Introduction to Convolutional Neural Networks

Jon Shlens
Google Brain
29 July 2018

Agenda

1. Challenges and inspiration from vision

2. Convolutional neural networks

3. Modern developments

    - architectures, meta-learning, normalization, transfer learning

4. Towards understanding higher-level visual features

5. Opportunities and conclusions

Agenda

1. **Challenges and inspiration from vision**
2. Convolutional neural networks
3. Modern developments
    - architectures, meta-learning, normalization, transfer learning
4. Towards understanding higher-level visual features
5. Opportunities and conclusions

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

PROJECT MAC

Artificial Intelligence Group
Vision Memo. No. 100.

July 7, 1966

THE SUMMER VISION PROJECT

Seymour Papert

The summer vision project is an attempt to use our summer workers effectively in the construction of a significant part of a visual system. The particular task was chosen partly because it can be segmented into sub-problems which will allow individuals to work independently and yet participate in the construction of a system complex enough to be a real landmark in the development of "pattern recognition".

Examples of artificial vision in action

- fine-grain classification

hibiscus

dahila

- generalization

meal

meal

- sensible errors

snake

dog

*** Trained a model for whole image recognition using Inception-v3 architecture.*

“So how exactly the computer sees? – The thing is most of computer vision researchers do not really understand how the computers see.

It’s like alchemy and chemistry. Alchemy came first and chemistry came then. And right now we are in the alchemy stage of computer vision, where it works but we are not sure why. And it is the chemistry stage that I look forward to.”

- Bill Freeman

“How Computer Vision Is Finally Taking Off, After 50 Years”

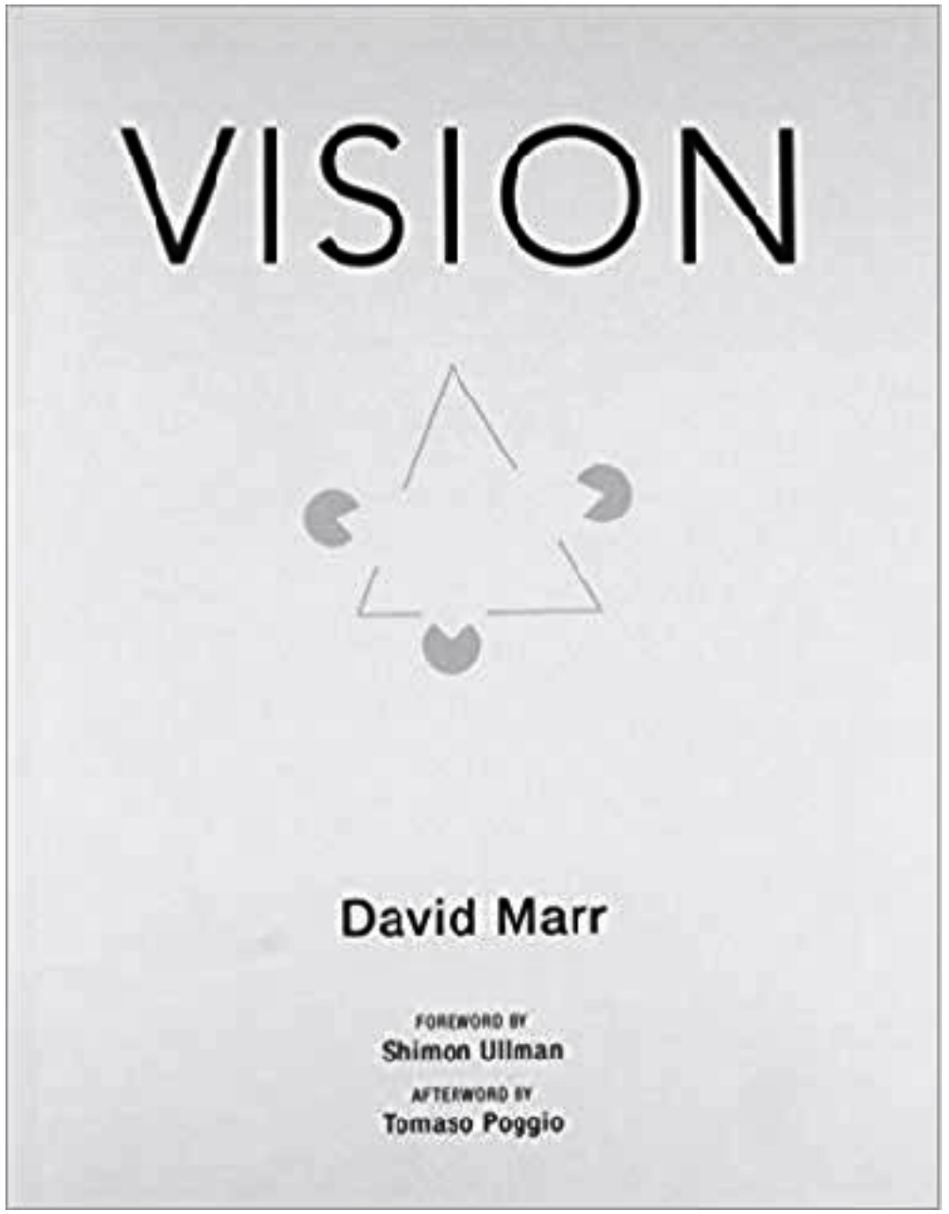
https://www.youtube.com/watch?v=eQLcDmfmGB0

Marr's Levels of Analysis

Computational
- What does the system do?
- What problems does it solve?
- Why does it do these things?

Algorithmic / Representational
- How does the system do what it does?
- What representations does it use?
- What processes does it employ to build and manipulate the representations?

Implementational / Physical
- How is the system physically realized?



Vision: A Computational Investigation into the Human Representation and Processing of Visual Information
D Marr (1982)

The structure of images is complex

invariances
scale
translation
cropping
dilation
homogeneity

perceptual sensitivity
color
edges
orientations

Natural image statistics and neural representation
E. Simoncelli, B. Olshausen (2001)

Extracting semantics is challenging.

challenges

occlusion

deformation

illumination

viewpoint

object pose

scale

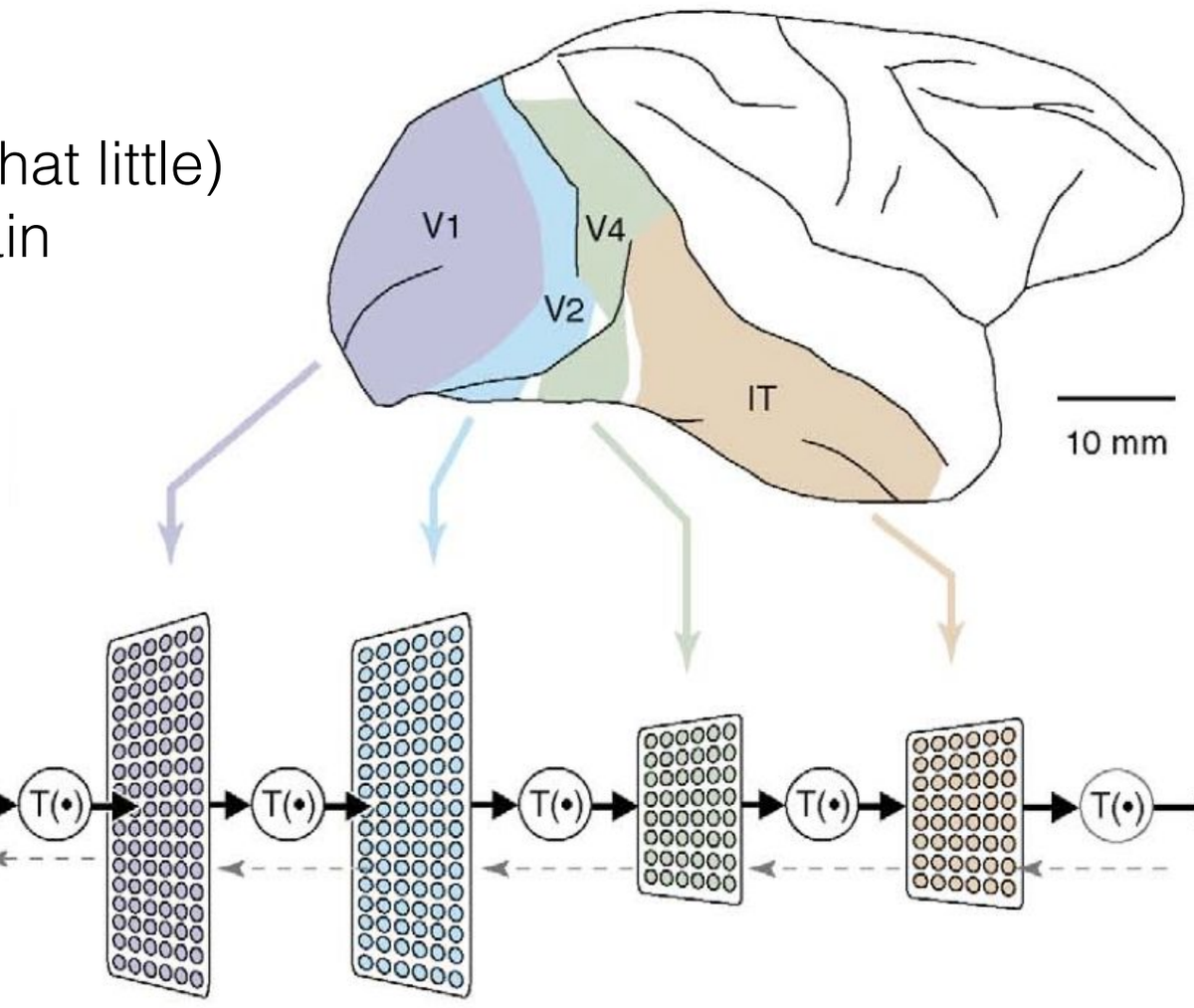
Webster's (3rd ed.) ~ 54,000 words

human (college) ?

Growth of a Functionally Important Lexicon
E Zechmeister, A Chronis, et al (1995)
Visual Object Recognition
K Grauman, B Leibe (2011)

Inspired by the human visual pathway

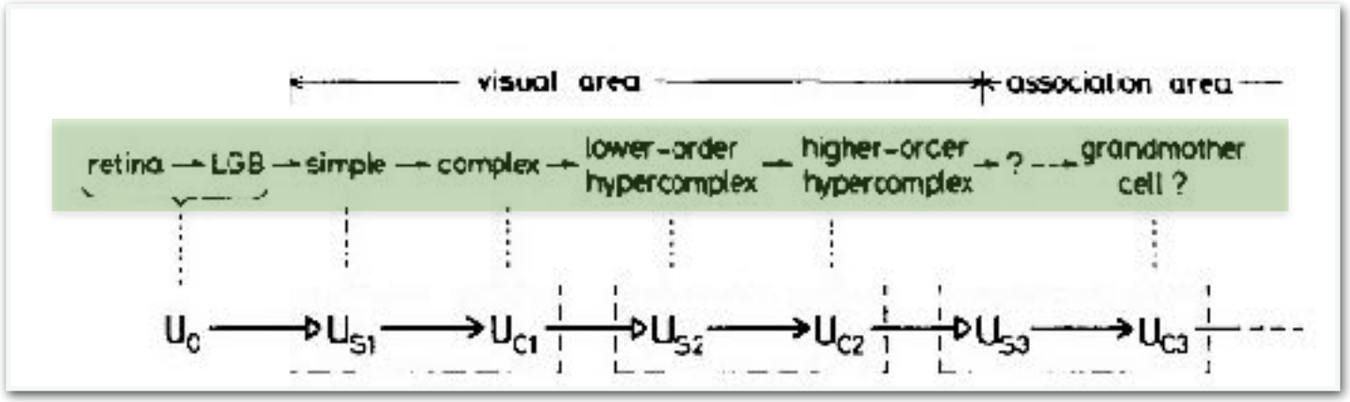
Loosely inspired by (what little) we know about the brain



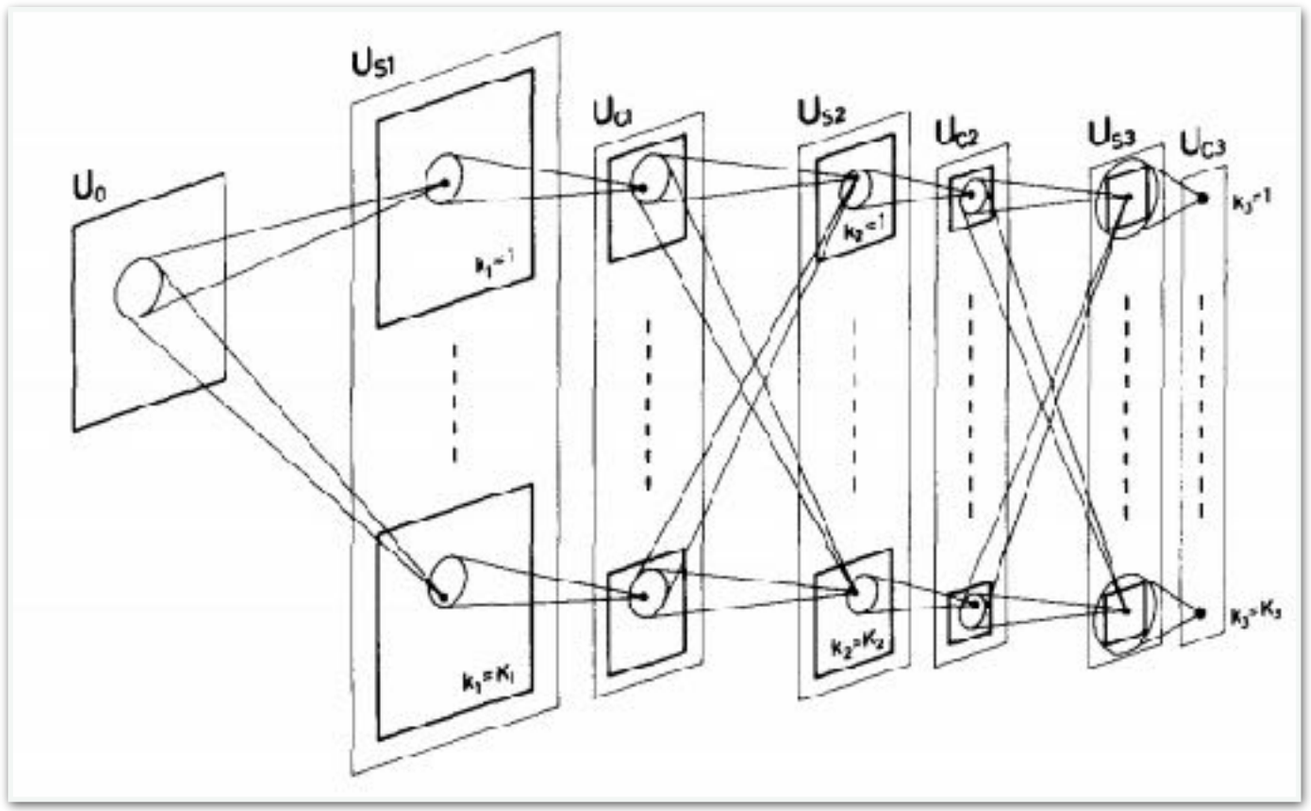
Receptive fields, binocular interaction and functional architecture in the cat's visual cortex
DH Hubel, TN Wiesel (1962)

Untangling invariant object recognition
J DiCarlo and D Cox (2007)

Hierarchical and localized neural networks

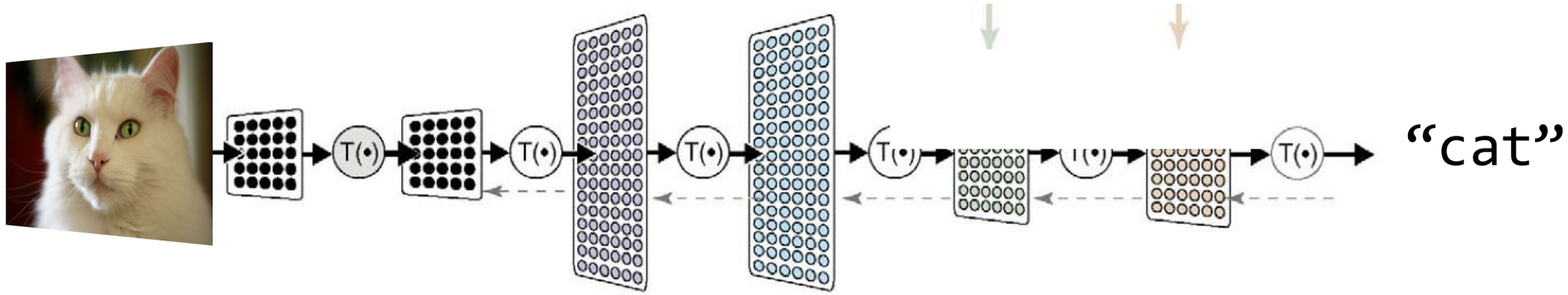


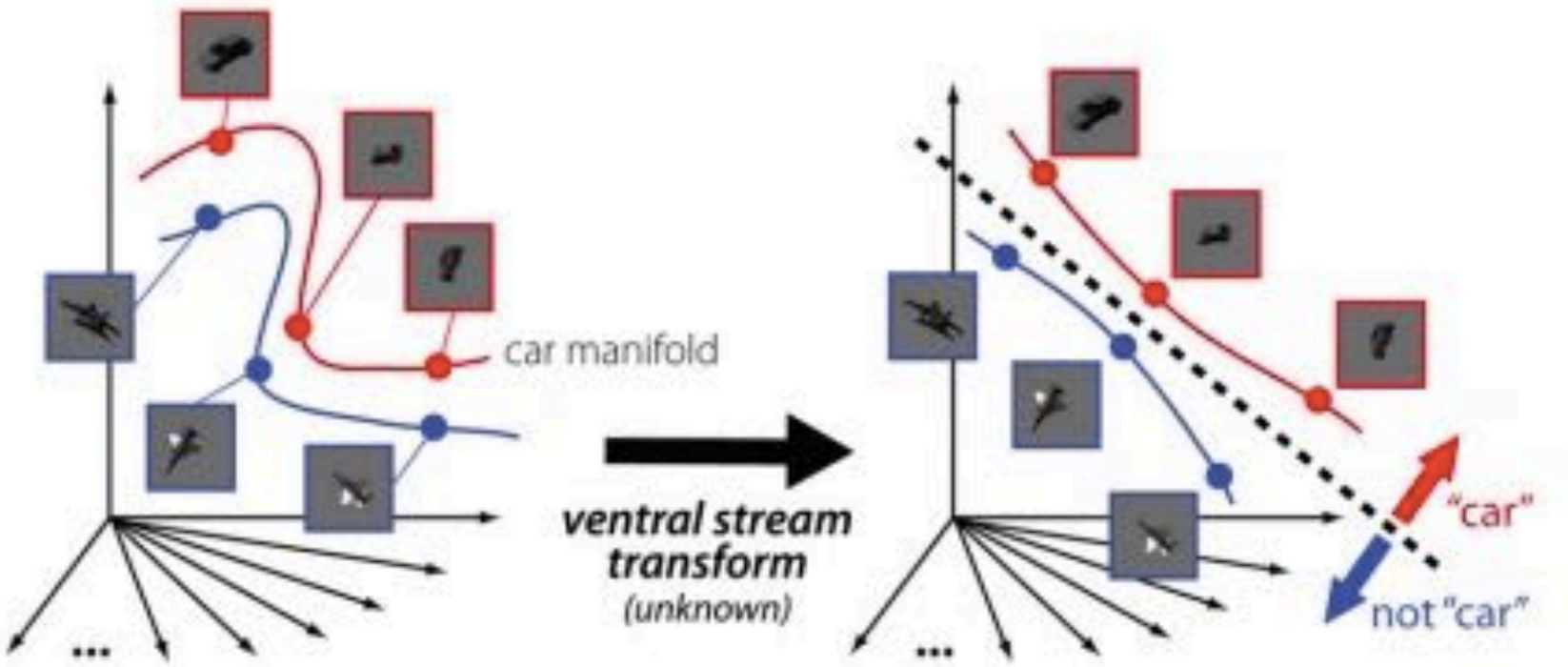
← *ventral pathway*



Neocognitron: A Self-organizing Neural Network Model for a Mechanism of Pattern Recognition
K Fukushima (1980)

Theories of biological image recognition





How Does the Brain Solve Visual Object Recognition?
James J. DiCarlo, Davide Zoccolan, Nicole C. Rust (2012)
Untangling Invariant Object Recognition
J DiCarlo and D Cox (2007)
Performance optimized hierarchical models predict neural responses in higher visual cortex
D Yamins, H Hong, C Cadieu, E Solomon, D Seibert, and J. DiCarlo (2014)

Agenda

1. Challenges and inspiration from vision

2. **Convolutional neural networks**

3. Modern developments

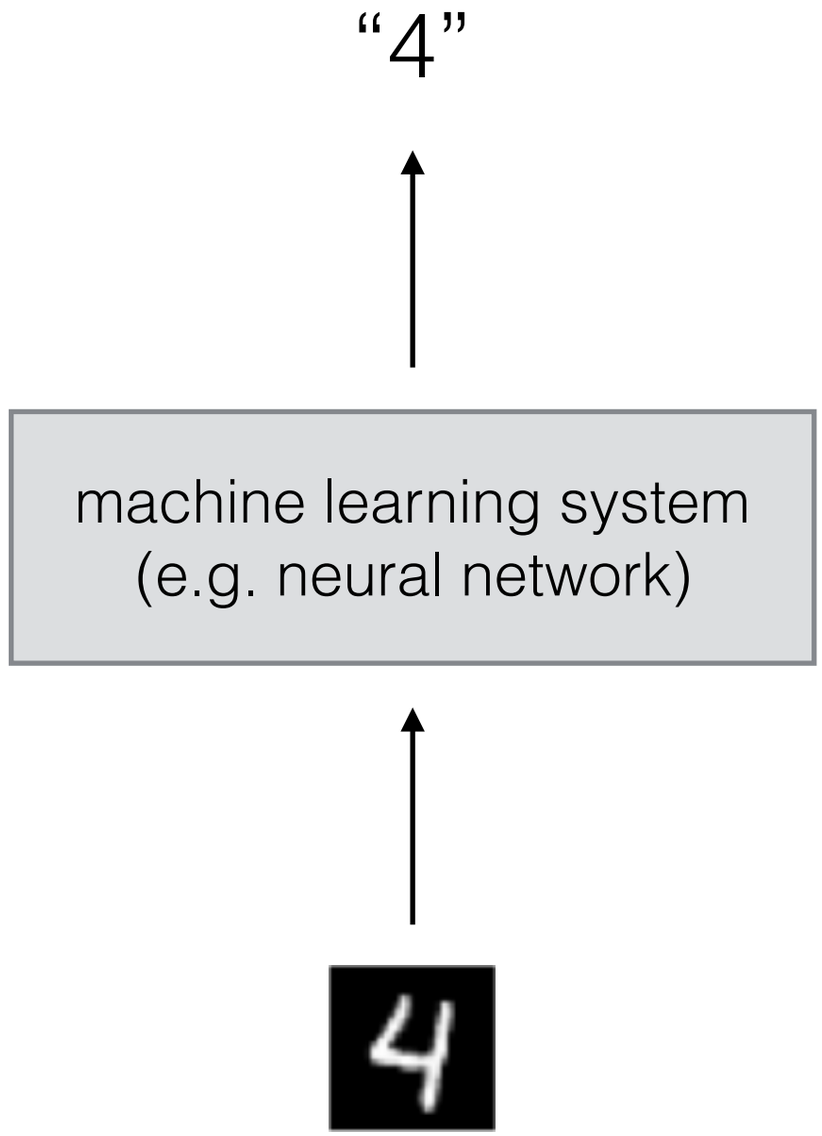
    - architectures, meta-learning, normalization, transfer learning

4. Towards understanding higher-level visual features

5. Opportunities and conclusions

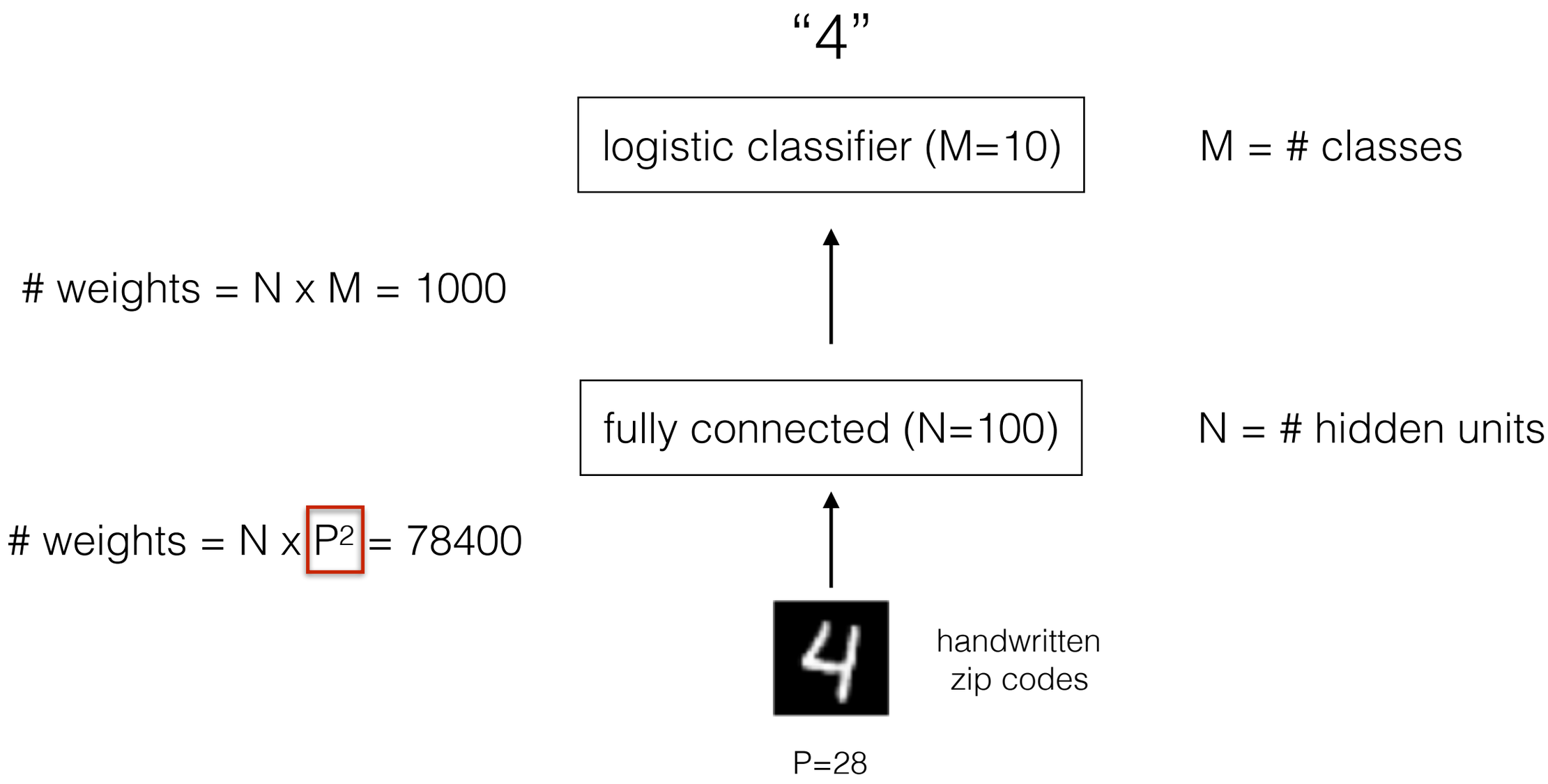
E. Coli of image recognition

http://yann.lecun.com/exdb/mnist/



The MNIST database of handwritten digits
Y LeCun, C Cortes, C Burges (1998)

Multi-layer perceptron on MNIST.



