer 2)



ZF (Layer 2)

Zeiler-Fergus (ZF): Drop out

Regularization with more **dropout**: introduced in the input layer.



Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. R. (2012). [Improving neural networks by preventing co-adaptation of feature detectors](). *arXiv preprint arXiv:1207.0580*.

Zeiler-Fergus (ZF): Results

Error %	Val Top-1	Val Top-5	Test Top-5
(Gunji et al., 2012)	-	-	26.2
(Krizhevsky et al., 2012), 1 convnet	40.7	18.2	−−
(Krizhevsky et al., 2012), 5 convnets	38.1	16.4	16.4
(Krizhevsky et al., 2012)*, 1 convnets	39.0	16.6	−−
(Krizhevsky et al., 2012)*, 7 convnets	36.7	15.4	15.3
Our replication of (Krizhevsky et al., 2012), 1 convnet	40.5	18.1	−−
1 convnet as per Fig. 3	38.4	16.5	−−
5 convnets as per Fig. 3 – (a)	36.7	15.3	15.3
1 convnet as per Fig. 3 but with layers 3,4,5: 512,1024,512 maps – (b)	37.5	16.0	16.1
6 convnets, (a) & (b) combined	**36.0**	**14.7**	**14.8**

Table 2. ImageNet 2012 classification error rates. The $*$ indicates models that were trained on both ImageNet 2011 and 2012 training sets.

Zeiler-Fergus (ZF): Results

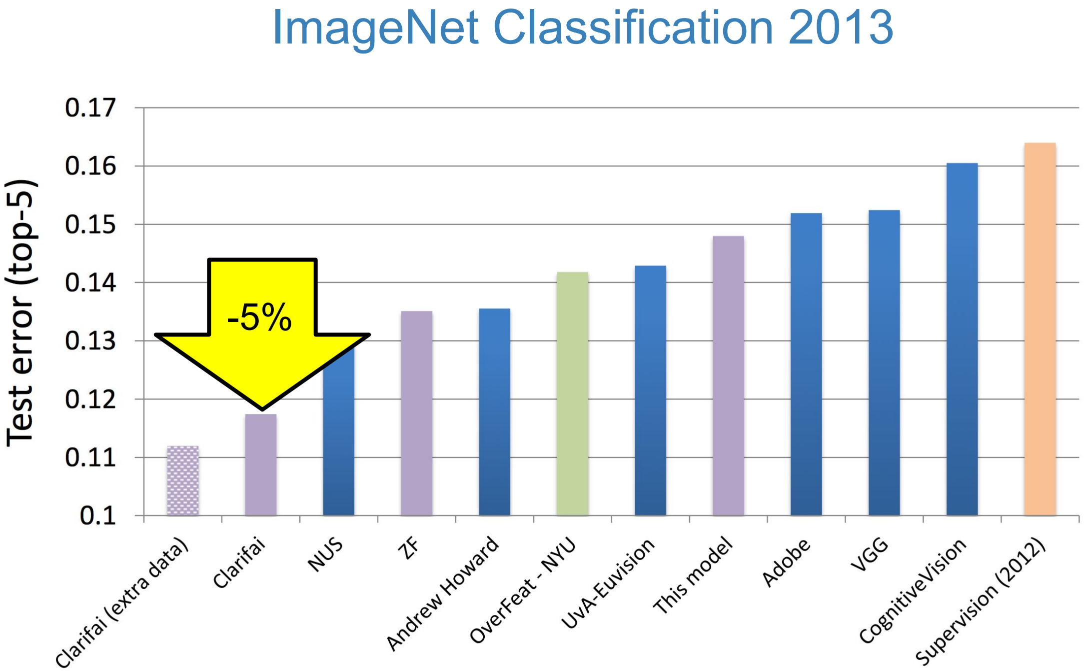
Error %	Val Top-1	Val Top-5	Test Top-5
(Gunji et al., 2012)	-	-	26.2
(Krizhevsky et al., 2012), 1 convnet	40.7	18.2	−−
(Krizhevsky et al., 2012), 5 convnets	38.1	16.4	16.4
(Krizhevsky et al., 2012)*, 1 convnets	39.0	16.6	−−
(Krizhevsky et al., 2012)*, 7 convnets	36.7	15.4	15.3
Our replication of (Krizhevsky et al., 2012), 1 convnet	40.5	18.1	−−
1 convnet as per Fig. 3	38.4	16.5	−−
5 convnets as per Fig. 3 – (a)	36.7	15.3	15.3
1 convnet as per Fig. 3 but with layers 3,4,5: 512,1024,512 maps – (b)	37.5	16.0	16.1
6 convnets, (a) & (b) combined	**36.0**	**14.7**	**14.8**

Table 2. ImageNet 2012 classification error rates. The $*$ indicates models that were trained on both ImageNet 2011 and 2012 training sets.

E2E: Classification: ImageNet ILSRVC

IMAGENET

clarifai



Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., ... & Fei-Fei, L. (2015). Imagenet large scale visual recognition challenge. *arXiv preprint arXiv:1409.0575*. [web]

E2E: Classification

AlexNet

image
conv-64
conv-192
conv-384
conv-256
conv-256
FC-4096
FC-4096
FC-1000

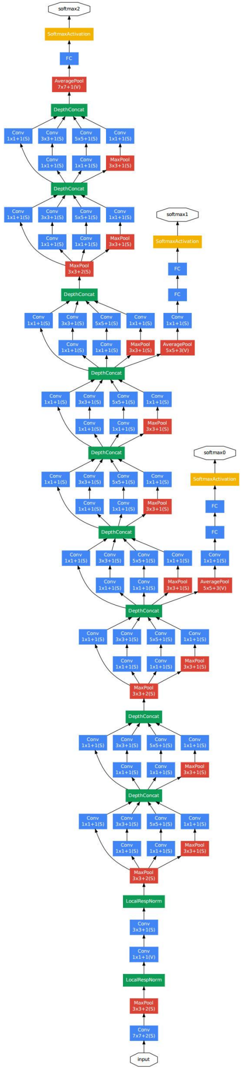


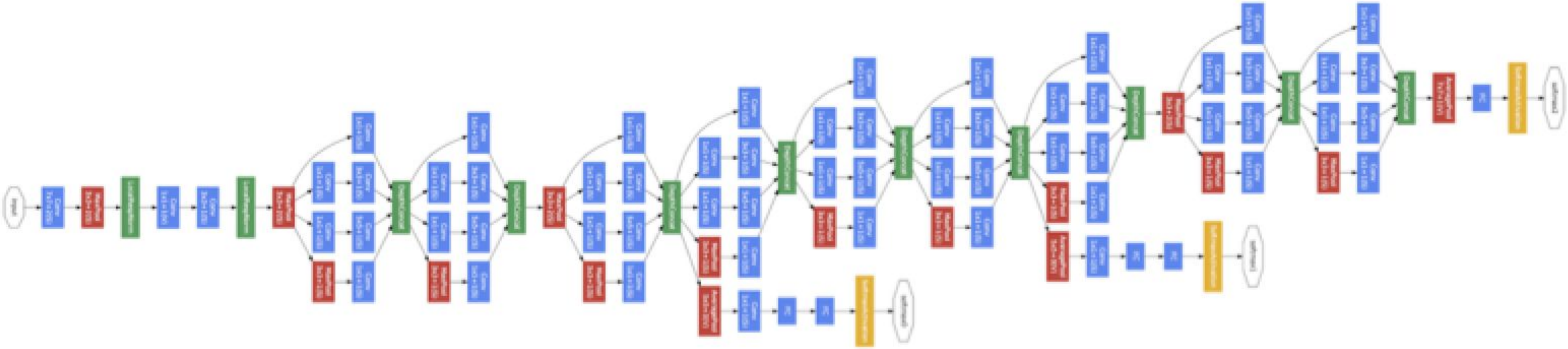
image
conv-64
conv-64
maxpool
conv-128
conv-128
maxpool
conv-256
conv-256
conv-256
conv-256
maxpool
conv-512
conv-512
conv-512
conv-512
maxpool
conv-512
conv-512
conv-512
conv-512
maxpool
FC-4096
FC-4096
FC-1000
softmax

E2E: Classification: GoogLeNet

Movie: Inception (2010)

