6.819/6.869 Advances in Computer Vision

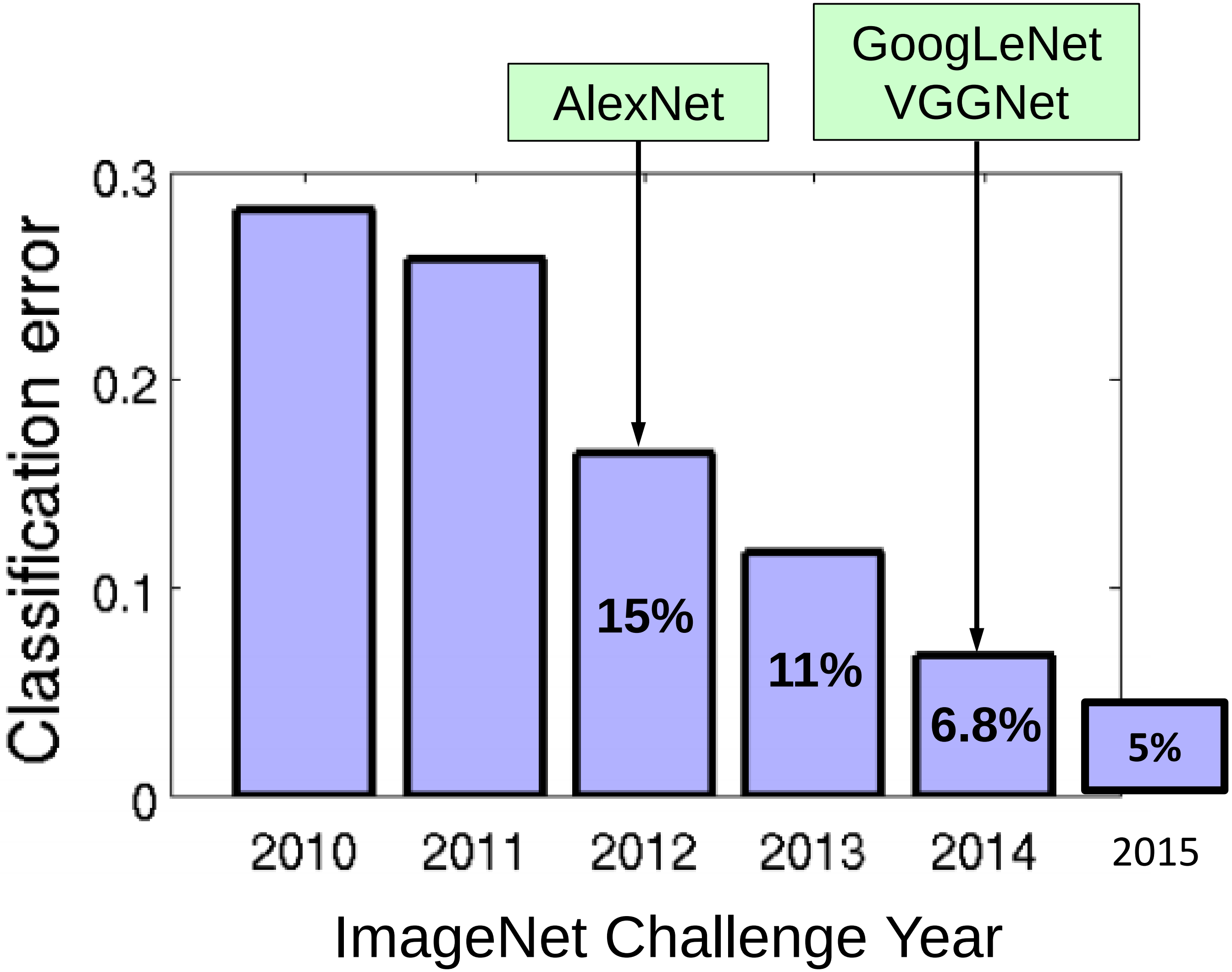
Aditya Khosla

Image by
kirkh.deviantart.com

Today's class

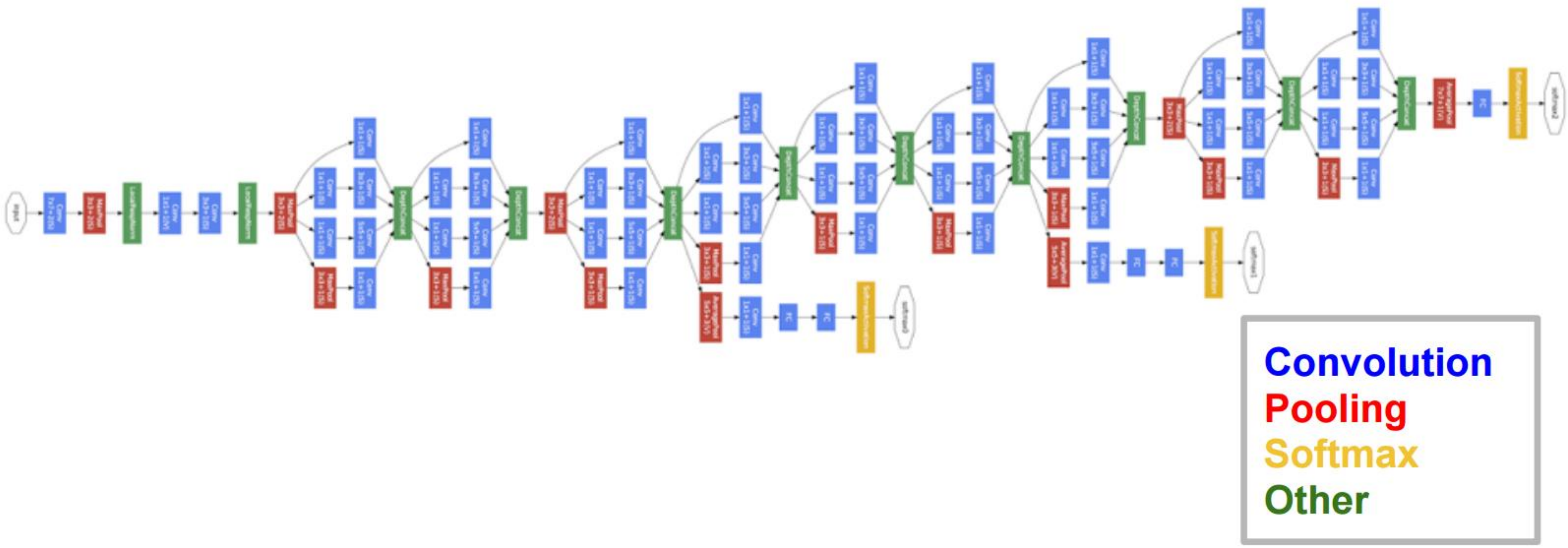
- Part 1: From state-of-the-art to state-of-the-artest
  - Fine-tuning
  - Data augmentation
- Part 2: Applications
  - Detection, segmentation, …
- Part 3: Learning sequences
  - RNNs/LSTMs

Object recognition



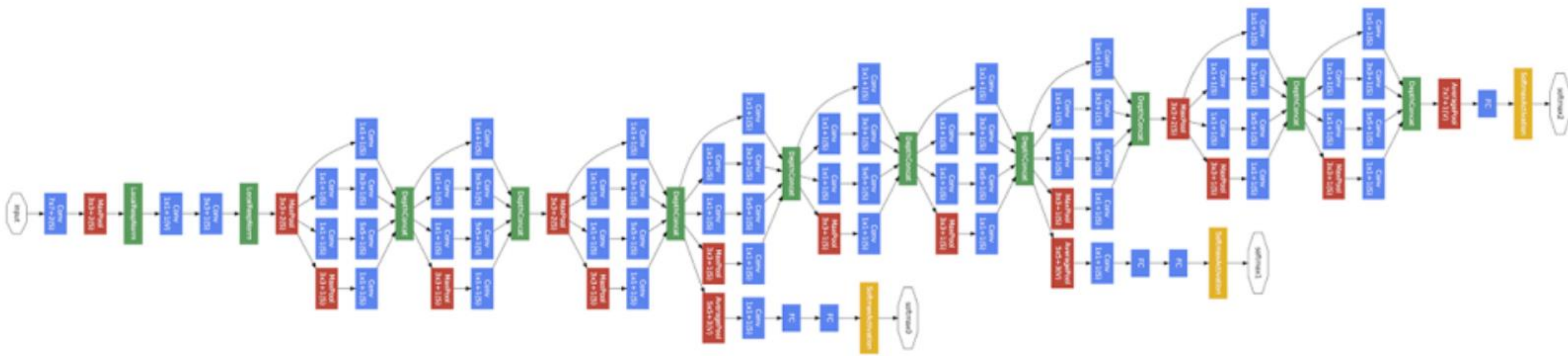

WE NEED TO GO
DEEPER

GoogLeNet



Credit: Szegedy et al

GoogLeNet vs AlexNet

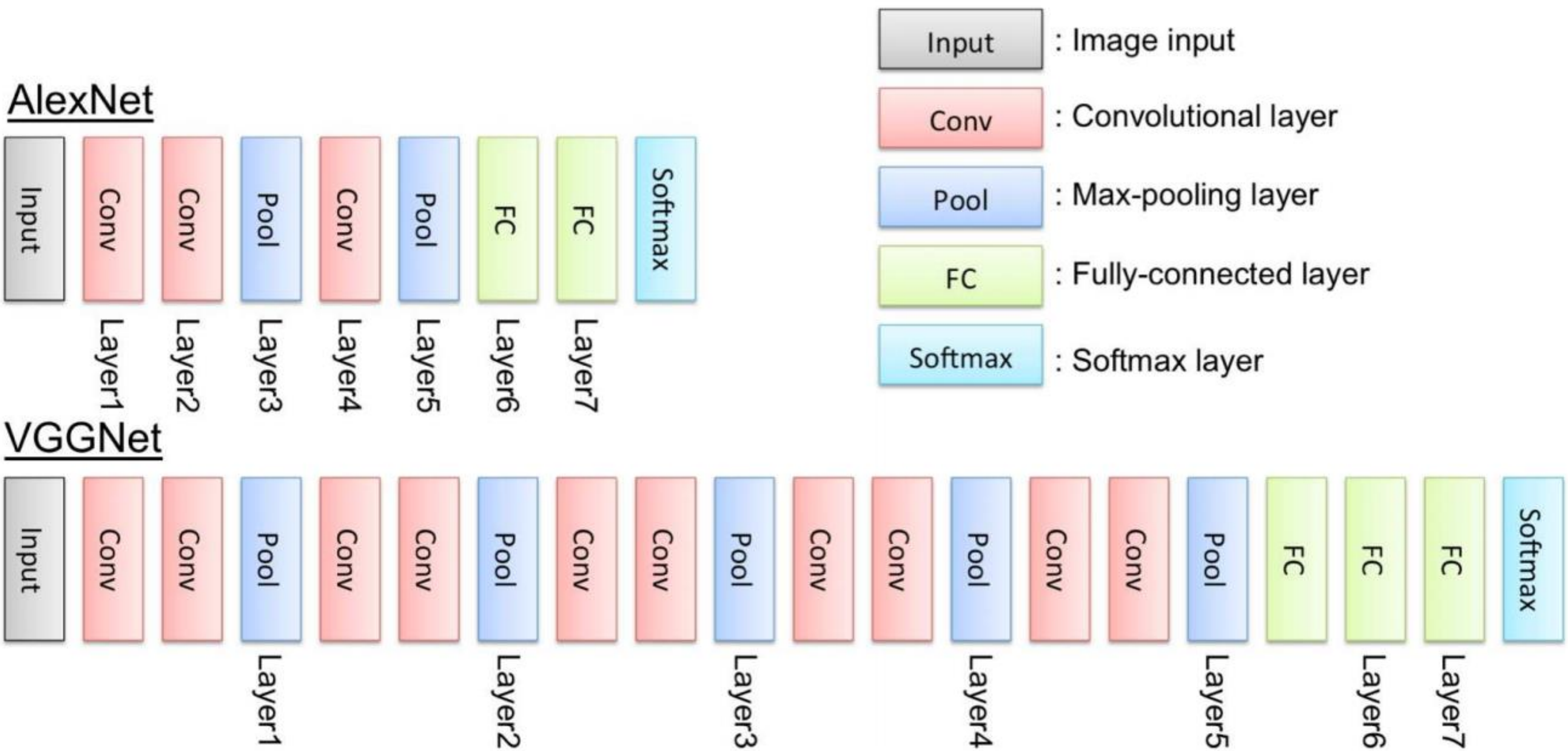


GoogLeNet

AlexNet

Convolution
Pooling
Softmax
Other

Credit: Szegedy et al


AlexNet
Input
Conv
Conv
Pool
Conv
Pool
FC
FC
Softmax
Layer1
Layer2
Layer3
Layer4
Layer5
Layer6
Layer7
VGGNet
Input
Conv
Conv
Pool
Conv
Conv
Pool
Conv
Conv
Pool
Conv
Conv
Pool
Conv
Conv
Pool
FC
FC
FC
Softmax
Layer1
Layer2
Layer3
Layer4
Layer5
Layer6
Layer7
Input : Image input
Conv : Convolutional layer
Pool : Max-pooling layer
FC : Fully-connected layer
Softmax : Softmax layer

GoogLeNet

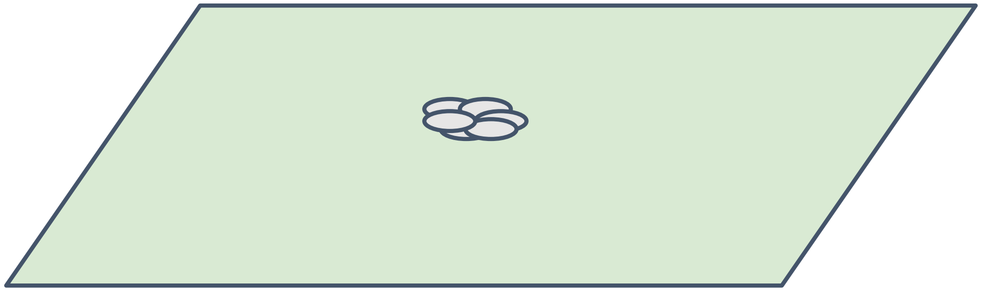
- Power and Memory use considerations are important for practical use.

- Image data is mostly sparse and clustered.