- Note that weights grow as the square of the number of pixels.
- Consider that the camera uses P = 2000, then the number of weights would be **4 million.**

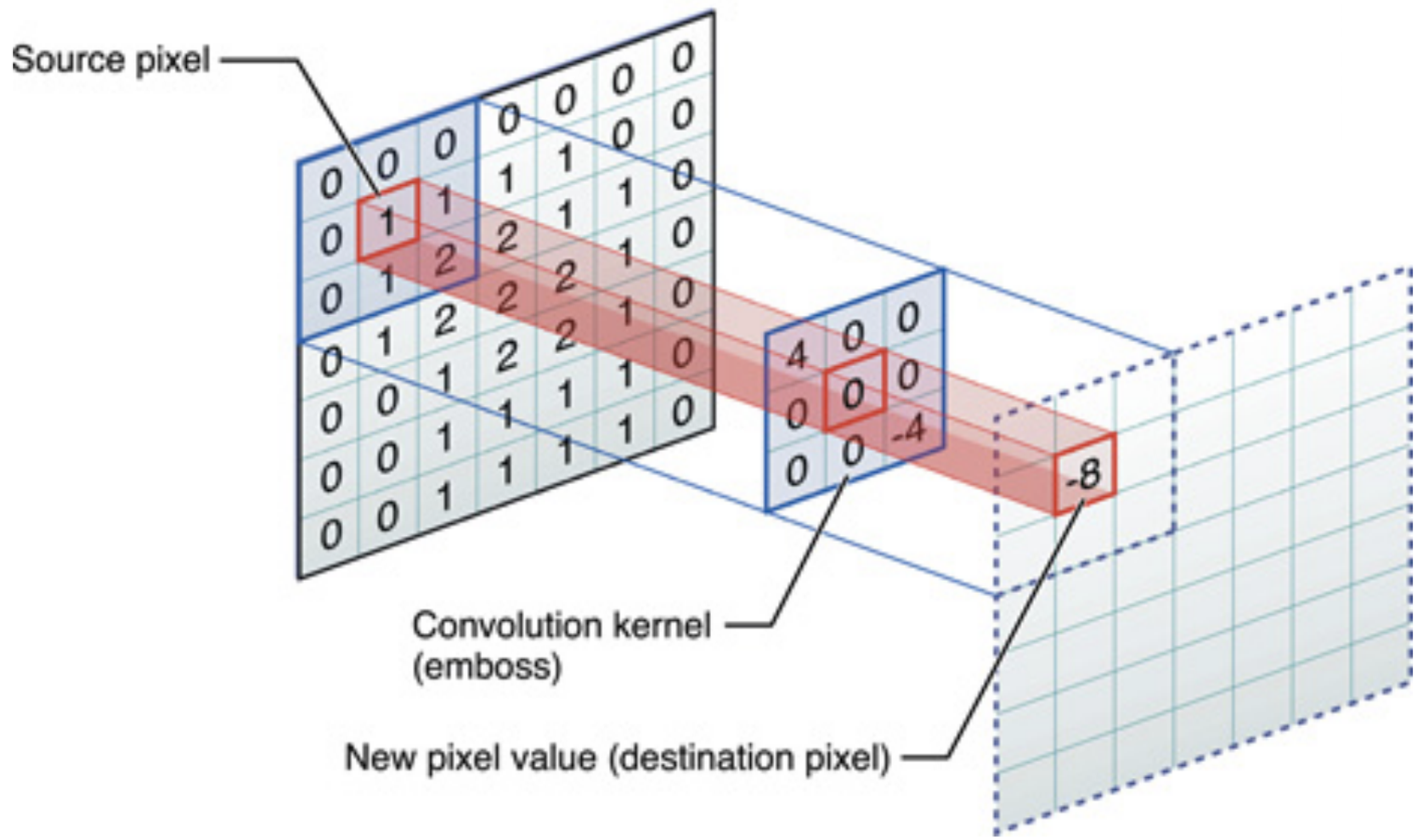
Natural image statistics obey invariances.

…
translation
cropping
dilation
contrast
rotation
scale
brightness
…

Statistics of natural images: Scaling in the woods
D Ruderman and W Bialek (1994)
Natural image statistics and neural representation
E Simoncelli and B Olshausen (2001)

Translation invariance —> convolutions

- Convolutional kernels are a *spatially localized* receptive field whose weights are *shared* across spatial locations.



original

filter (5 x 5)

0	0	0	0	0
0	0	0	0	0
0	0	1	0	0
0	0	0	0	0
0	0	0	0	0

identity

https://docs.gimp.org/en/plug-in-convmatrix.html

original

filter (5 x 5)

0	0	0	0	0
0	1	1	1	0
0	1	1	1	0
0	1	1	1	0
0	0	0	0	0

blur

https://docs.gimp.org/en/plug-in-convmatrix.html

original

filter (5 x 5)

0	0	0	0	0
0	0	-1	0	0
0	-1	4	-1	0
0	0	-1	0	0
0	0	0	0	0

sharpen

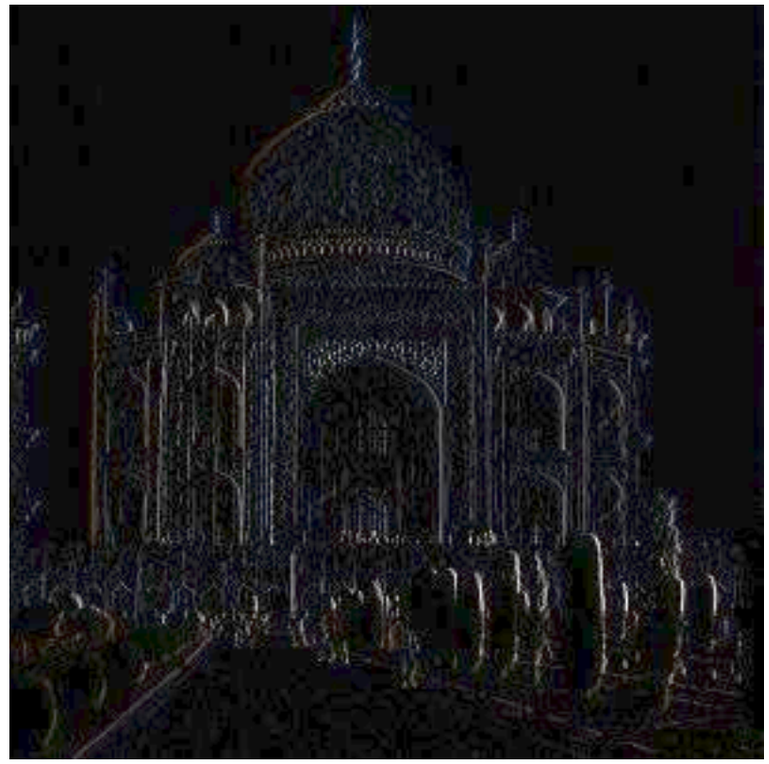
https://docs.gimp.org/en/plug-in-convmatrix.html

original

filter (5 x 5)

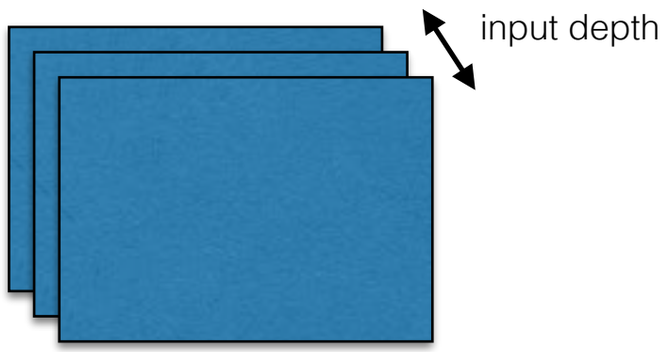
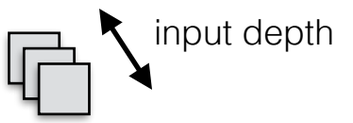
0	0	0	0	0
0	0	0	0	0
0	-1	1	0	0
0	0	0	0	0
0	0	0	0	0

vertical edge detector



https://docs.gimp.org/en/plug-in-convmatrix.html

Generalizing convolutions in depth.

example	input activations	filter bank	output activations
grayscale image			
RGB image			

Generalizing convolutions in depth.

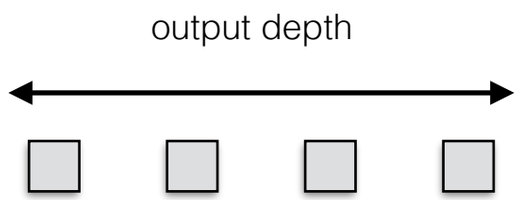
example

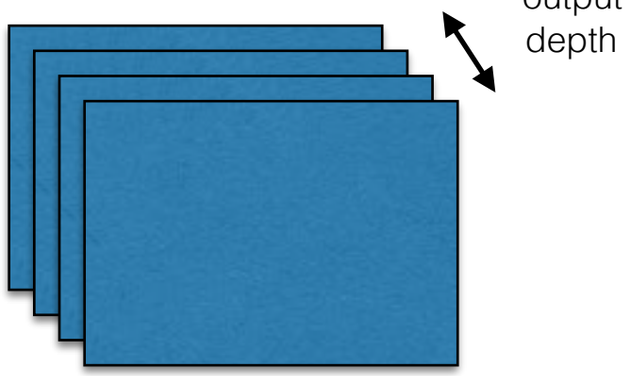
input activations

filter bank

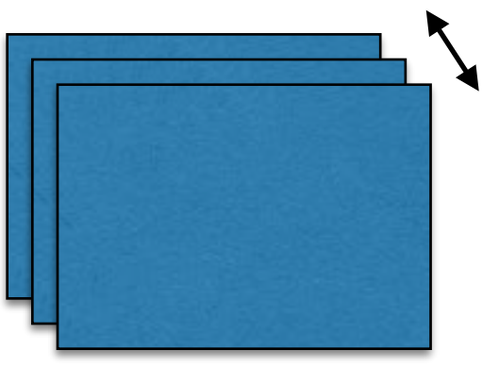
output activations

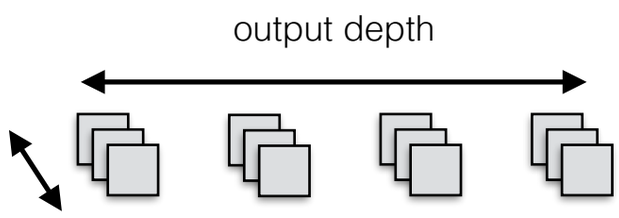
edge detector filter bank

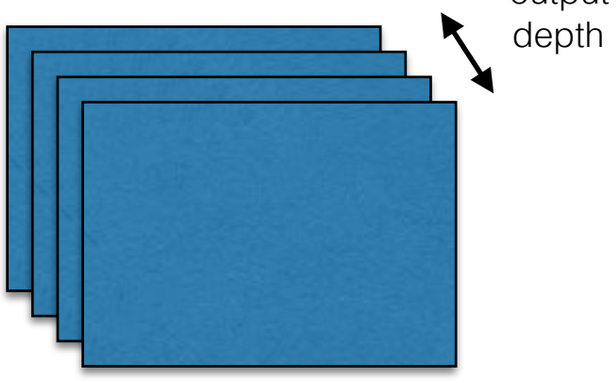




convolutional network

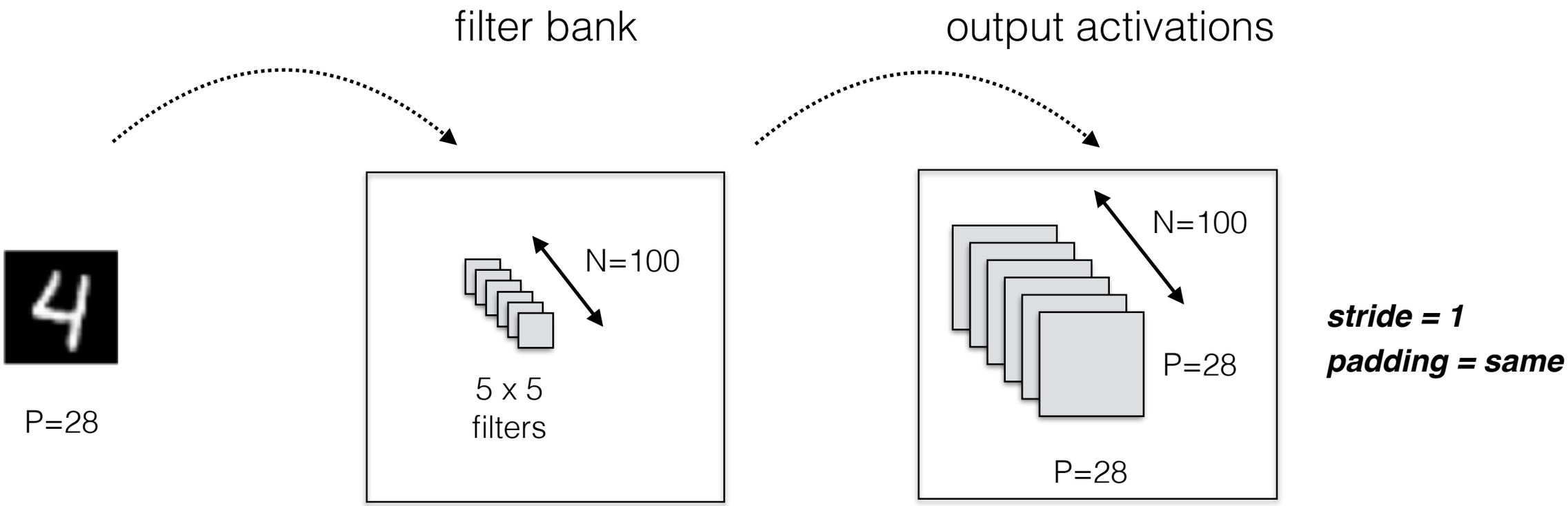






- **input** and **output depth** are arbitrary parameters and not equal.
- Convolutional neural networks operate with depths up to 1024.