E2E: Classification: GoogLeNet

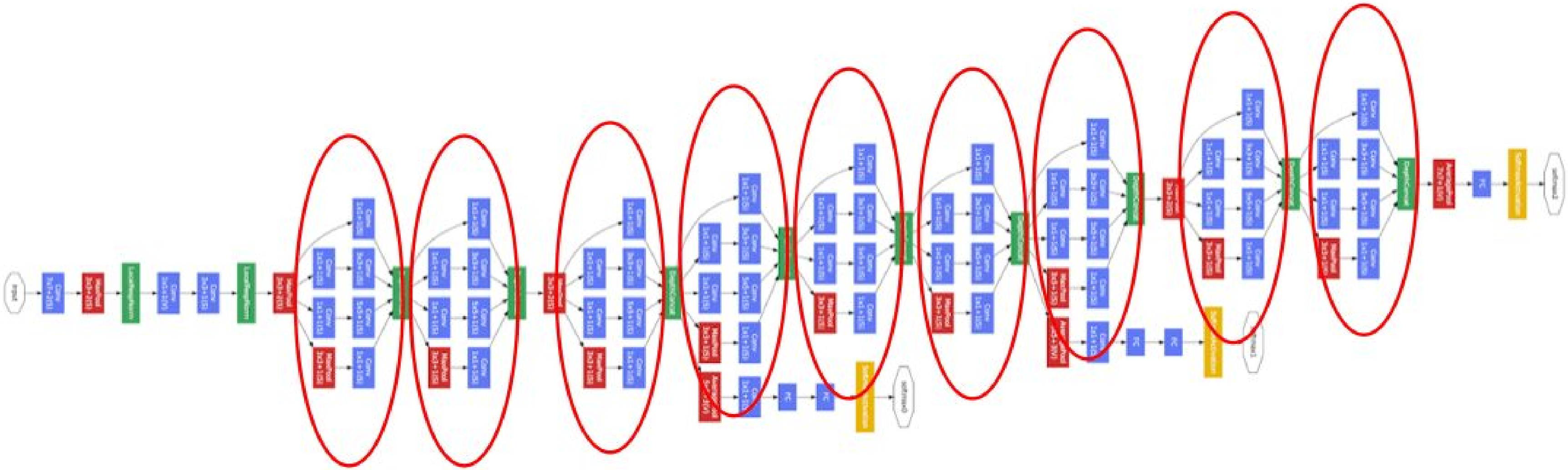
- 22 layers, but 12 times fewer parameters than AlexNet.



Convolution
Pooling
Softmax
Other

Szegedy, Christian, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. "Going deeper with convolutions." CVPR 2015. [video] [slides] [poster]

E2E: Classification: GoogLeNet

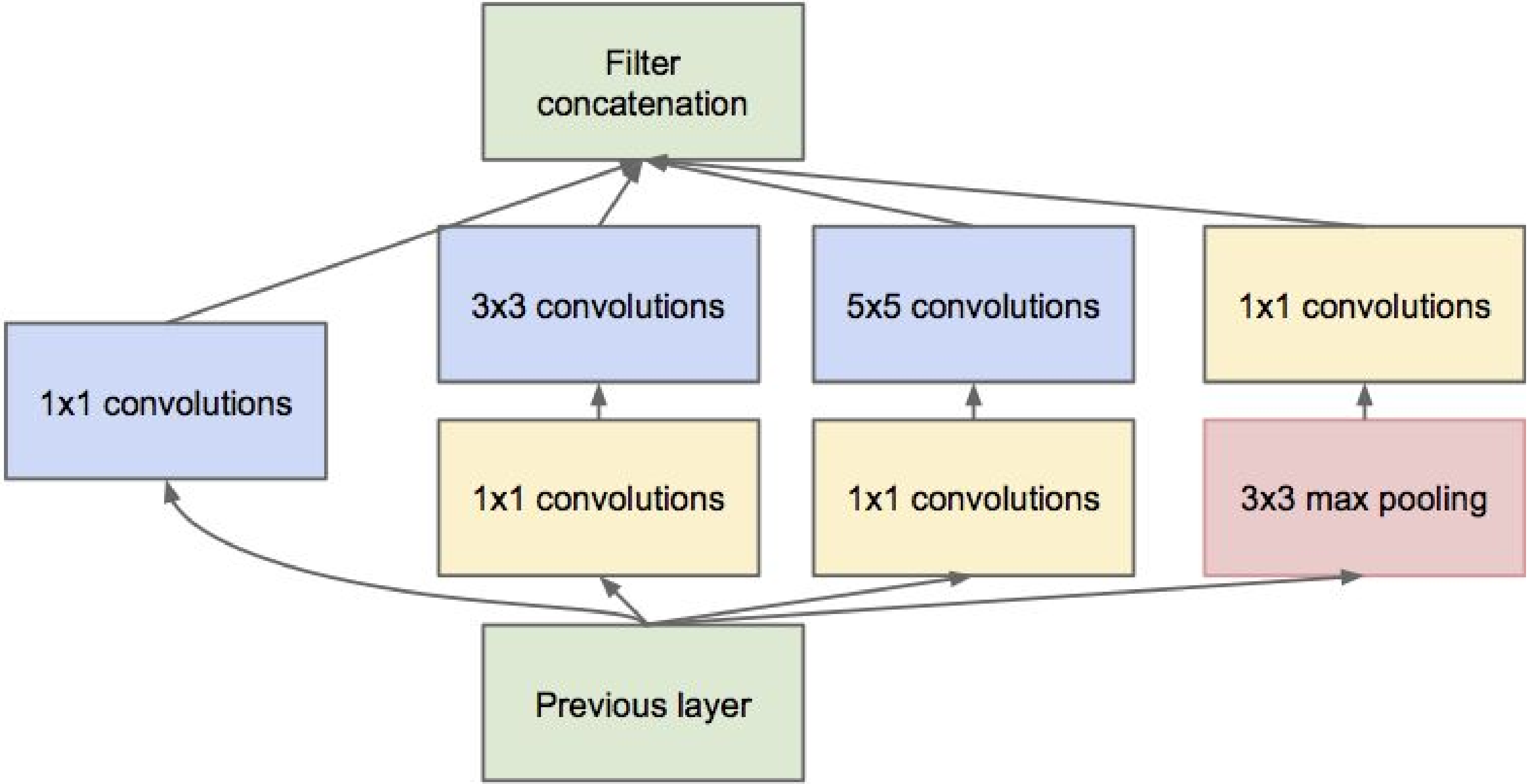


9 **Inception** modules

Network in a network in a network...