- Hebbian Principle:

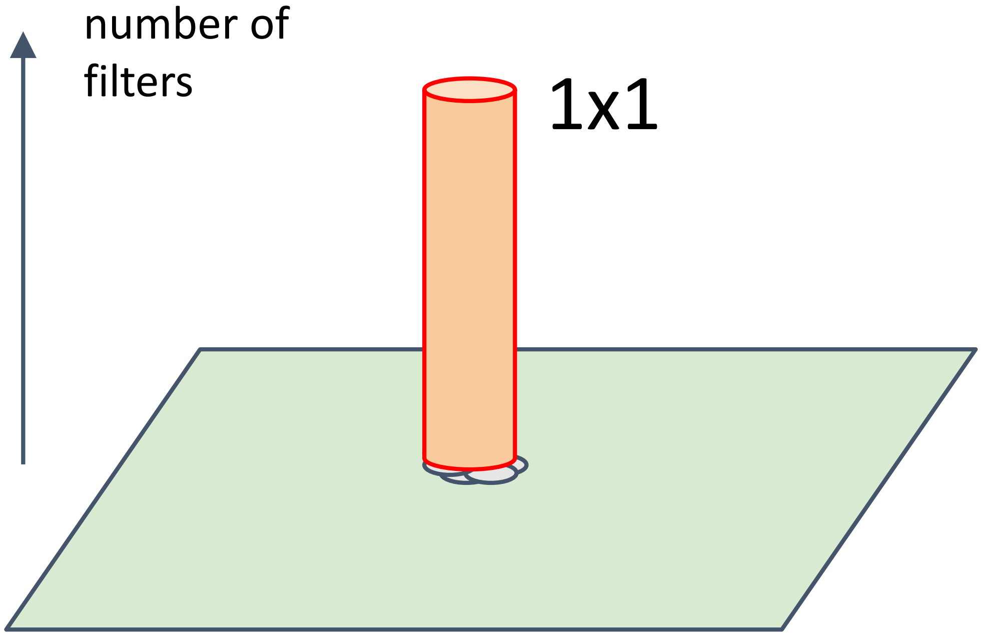
“Neurons that fire together, wire together”