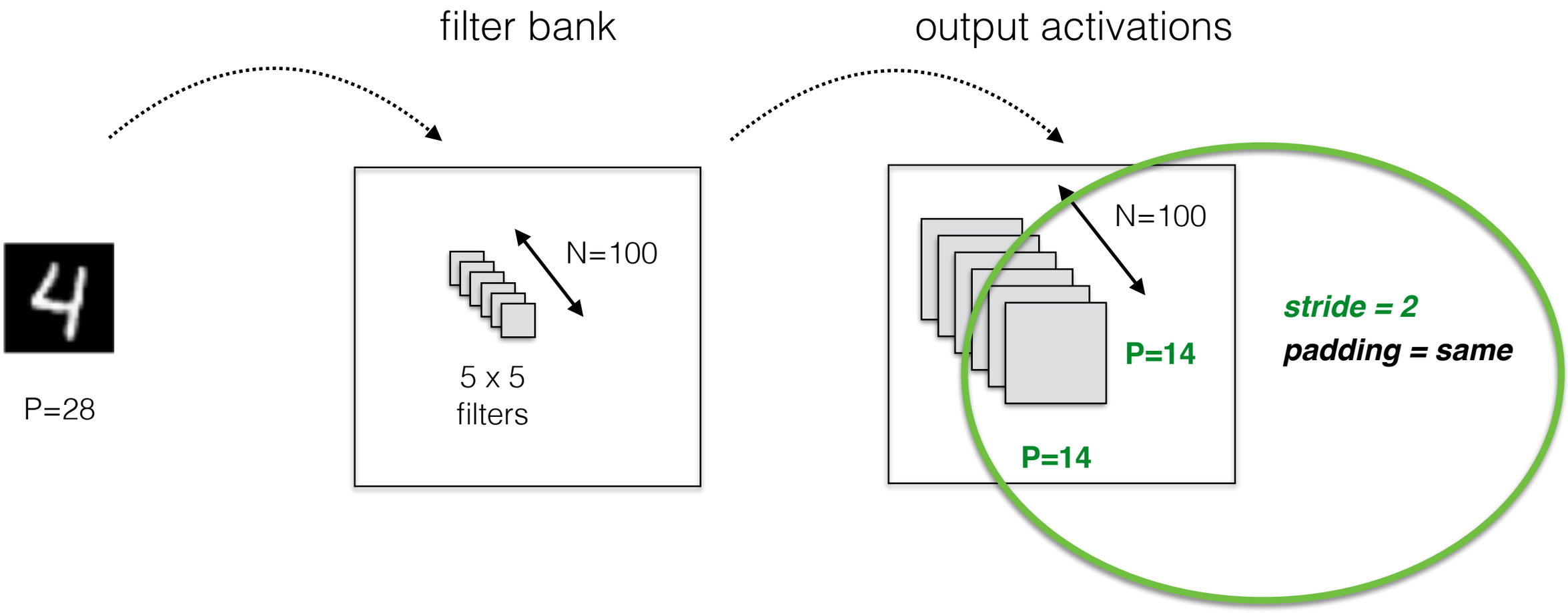
Thinking about convolutional parameters



padding: what to do at the edges: valid or same

stride: # pixels to shift when applying a kernel

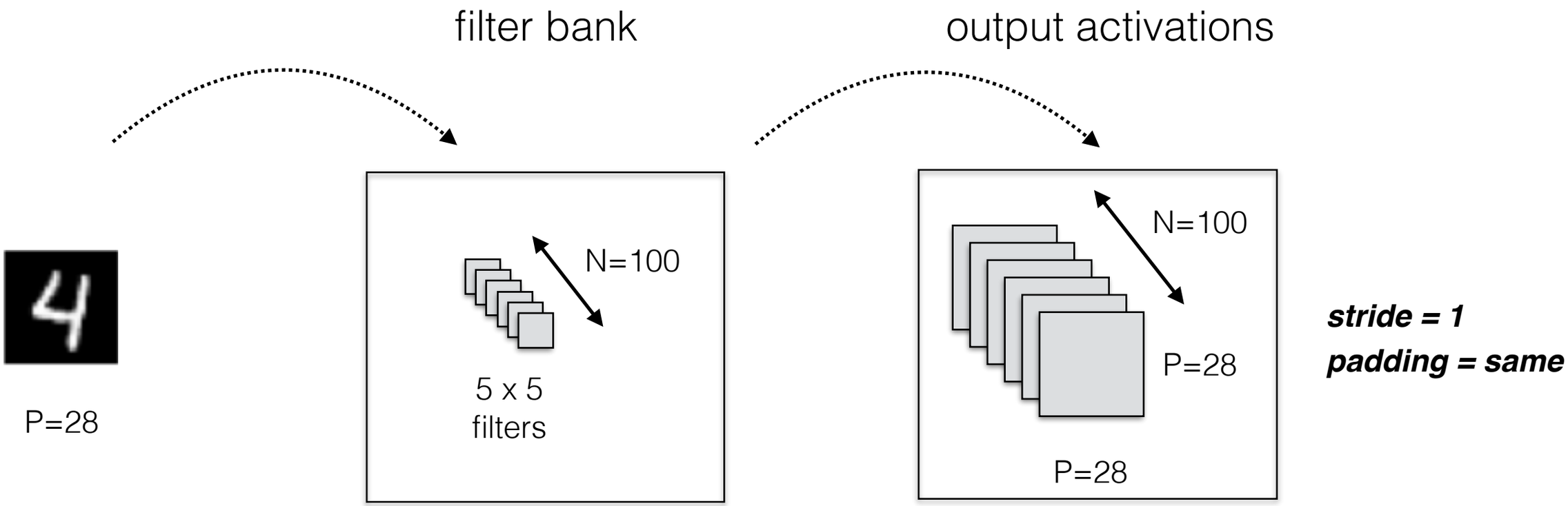
Thinking about convolutional parameters



padding: what to do at the edges: valid or same

stride: # pixels to shift when applying a kernel

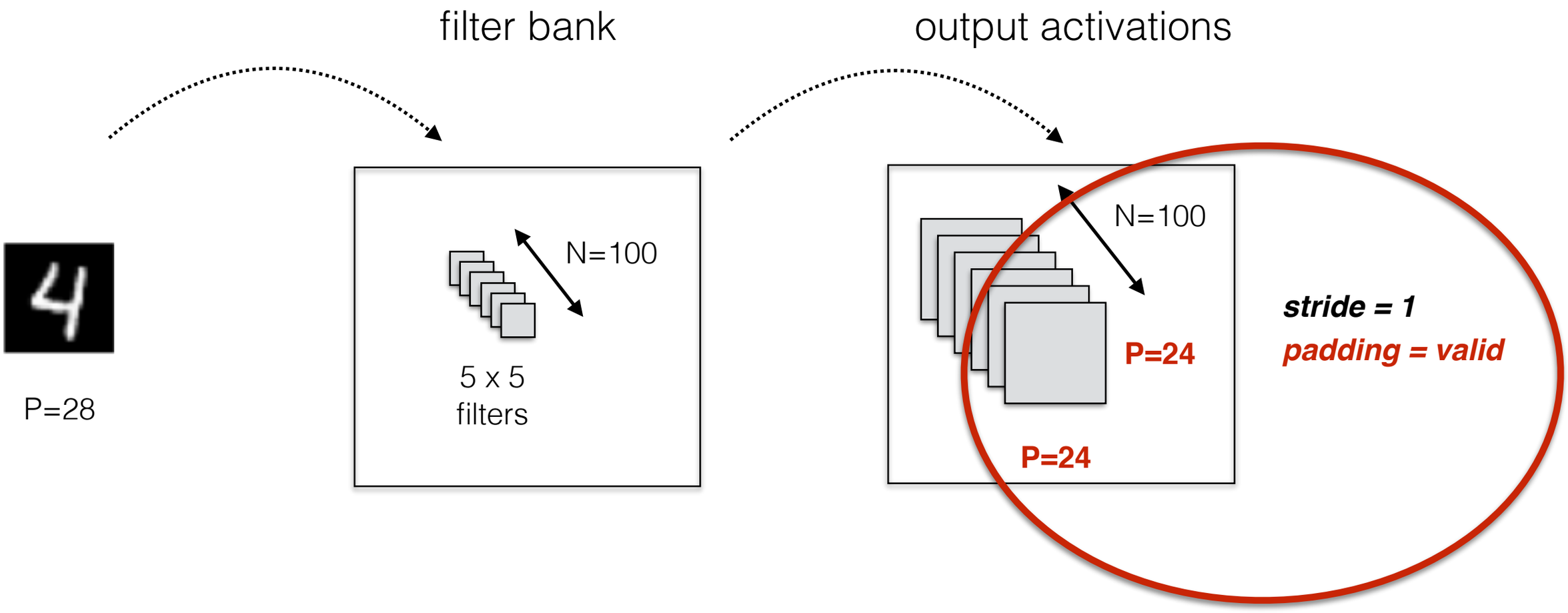
Thinking about convolutional parameters



padding: what to do at the edges: valid or same

stride: # pixels to shift when applying a kernel

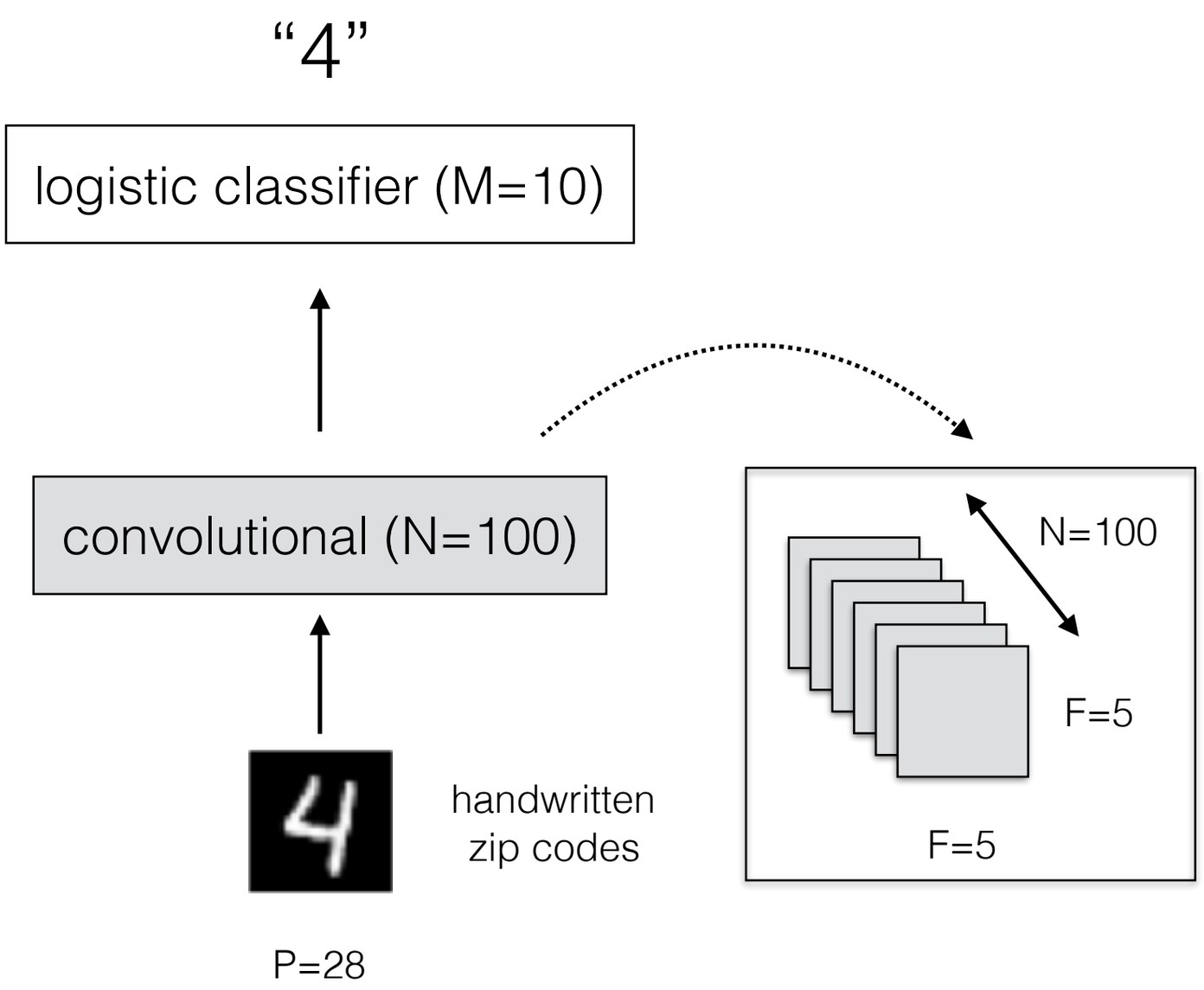
Thinking about convolutional parameters



padding: what to do at the edges: valid or same

stride: # pixels to shift when applying a kernel

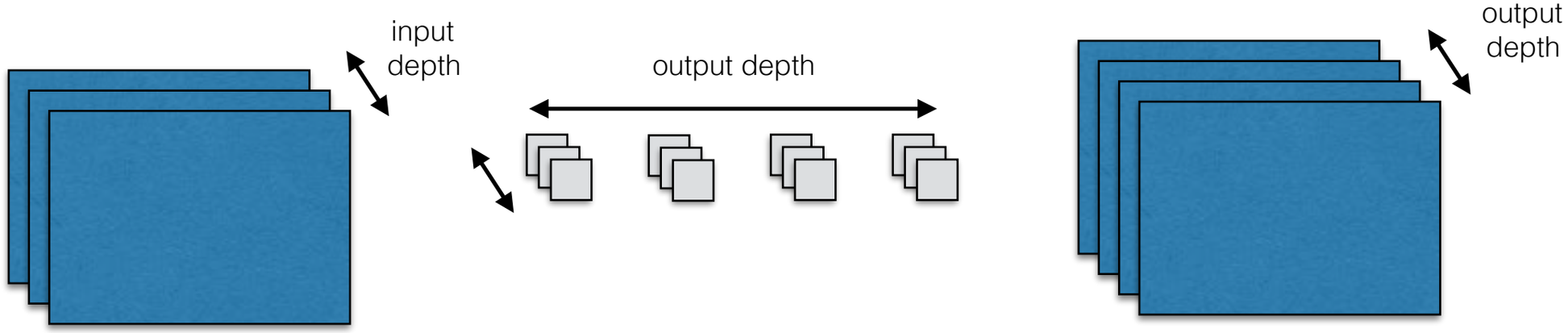
Convolutional neural network on MNIST.



- The number of model parameters is independent of image size.

Thinking about computational cost

convolutional network



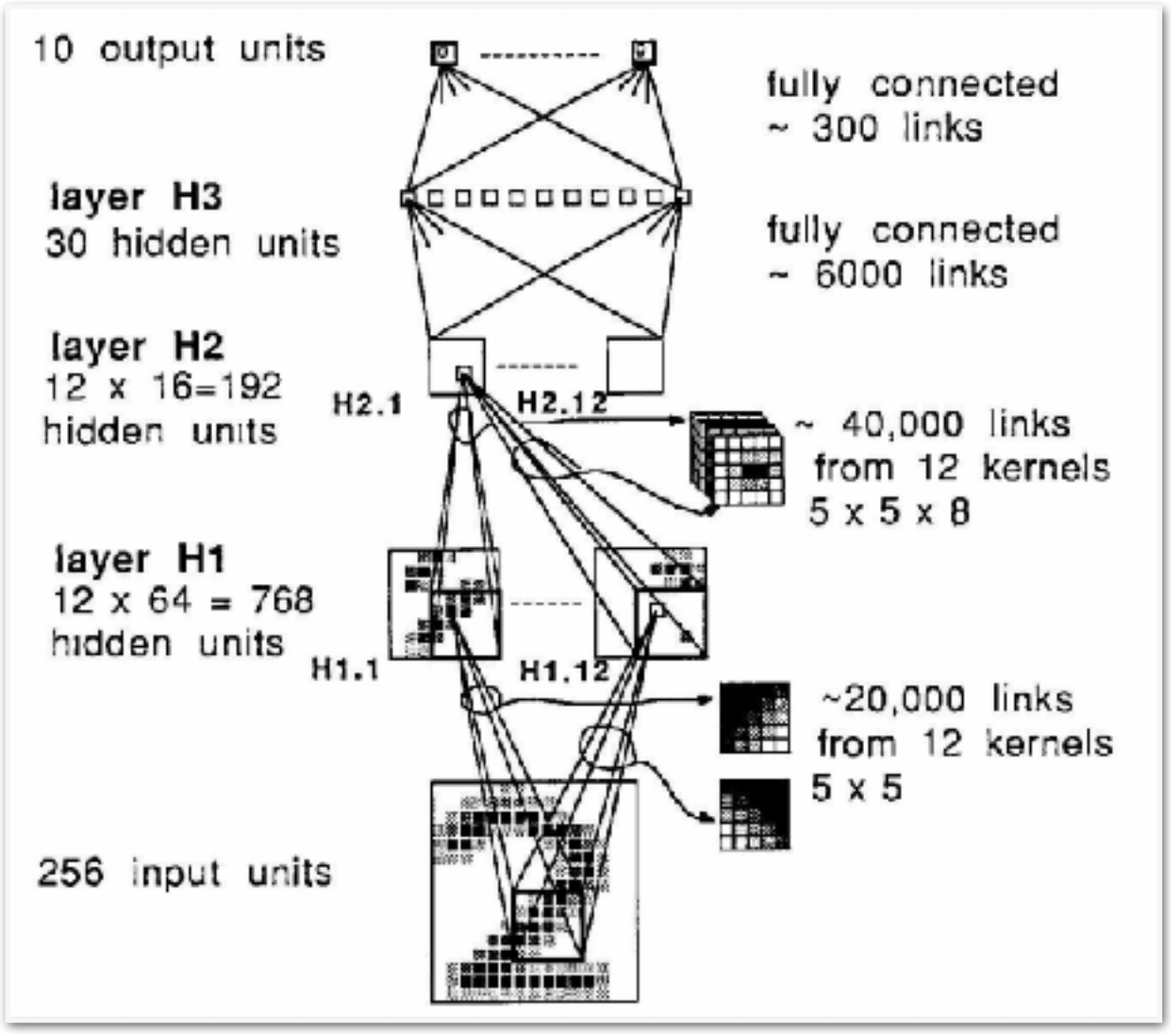
- How many parameters are in a single layer?

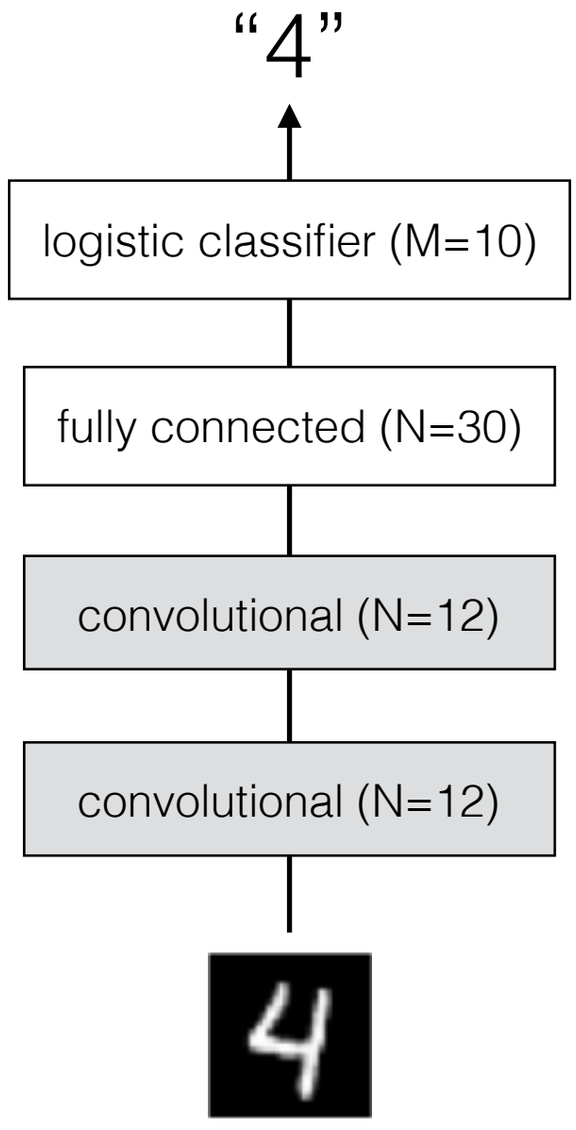
```
(filter width x filter height) x (input depth) x (output depth)
```

- How much computational cost in a single layer?

```
(filter width x filter height) x (input depth) x (output depth)
      x (input width / stride) x (input height / stride )
```

A first, modern convolutional neural network





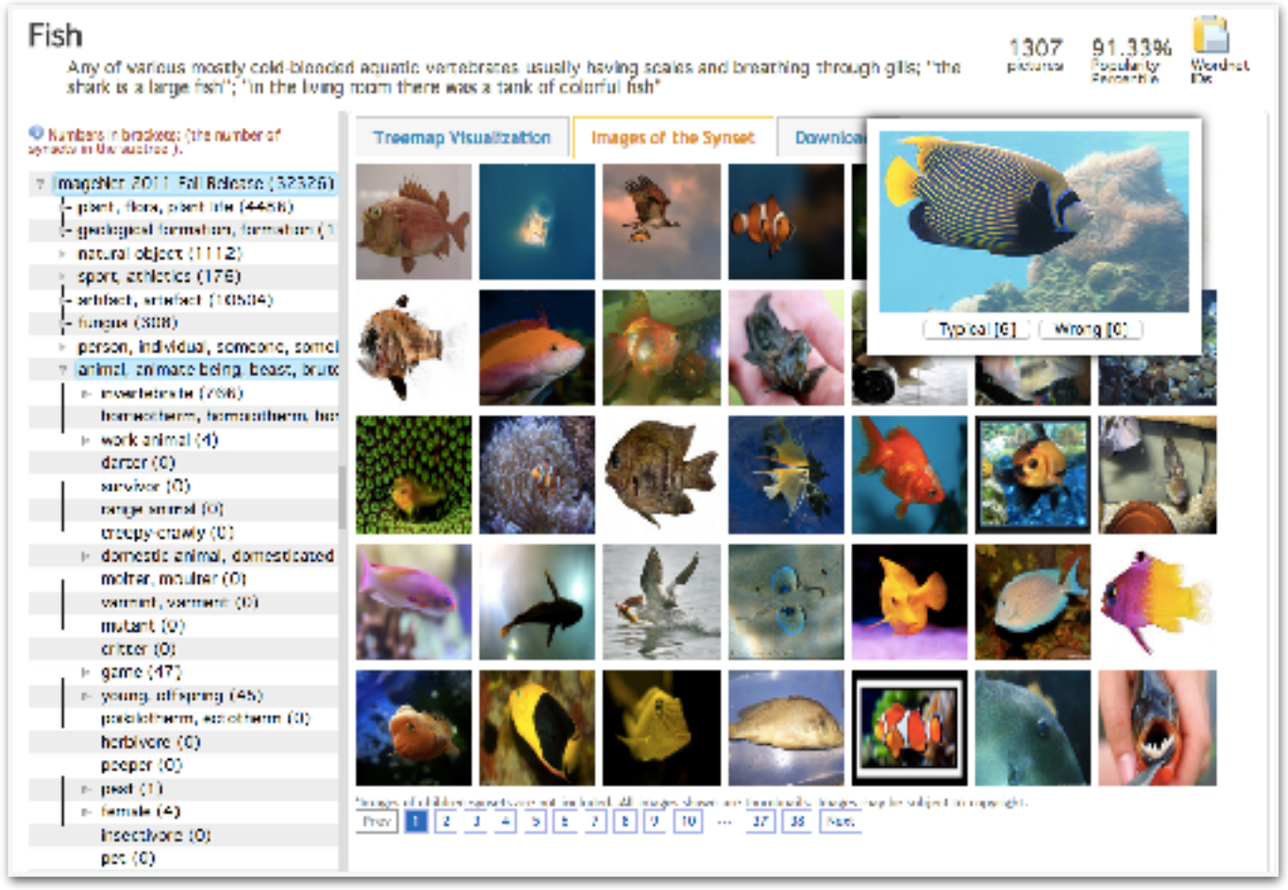
Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition
K Fukushima, and M Sei (1982)
Backpropagation applied to handwritten zip code recognition
Y LeCun et al (1989)

The computer vision competition: IMAGENET

Large scale academic competition focused on predicting 1000 object classes (~1.2M images).

classes

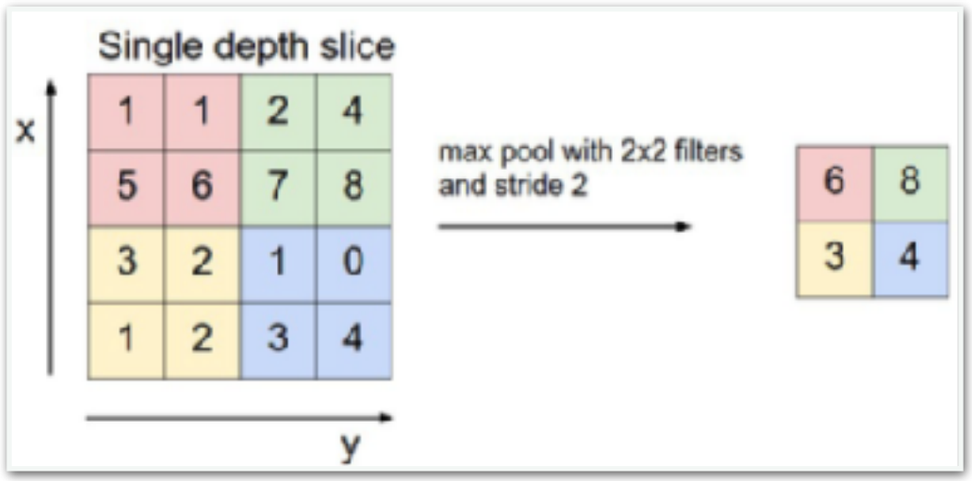
- electric ray
- barracuda
- coho salmon
- tench
- goldfish
- sawfish
- smalltooth sawfish
- guitarfish
- stingray
- roughtail stingray
- ...



Imagenet: A large-scale hierarchical image database
J Deng et al (2009)

Scaling to higher resolution images

- Computational demand grows as **quadratically** as the image size.
  - **Spatial pooling** builds invariance across spatial dimensions (and save compute!).

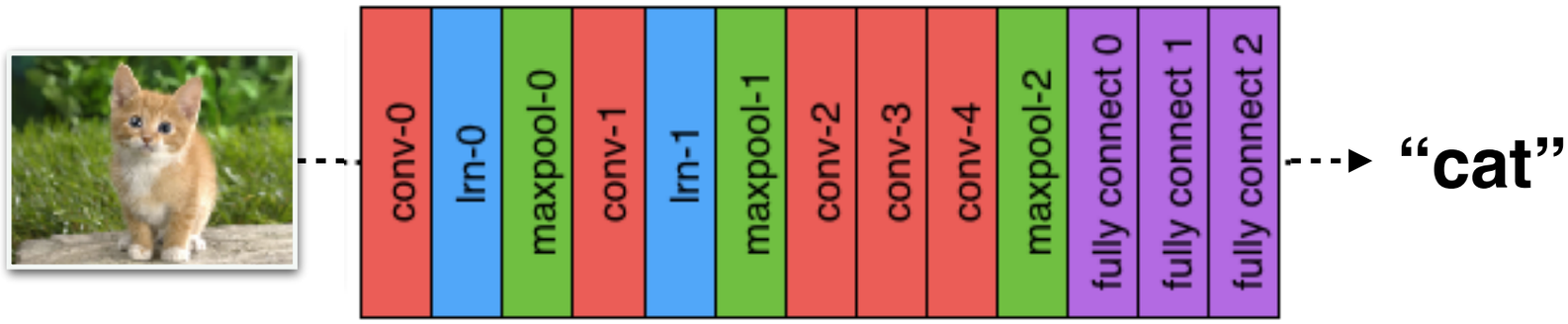


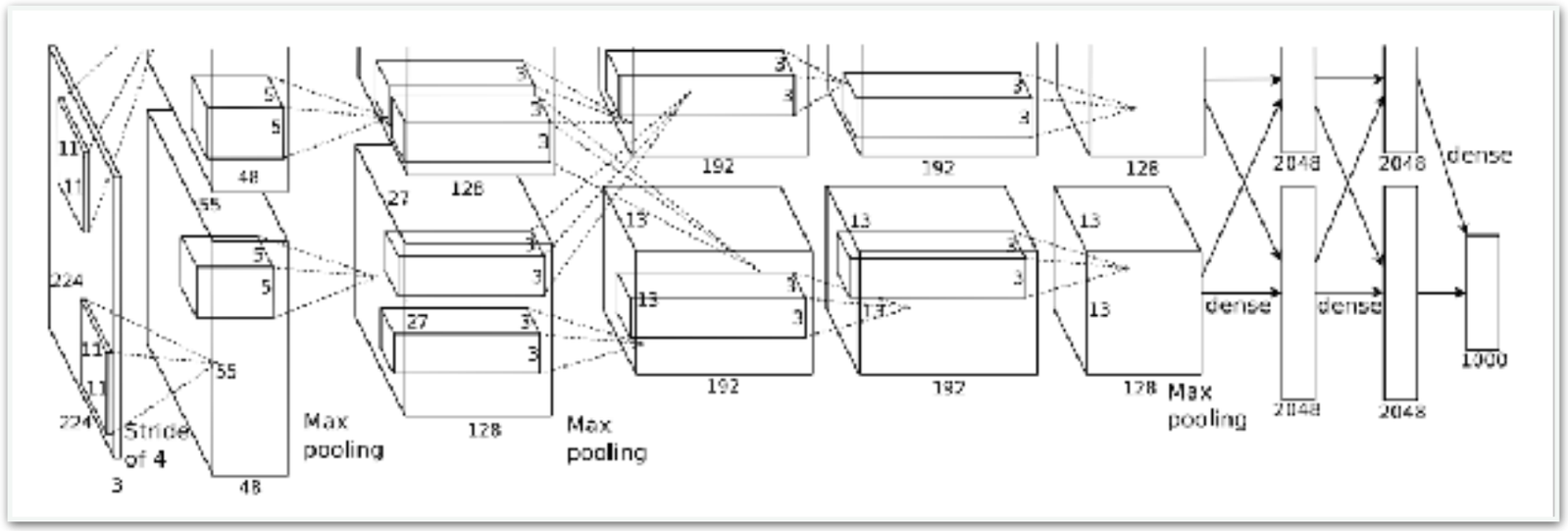
Courtesy of R Zemel

  - **Regularization** mitigates overfitting (e.g. weight decay, dropout).
  - **Normalization** empirically accelerates training and makes better model images.

What is the Best Multi-Stage Architecture for Object Recognition?
K Jarrett, K Kavukcuoglu, MA Ranzato and Y LeCun (2009)

Convolutional neural networks, enlarged.





- Deeper and larger (70K —> 60M params) version of LeNet.
- Achieving scale in both compute and data is critical.

ImageNet Classification with Deep Convolutional Neural Networks
A Krizhevsky I Sutskever, G Hinton (2012)

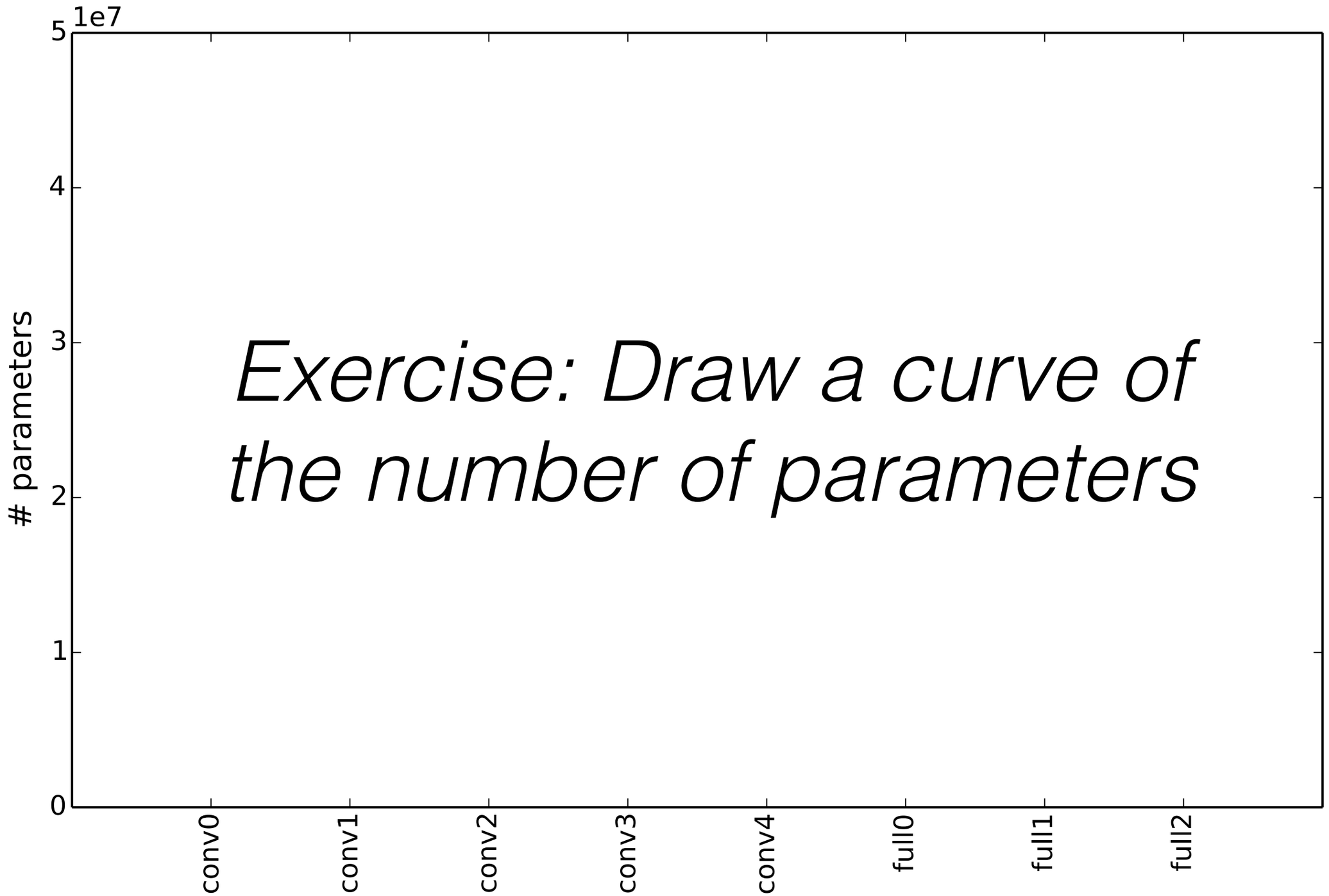
	kernel	*stride*	*input*	*output*
conv-0	11 x 11	4	**224x224x3**	56x56x96
lrn-0			56x56x96	56x56x96
maxpool-0	3 x 3	2	56x56x96	28x28x96
conv-1	5 x 5	1	28x28x96	28x28x256
lrn-1			28x28x256	28x28x256
maxpool-1	3 x 3	2	28x28x256	14x14x256
conv-2	3 x 3	1	14x14x256	14x14x384
conv-3	3 x 3	1	14x14x384	14x14x384
conv-4	3 x 3	1	14x14x384	14x14x256
maxpool-2	3 x 3	2	14x14x256	7x7x256
fully connect 0			7x7x256	1x1x4096
fully connect 1			1x1x4096	1x1x4096
fully connect 2			1x1x4096	1x1x**1000**