Convolution
Pooling
Softmax
Other

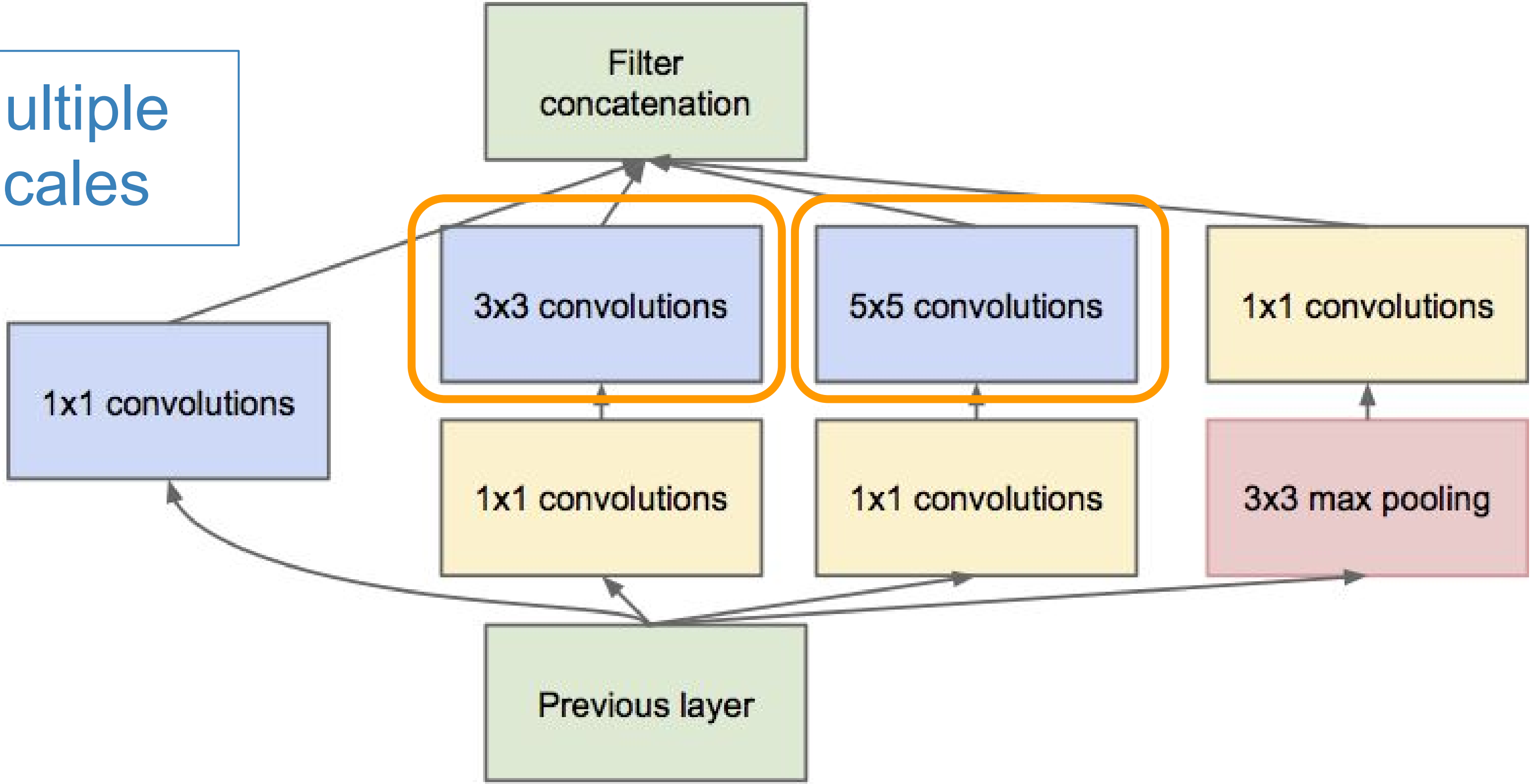
E2E: Classification: GoogLeNet



Lin, Min, Qiang Chen, and Shuicheng Yan. "Network in network." *ICLR 2014.*

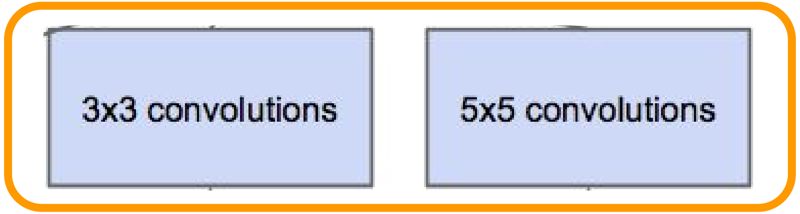
E2E: Classification: GoogLeNet

Multiple scales

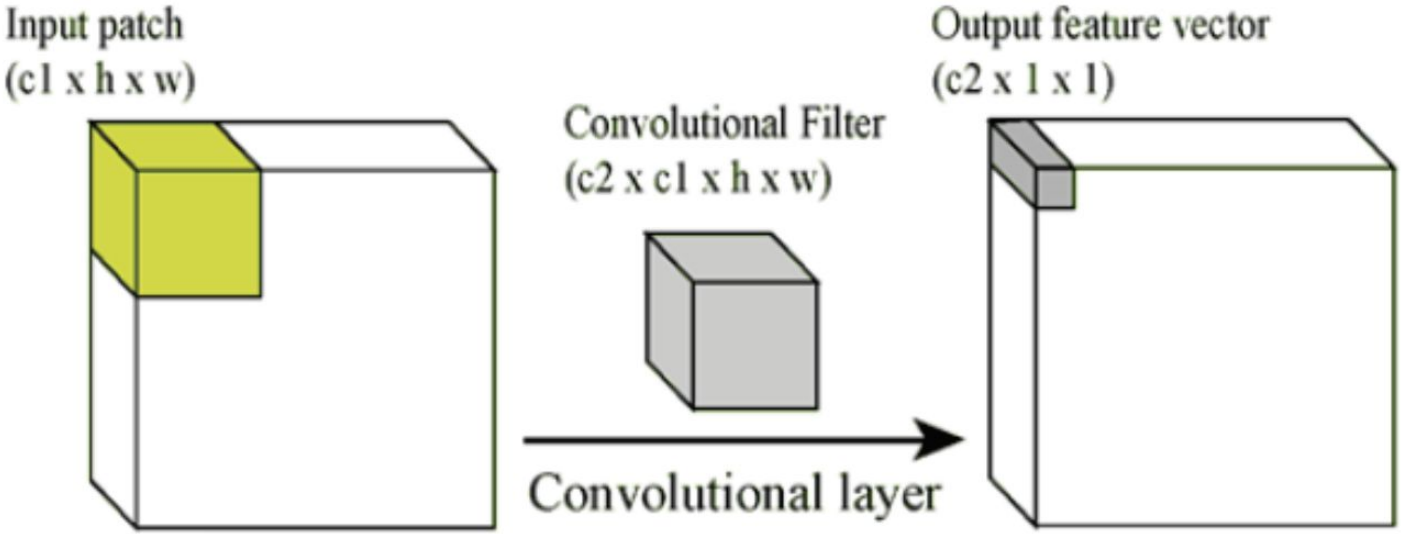


Lin, Min, Qiang Chen, and Shuicheng Yan. "Network in network." *ICLR 2014.*

E2E: Classification: GoogLeNet (NiN)



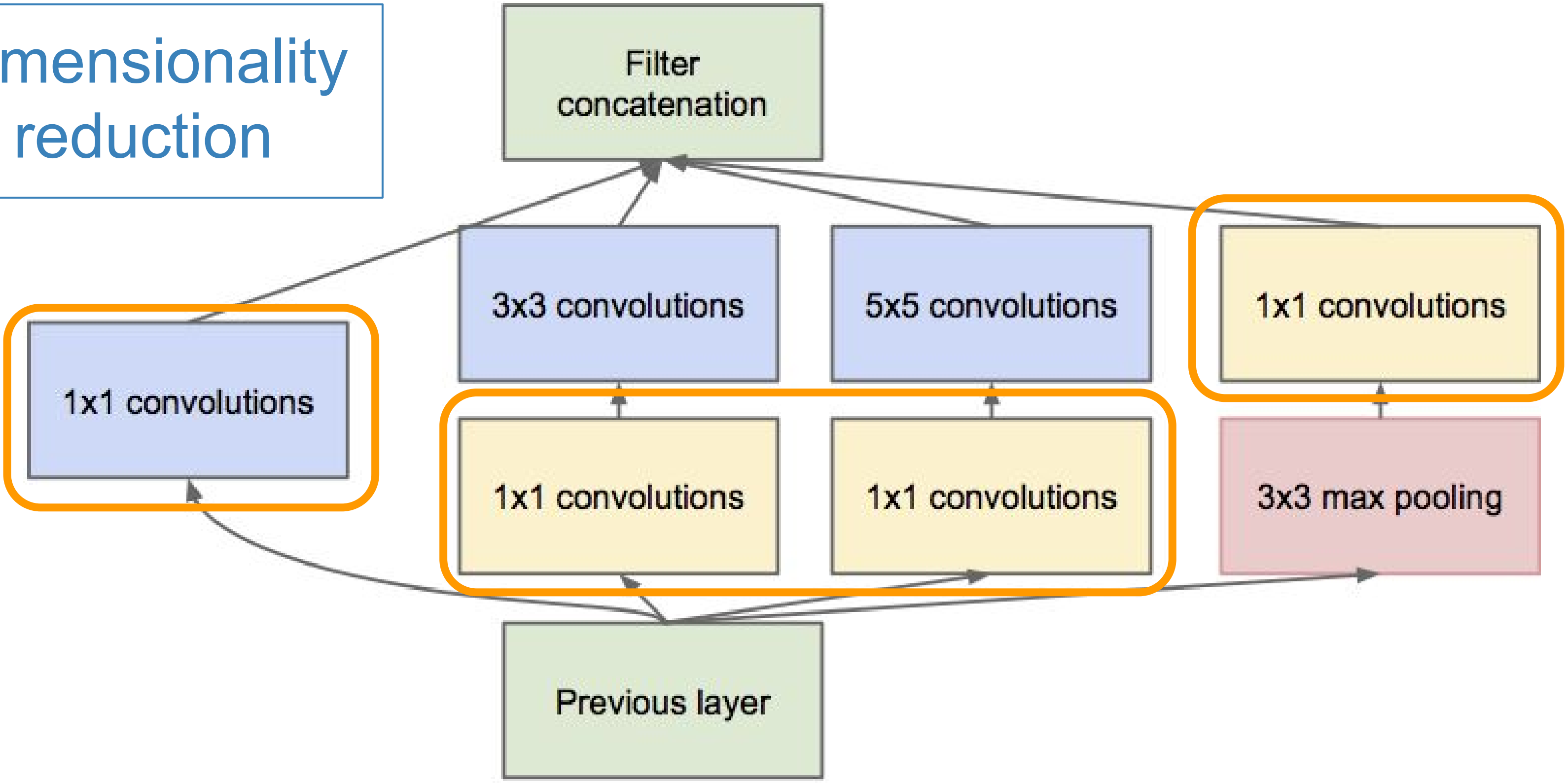
3x3 and 5x5 convolutions deal with different scales.