In images, correlations tend to be local



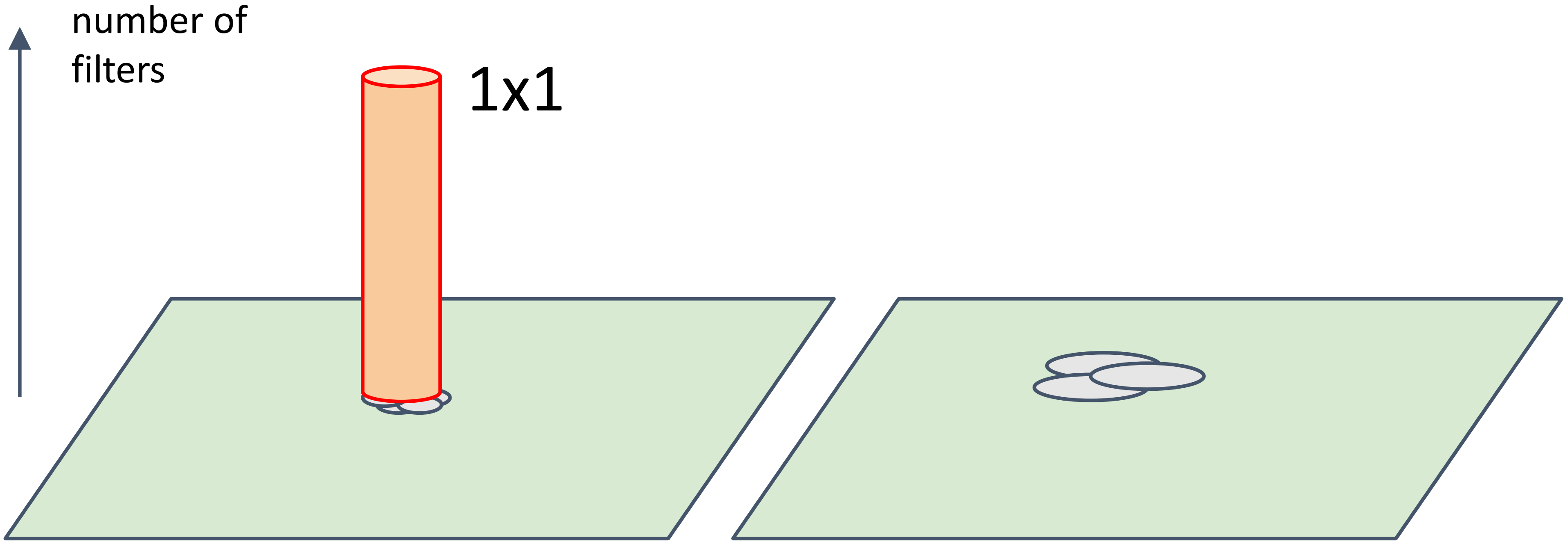
Credit: Szegedy et al

Cover very local clusters by 1x1 convolutions



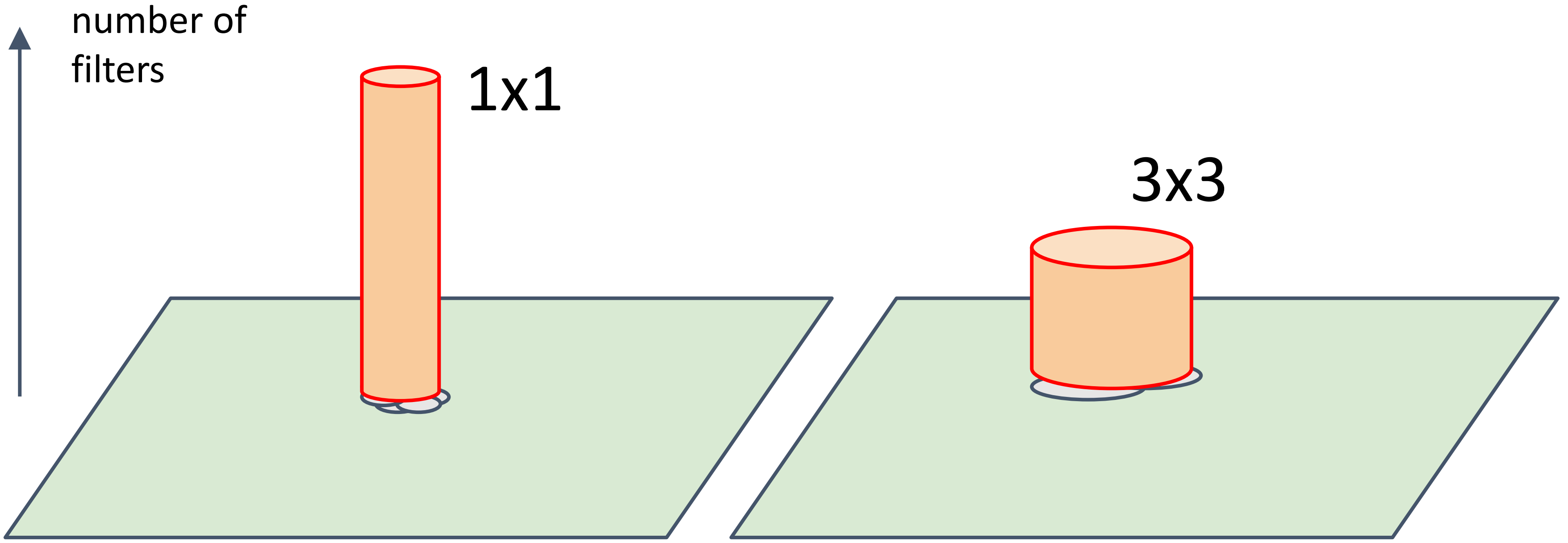
Credit: Szegedy et al

Less spread out correlations



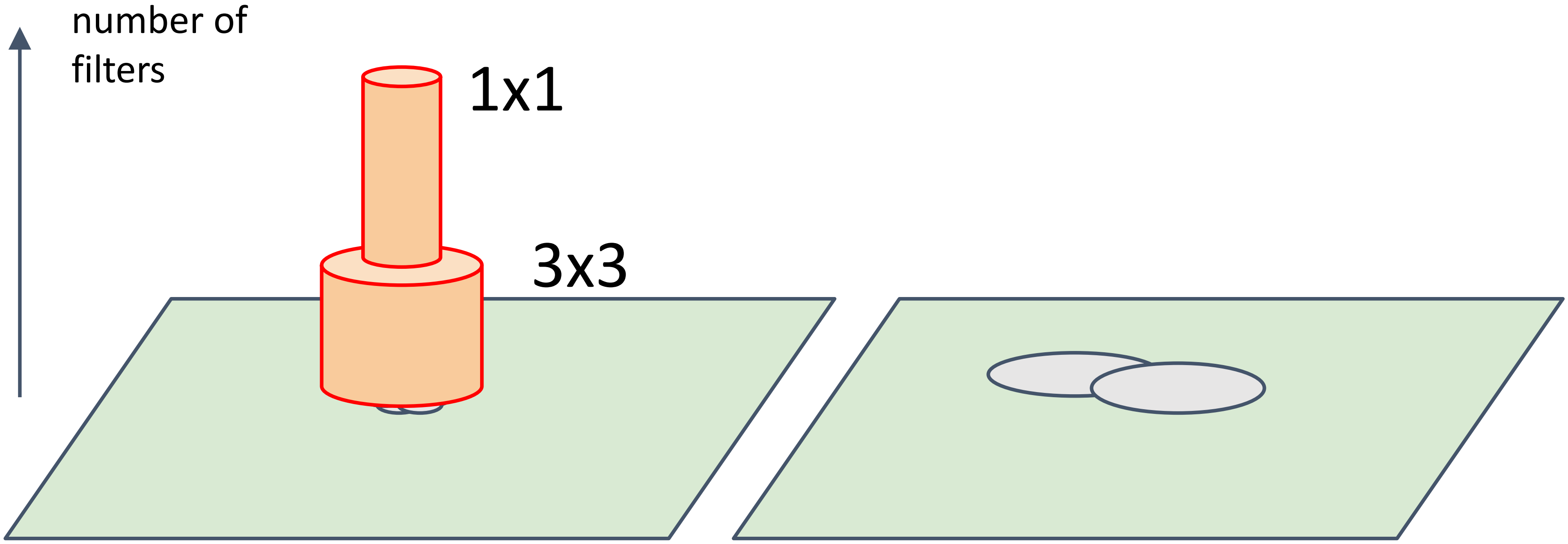
Credit: Szegedy et al

Cover more spread out clusters by 3x3 convolutions



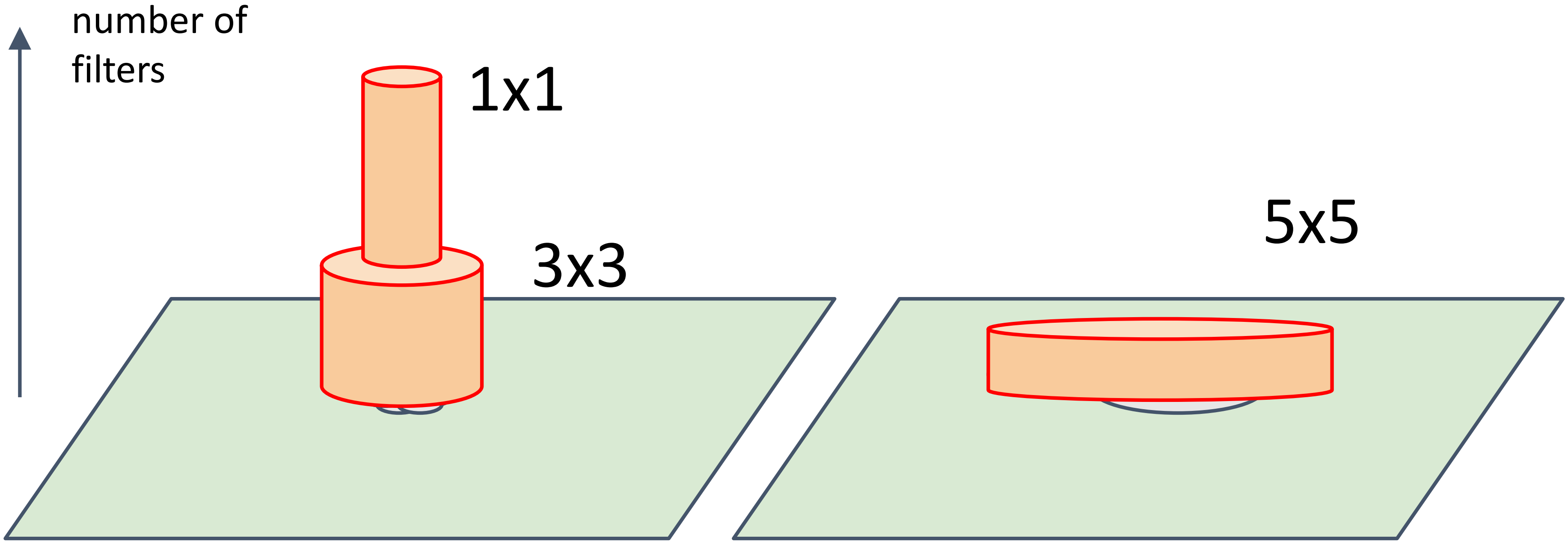
Credit: Szegedy et al

Cover more spread out clusters by 5x5 convolutions



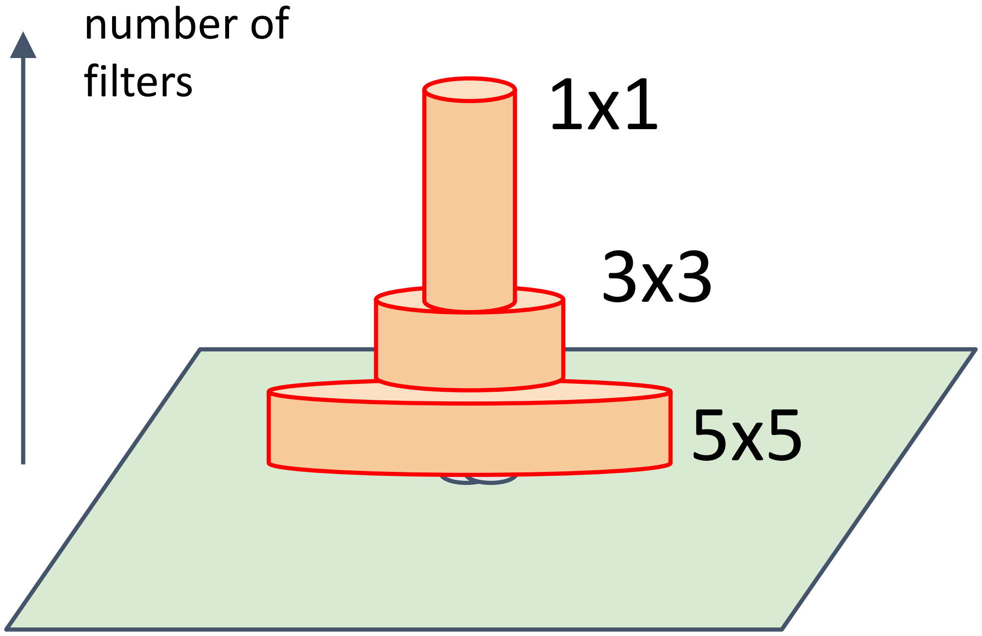
Credit: Szegedy et al

Cover more spread out clusters by 5x5 convolutions



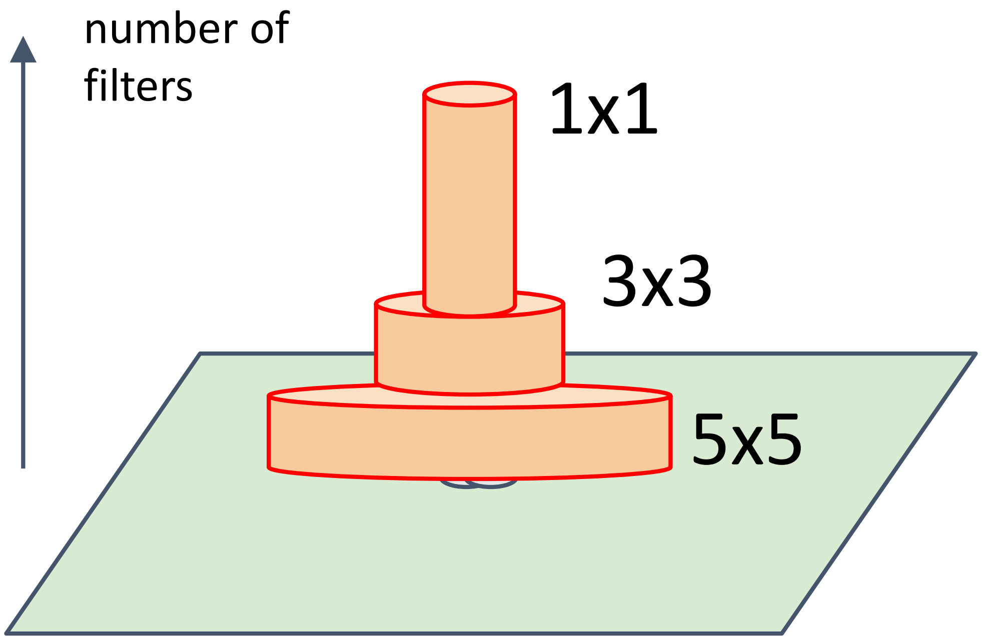
Credit: Szegedy et al

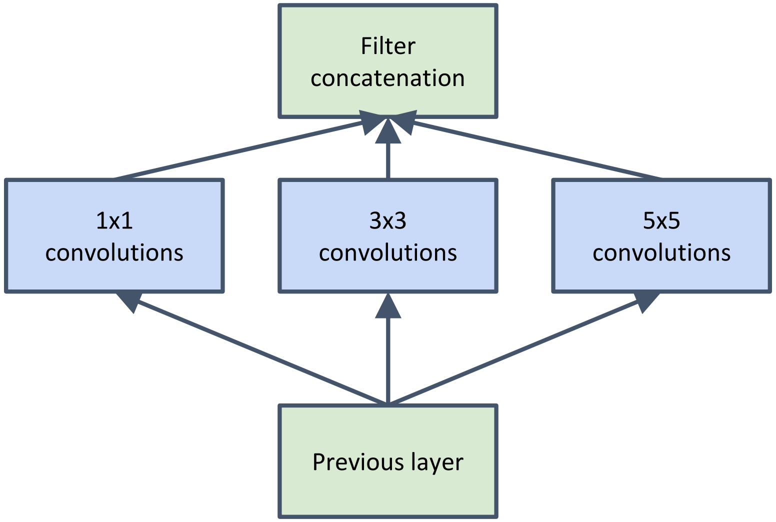
A heterogeneous set of convolutions



Credit: Szegedy et al

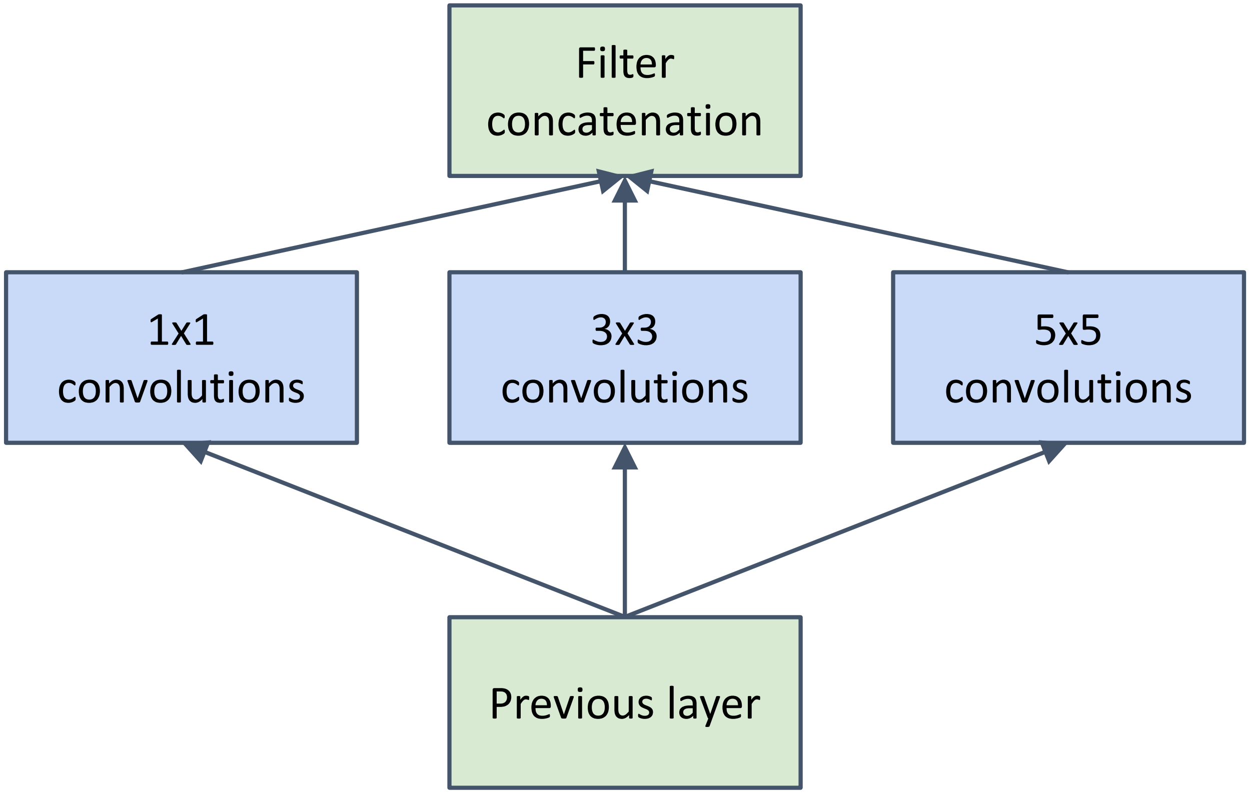
Schematic view (naive version)





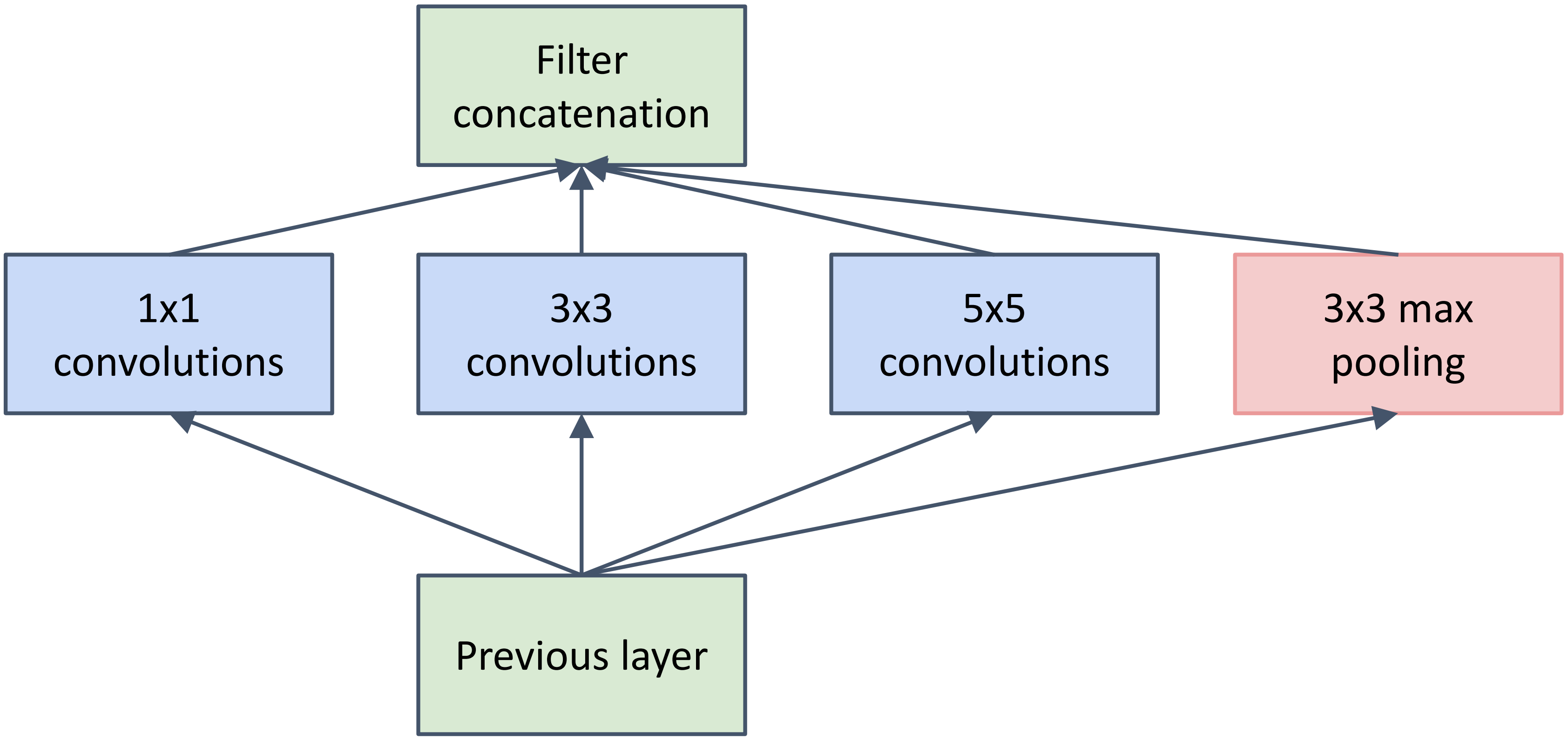
Credit: Szegedy et al

Naive idea



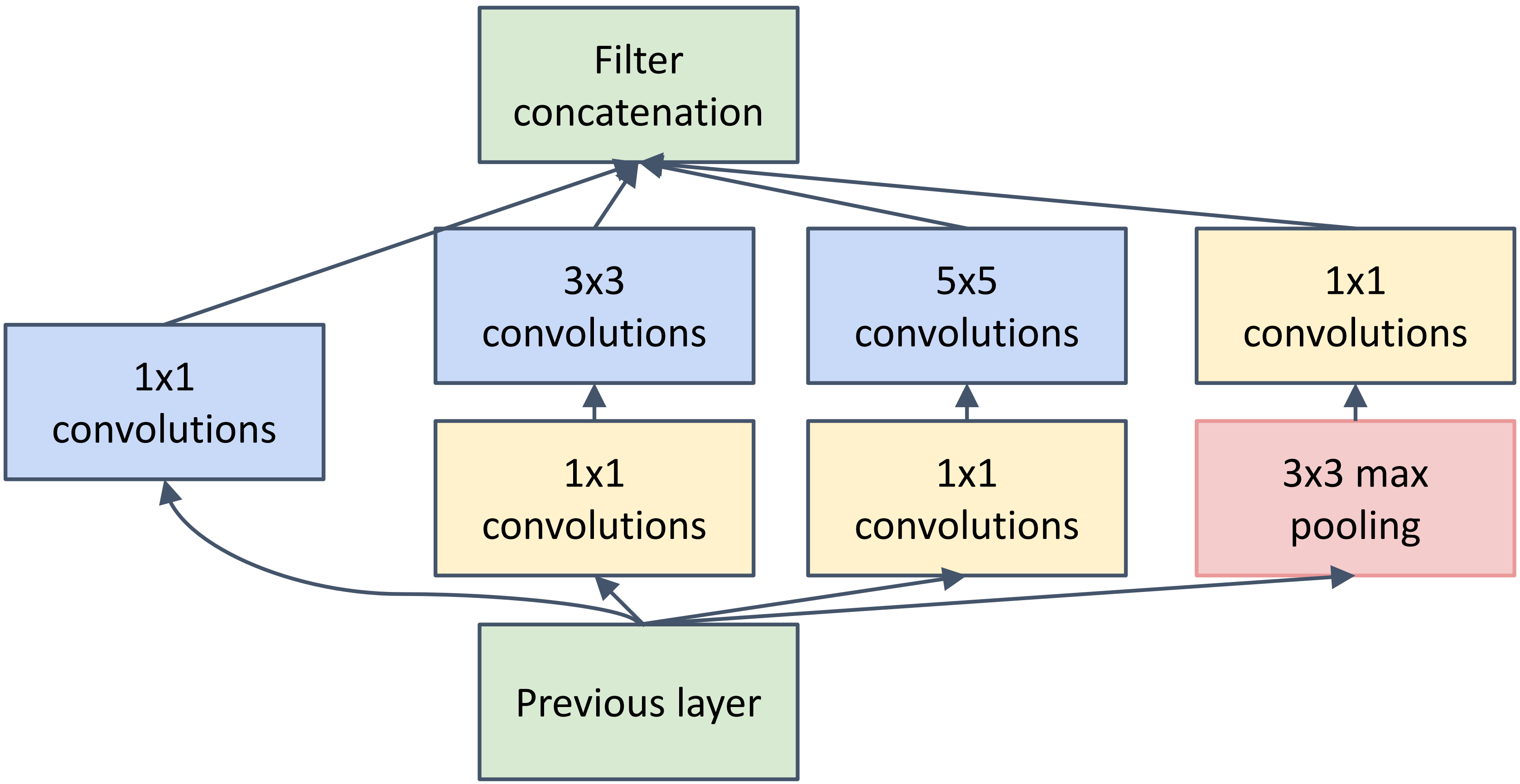
Credit: Szegedy et al

Naive idea (**does not work!**)