“cat”

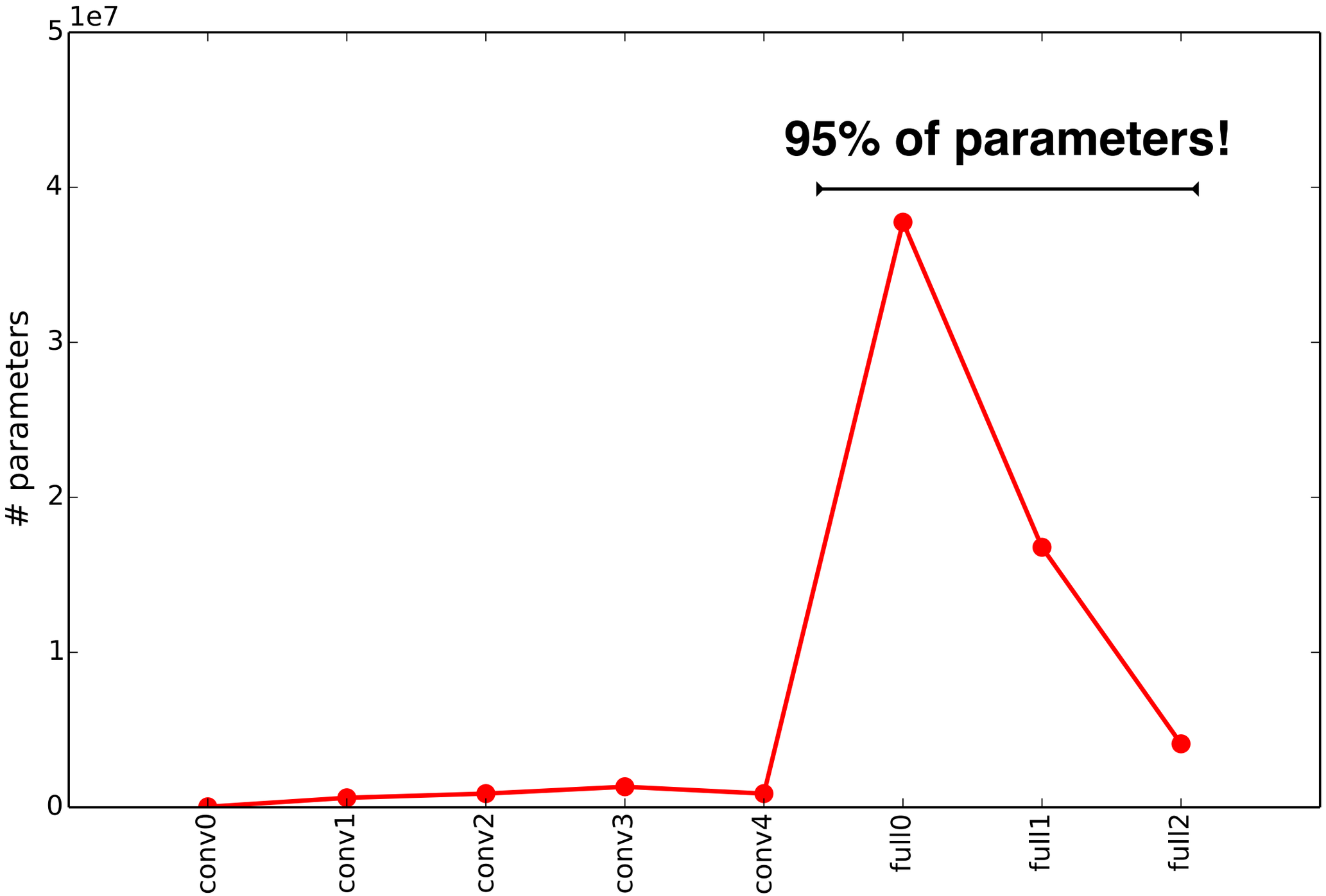
1000 ImageNet classes

https://medium.com/@smallfishbigsea/a-walk-through-of-alexnet-6cbd137a5637

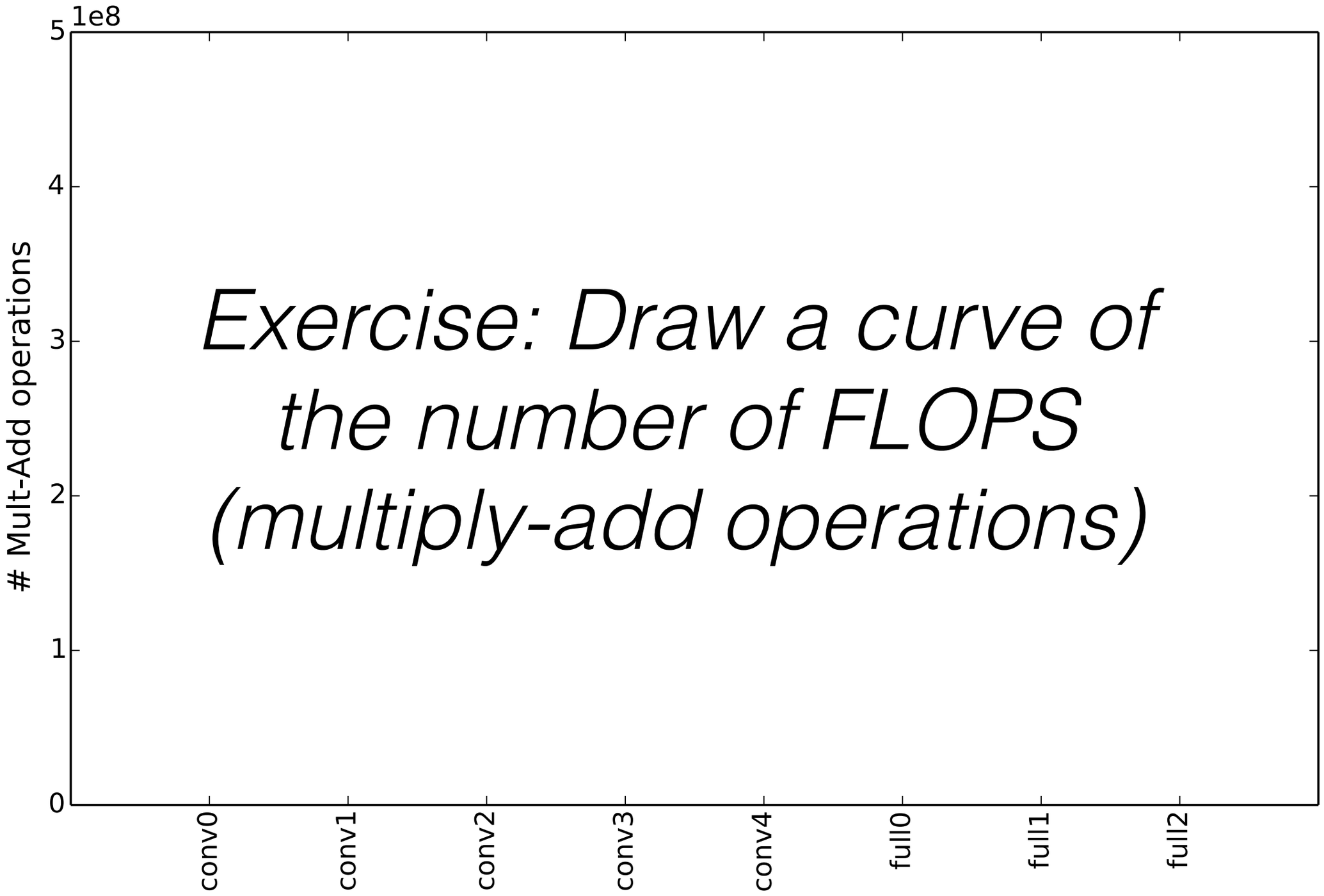
Exercise: parameters



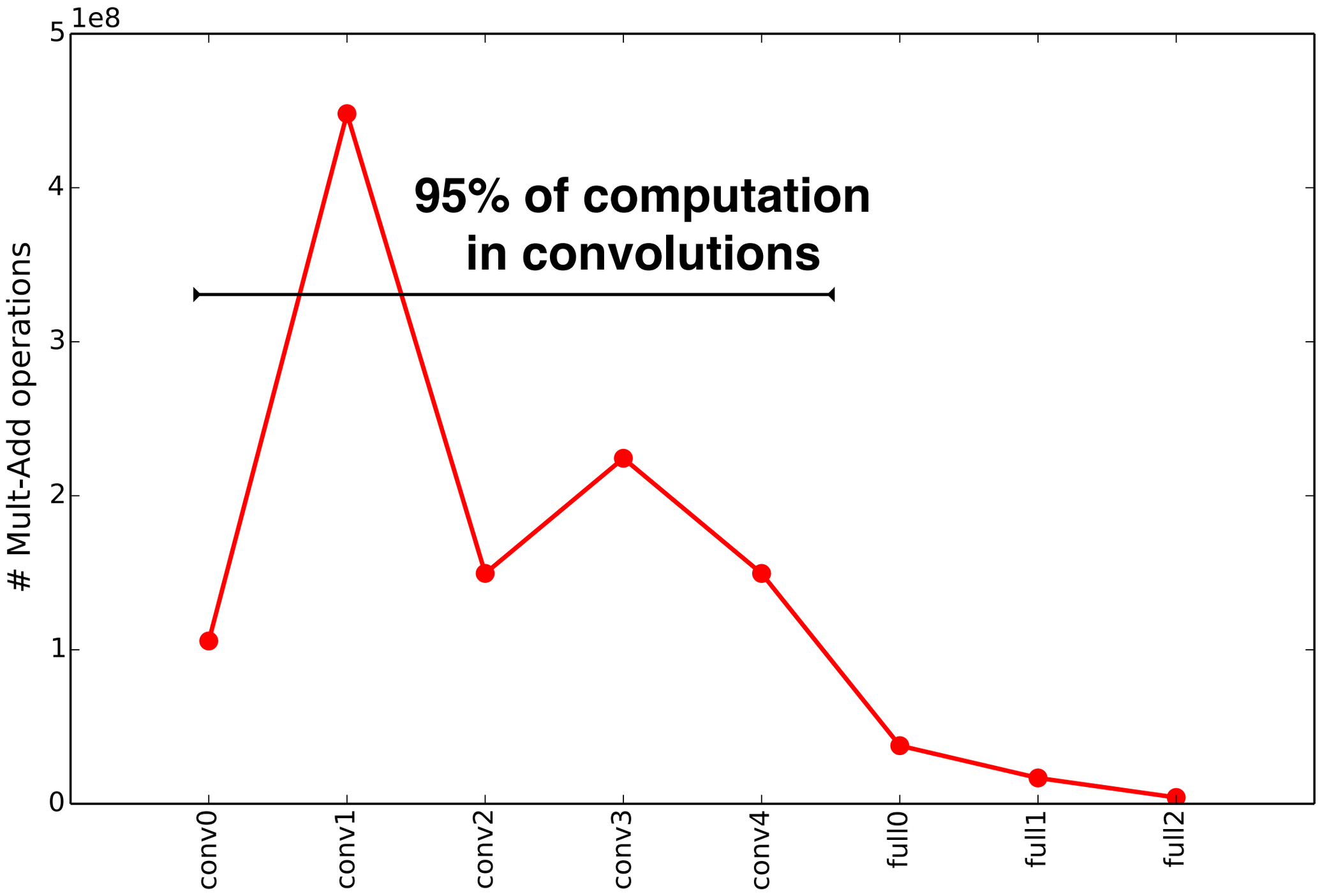
Exercise: parameters



Exercise: computation for a single image



Exercise: computation



Agenda

1. Challenges and inspiration from vision

2. Convolutional neural networks

3. **Modern developments**

   - architectures, meta-learning, normalization, transfer learning

4. Towards understanding higher-level visual features

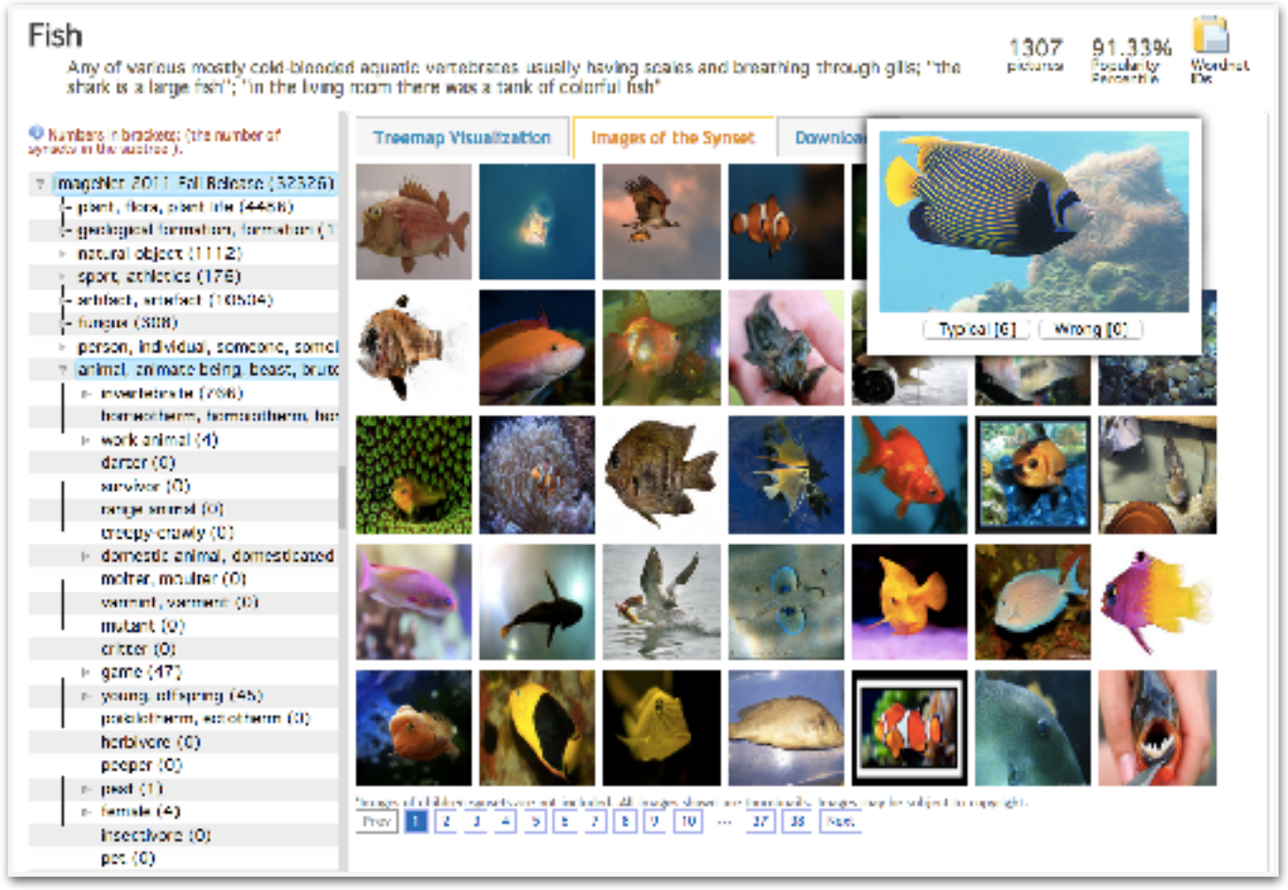
5. Opportunities and conclusions

The computer vision competition: IMAGENET

Large scale academic competition focused on predicting 1000 object classes (~1.2M images).

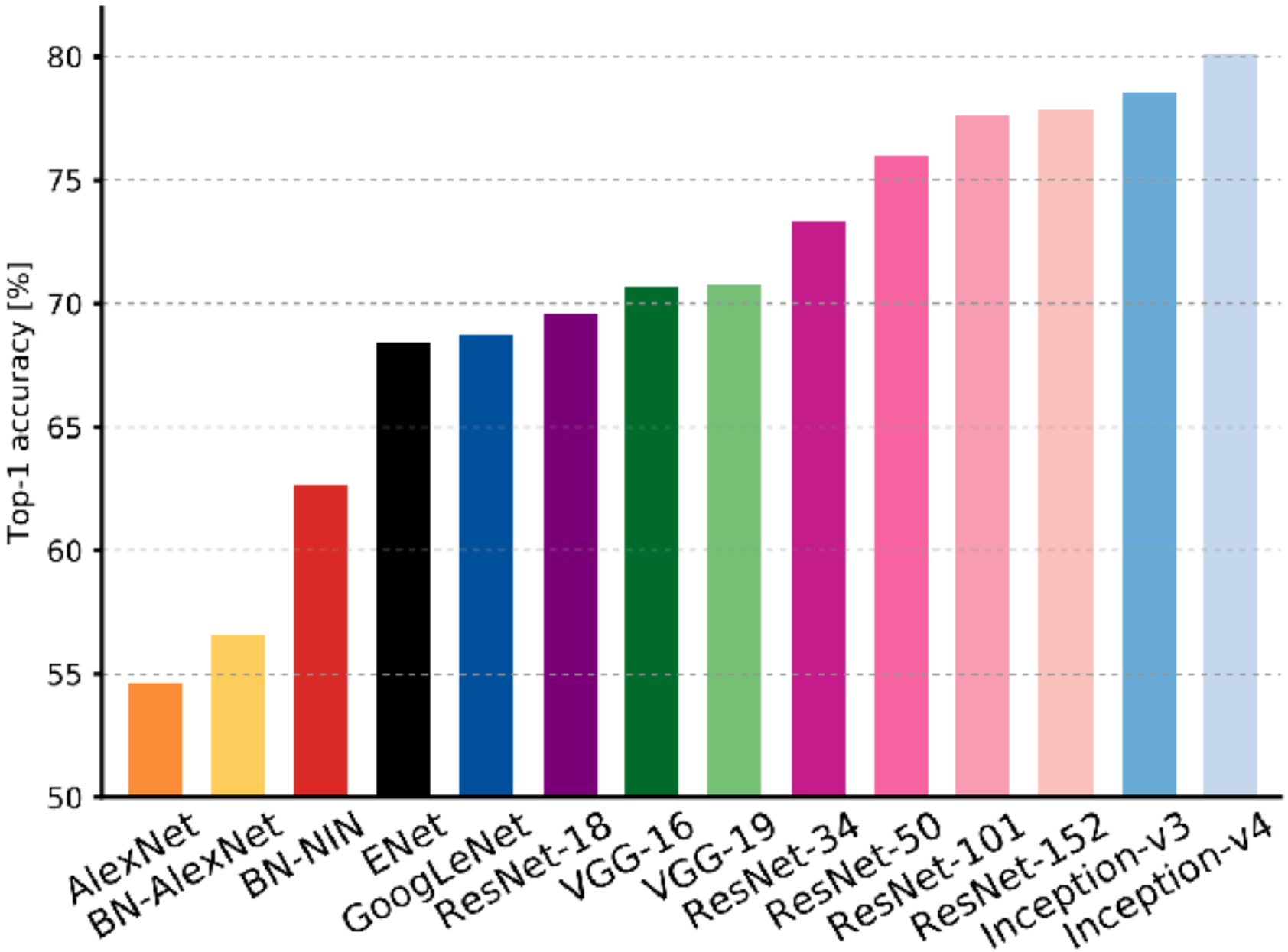
classes

- electric ray
- barracuda
- coho salmon
- tench
- goldfish
- sawfish
- smalltooth sawfish
- guitarfish
- stingray
- roughtail stingray
- ...



Imagenet: A large-scale hierarchical image database
J Deng et al (2009)

Steady progress in image recognition.



An Analysis of Deep Neural Network Models for Practical Applications

A Canziani, A Paszke, E Culurciello (2016)

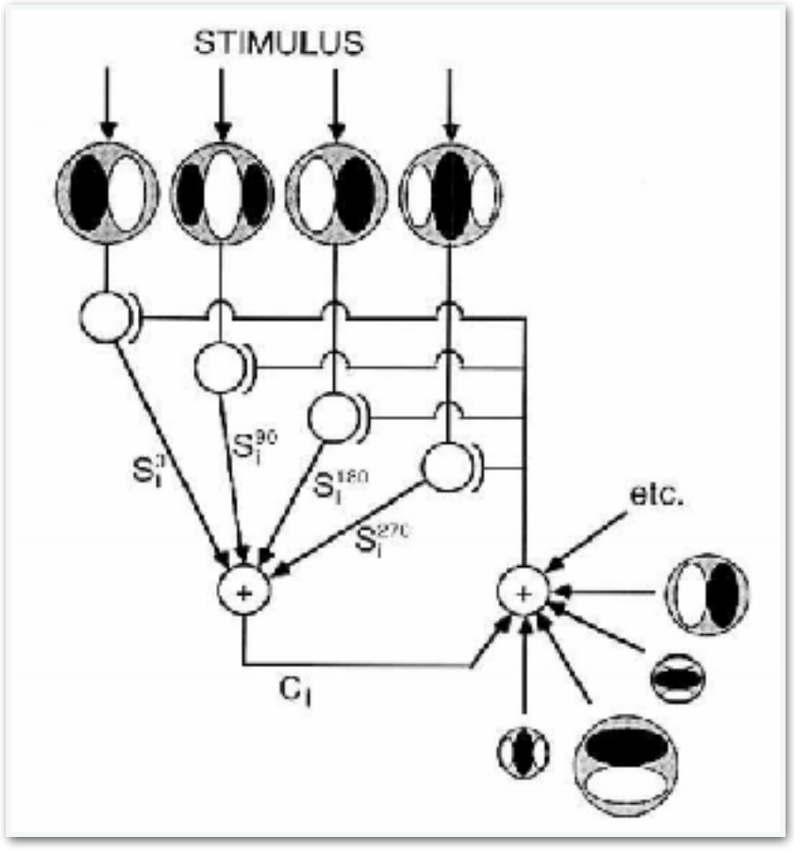
Trends in network architecture.

1. Normalization methods are an important ingredient for achieving state-of-the-art performance

2. Deeper and larger networks lead to better predictive performance.

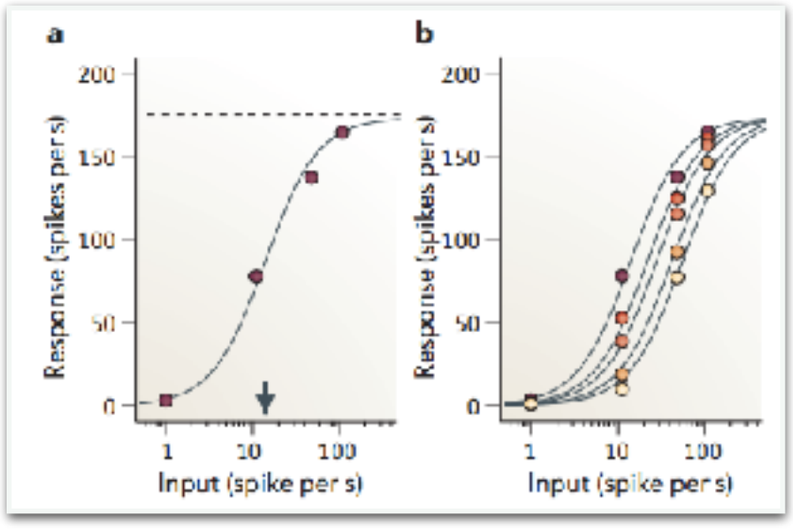
3. Multi-scale architectures provide great predictive performance while minimizing computational demand.

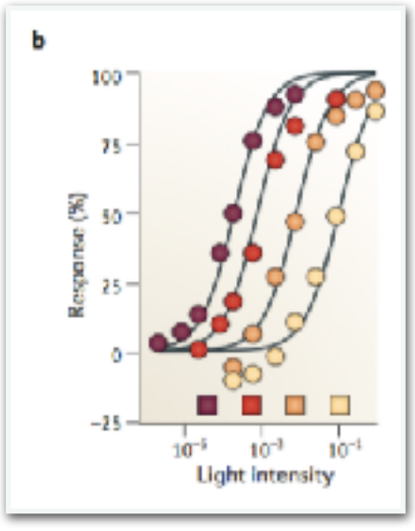
Normalization as a canonical computation.



divisive normalization in cortical neurons
(originally from D Heeger, 1992)

response of olfactory neurons in fruit fly
(originally from Olsen et al, 2010)





response of turtle photoreceptor
(originally from R Normann et al, 1979)

Normalization of cell responses in cat striate cortex
D Heeger (1992)
Normalization as a canonical neuronal computation
M Caradini and D Heeger (2012)

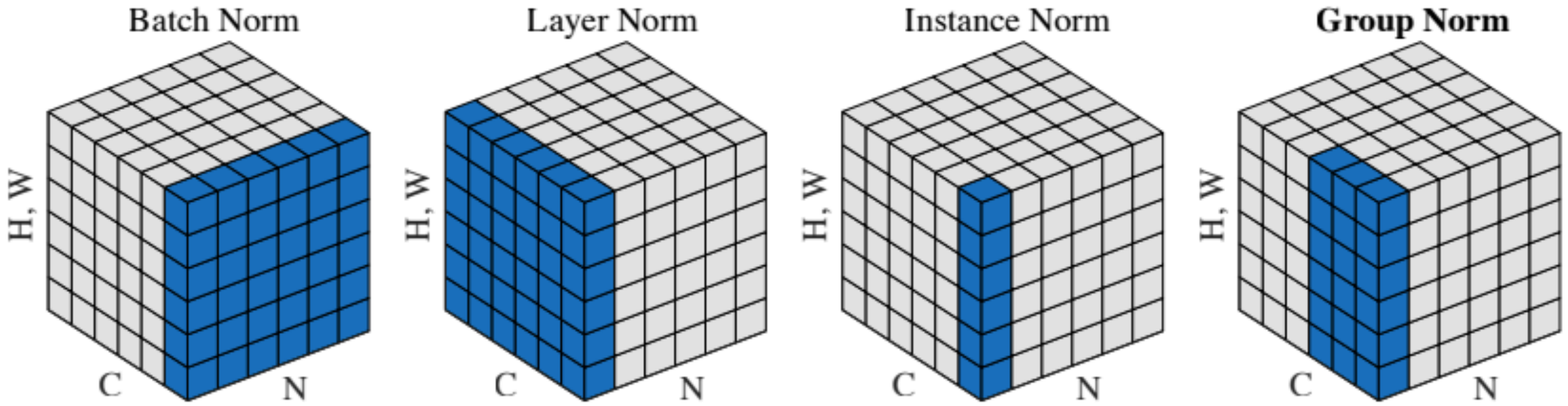
Normalization achieves state-of-the-art performance

- Many variations, none of which is strictly biological.
- Almost all vision models employ some form of normalization throughout a network representation.