Lin, Min, Qiang Chen, and Shuicheng Yan. "Network in network." *ICLR 2014. [Slides]*

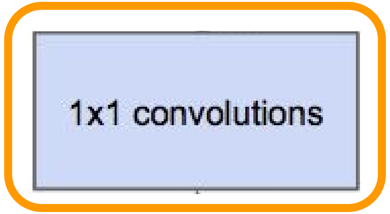
E2E: Classification: GoogLeNet

Dimensionality reduction

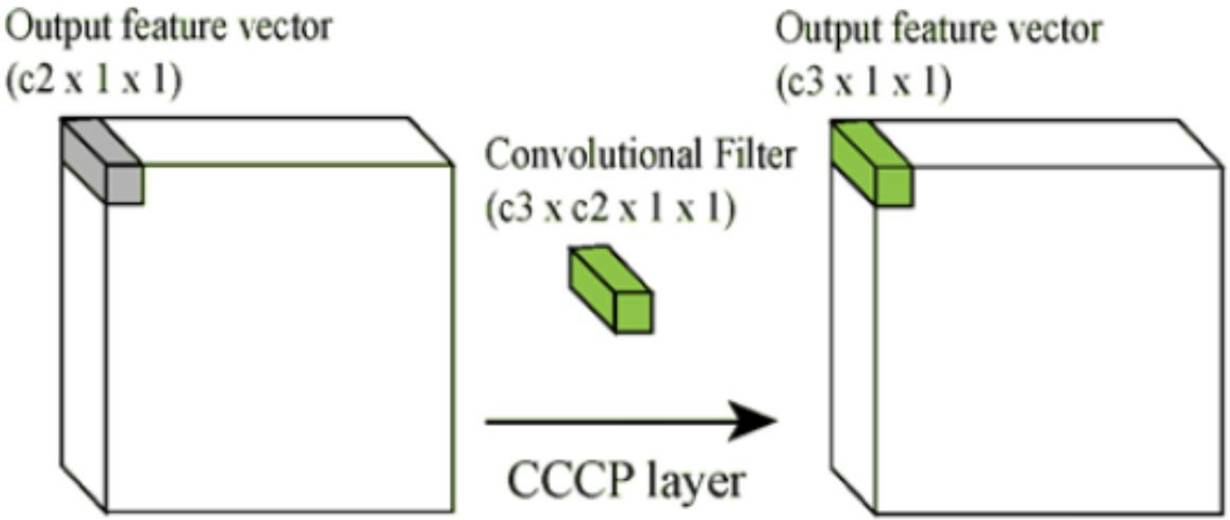


Lin, Min, Qiang Chen, and Shuicheng Yan. "Network in network." *ICLR 2014.*

E2E: Classification: GoogLeNet (NiN)



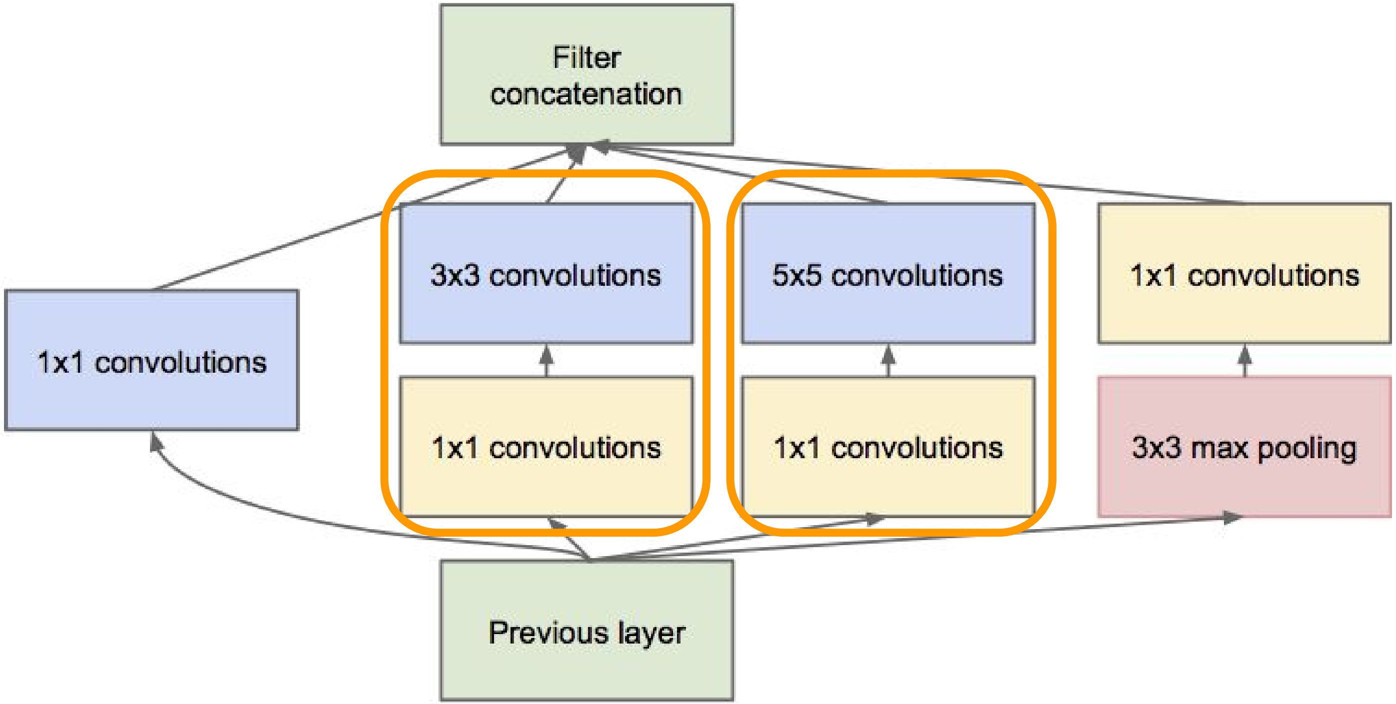
1x1 convolutions does <u>dimensionality reduction</u> (c3<c2) and accounts for rectified linear units (ReLU).