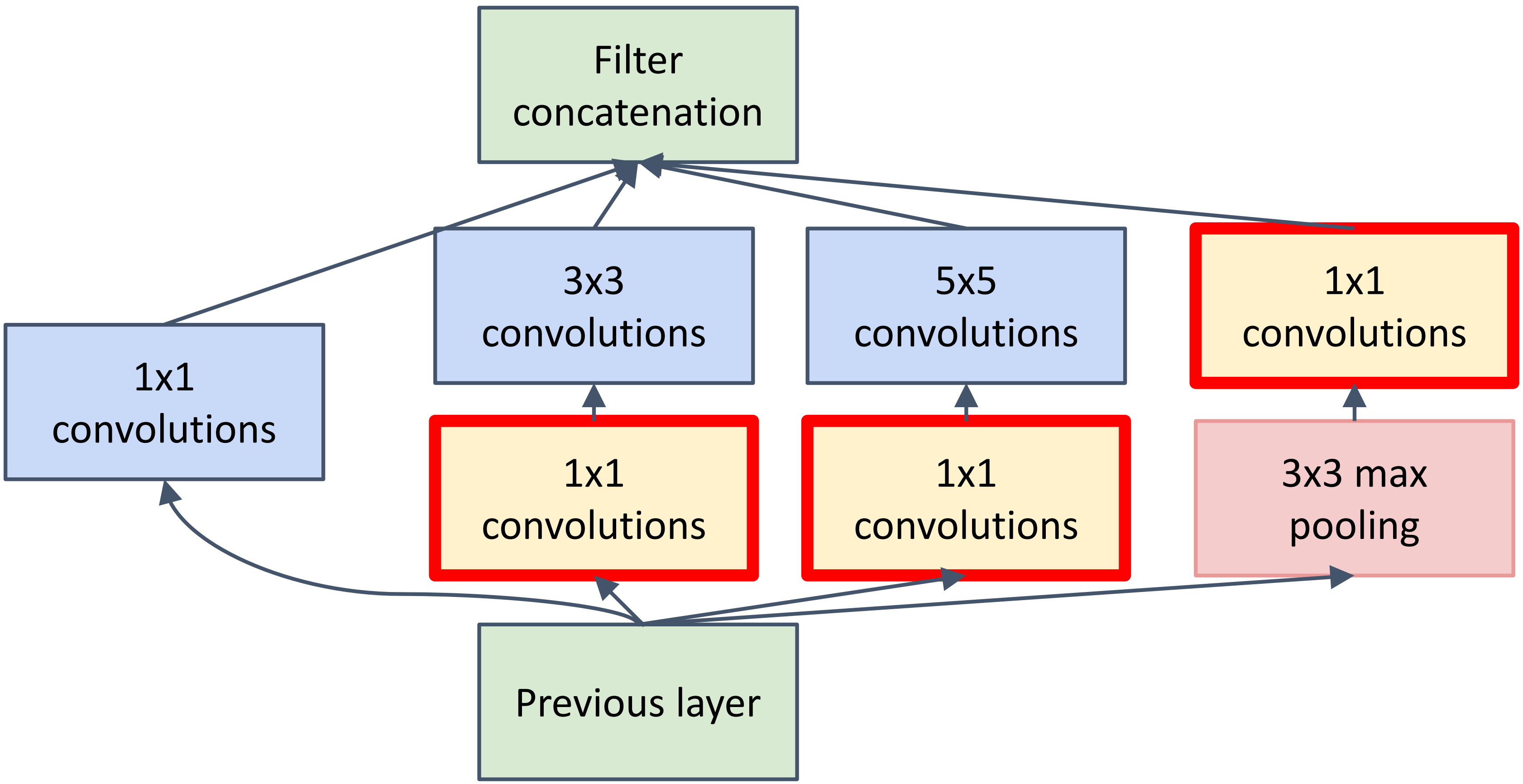
Credit: Szegedy et al

Inception module



Credit: Szegedy et al

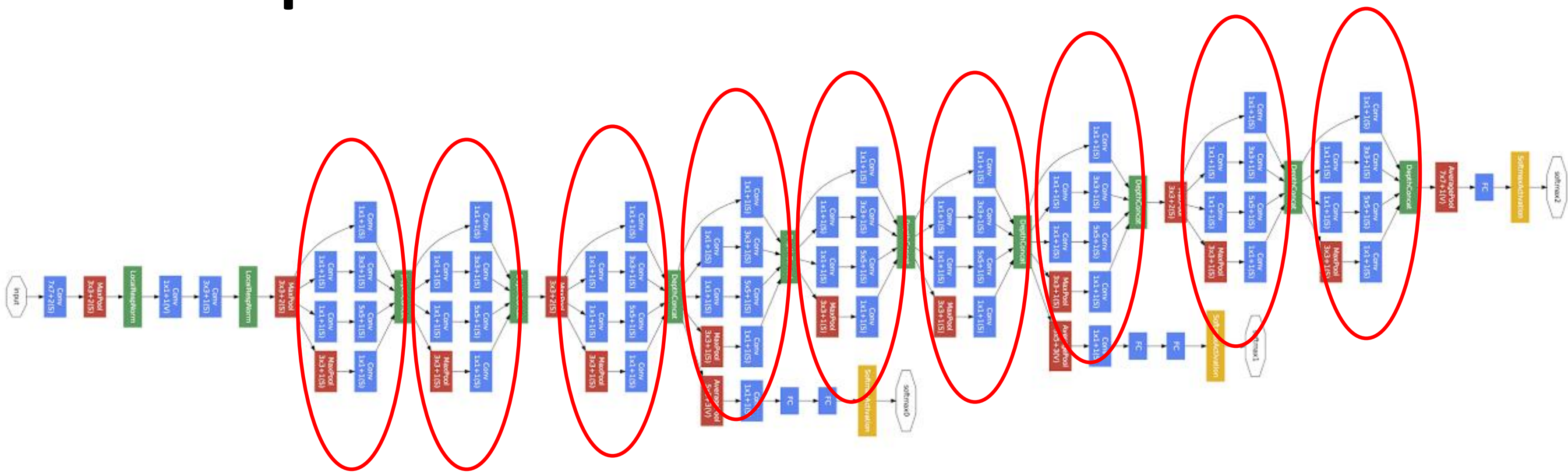
Inception module



Dimensionality reduction!

Credit: Szegedy et al

Inception

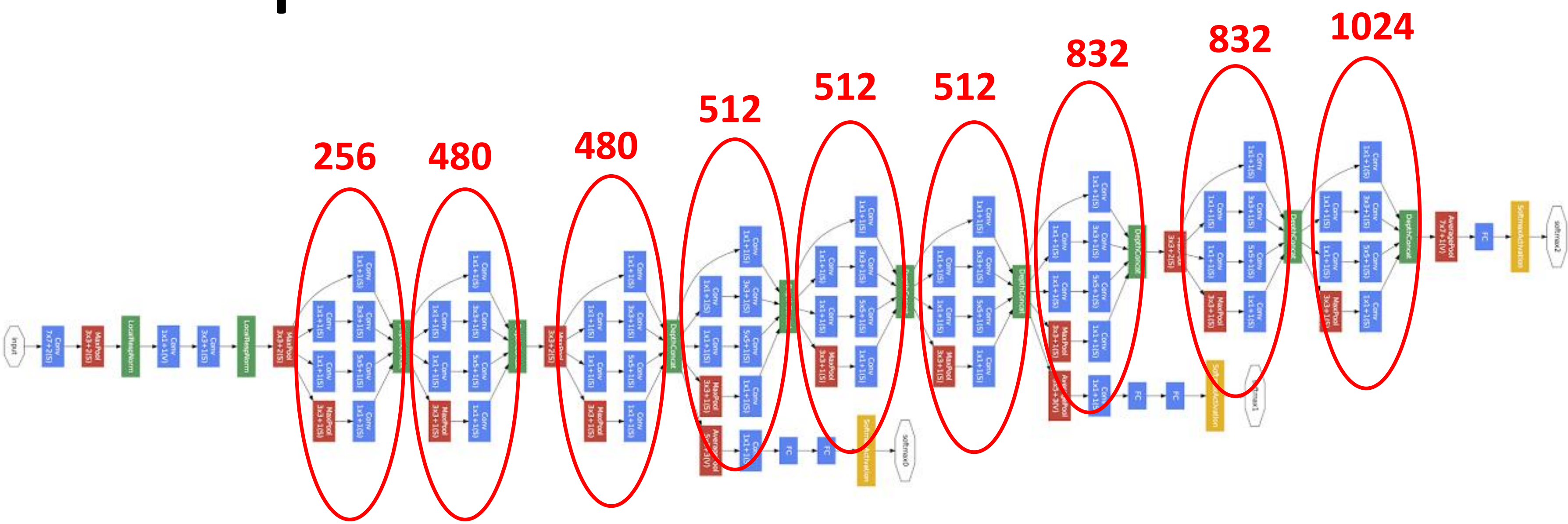


9 **Inception** modules

Convolution
Pooling
Softmax
Other

Credit: Szegedy et al

Inception



Width of inception modules ranges from 256 filters (in early modules) to 1024 in top inception modules.

Can remove fully connected layers on top completely

Number of parameters is reduced to 5 million

Computional cost is increased by less than 2X compared to AlexNet. (<1.5Bn operations/evaluation)

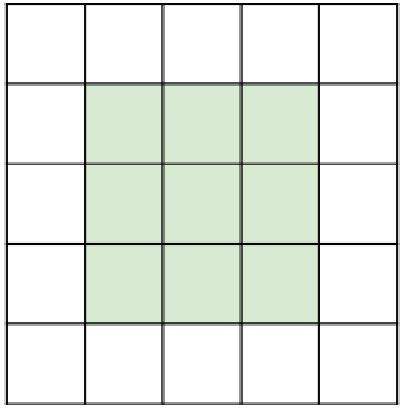
Credit: Szegedy et al

The power of small filters

(and stride 1)

Suppose we stack two CONV layers with receptive field size 3x3
=> Each neuron in 1st CONV sees a 3x3 region of input.

1st CONV neuron view of the input:



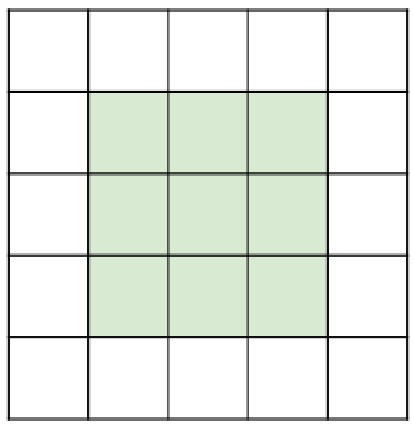
Credit: Fei-Fei

The power of small filters

Suppose we stack two CONV layers with receptive field size 3x3
=> Each neuron in 1st CONV sees a 3x3 region of input.

Q: What region of input does each neuron in 2nd CONV see?

2nd CONV neuron
view of 1st conv:



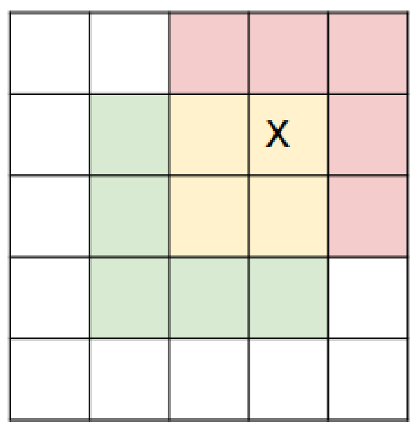
Credit: Fei-Fei

The power of small filters

Suppose we stack two CONV layers with receptive field size 3x3
=> Each neuron in 1st CONV sees a 3x3 region of input.

Q: What region of input does each neuron in 2nd CONV see?

2nd CONV neuron
view of input:



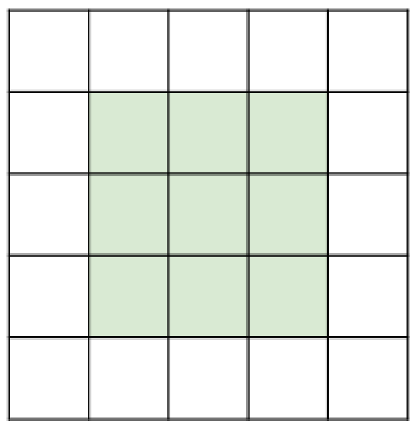
Answer: [5x5]

Credit: Fei-Fei

The power of small filters

Suppose we stack **three** CONV layers with receptive field size 3x3
Q: What region of input does each neuron in 3rd CONV see?

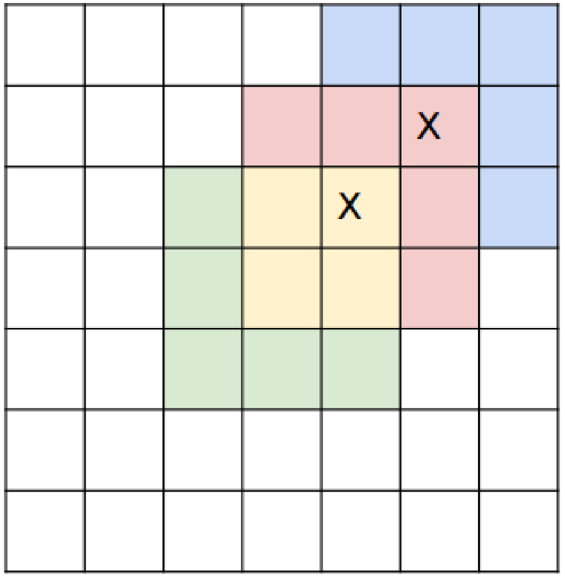
3rd CONV neuron
view of 2nd CONV:



Credit: Fei-Fei

The power of small filters

Suppose we stack **three** CONV layers with receptive field size 3x3
Q: What region of input does each neuron in 3rd CONV see?



Answer: [7x7]

Credit: Fei-Fei

The power of small filters

Suppose input has depth C & we want output depth C as well

1x CONV with 7x7 filters

Number of weights:

C*(7*7*C)
= **49 C^2**

3x CONV with 3x3 filters

Number of weights:

C*(3*3*C) + C*(3*3*C) + C*(3*3*C)
= 3 * 9 * C^2
= **27 C^2**

Credit: Fei-Fei

VGGnet

[Very Deep Convolutional Networks for Large-Scale Image Recognition, Simonyan et al., 2014]

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					

Table 2: **Number of parameters** (in millions).