- Accelerate training efficiency up to 20-fold.
- Train models previously untrainable
- Boost cross-validated performance

Batch Normalization
S Ioffe and C Szegedy (2015)
Layer Normalization
J Ba, J Kiros, G Hinton (2016)
Instance Normalization
D Ulyanov, A Vedaldi, V Lempitsky (2016)
Normalizing the Normalizers
M Ren, R Liao, R Urtasun, F Sinz, R Zemel (2016)
Group Normalization
Y Wu, K He (2018)

Survey of normalization methods



Consider the activations $\vec{x}$ where each dimension x_c is a channel associated with a group G

$$\vec{x} = \{x_c\}$$

Group Normalization
Y Wu, K He (2018)

Group normalization

1. Calculate the mean μ and variance σ^2 within each group of channels **and normalize.**

$$\mu = \frac{1}{|G|}\sum_{i \in G} x_c$$

$$\sigma^2 = \frac{1}{|G|}\sum_{i \in G}(x_c - \mu)^2$$

$$\longrightarrow \quad \bar{x}_c = \frac{x_c - \mu}{\sqrt{\sigma^2 + \epsilon}}$$

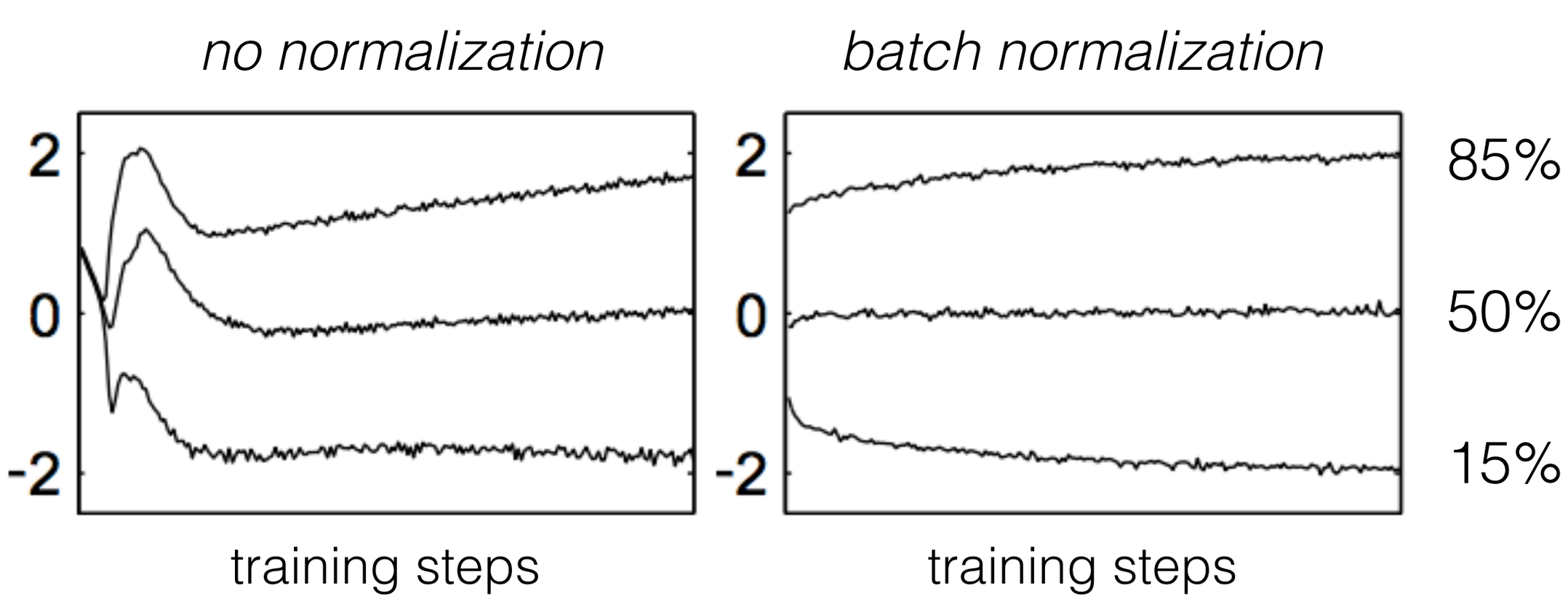
2. Learn the mean and variance (γ, β) of each layer as parameters

$$y_c = \gamma \bar{x}_c + \beta$$

Group Normalization
Y Wu, K He (2018)

Normalization stabilizes activations during training

- Distribution of hidden layer activations during training a convolutional network on MNIST.



Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift
S Ioffe and C Szegedy (2015)
How Does Batch Normalization Help Optimization?
S Santurkar, D Tsipras, A Ilyas, A Madry (2018)

Towards greater network depth and scale.

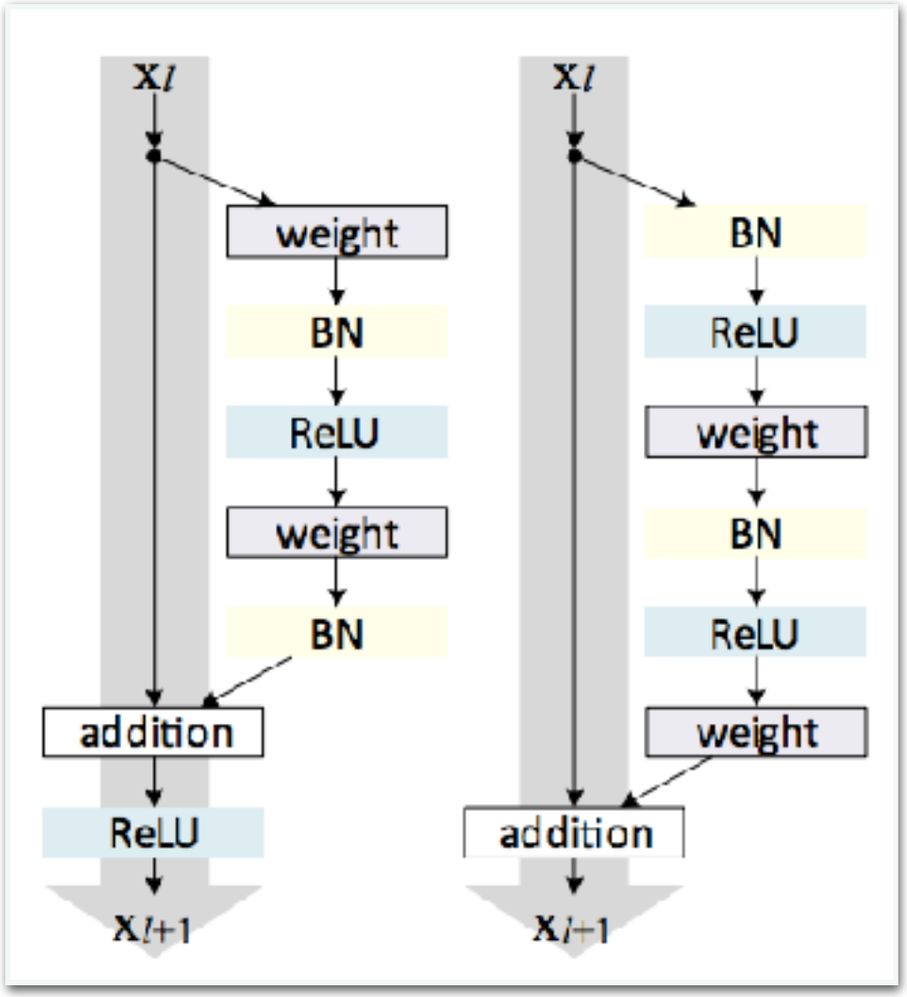
LeNet (1998)

Thanks to Joseph Cohen for diagrams

Animation by Dan Mané

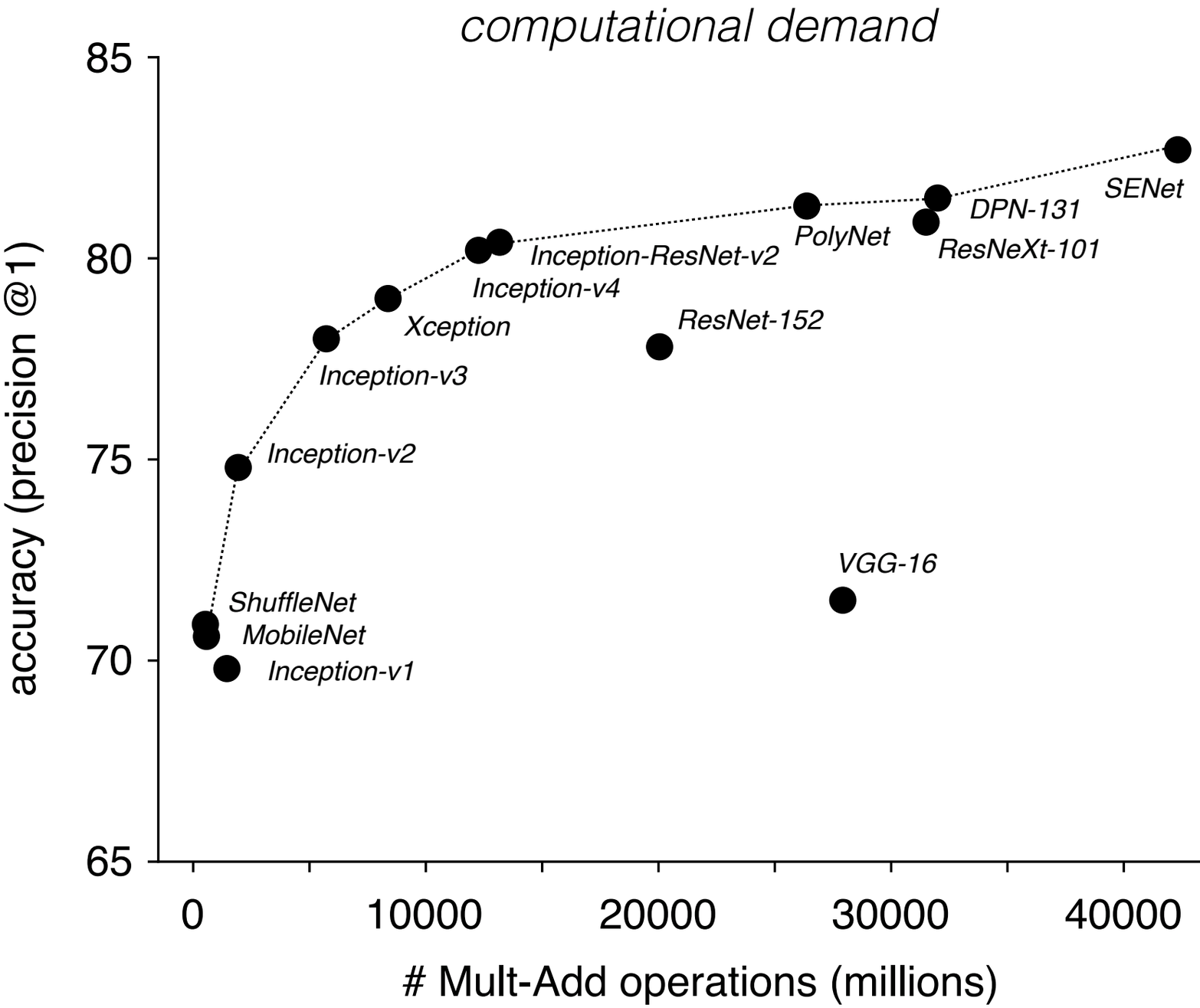
Vanishing gradient problem motivate architectures

- Additive pathways mitigate vanishing gradients.



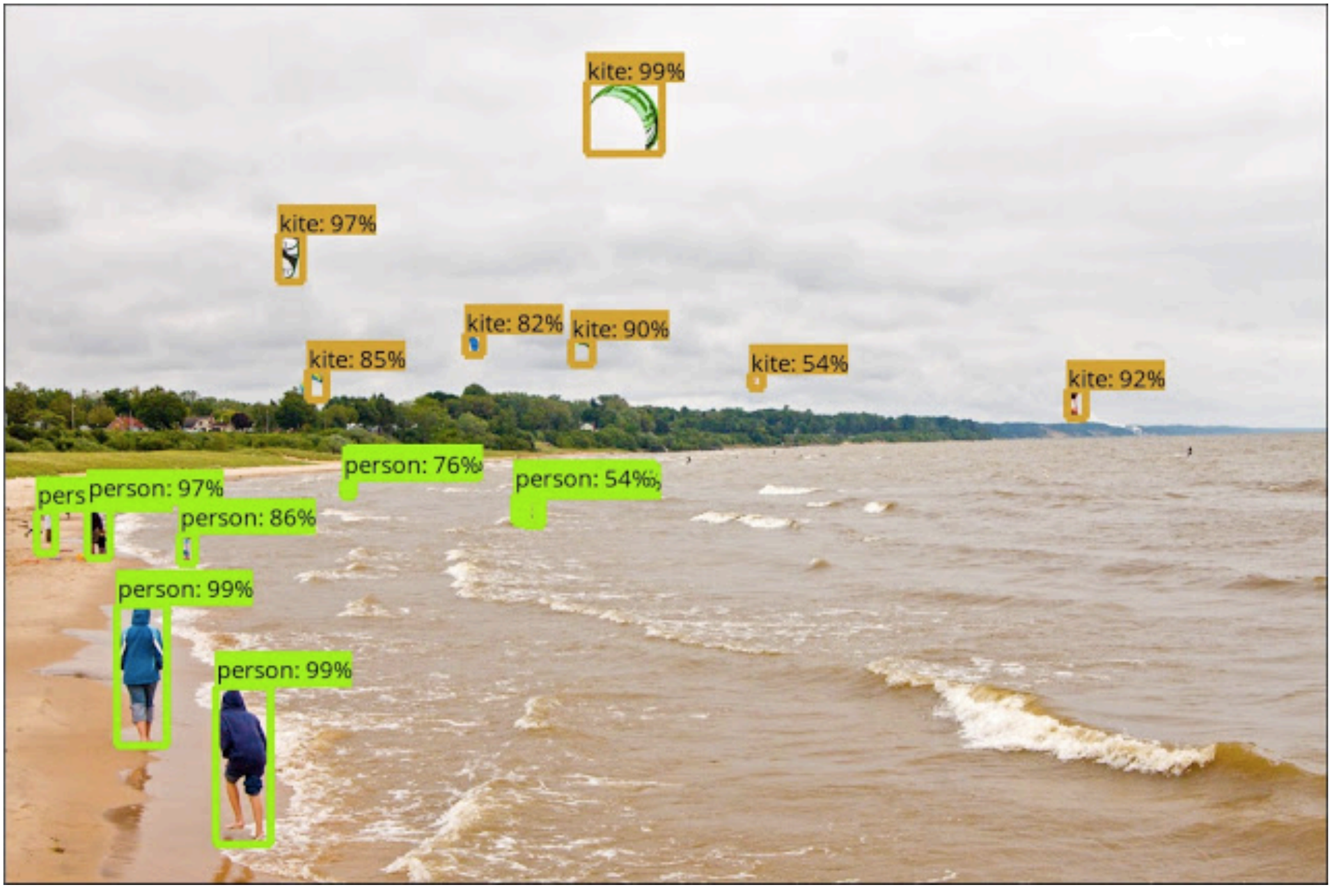
Gradient Flow in Recurrent Nets: the Difficulty of Learning Long-Term Dependencies
S Hochreiter et al (2001)
Deep residual learning for image recognition
K He, X Zhang, S Ren, and J Sun (2015)

The world of ImageNet architectures



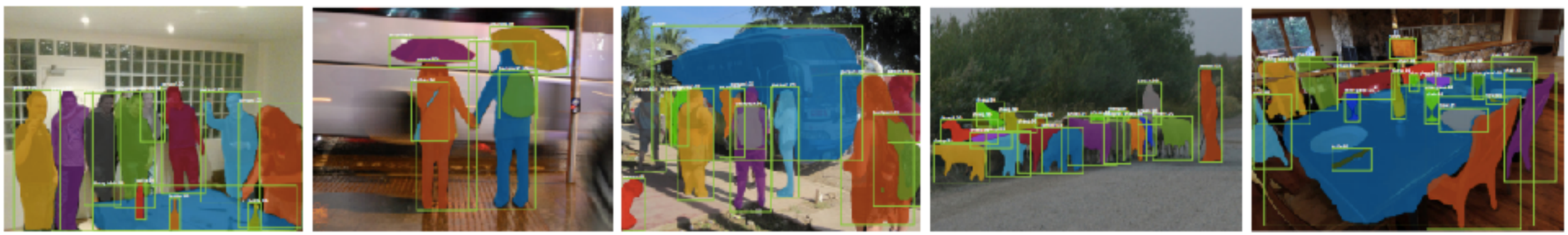
Learning Transferable Architectures for Scalable Image Recognition
B Zoph, V Vasudevan, J Shlens, Q Le (2017)

Vision tasks necessitate rich image features

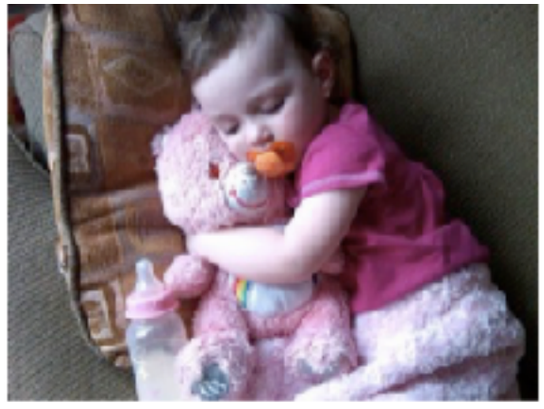


Speed/accuracy trade-offs for modern convolutional object detectors.
J. Huang, V. Rathod et al (2017)

Transfer image features to other visual problems



Mask R-CNN
K He, G Gkioxari, P Dollár, R Girshick (2017)



"A close up of a child holding a stuffed animal."

"Two pizzas sitting on top of a stove top open."

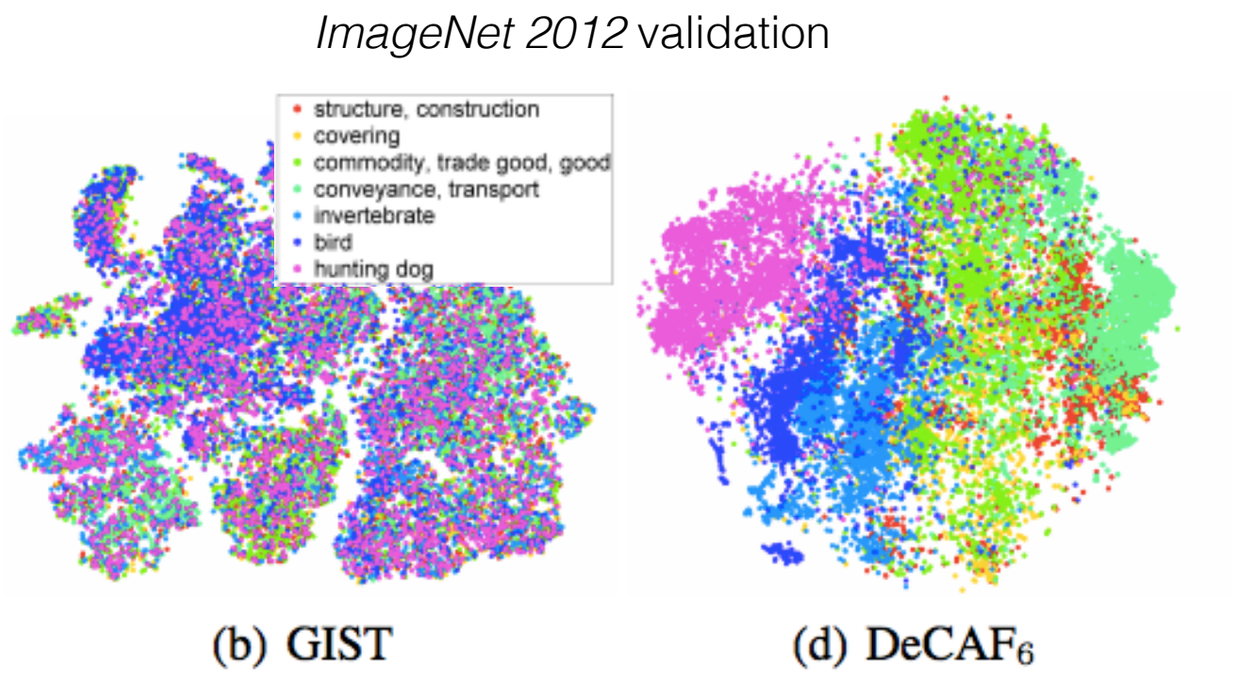
"A man flying through the air while riding a snowboard."

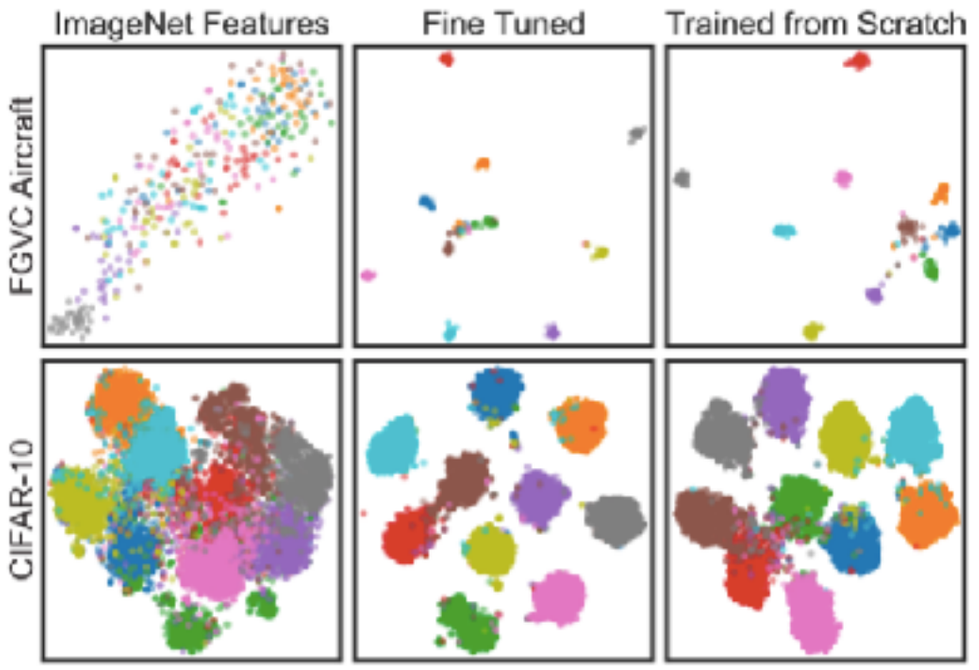
Show and Tell: A Neural Image Caption Generator
O Vinyals, A Toshev, S Bengio, D Erhan (2015)

Image features capture statistical structure

A Neural Algorithm of Artistic Style
L. Gatys, A. Ecker, M. Bethge (2015)
Exploring the structure of a real-time, arbitrary neural artistic stylization network
G Ghiasi, H Lee, M Kudlur, V Dumoulin and J Shlens (2017)

Architectures transfer across problems.