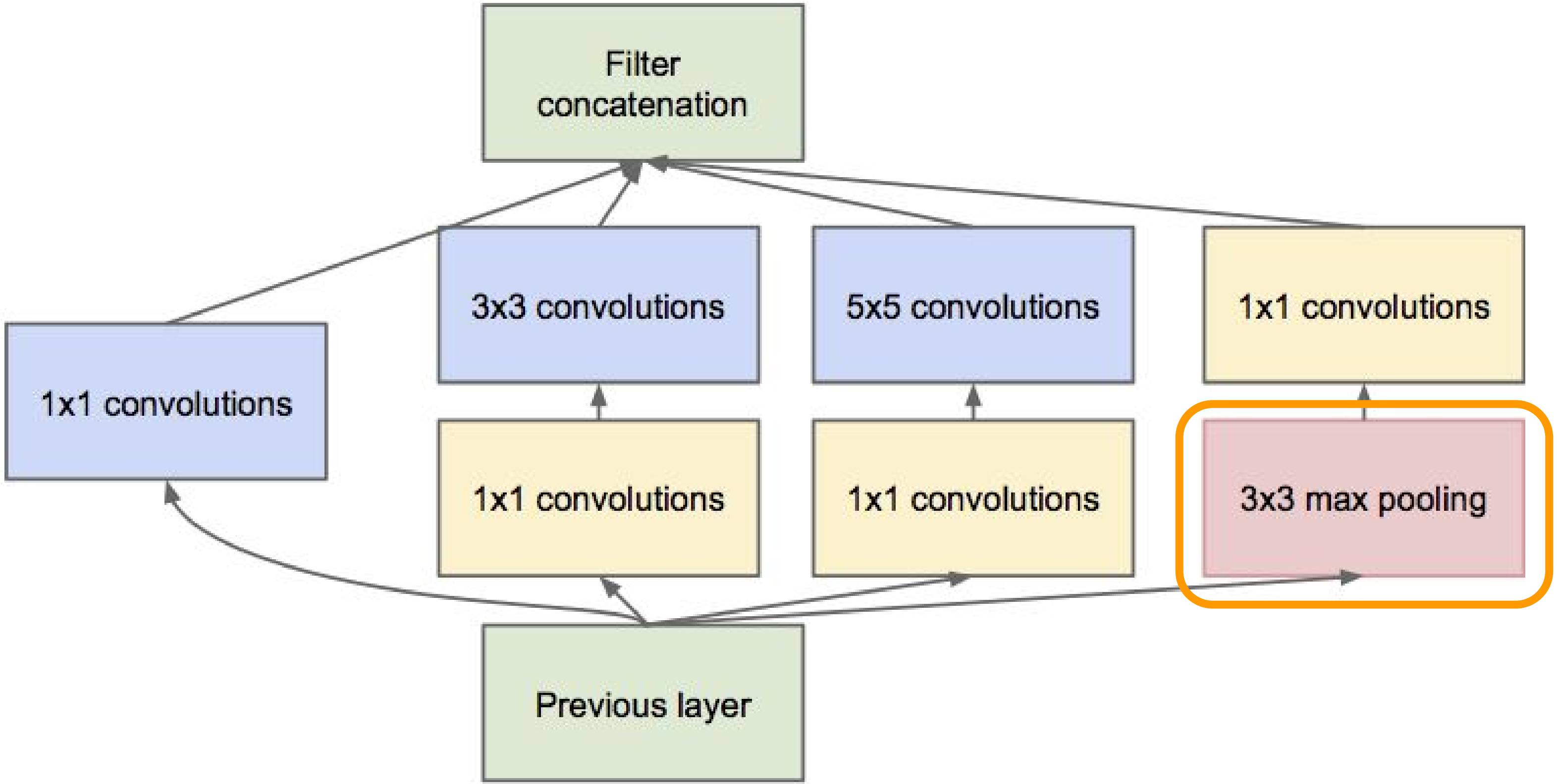
Lin, Min, Qiang Chen, and Shuicheng Yan. ["Network in network."](#) *ICLR 2014. [Slides]*

E2E: Classification: GoogLeNet

In GoogLeNet, **the Cascaded 1x1 Convolutions** compute reductions <u>before</u> the expensive 3x3 and 5x5 convolutions.

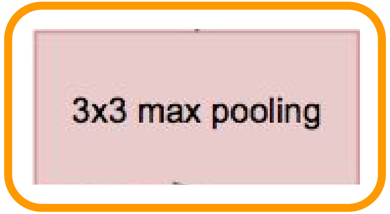


E2E: Classification: GoogLeNet

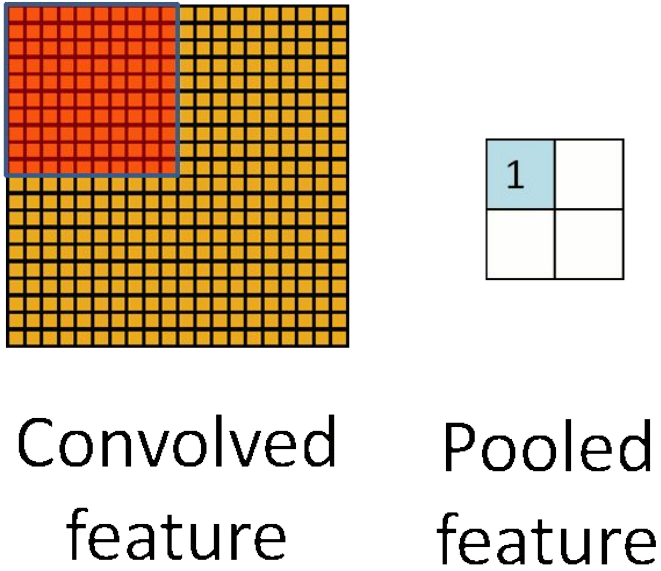


Lin, Min, Qiang Chen, and Shuicheng Yan. ["Network in network."](#) *ICLR 2014.*

E2E: Classification: GoogLeNet

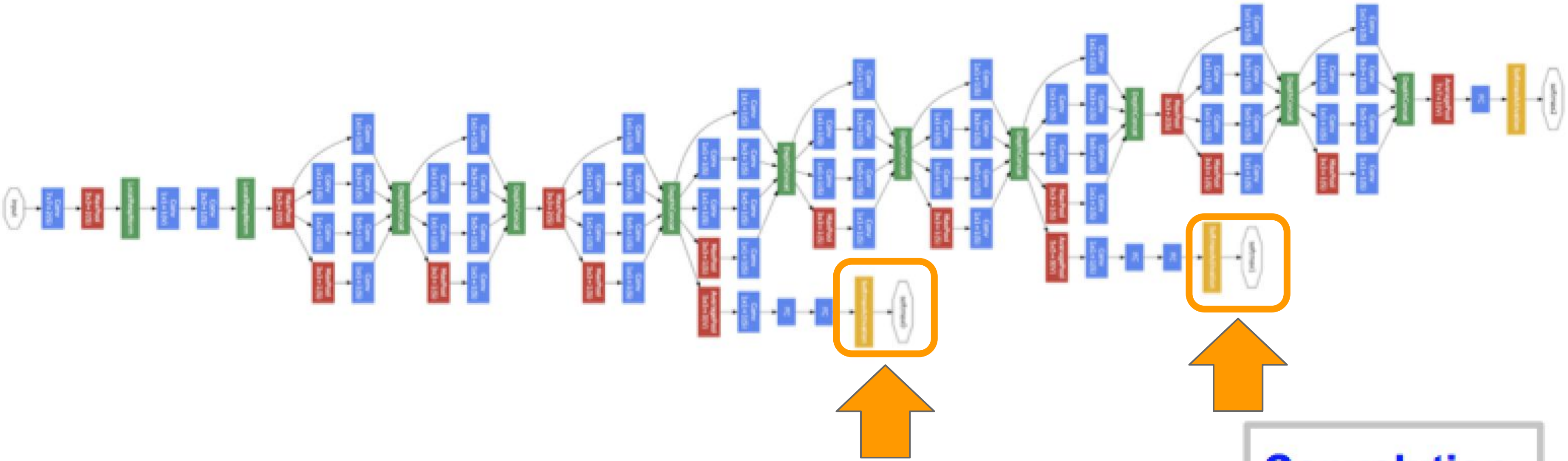


They somewhat spatial invariance, and has proven a benefitial effect by adding an alternative parallel path.



E2E: Classification: GoogLeNet

Two Softmax Classifiers at intermediate layers combat the vanishing gradient while providing regularization at training time.