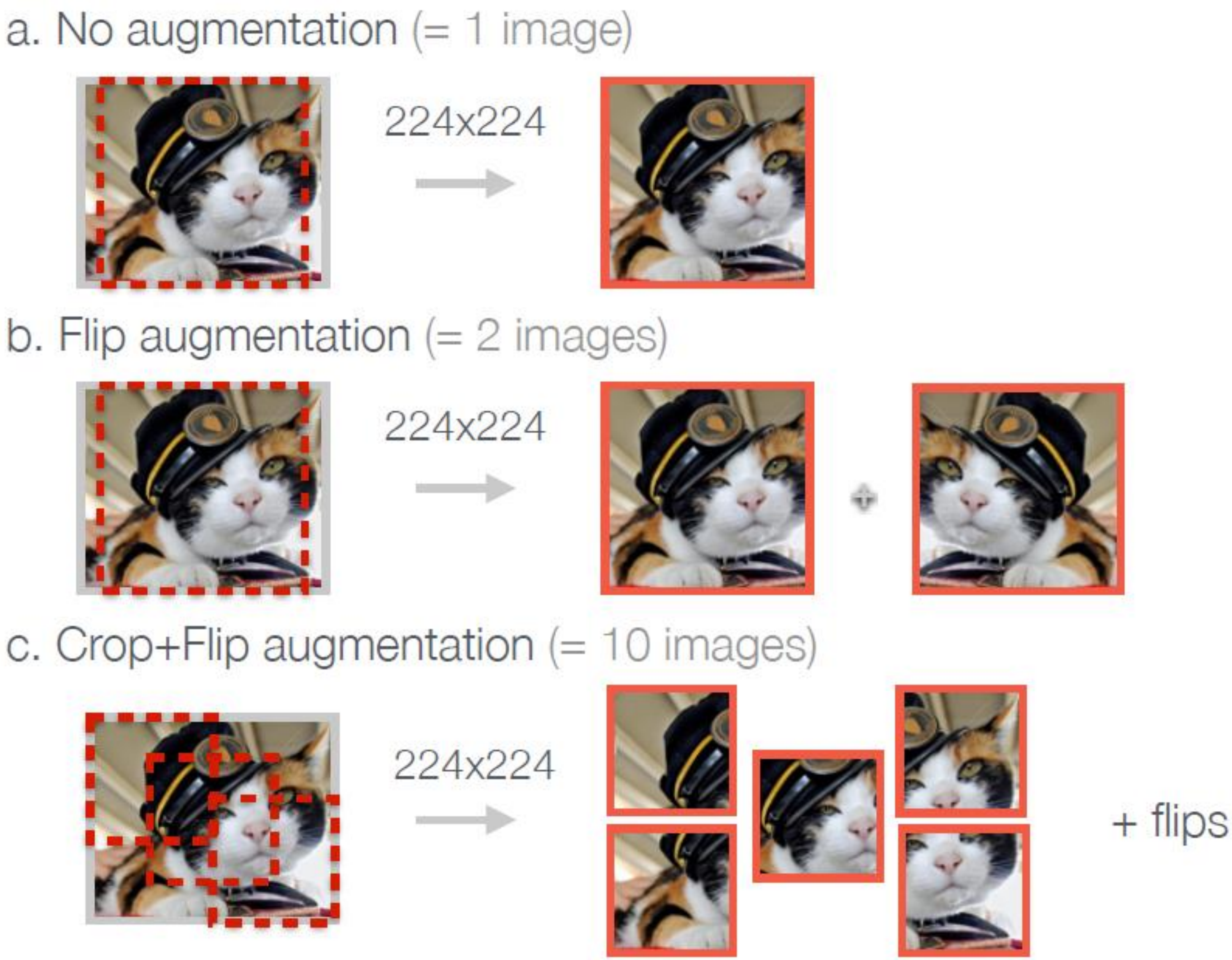
Network	A,A-LRN	B	C	D	E
Number of parameters	133	133	134	138	144

ConvNet config. (Table 1)	smallest image side		top-1 val. error (%)	top-5 val. error (%)
	train (S)	test (Q)		
A	256	256	29.6	10.4
A-LRN	256	256	29.7	10.5
B	256	256	28.7	9.9
C	256	256	28.1	9.4
	384	384	28.1	9.3
	[256;512]	384	27.3	8.8
D	256	256	27.0	8.8
	384	384	26.8	8.7
	[256;512]	384	25.6	8.1
E	256	256	27.3	9.0
	384	384	26.9	8.7
	[256;512]	384	**25.5**	**8.0**

=> Evidence that using 3x3 instead of 1x1 works better

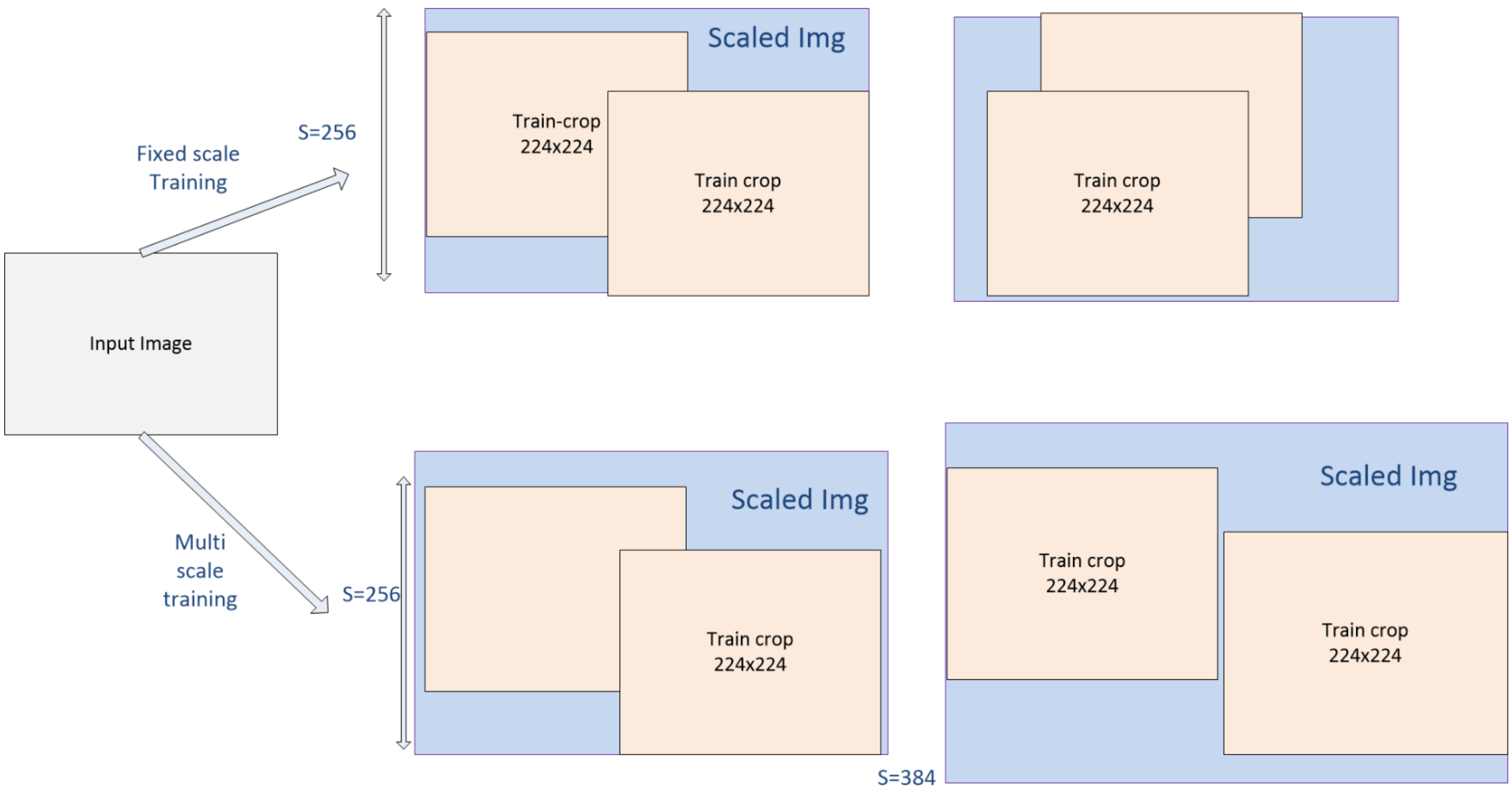
Credit: Fei-Fei

Data augmentation



Data augmentation

- For both training and testing

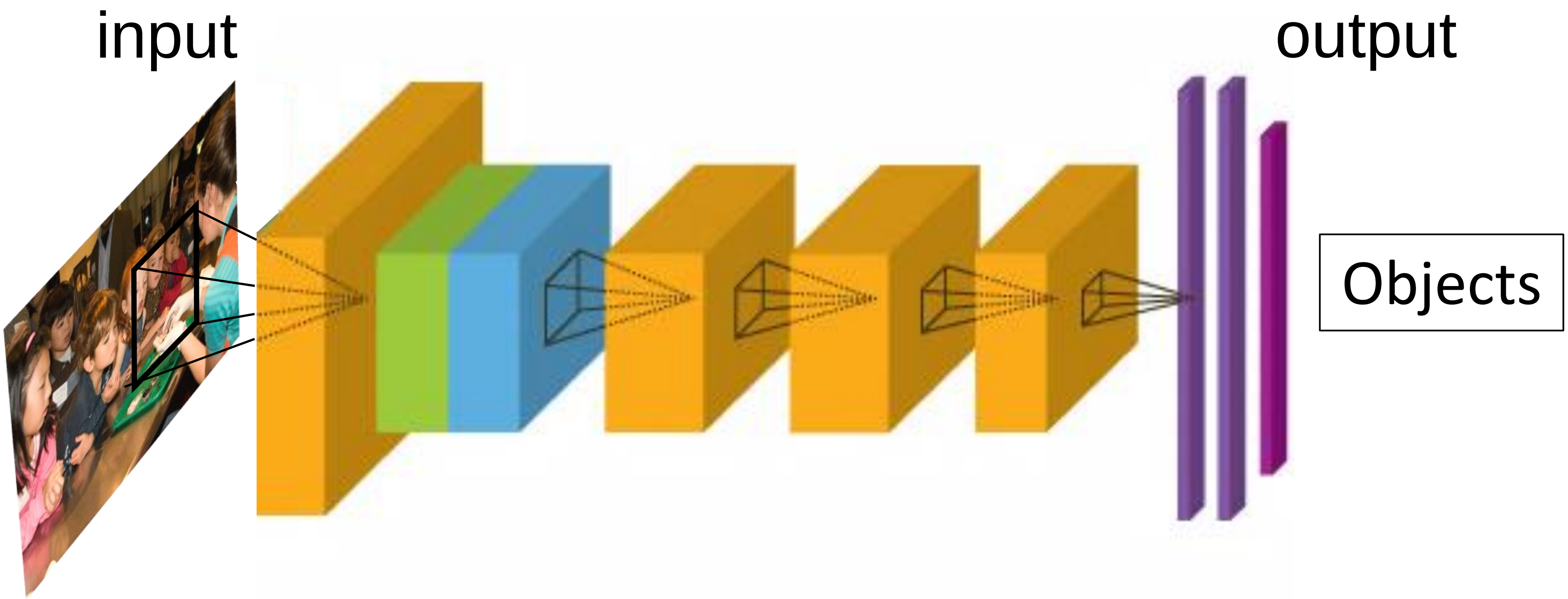


Classification results on ImageNet 2012

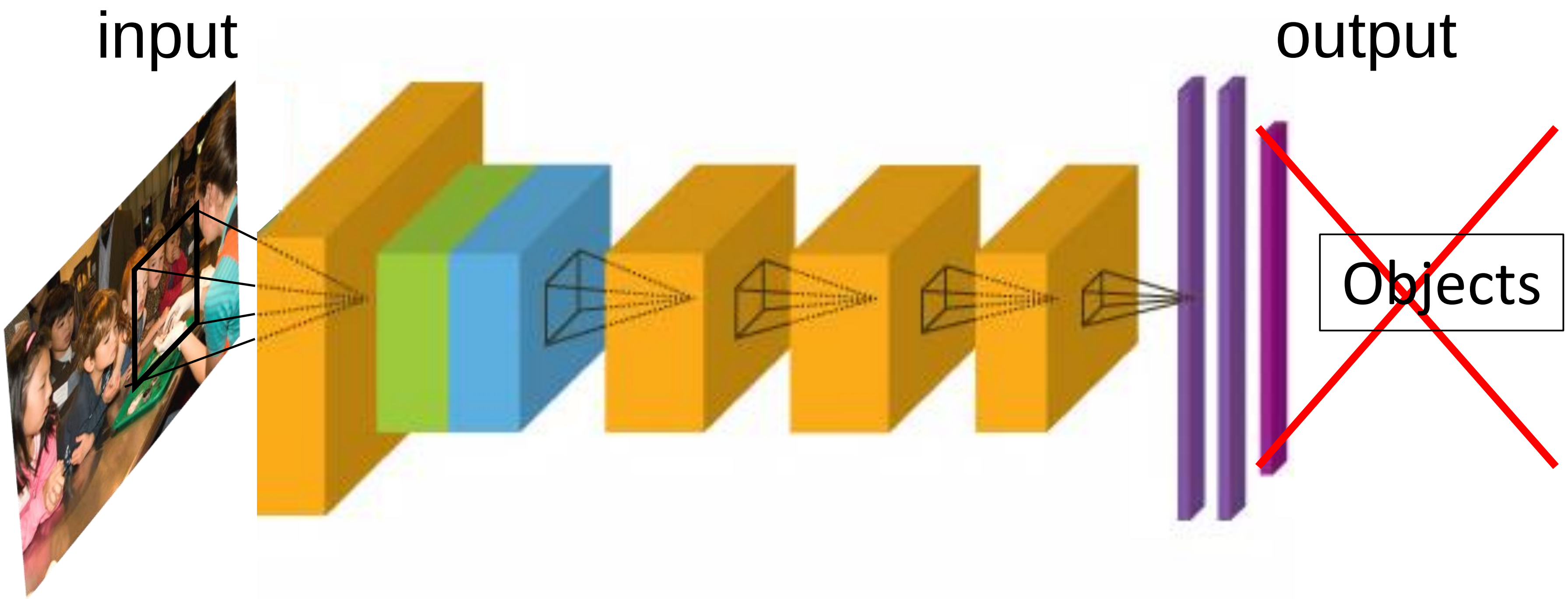
Number of Models	Number of Crops	Computational Cost	Top-5 Error	Compared to Base
1	1 (center crop)	1x	10.07%	-
1	10*	10x	9.15%	-0.92%
1	144 (Our approach)	144x	7.89%	-2.18%
7	1 (center crop)	7x	8.09%	-1.98%
7	10*	70x	7.62%	-2.45%
7	144 (Our approach)	1008x	6.67%	-3.41%

*Cropping by [Krizhevsky et al 2014]