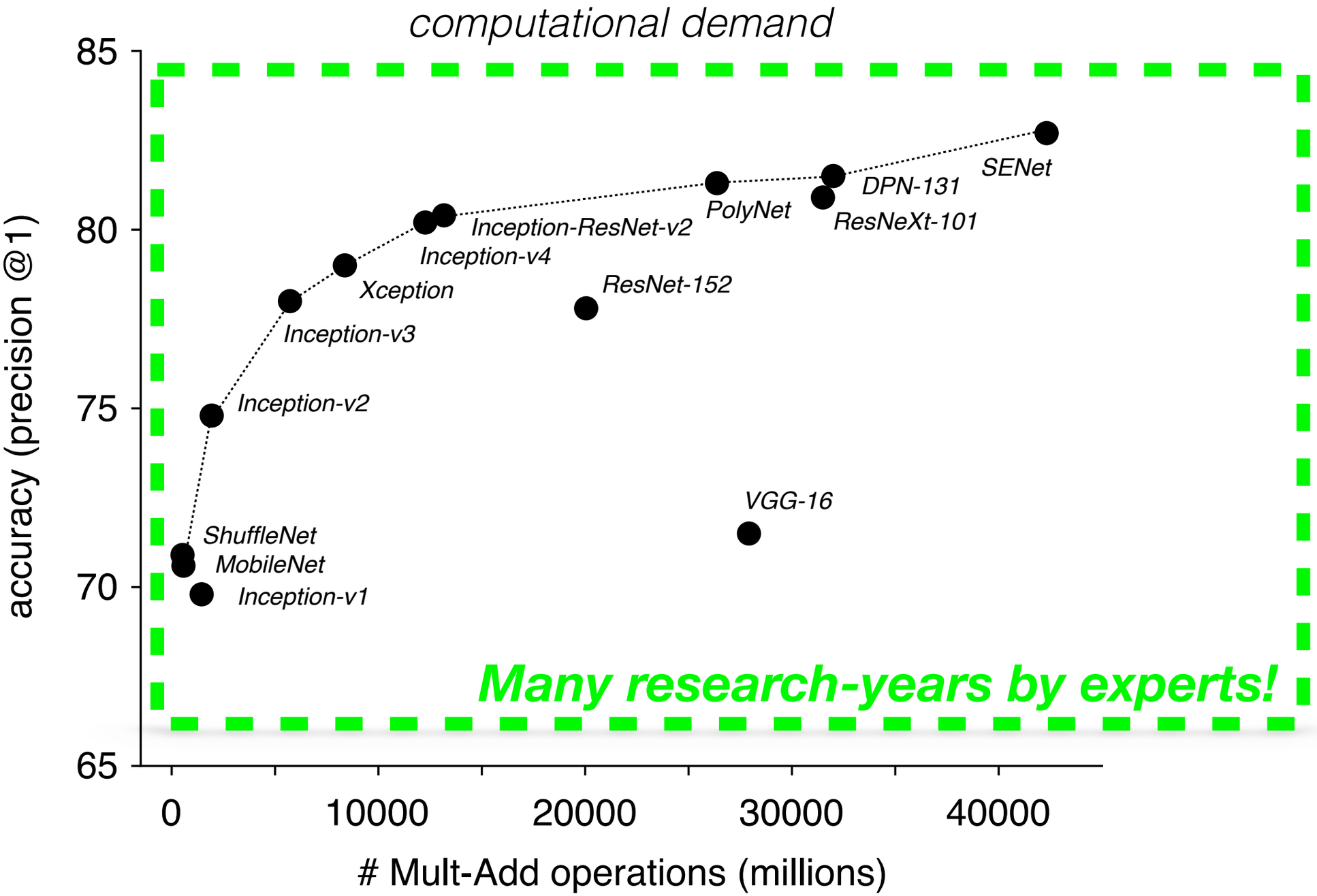


- Low dimensional visualizations of network features indicate that features are generically useful across problems.

DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition
J Donahue, Y Jia, O Vinyals, J Hoffman, N Zhang, E Tzeng, T Darrell. (2013)

Do Better ImageNet Models Transfer Better?
S Kornblith, J Shlens and Q Le (2018)

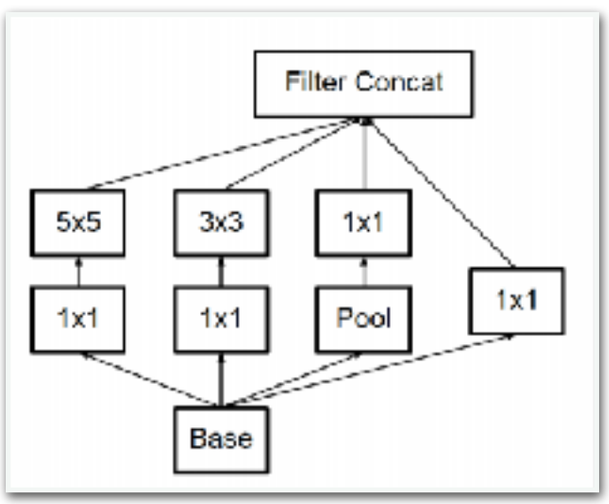
The world of ImageNet architectures



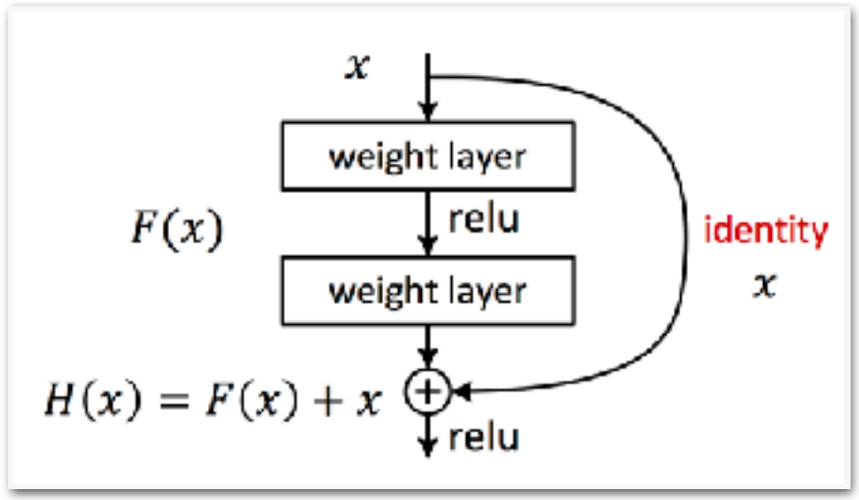
Towards engineering good network motifs

- A critical aspect of improving machine learning is the design of architecture *motifs*.

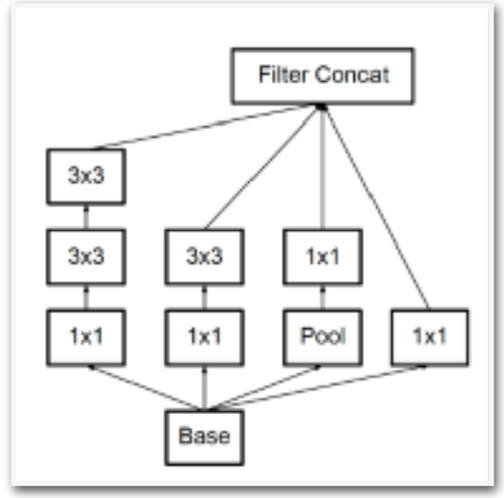
Inception-v1



ResNet



Inception-v3

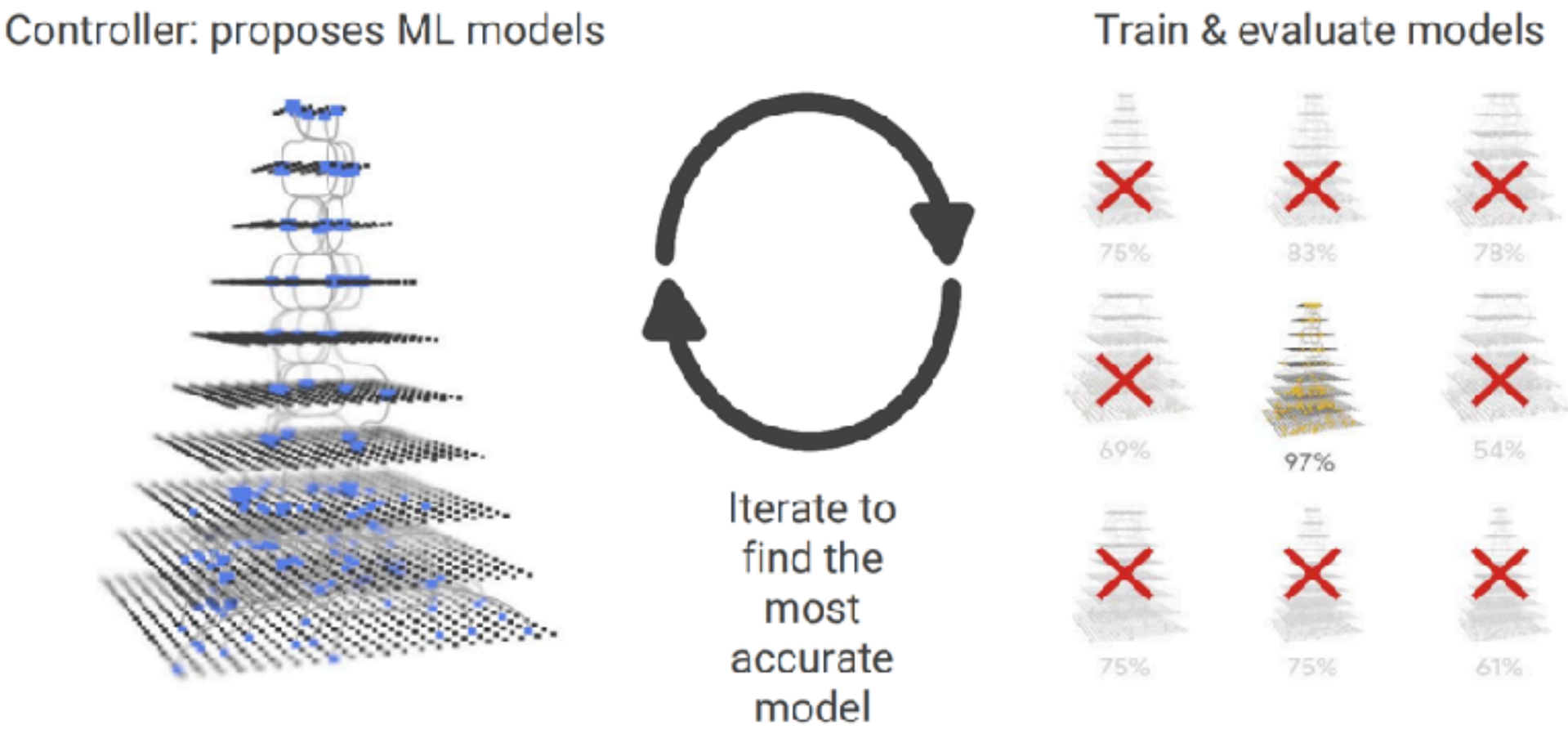


• • •

Courtesy of Kaiming He

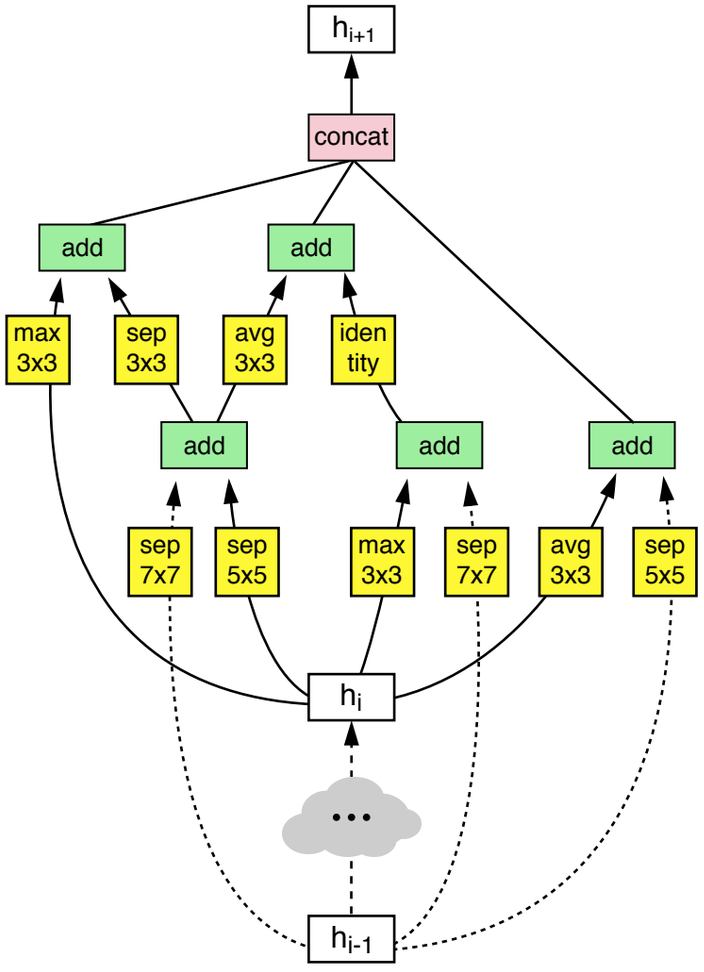
- A few humans are expert … but can we systematize the process of discovery?

Perhaps computers are better then humans?

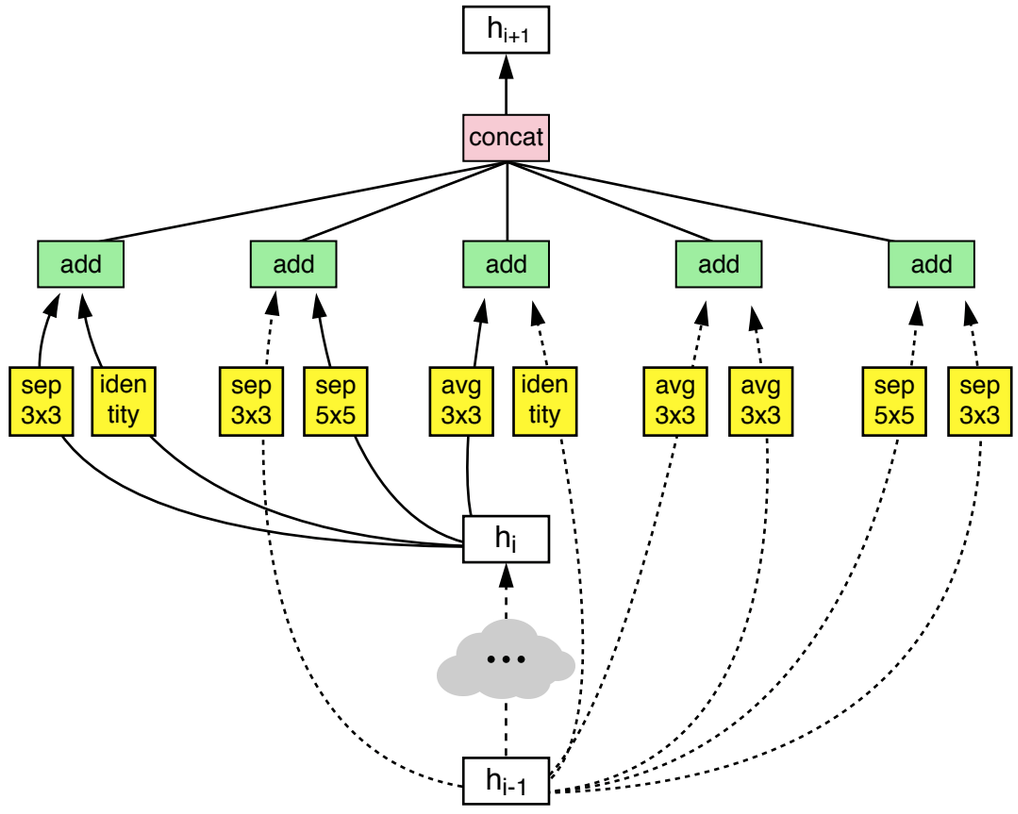


Neural Architecture Search with Reinforcement Learning
B Zoph, Q Le (2016)
Learning Transferable Architectures for Scalable Image Recognition
B Zoph, V Vasudevan, J Shlens, Q Le (2017)
Progressive Neural Architecture Search
C Liu, B Zoph, J Shlens, W Hua, LJ Li, L Fei-Fei, A Yuille, J Huang, K Murphy (2018)
DARTS: Differentiable Architecture Search
H Liu, K Simonyan, Y Yang (2018)

Learned network motifs



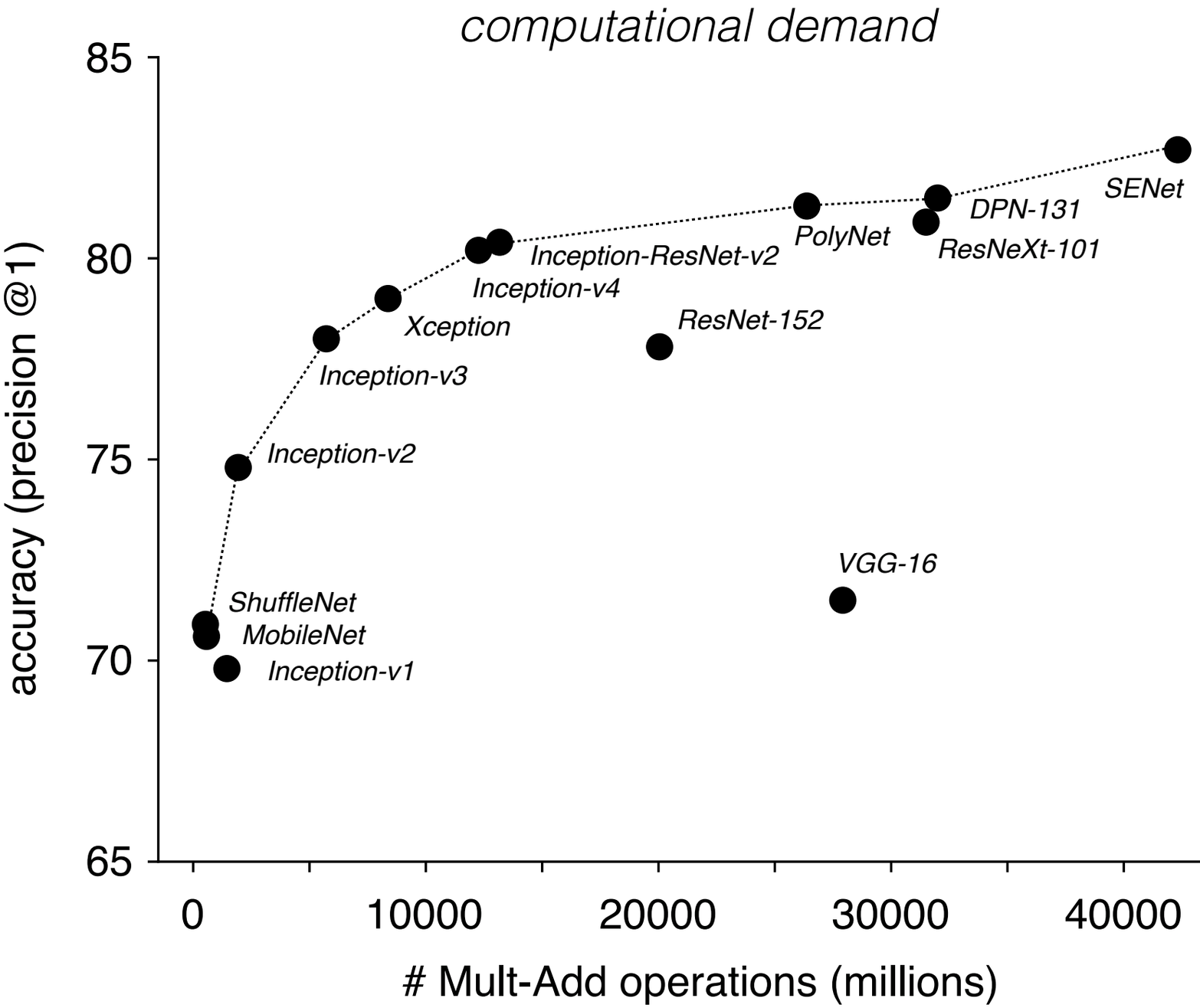
reduction cell



normal cell

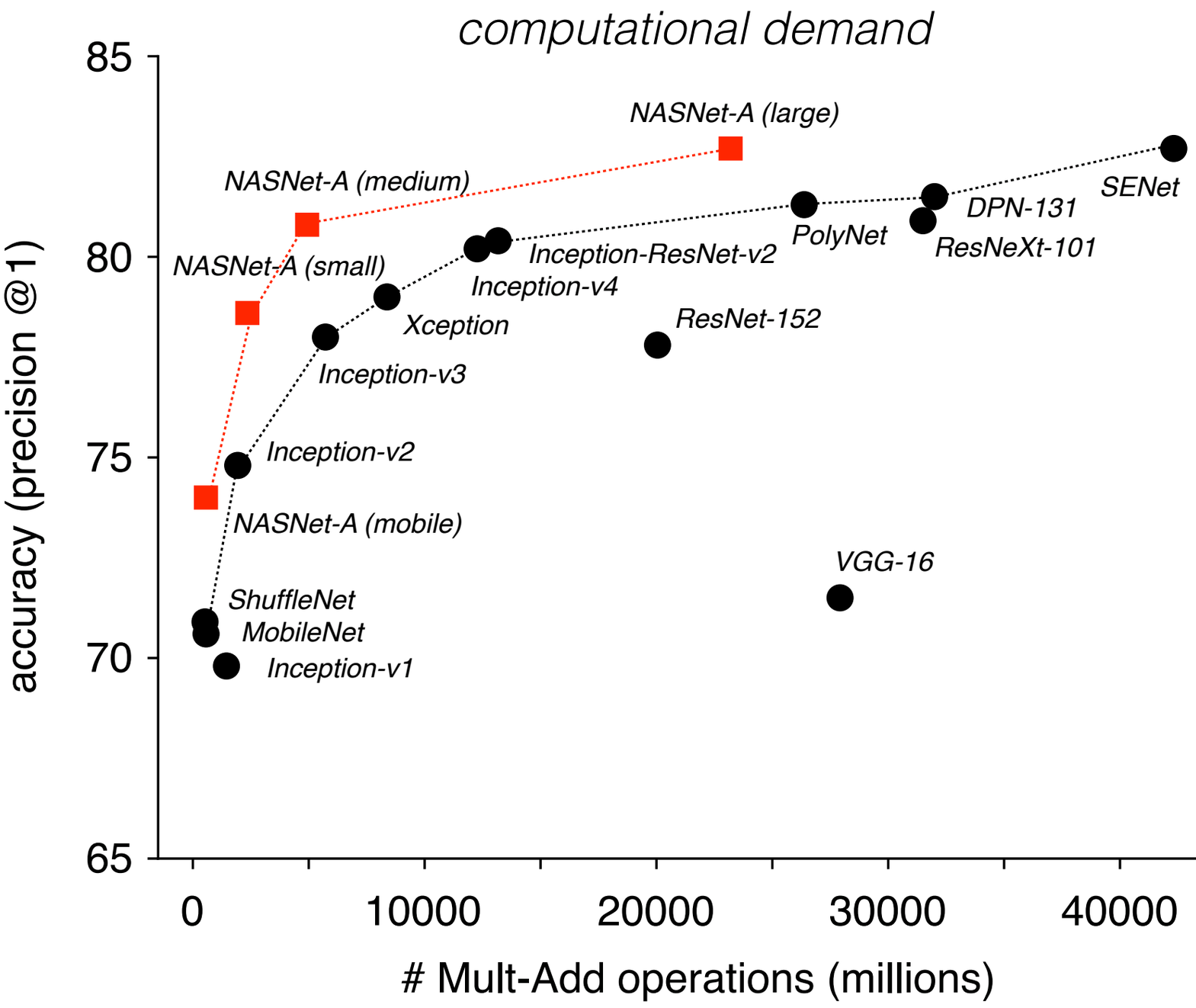
Learning Transferable Architectures for Scalable Image Recognition
B Zoph, V Vasudevan, J Shlens, Q Le (2017)

The world of human-invented architectures.