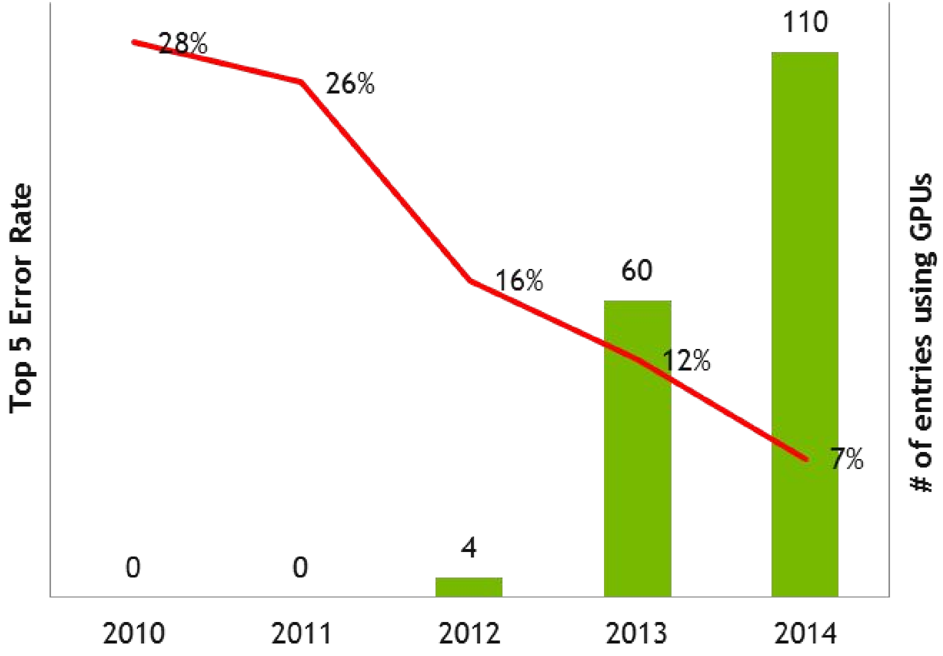


...and no fully connected layers needed !

E2E: Classification: GoogLeNet

type	output size	depth	#1×1	#3×3 reduce	#3×3	#5×5 reduce	#5×5	pool proj (p)	params	FLOPS
conv1 (7×7×3, 2)	112×112×64	1							9K	119M
max pool + norm	56×56×64	0						m 3×3, 2		
inception (2)	56×56×192	2		64	192				115K	360M
norm + max pool	28×28×192	0						m 3×3, 2		
inception (3a)	28×28×256	2	64	96	128	16	32	m, 32p	164K	128M
inception (3b)	28×28×320	2	64	96	128	32	64	L_2, 64p	228K	179M
inception (3c)	14×14×640	2	0	128	256,2	32	64,2	m 3×3,2	398K	108M
inception (4a)	14×14×640	2	256	96	192	32	64	L_2, 128p	545K	107M
inception (4b)	14×14×640	2	224	112	224	32	64	L_2, 128p	595K	117M
inception (4c)	14×14×640	2	192	128	256	32	64	L_2, 128p	654K	128M
inception (4d)	14×14×640	2	160	144	288	32	64	L_2, 128p	722K	142M
inception (4e)	7×7×1024	2	0	160	256,2	64	128,2	m 3×3,2	717K	56M
inception (5a)	7×7×1024	2	384	192	384	48	128	L_2, 128p	1.6M	78M
inception (5b)	7×7×1024	2	384	192	384	48	128	m, 128p	1.6M	78M
avg pool	1×1×1024	0								
fully conn	1×1×128	1							131K	0.1M
L2 normalization	1×1×128	0								
total									7.5M	1.6B

E2E: Classification: GoogLeNet



NVIDIA, [“NVIDIA and IBM CLoud Support ImageNet Large Scale Visual Recognition Challenge”]() (2015)

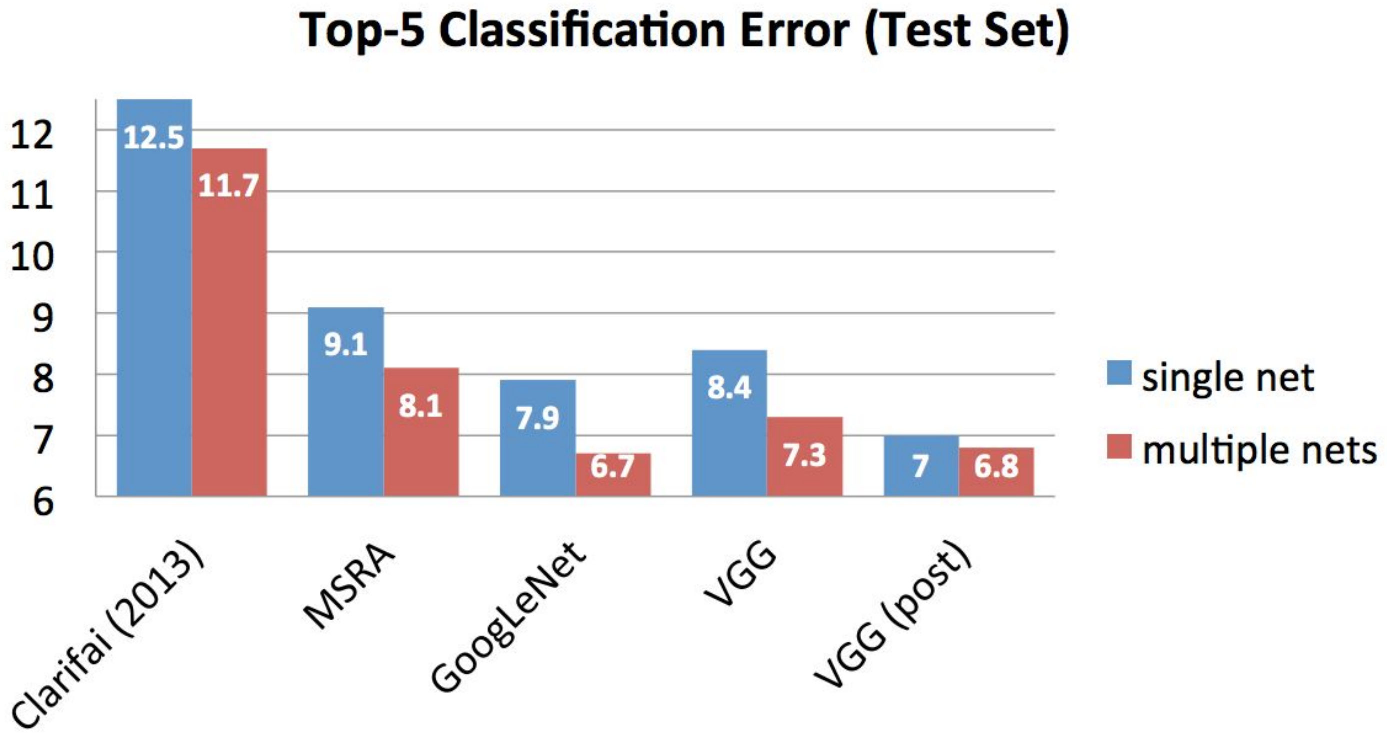
E2E: Classification: GoogLeNet

Szegedy, Christian, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. "Going deeper with convolutions." CVPR 2015. [video] [slides] [poster]

E2E: Classification: VGG

Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." *International Conference on Learning Representations (2015)*. [video] [slides] [project]

E2E: Classification: VGG

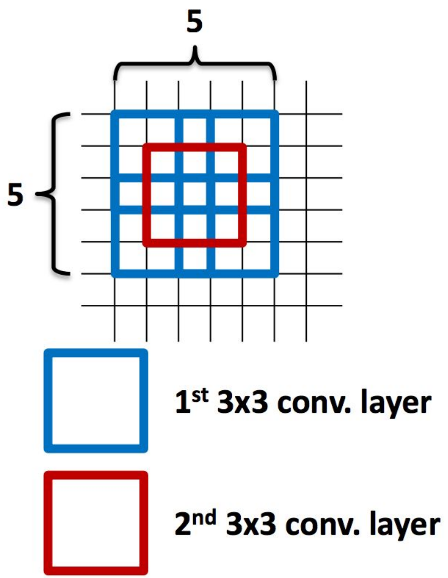


Simonyan, Karen, and Andrew Zisserman. ["Very deep convolutional networks for large-scale image recognition."] *International Conference on Learning Representations (2015)*. [video] [slides] [project]

E2E: Classification: VGG: 3x3 Stacks

Why 3x3 layers?