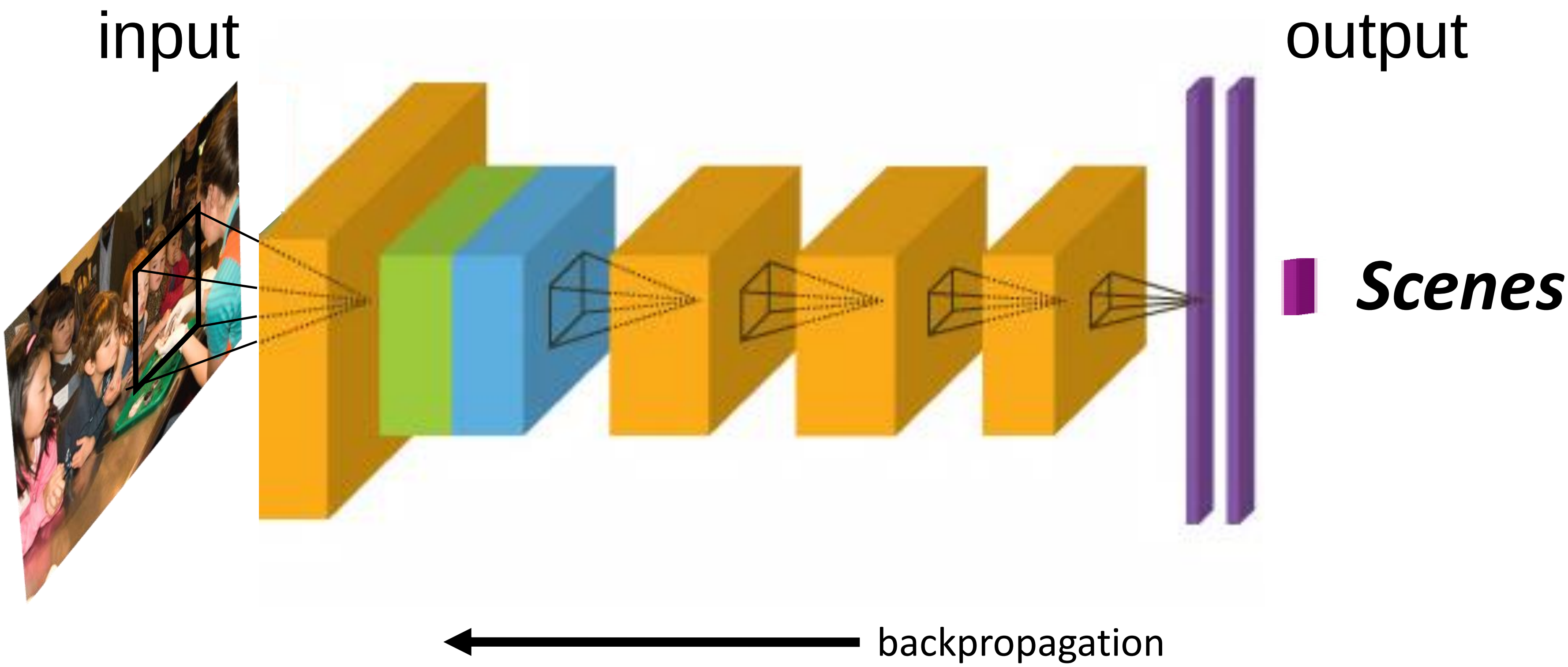
Fine-tuning



Fine-tuning



Fine-tuning



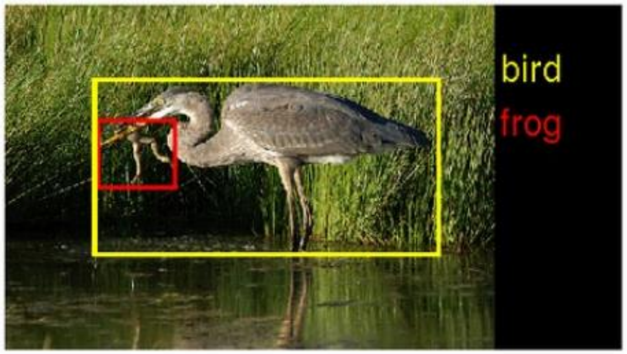
Visual Classification

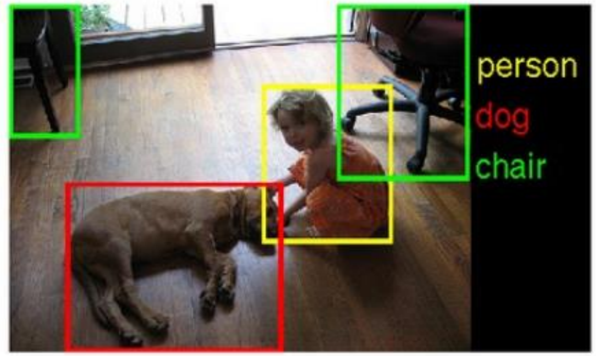
Results of MIT 67 Scene Classification

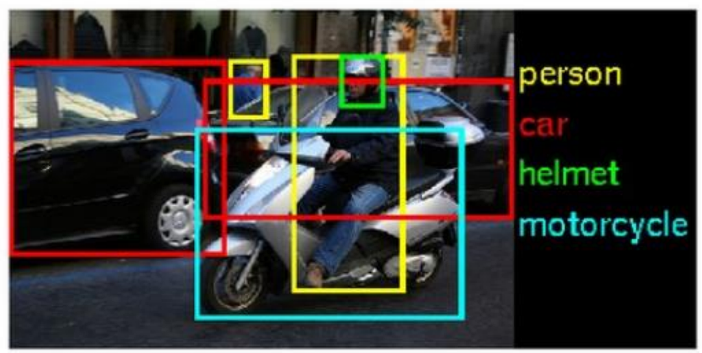
Method	mean Accuracy
ROI + Gist[36]	26.1
DPM[30]	30.4
Object Bank[24]	37.6
RBow[31]	37.9
BoP[21]	46.1
miSVM[25]	46.4
D-Parts[40]	51.4
IFV[21]	60.8
MLrep[9]	64.0
CNN-SVM	58.4
CNNaug-SVM	69.0
CNN(AlexConvNet)+multiscale pooling [16]	68.9

Using a CNN off-the-shelf representation with linear SVMs training significantly outperforms a majority of the baselines.

Detection







Model must output:

A set of detections

Each detection has:

- confidence
- class (integer)
- x1,y1,x2,y2 bounding box coordinates

Credit: Fei-Fei

R-CNN

Rich feature hierarchies for accurate object detection and semantic segmentation
[Ross Girshick, Jeff Donahue, Trevor Darrell, Jitendra Malik]

Idea: *Turn a Detection Problem into an Image Classification problem (but over image regions).*

Content of every labeled bounding box for is a positive example for a class.

Every other bounding box in the image is a special **negative class**.

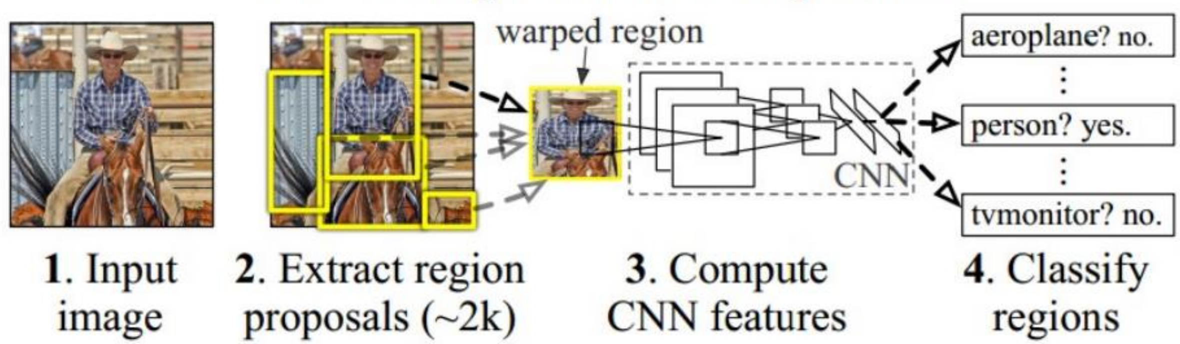
Credit: Fei-Fei

R-CNN

Rich feature hierarchies for accurate object detection and semantic segmentation
[Ross Girshick, Jeff Donahue, Trevor Darrell, Jitendra Malik]

***Idea**: Turn a Detection Problem into an Image Classification problem (but over image regions).*

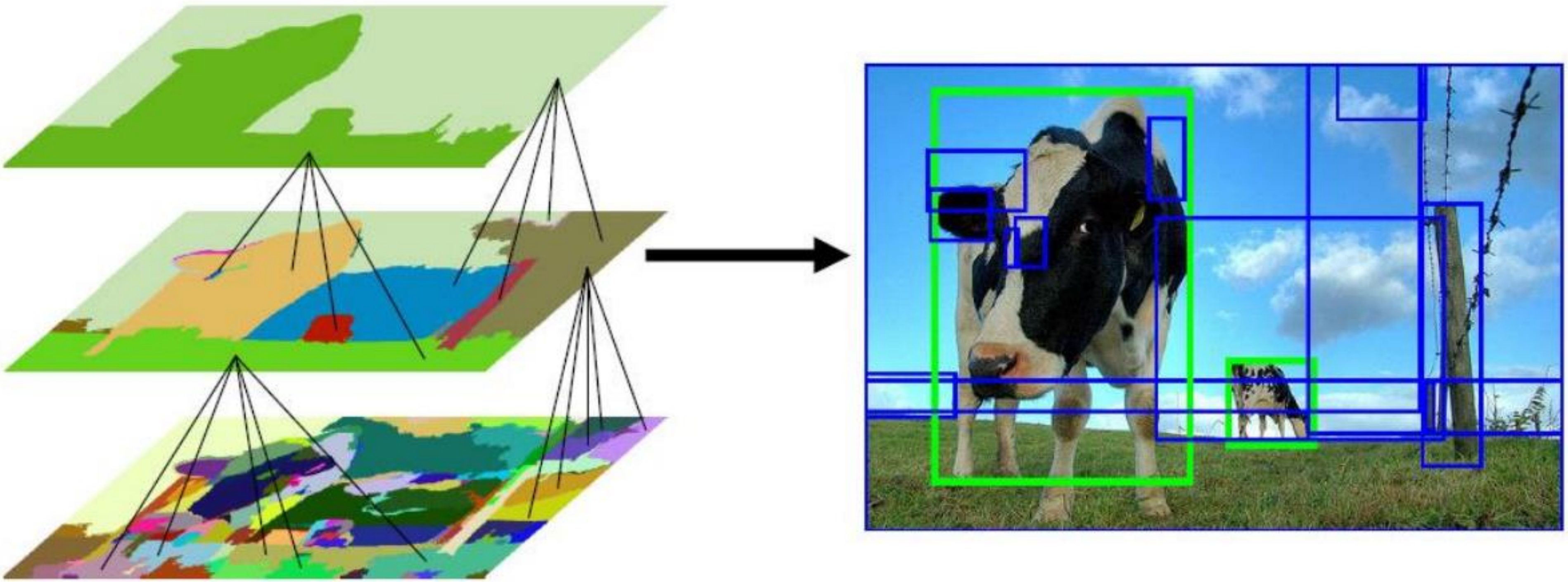
R-CNN: ***Regions with CNN features***



Credit: Fei-Fei

Selective Search for Object Recognition

[J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, A. W. M. Smeulders]



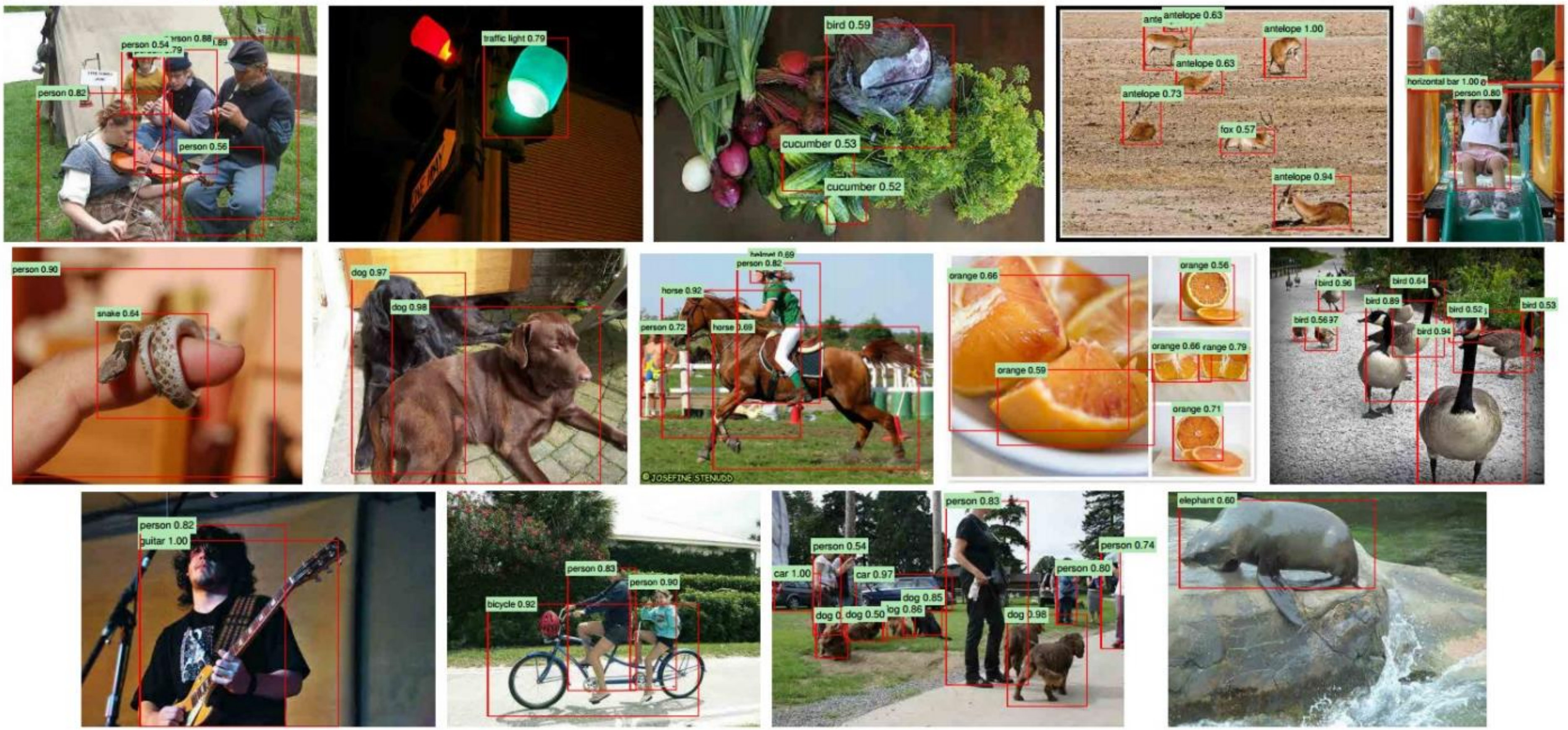
Gives on average ~2,000 candidate region proposals per image.