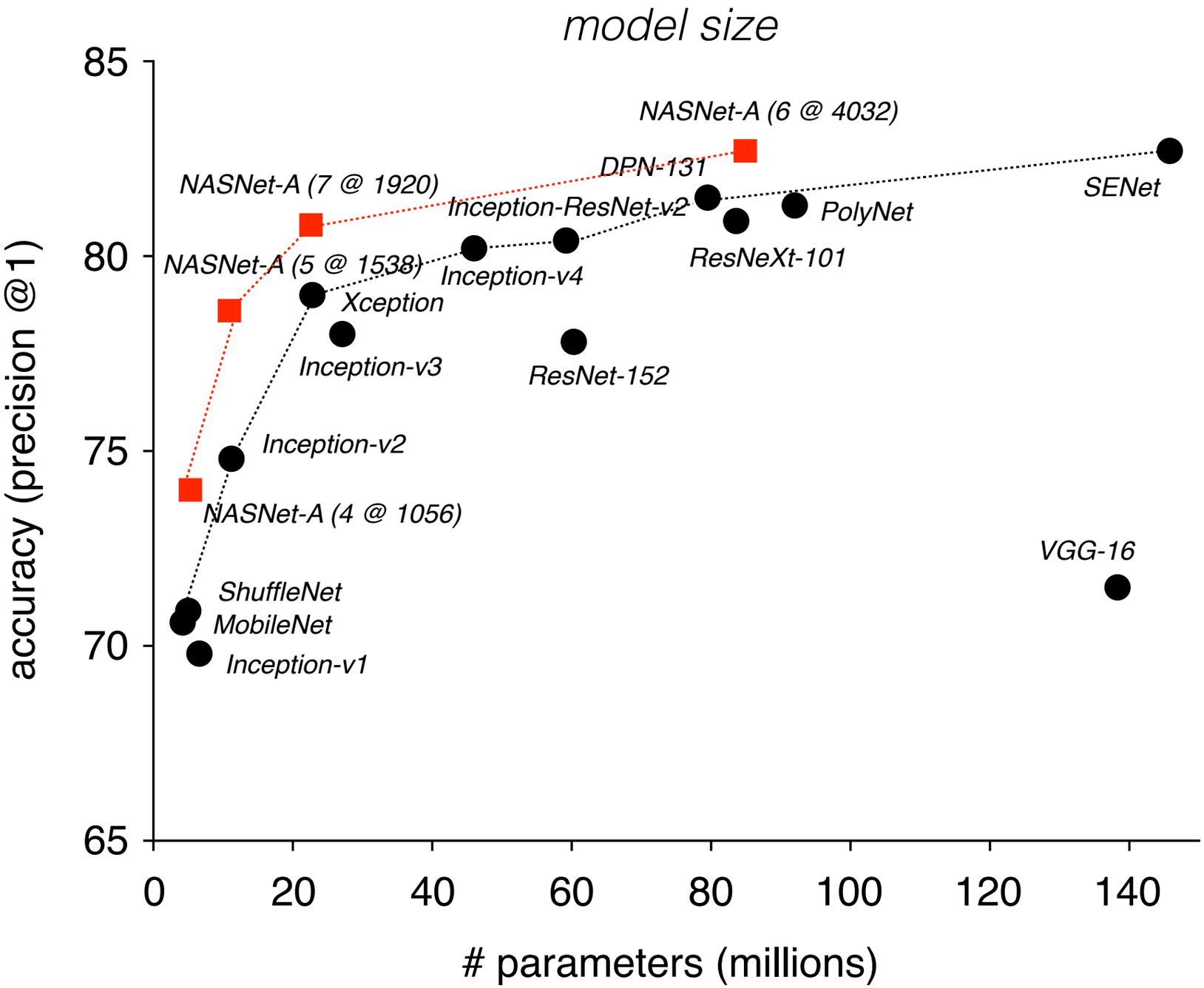
Learning Transferable Architectures for Scalable Image Recognition
B Zoph, V Vasudevan, J Shlens, Q Le (2017)

Learned architectures surpass human-invented



Learning Transferable Architectures for Scalable Image Recognition
B Zoph, V Vasudevan, J Shlens, Q Le (2017)

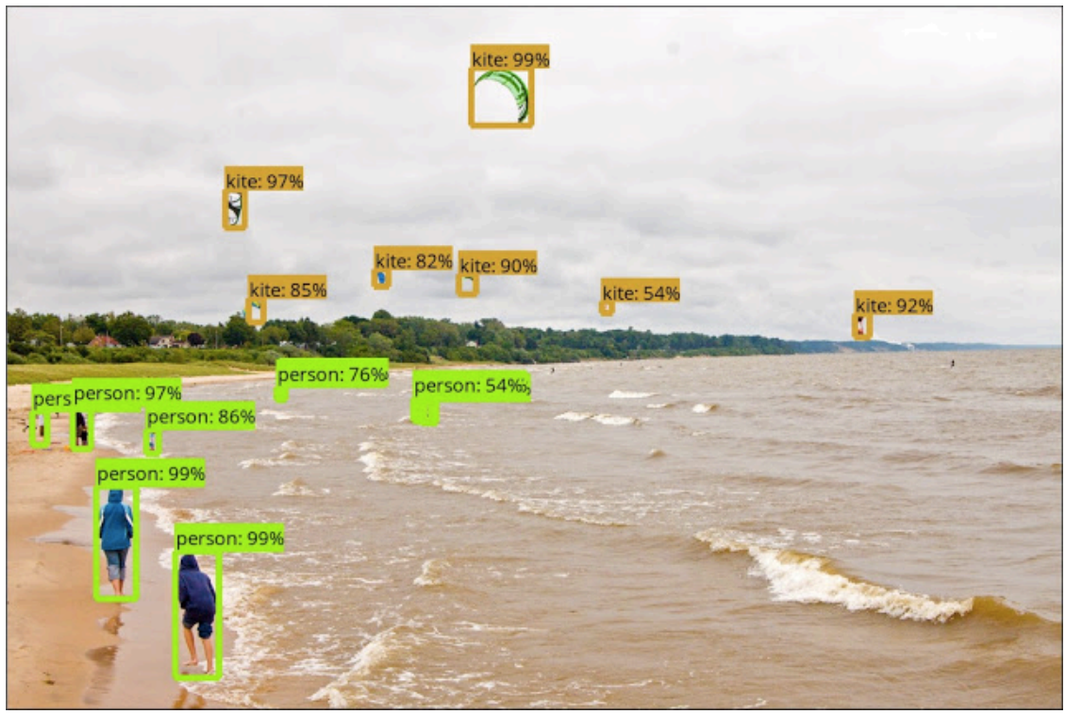
Learned architectures surpass human-invented



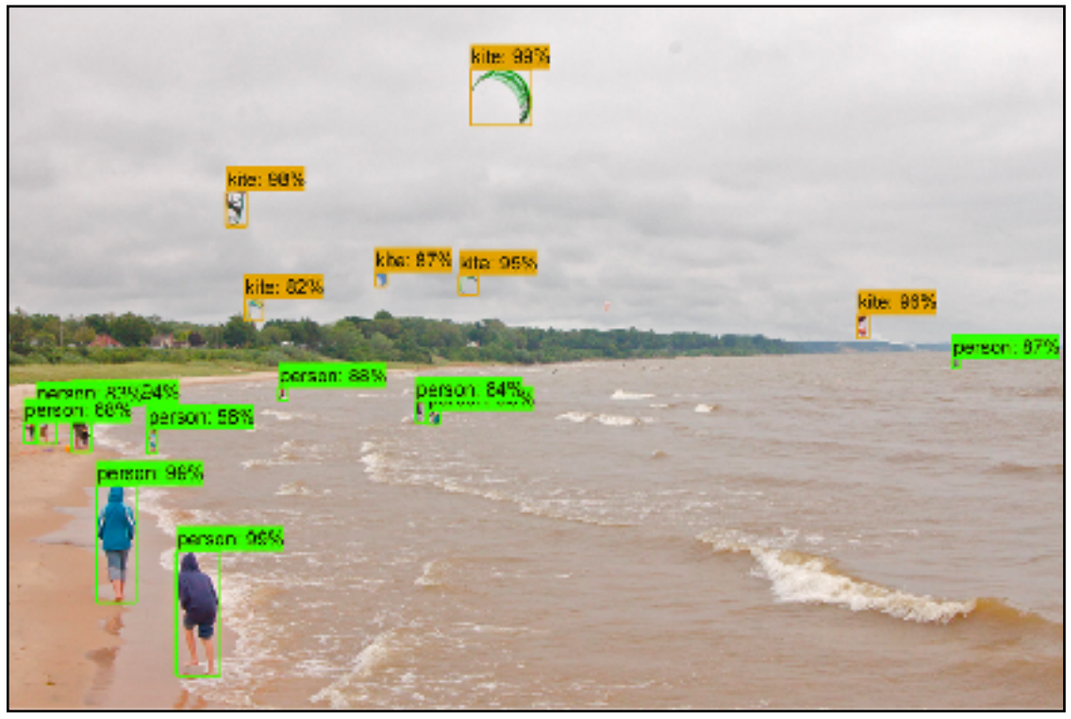
Learning Transferable Architectures for Scalable Image Recognition
B Zoph, V Vasudevan, J Shlens, Q Le (2017)

Vision tasks necessitate rich image features

previous state-of-the-art



NASNet image features



	performance (mAP)
ResNet-101-FPN *(focal loss)*	39.1%
Inception-ResNet-v2 *(TDM)*	36.8%
NASNet-A	**43.1%**

Learning Transferable Architectures for Scalable Image Recognition
B Zoph, V Vasudevan, J Shlens, Q Le (2017)

Agenda

1. Challenges and inspiration from vision

2. Convolutional neural networks

3. Modern developments

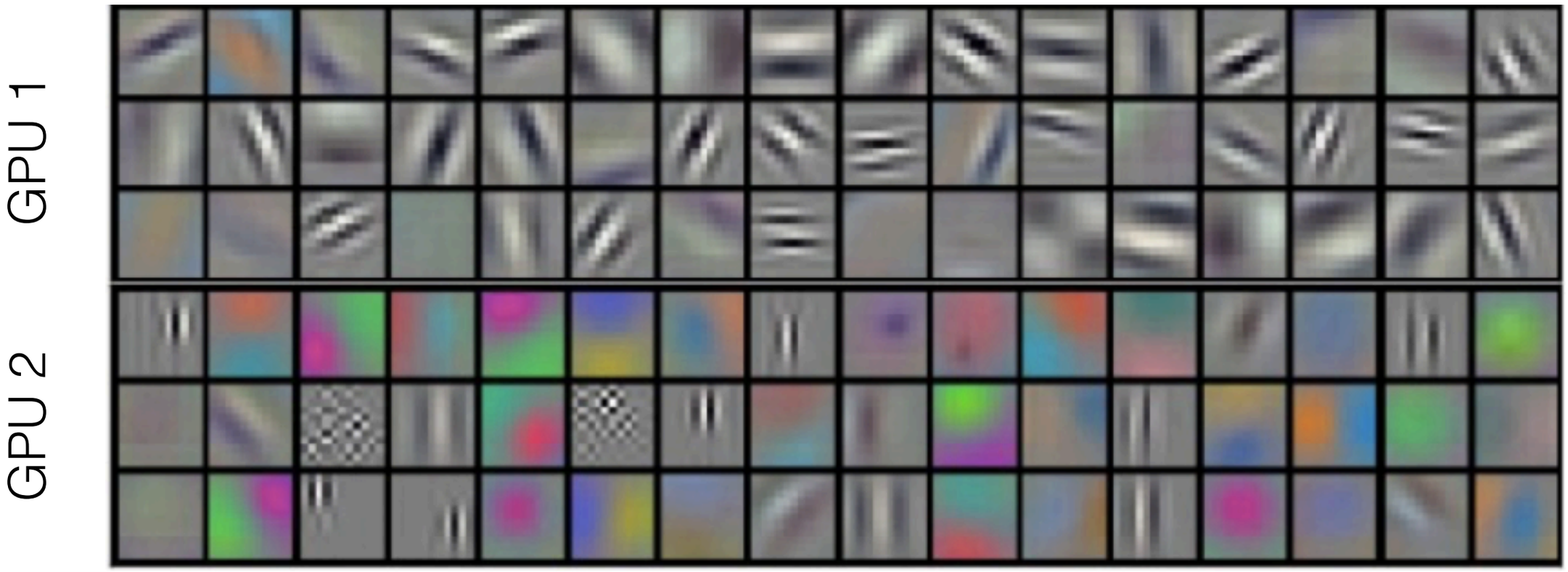
   - architectures, meta-learning, normalization, transfer learning

4. **Towards understanding higher-level visual features**

5. Opportunities and conclusions

How do we analyze low level features?

The weight vectors of first layer neurons form a Gabor function basis.

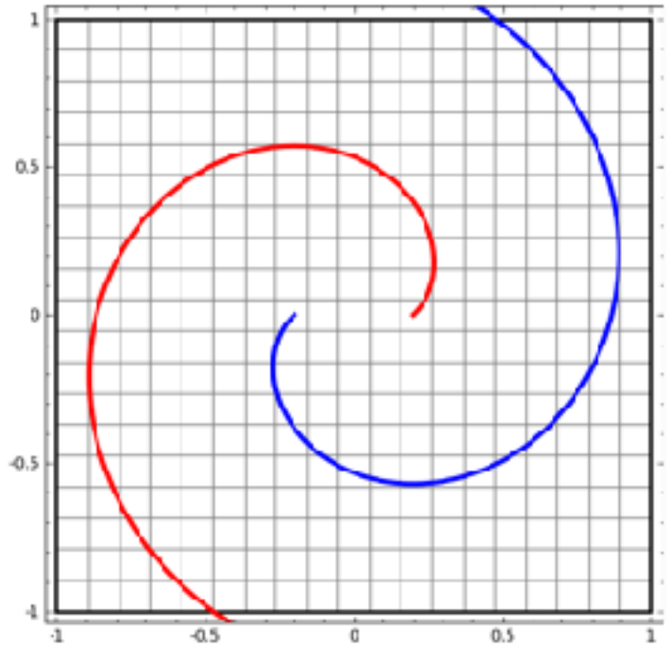


ImageNet Classification with Deep Convolutional Neural Networks
A Krizhevsky I Sutskever, G Hinton (2012)

How do we understand high level features?

- Nonlinearities distort and warp space and we can not just read off the weights.

- What *should* the statistics of a high dimensional representation look like at each layer?

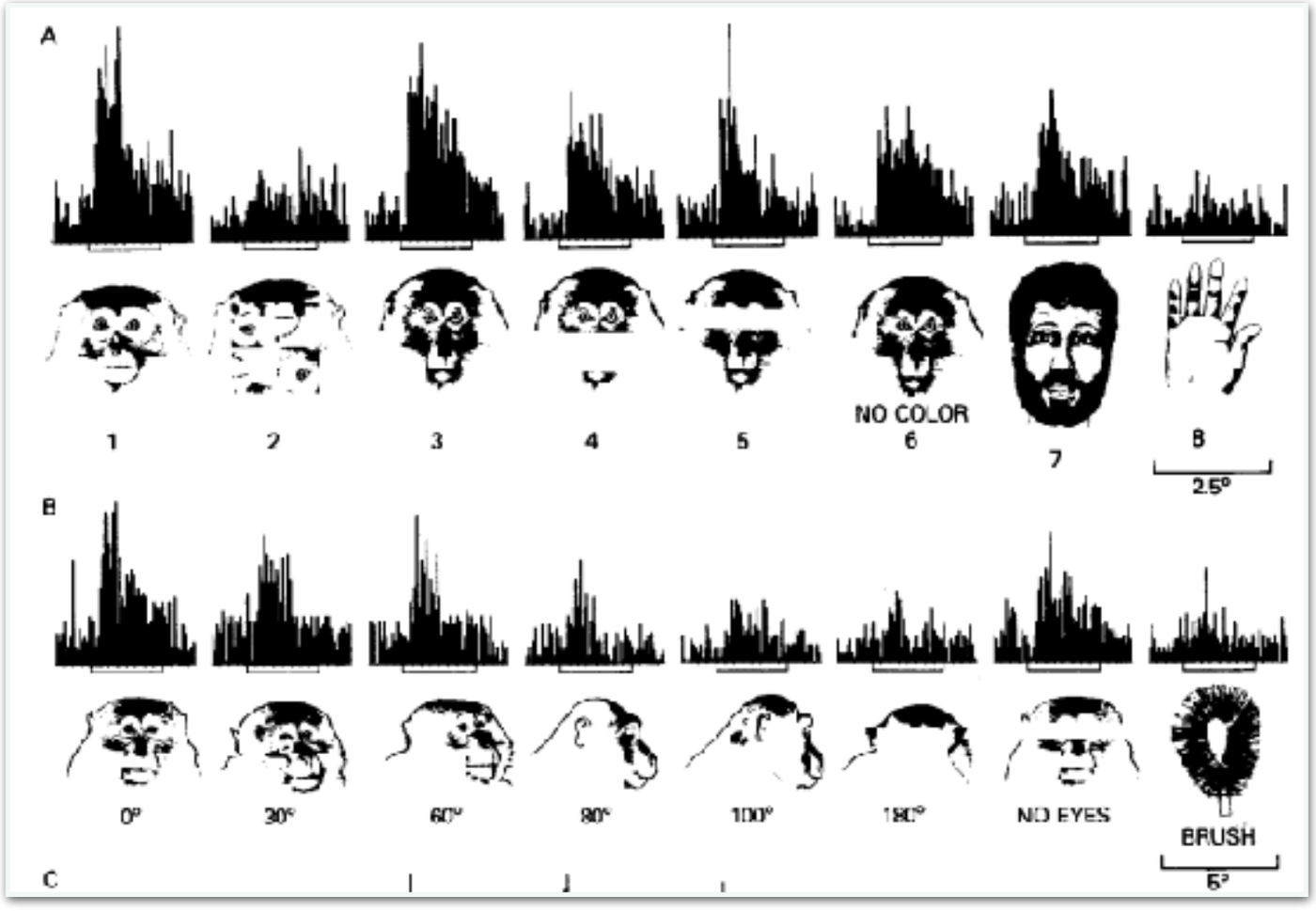


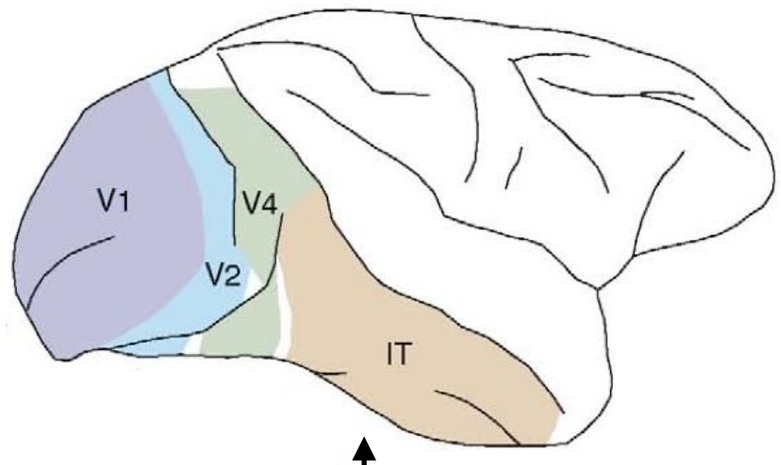
2-D feature space in 4 layer network

Neural networks, manifolds and topology
Chris Olah (2014)
http://colah.github.io/

How do we understand high level features?

This has been a topic in neuroscience for 30+ years.

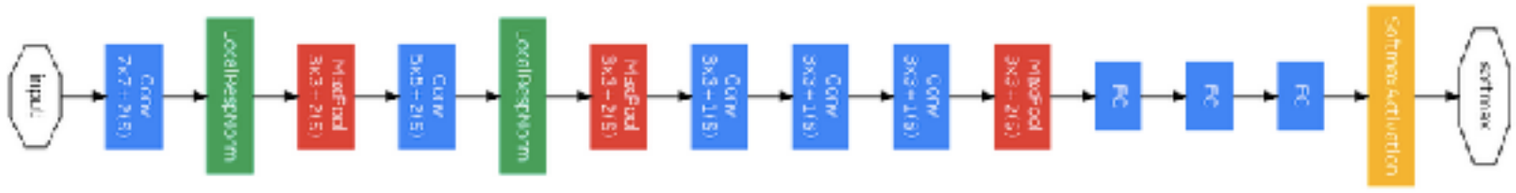




IT cortex

Stimulus-selective properties of inferior temporal neurons in the macaque
R Desimone, TD Albright, CG Gross, and C Bruce (1984)

Understanding higher level image features.



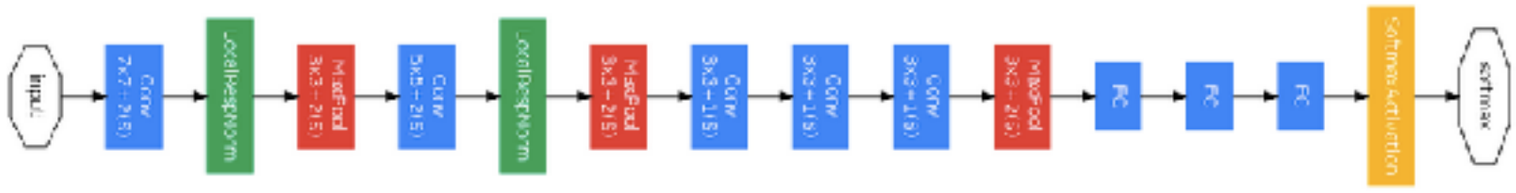
- What information do networks represent at each layer?

- Understanding these issues can teach us:

  a. How to build better machine learning architectures.

  b. How to understand the Marr's levels 2 and 3.

Understanding higher level image features.

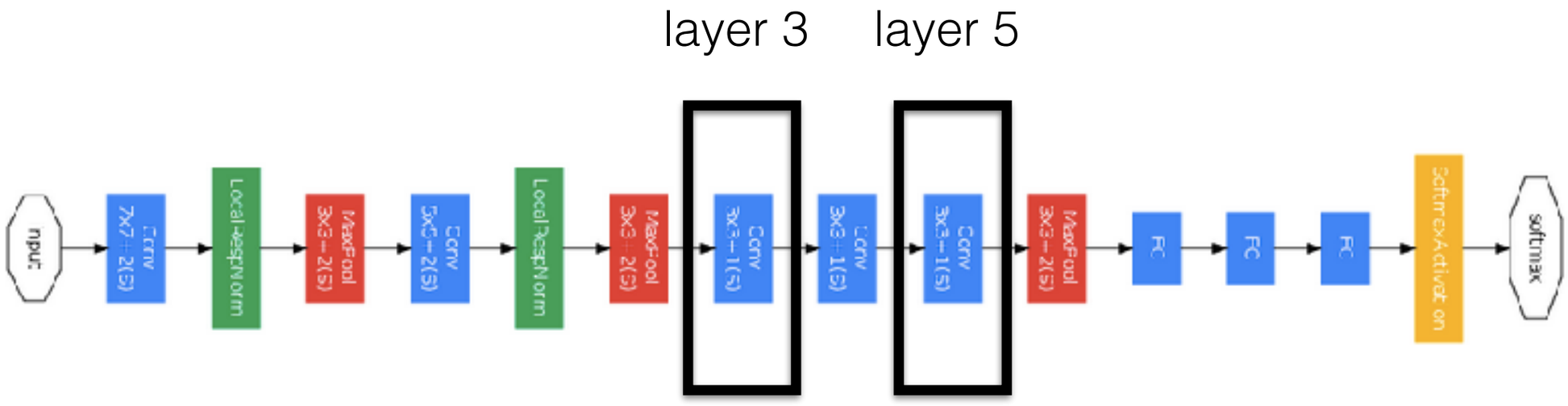


1. Visualizing which pixels elicit largest activations

2. Reconstructing images from network activity

3. Distort image pixels to amplify activation of features.

Find image pixels that elicit largest activation

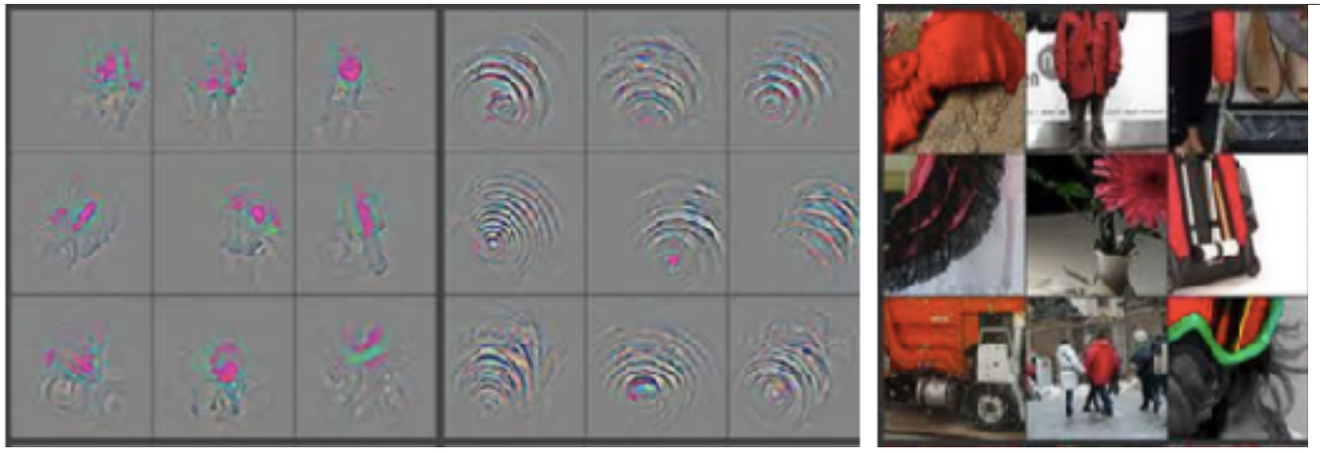


- Train a network for image classification on ImageNet.
- Examine individual neurons at middle layers of network and find responsible pixels.