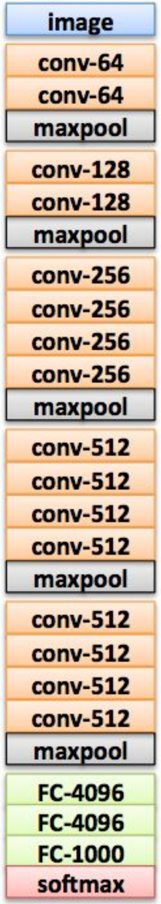
- Stacked conv. layers have a large receptive field
  - two 3x3 layers – 5x5 receptive field
  - three 3x3 layers – 7x7 receptive field
- More non-linearity
- Less parameters to learn
  - ~140M per net



Simonyan, Karen, and Andrew Zisserman. ["Very deep convolutional networks for large-scale image recognition."]() *International Conference on Learning Representations (2015)*. [video] [slides] [project]

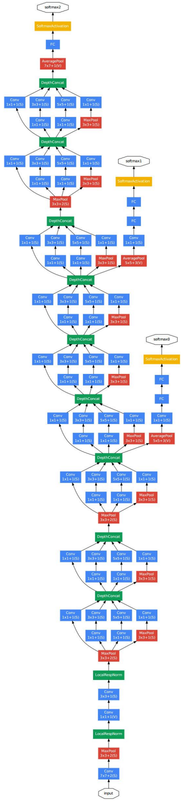
E2E: Classification: VGG

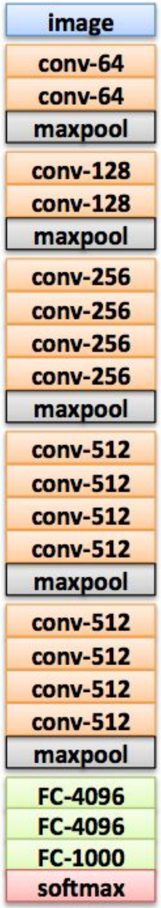
- No poolings between some convolutional layers.
- Convolution strides of 1 (no skipping).

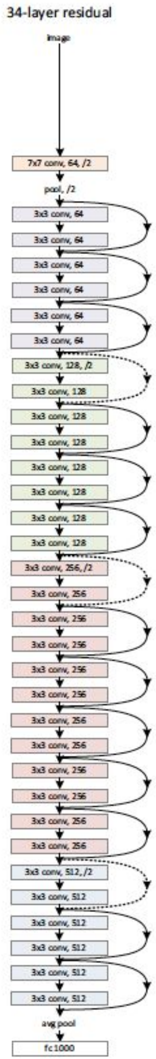


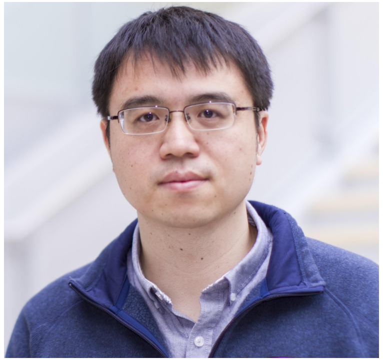
Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." *International Conference on Learning Representations (2015)*. [video] [slides] [project]

E2E: Classification



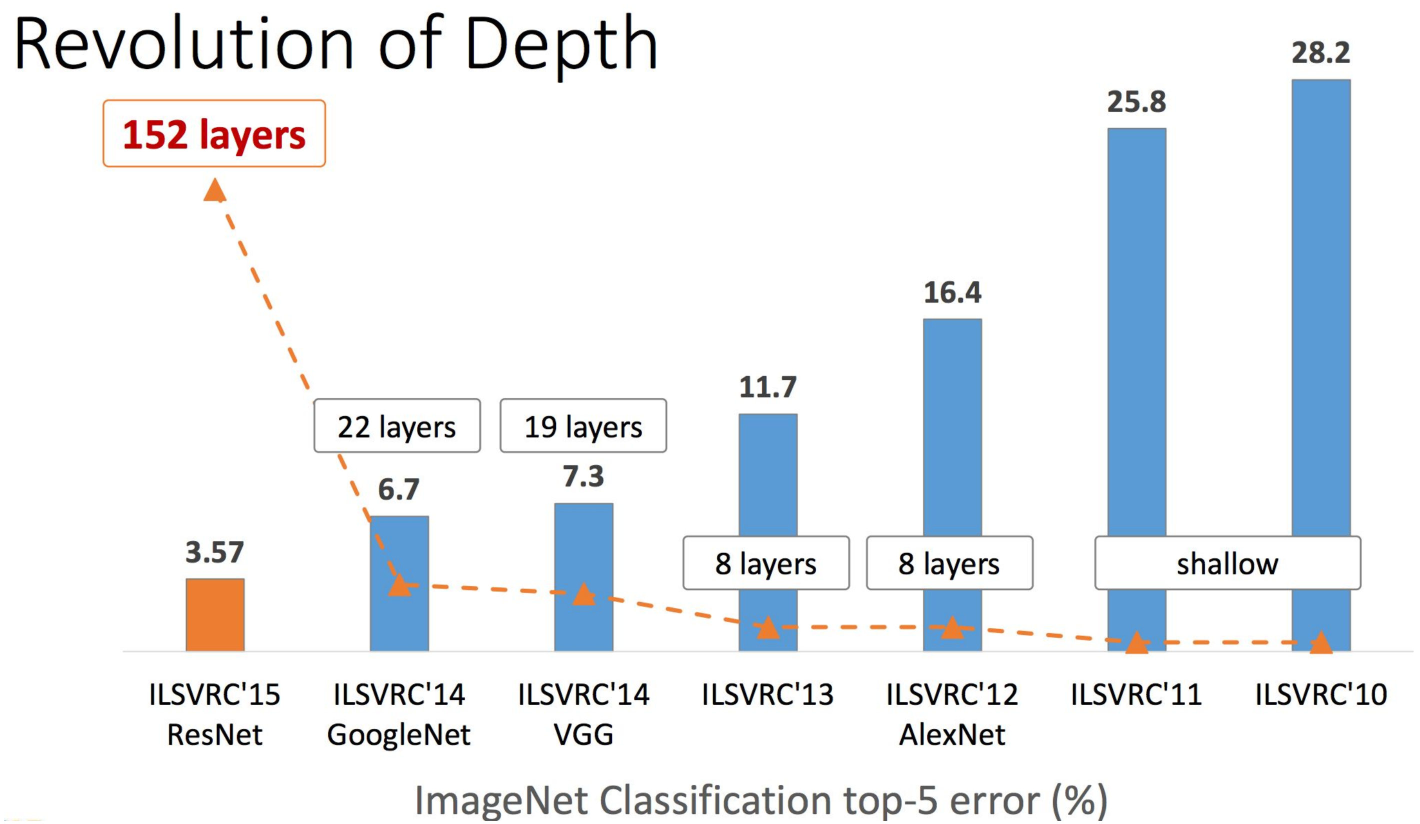






3.6% top 5 error…
with 152 layers !!

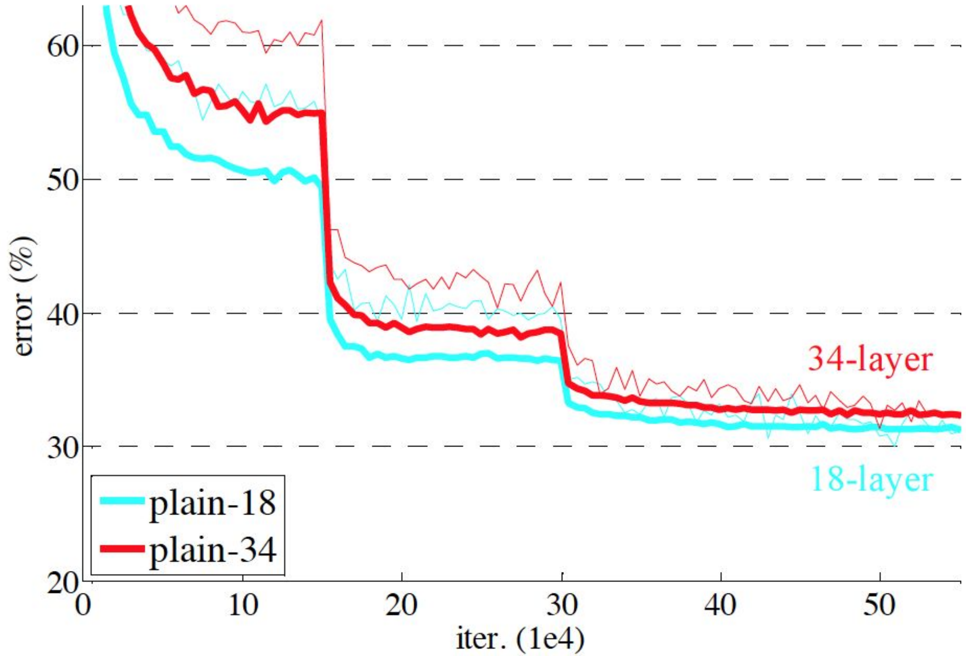
E2E: Classification: ResNet



He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep Residual Learning for Image Recognition." *arXiv preprint arXiv:1512.03385* (2015). [slides]

E2E: Classification: ResNet

- Deeper networks (34 is deeper than 18) are more difficult to train.