(This paradigm currently outperform the "sliding window" approach)

Credit: Fei-Fei

R-CNN Results

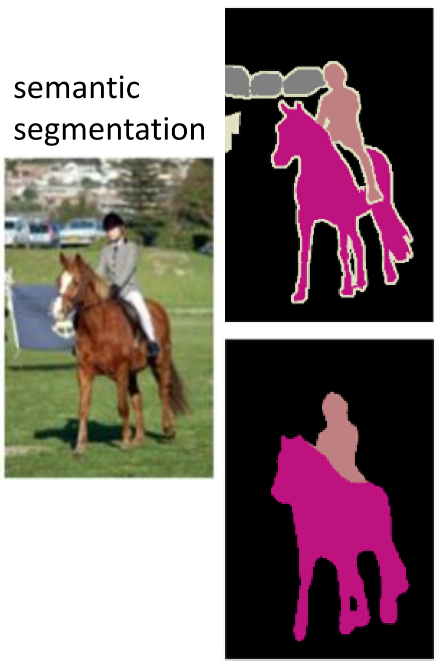
Rich feature hierarchies for accurate object detection and semantic segmentation
[Ross Girshick, Jeff Donahue, Trevor Darrell, Jitendra Malik]



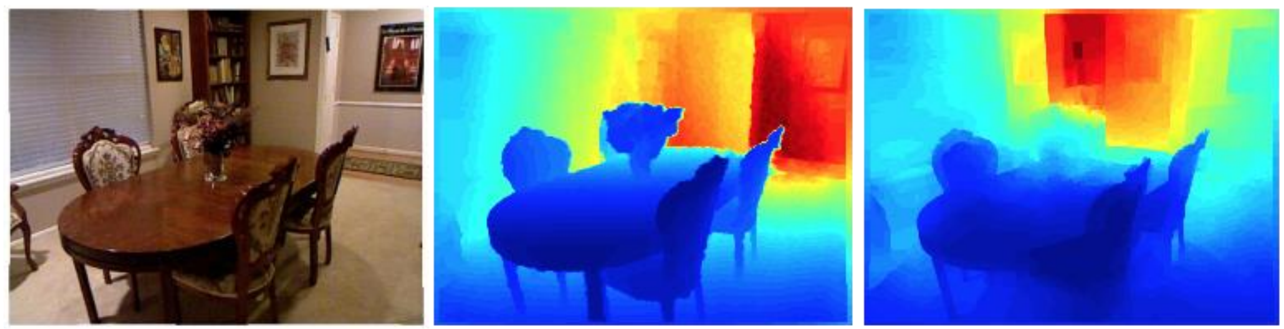
Credit: Fei-Fei

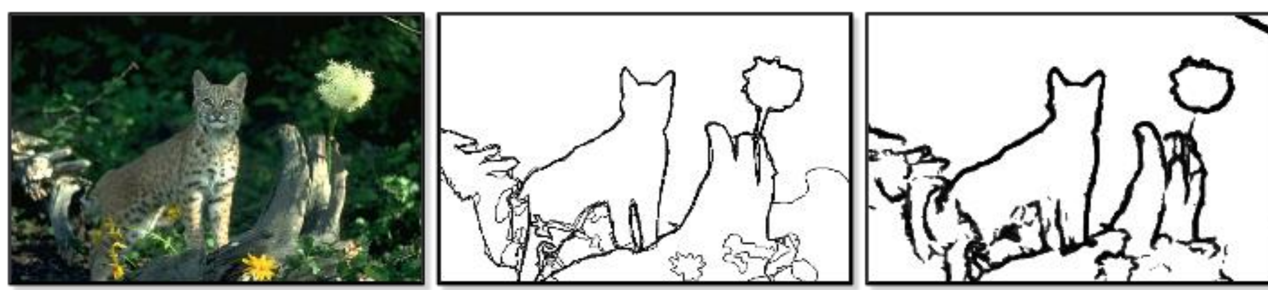
Pixels in, pixels out

semantic
segmentation



monocular depth estimation (Liu et al. 2015)

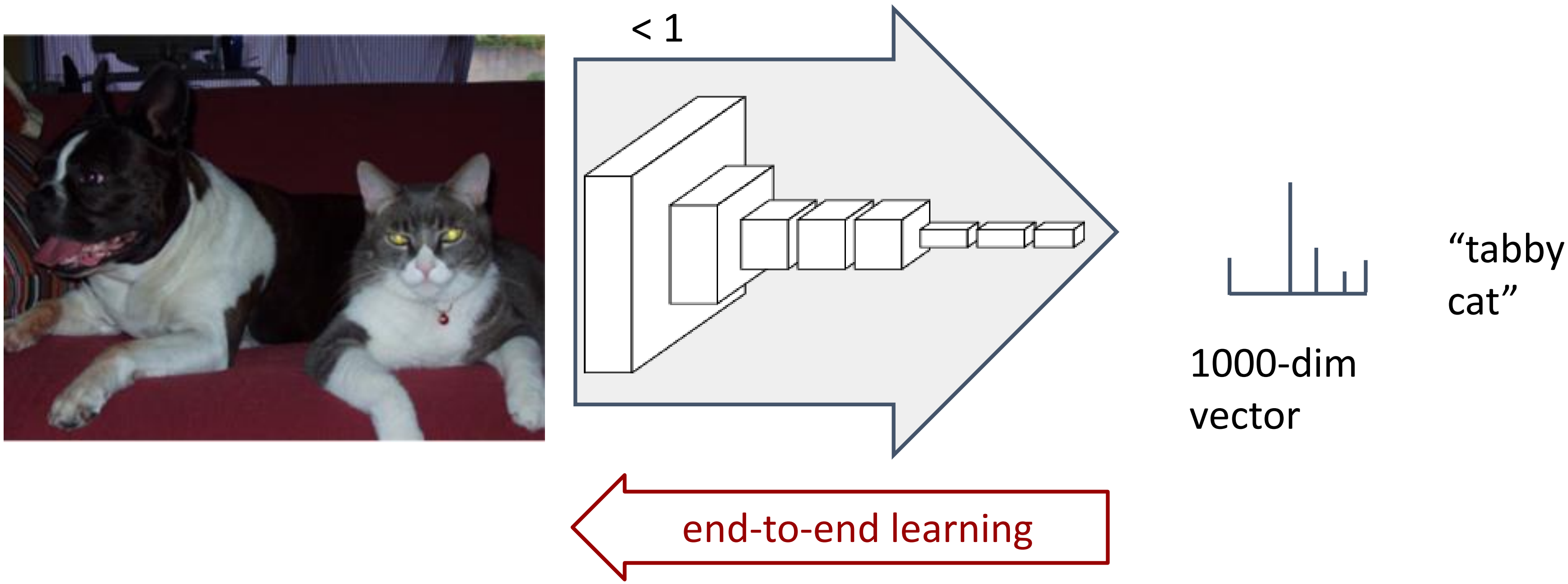




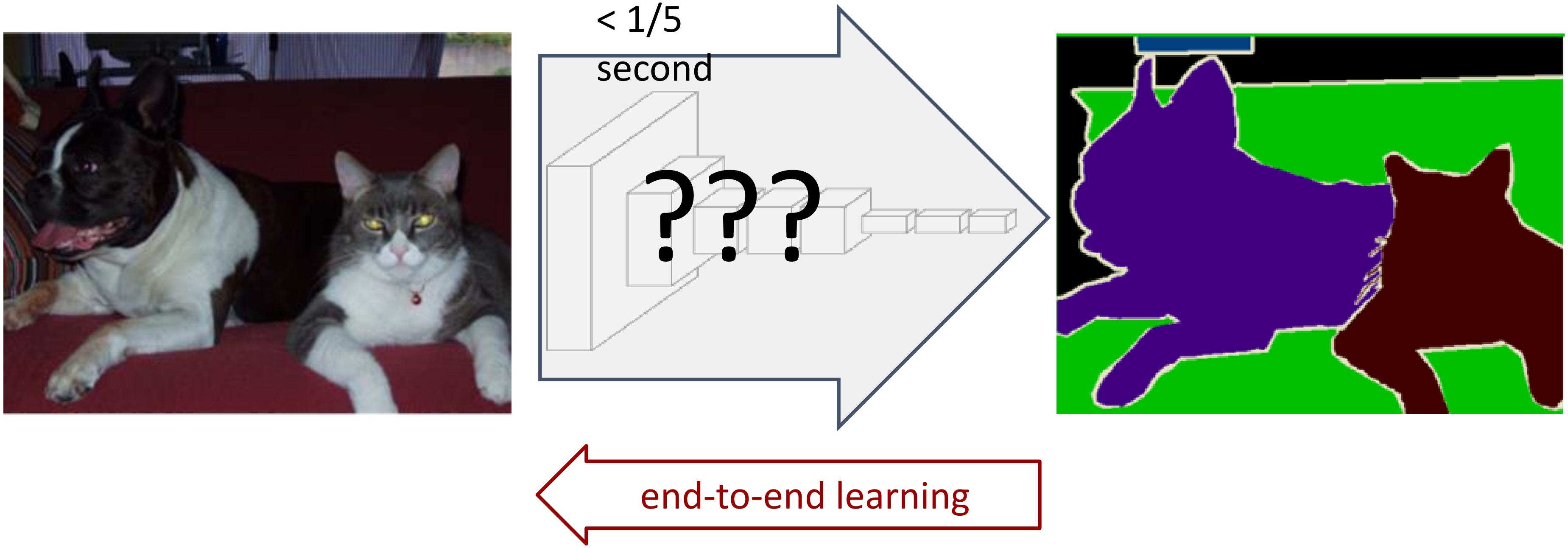
boundary prediction (Xie & Tu 2015)