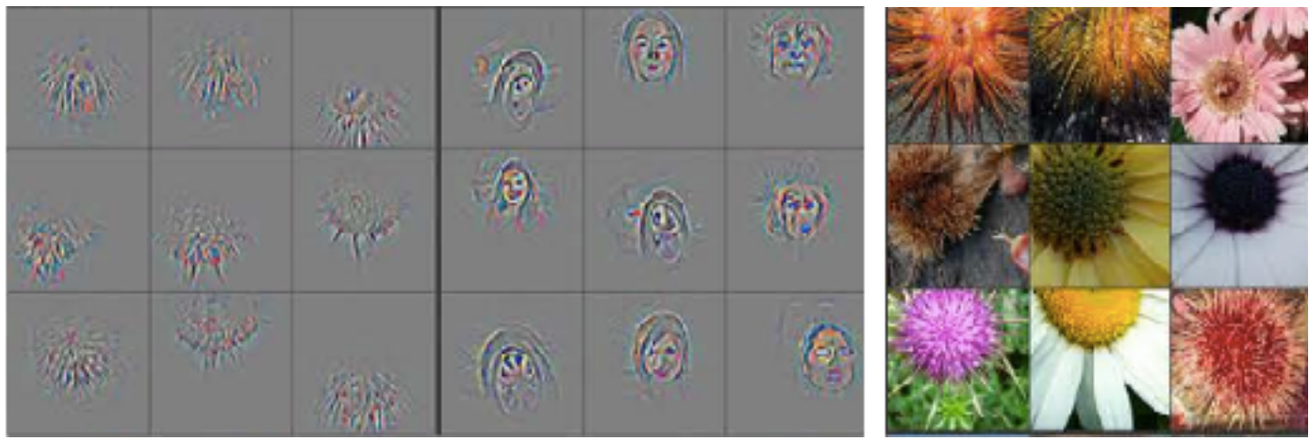
Visualizing and Understanding Convolutional Networks
M Zeiler and R Fergus (2013)

Find image pixels that elicit largest activation

layer 3

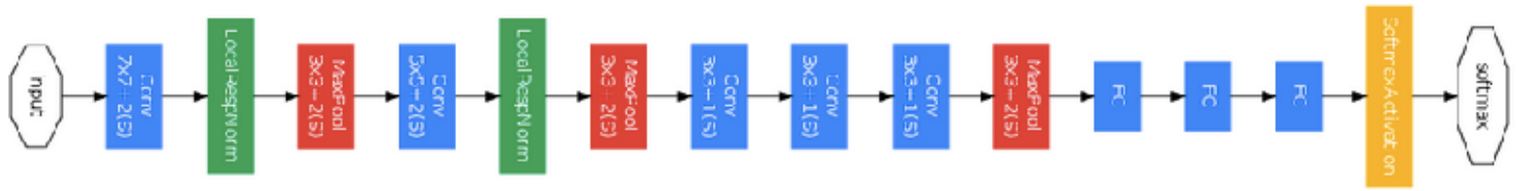


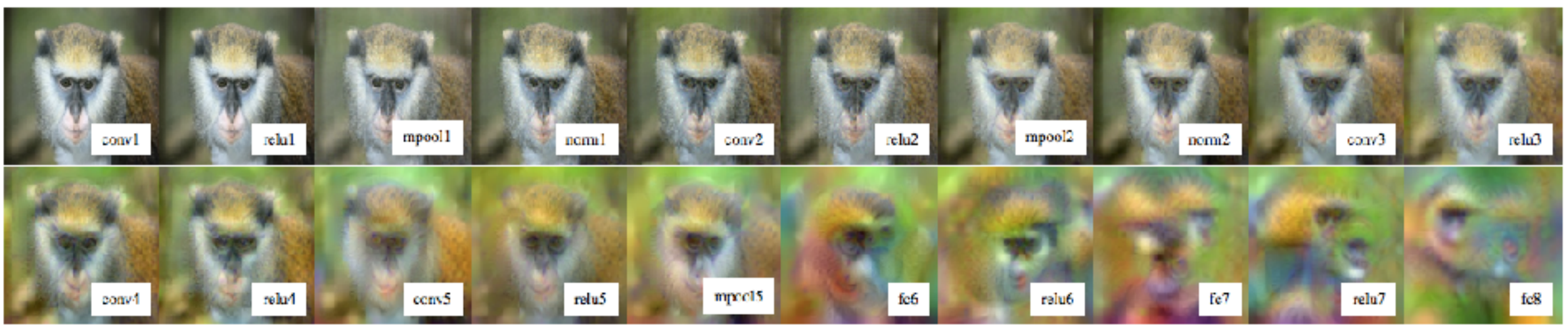
layer 5



Visualizing and Understanding Convolutional Networks
M Zeiler and R Fergus (2013)

Reconstruct image from network activations

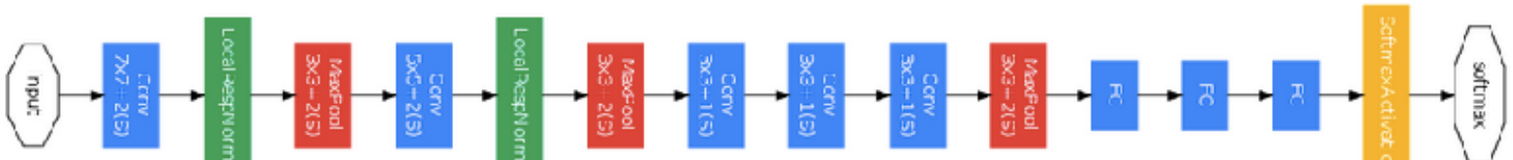




- How well can one reconstruct an image a given layer within a hierarchical convolutional network?

Understanding Deep Image Representations by Inverting Them.
A Mahendran and A Vedaldi (2014)

Distort pixels to amplify activation of features.



mid level

penultimate layer

- For a given image, amplify the activation of higher level features and visualize how it effects the image.

A. Mordvintsev, C. Olah and M. Tyka

http://googleresearch.blogspot.com/2015/06/inceptionism-going-deeper-into-neural.html

A. Mordvintsev, C. Olah and M. Tyka

http://googleresearch.blogspot.com/2015/06/inceptionism-going-deeper-into-neural.html

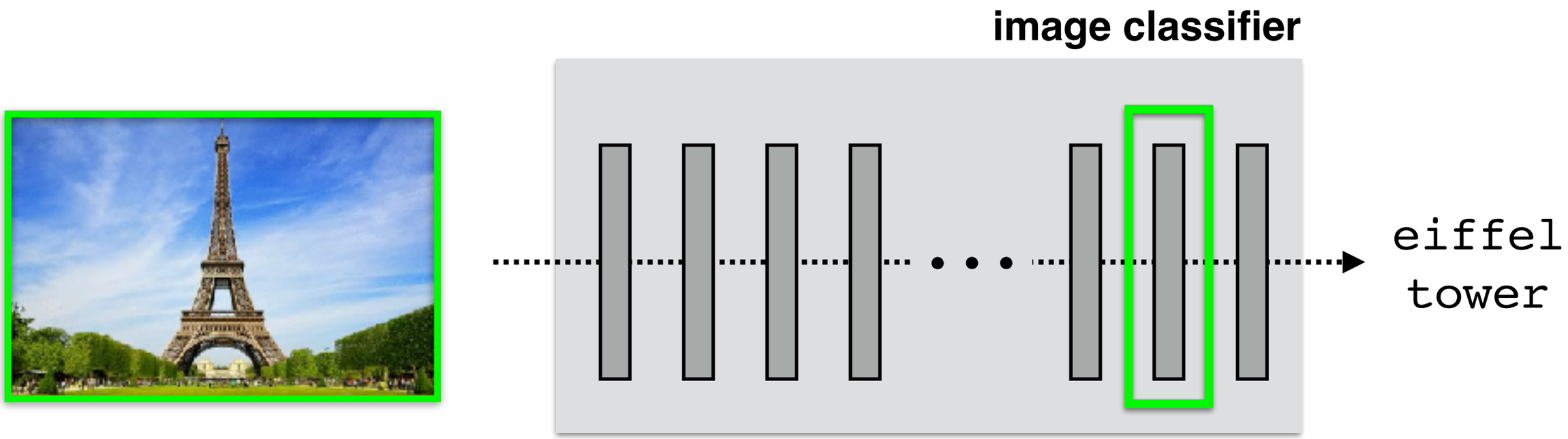
A. Mordvintsev, C. Olah and M. Tyka

http://googleresearch.blogspot.com/2015/06/inceptionism-going-deeper-into-neural.html

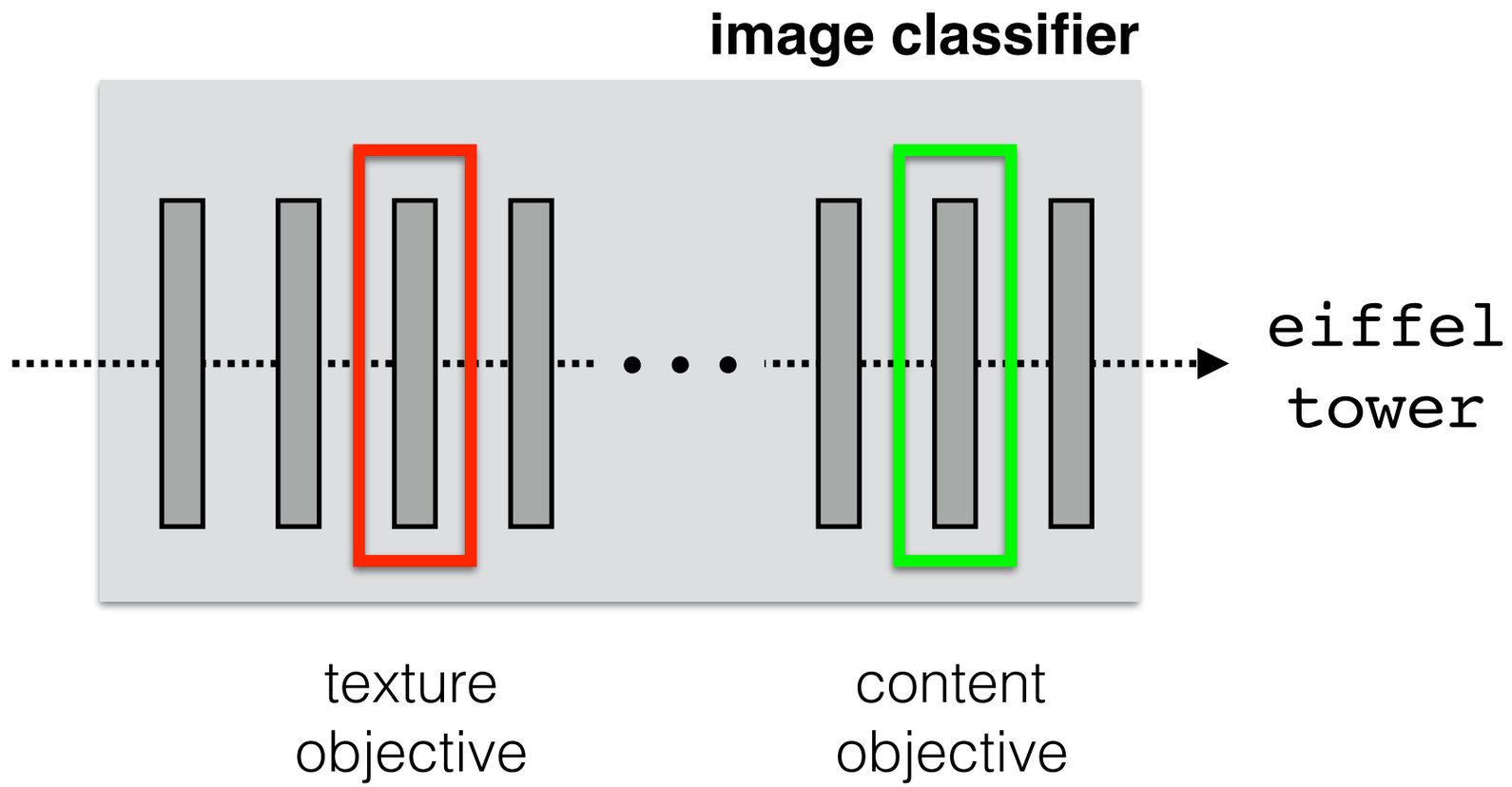
A. Mordvintsev, C. Olah and M. Tyka

http://googleresearch.blogspot.com/2015/06/inceptionism-going-deeper-into-neural.html

Higher level image features contain content



Mixing representations between images



A Neural Algorithm of Artistic Style
L. Gatys, A. Ecker, M. Bethge (2015)

https://github.com/kaishengtai/neuralart

A Neural Algorithm of Artistic Style
L. Gatys, A. Ecker, M. Bethge (2015)

Agenda

1. Challenges and inspiration from vision

2. Convolutional neural networks

3. Modern developments

    - architectures, meta-learning, normalization, transfer learning

4. Towards understanding higher-level visual features

5. **Opportunities and conclusions**

Networks make different types of errors

CNN errors



saltshaker	reel	hatchet
pill bottle	stethoscope	vase
water bottle	whistle	picher; ewer
lotion	ice lolly, lolly	coffeepot
hair spray	hair spray	mask
beer bottle	maypole	cup

human errors

toy terrier

Kerry blue terrier

Irish terrier

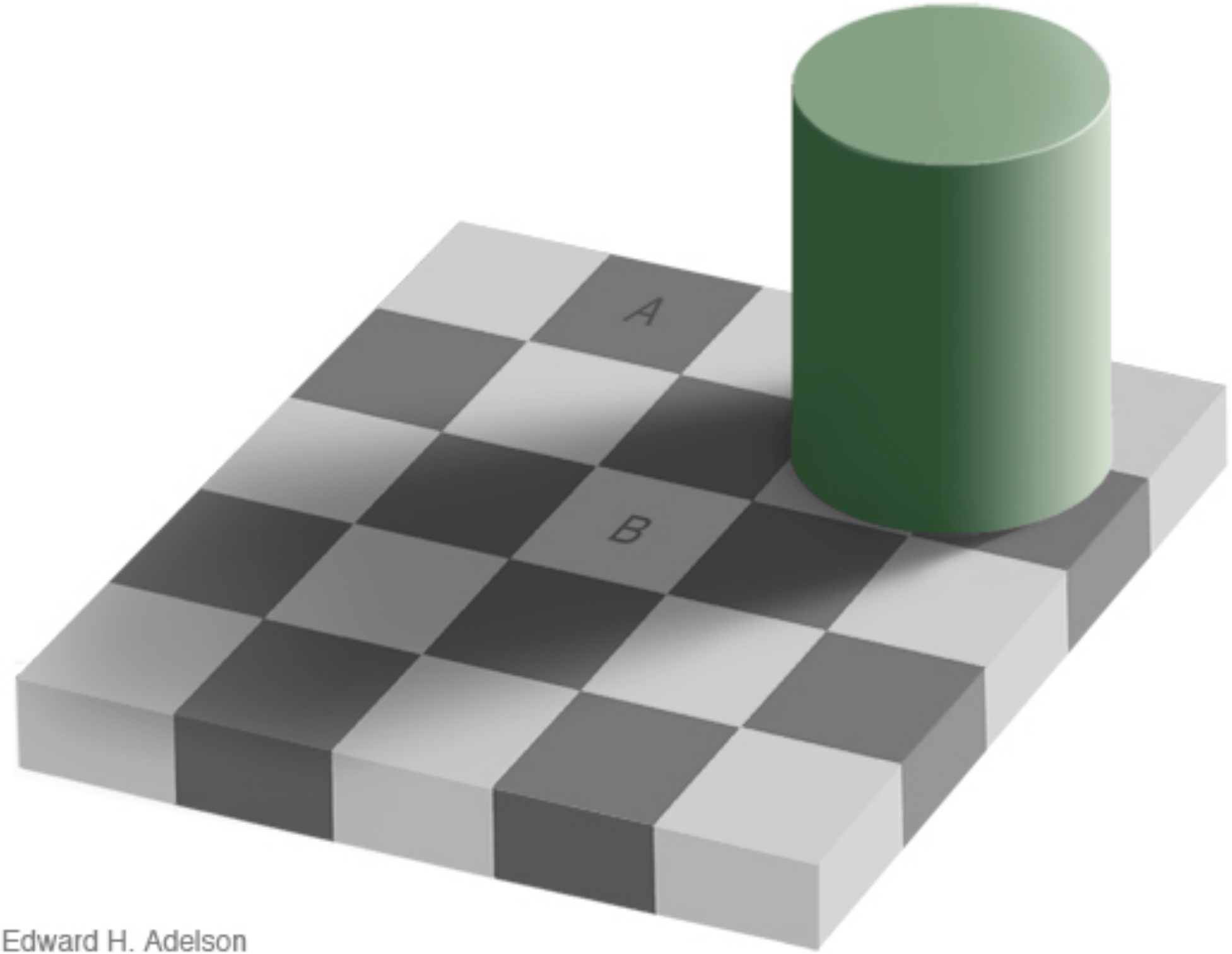
Norfolk terrier

Norwich terrier

- Best networks achieve an error rate of ~3.5% as of 2018
- One human achieved an error rate of ~5.1% as of 2014.

(trained on 500 images; tested on 1500 images) .

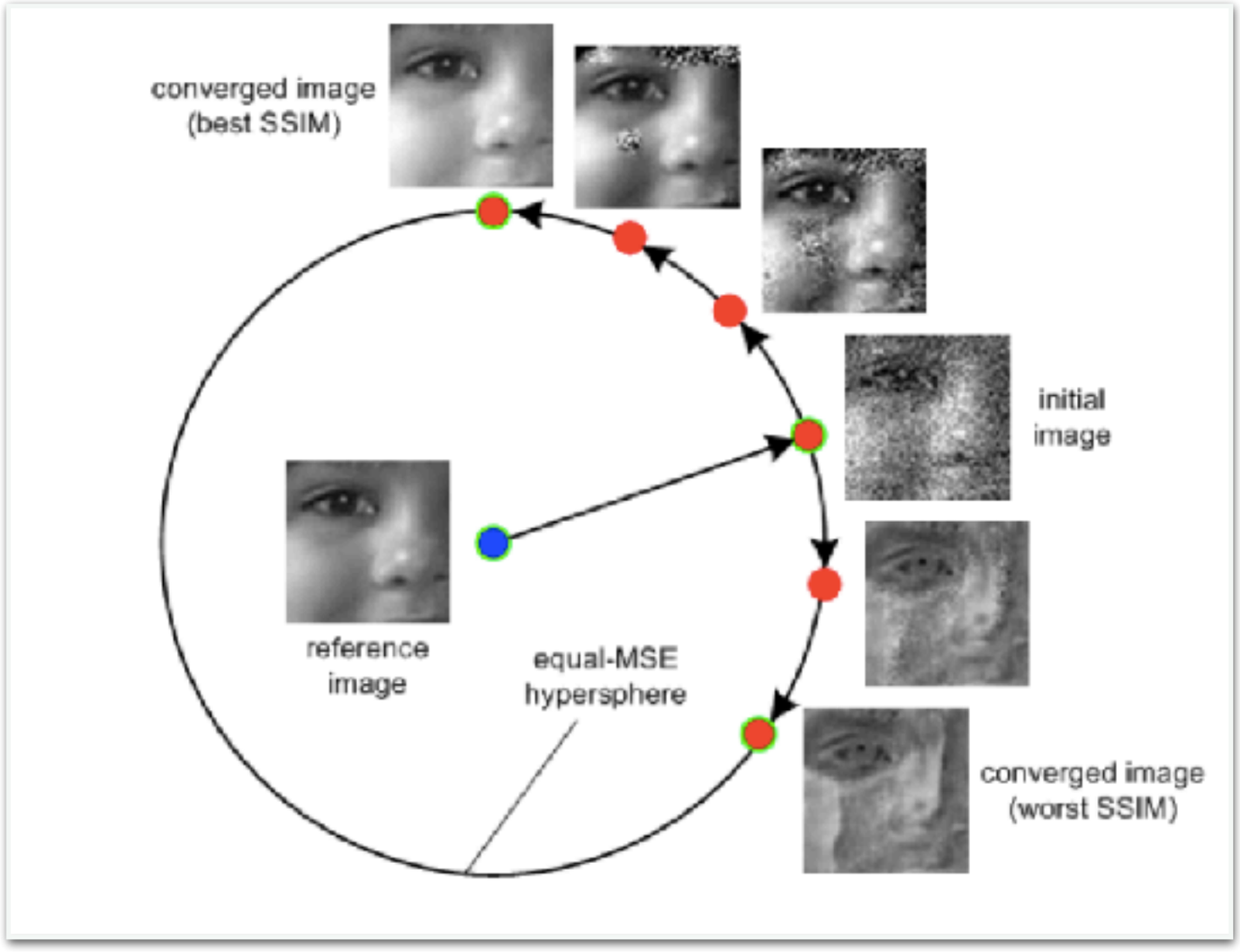
Andrej Karpathy
http://karpathy.github.io/



Edward H. Adelson

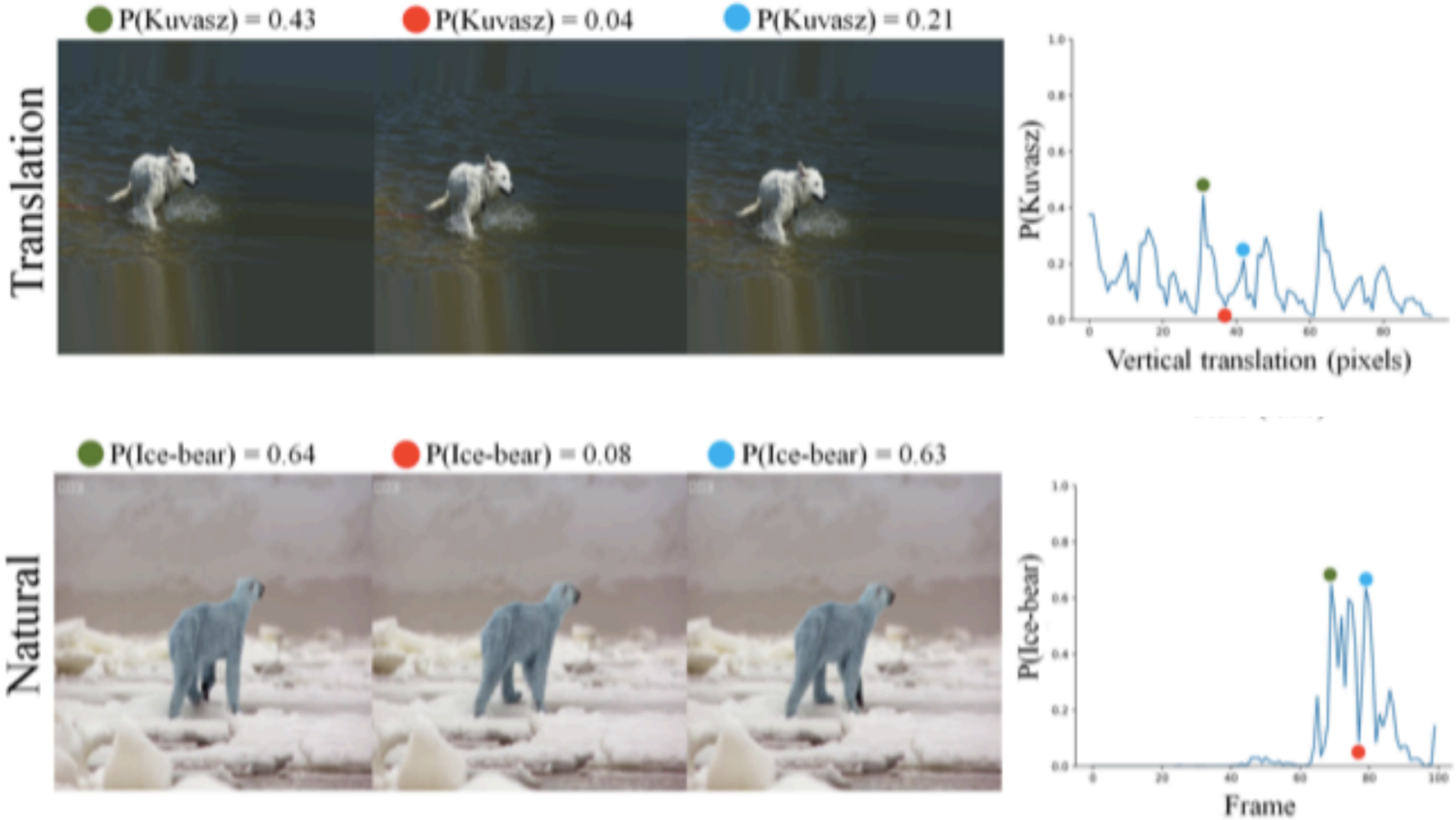
Hard to quantify perceptual image metrics.

- L_p losses do not correspond to perceptual distance.



)Image Quality Assessment: From Error Visibility to Structural Similarity
Z Wang et al (2003)
Mean squared error: Love it or leave it? A new look at signal fidelity measures
Z Wang and A Bovik (2009)

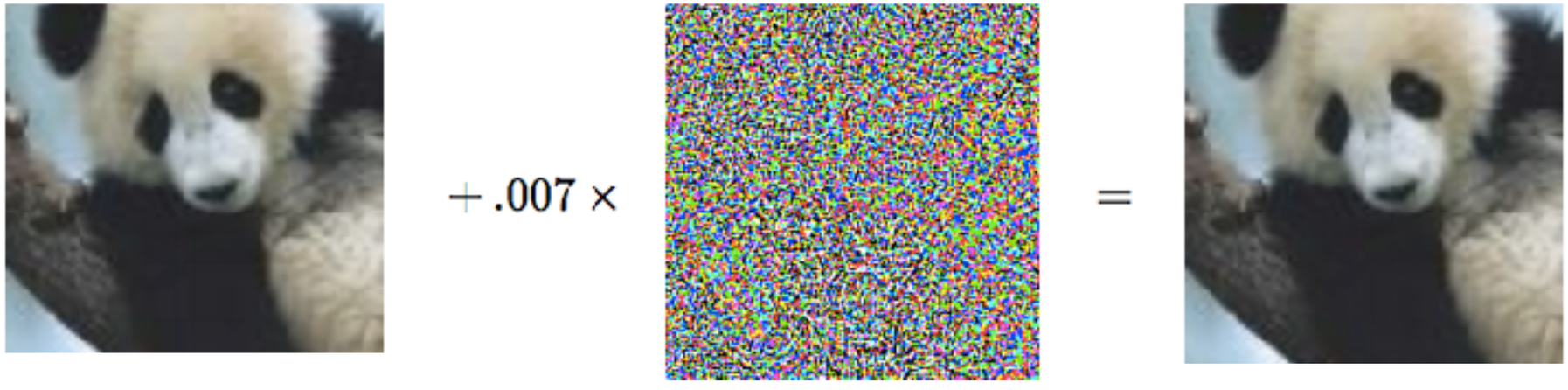
Perceptual spaces for deep networks?



- Image classification systems are far more brittle then the human perception.

Why do deep convolutional networks generalize so poorly to small image transformations?
A Azulay and Y Weiss (2018)

Perceptual spaces for deep networks?



- Maliciously generate slight deviations in images that profoundly effect an image classification system.

Intriguing Properties of Neural Networks
C Szegedy et al (2013)
Explaining and Harnessing Adversarial Examples
I Goodfellow, J Shlens and C Szegedy (2015)

Agenda

1. Challenge and inspiration for vision

2. Convolutional neural networks

3. Modern developments

    - architectures, meta-learning, normalization, transfer learning

4. Towards understanding higher-level visual features

5. Opportunities and conclusions

Take home messages.

- Think about computation when building network architectures.

- How do we better align network architectures with perception?

- Aiming for *chemistry* in the field of computer vision.

Thank you for your time.

Questions?