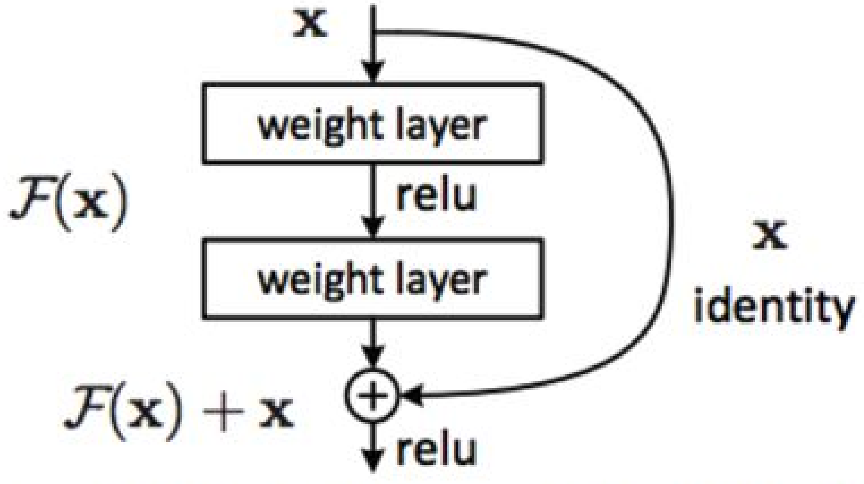


Thin curves: training error
Bold curves: validation error

He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. [“Deep Residual Learning for Image Recognition.”](#) *arXiv preprint arXiv:1512.03385* (2015). [[slides]](#)

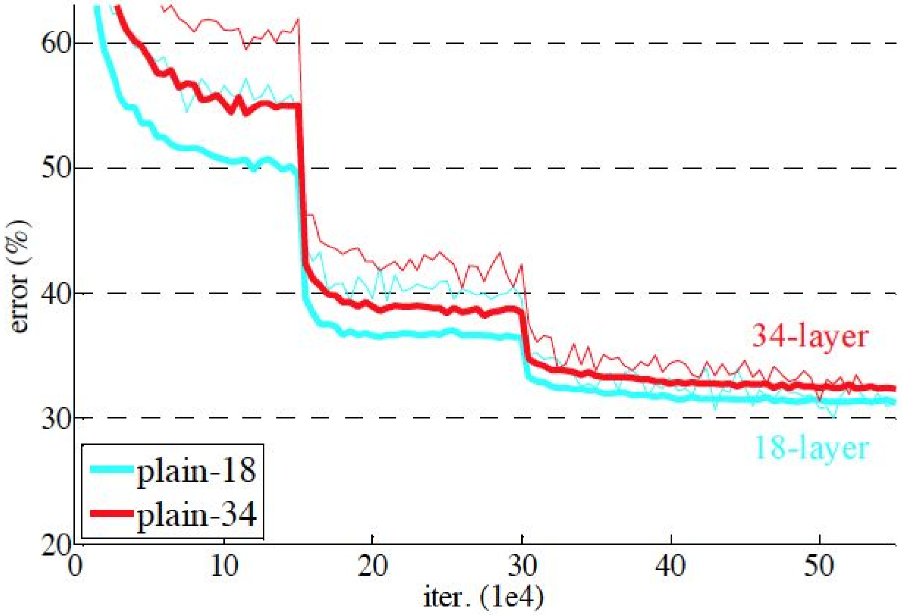
ResNet

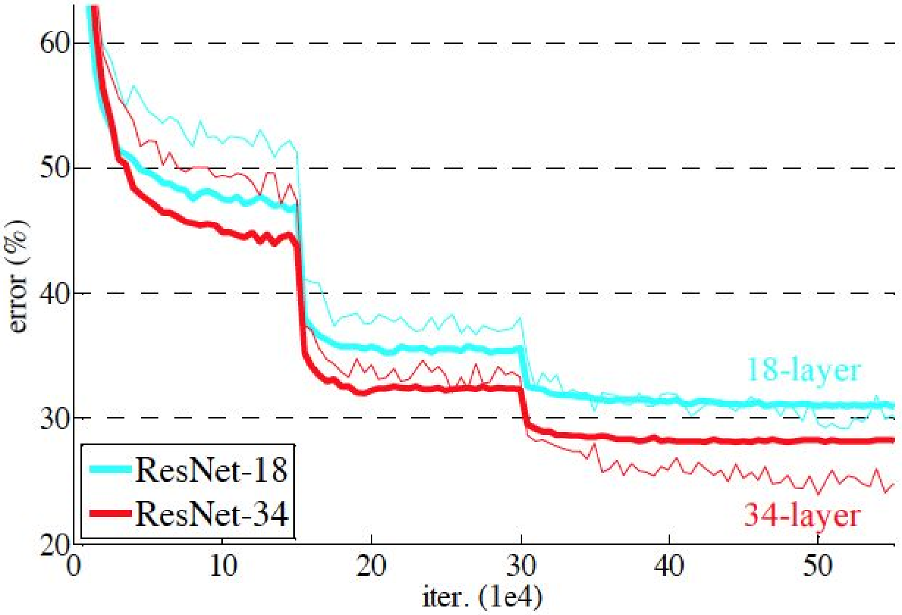
- Residual learning: reformulate the layers as learning residual functions with reference to the layer inputs, instead of learning unreferenced functions



He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep Residual Learning for Image Recognition." *arXiv preprint arXiv:1512.03385* (2015). [slides]

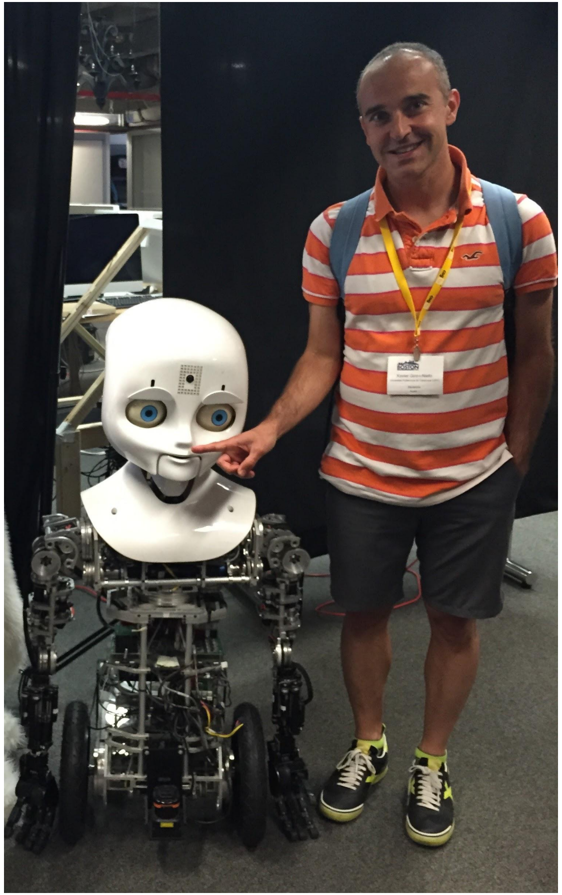
E2E: Classification: ResNet





He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep Residual Learning for Image Recognition." *arXiv preprint arXiv:1512.03385* (2015). [slides]

Thanks ! Q&A ?



Follow me at

/ProfessorXavi

@DocXavi

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Signal Theory
and Communications
Image Processing Group

https://imatge.upc.edu/web/people/xavier-giro