Credit: Long et al

ConvNets perform Classification

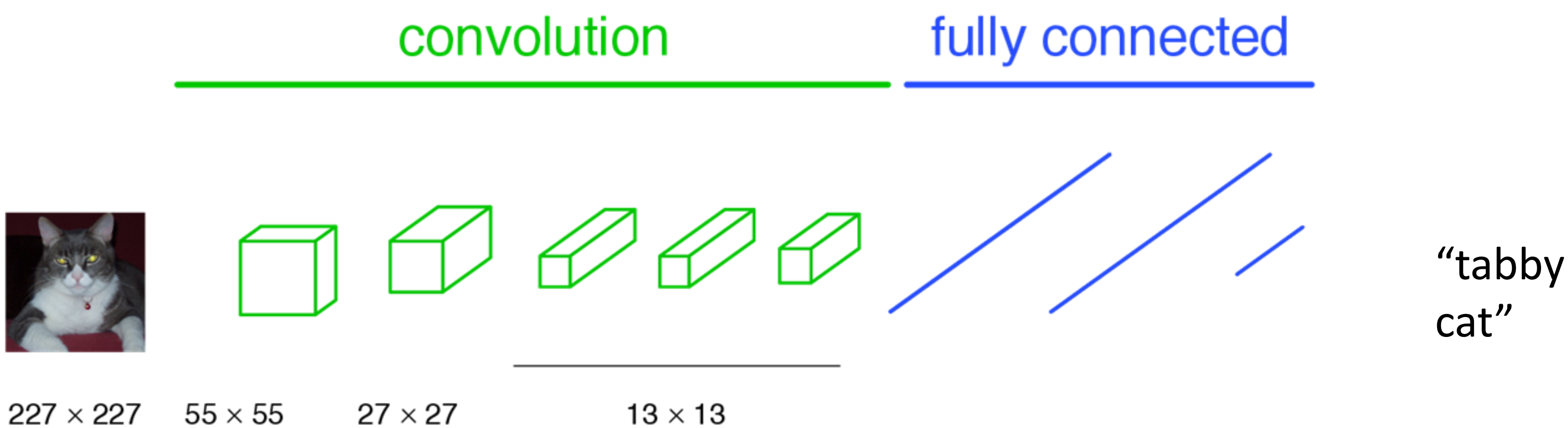


Credit: Long et al



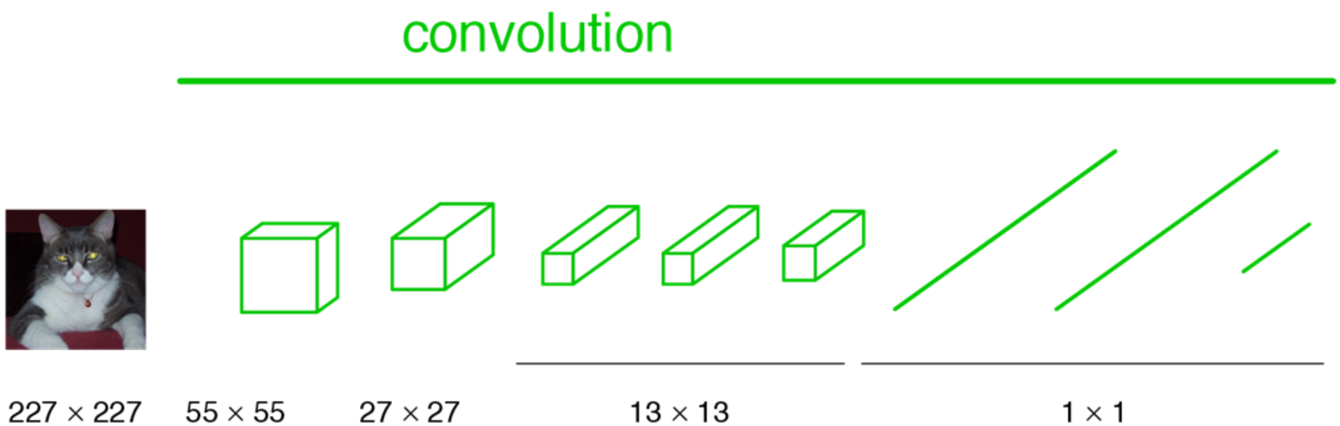
Credit: Long et al

A Classification Network



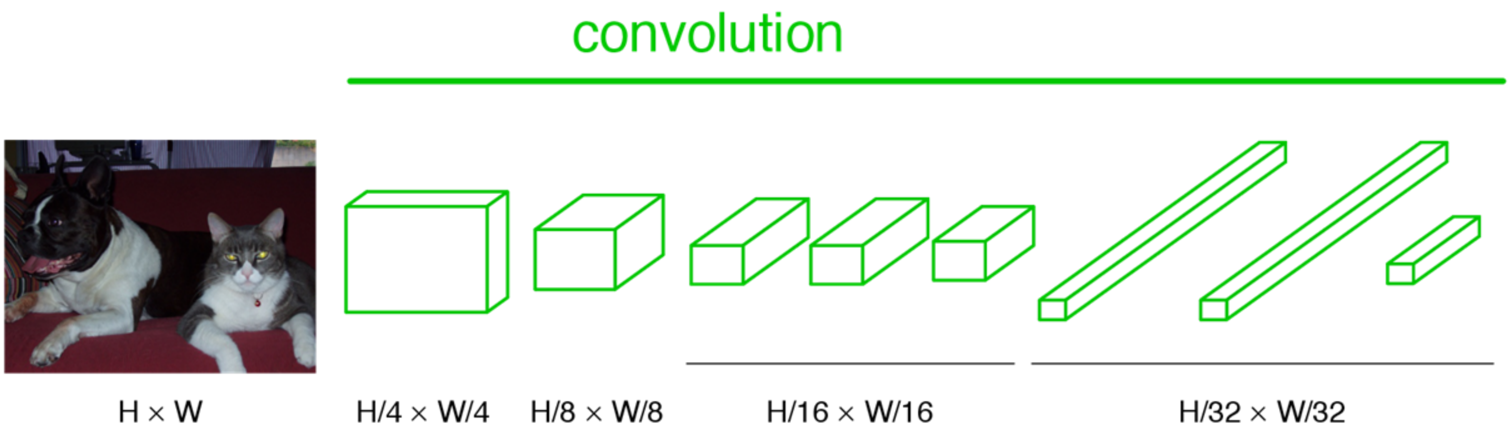
Credit: Long et al

Becoming fully convolutional



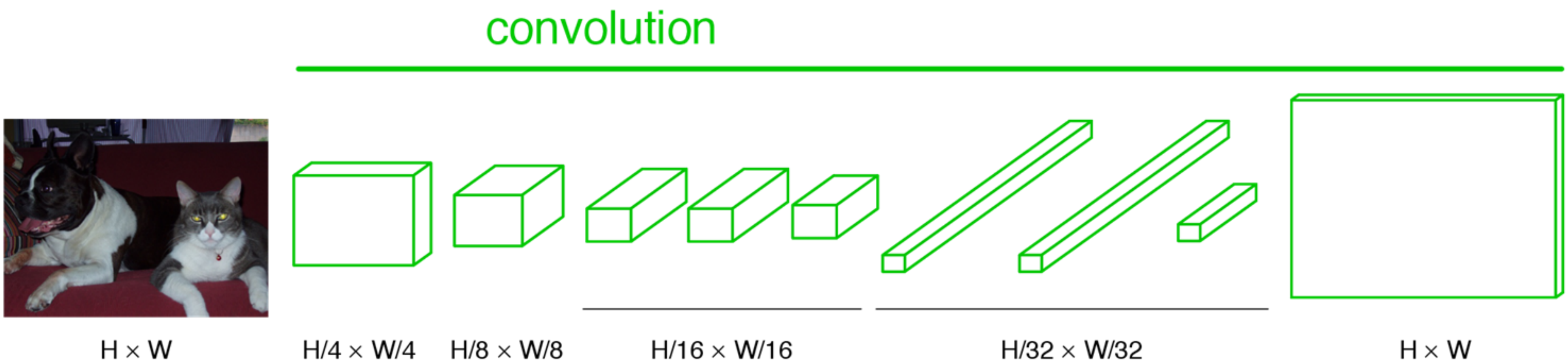
Credit: Long et al

Becoming fully convolutional



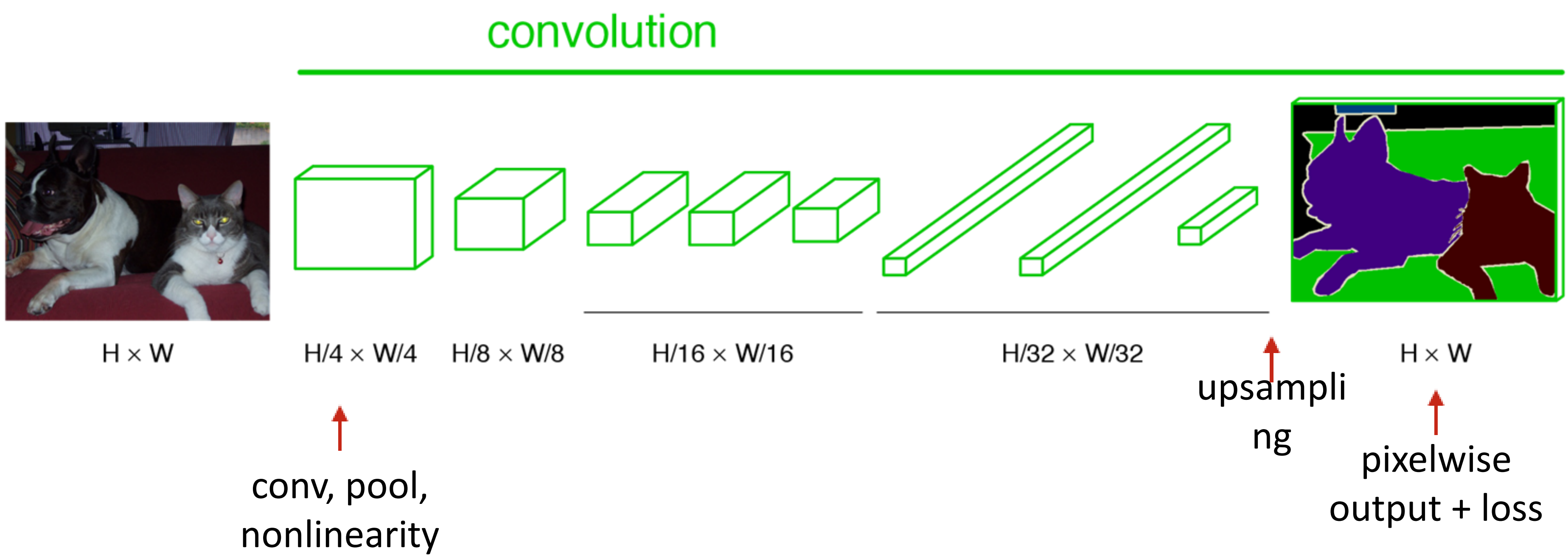
Credit: Long et al

Upsampling output



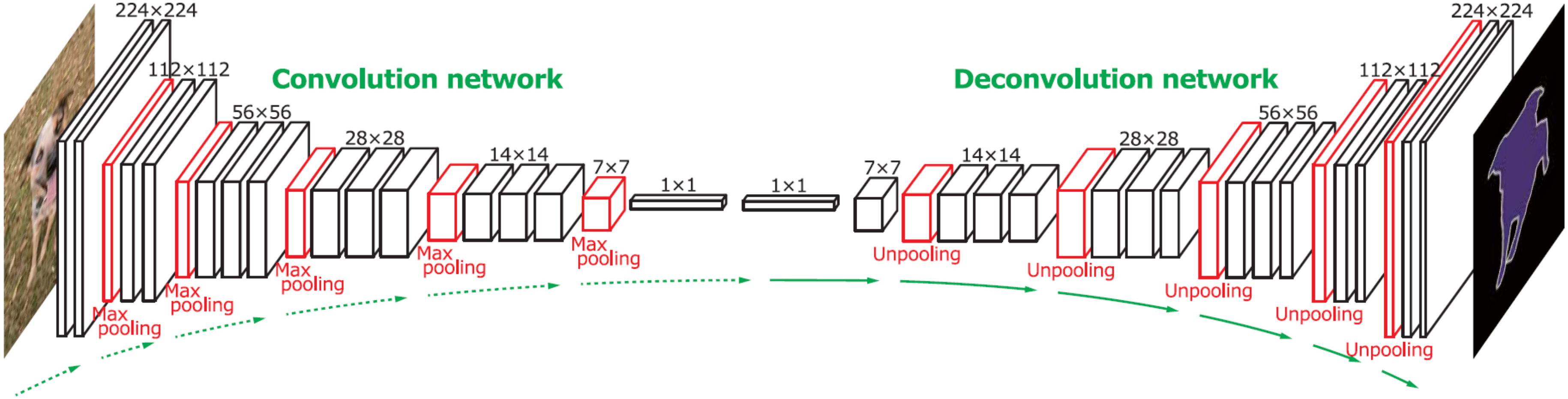
Credit: Long et al

End-to-end, Pixels-to-pixels network



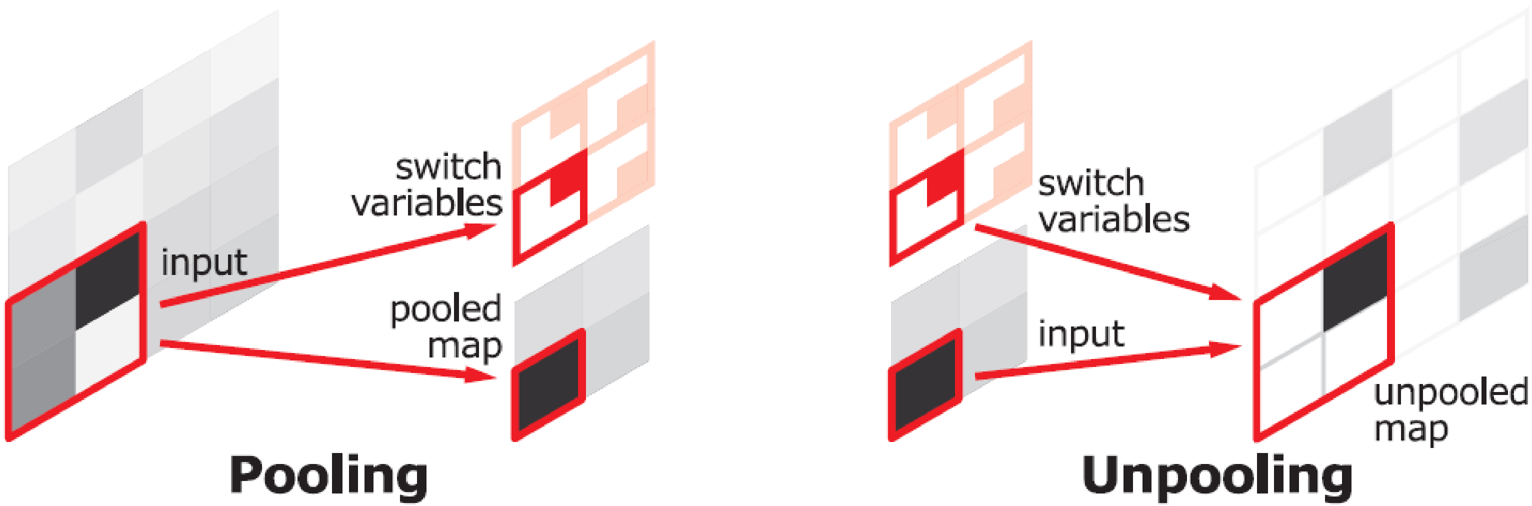
Credit: Long et al

Deconvolutional network



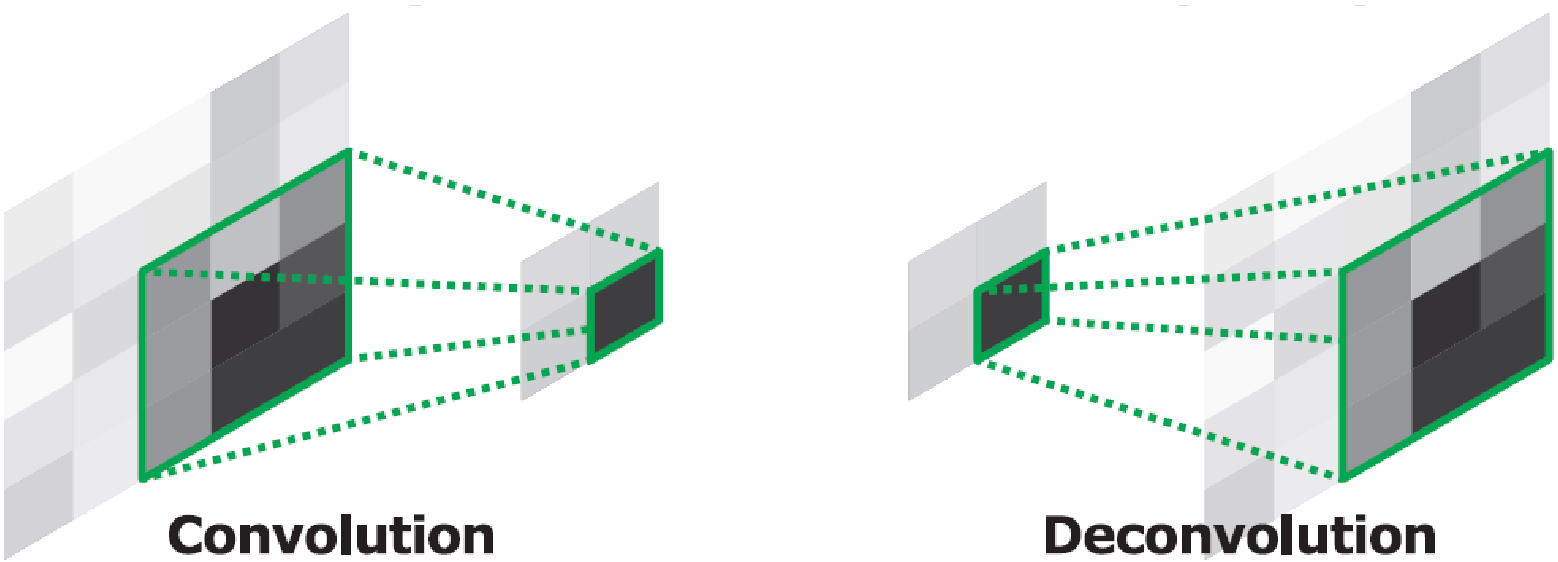
Credit: Noh et al

Upsampling



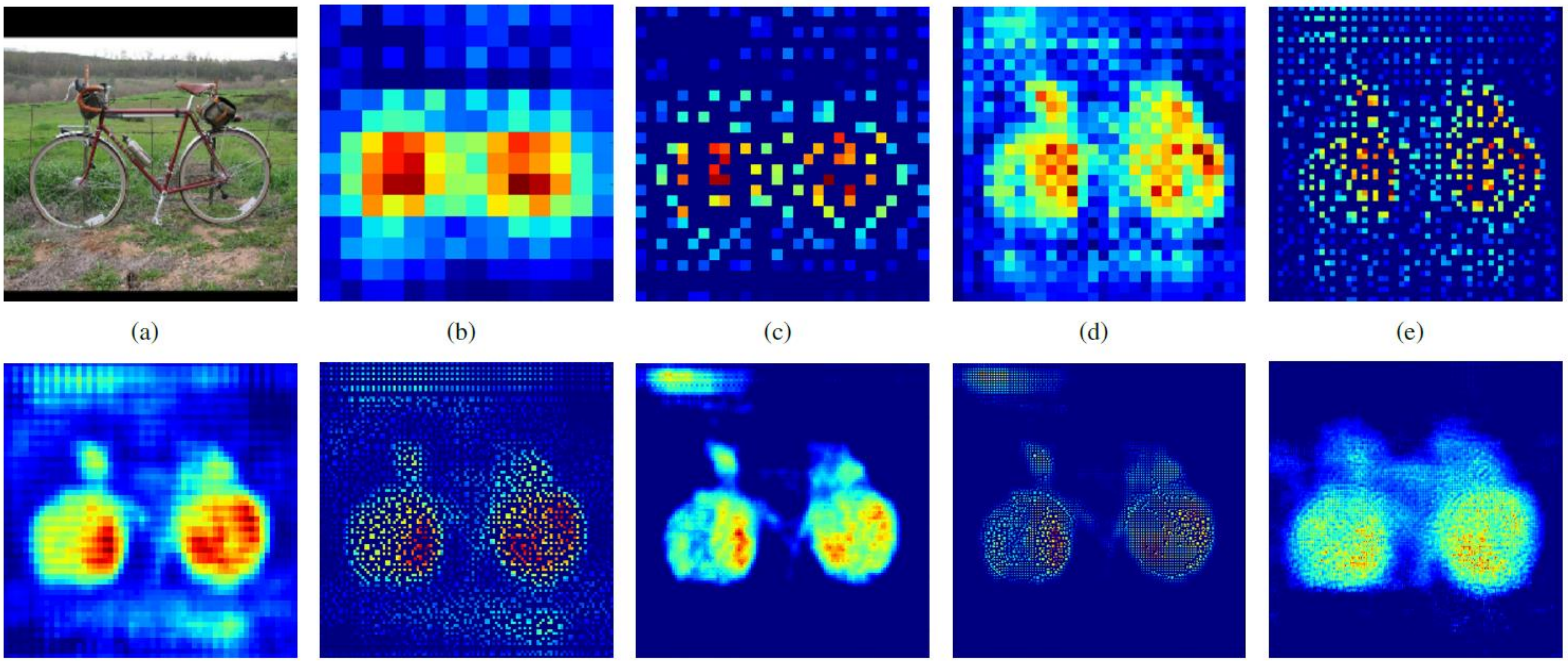
Credit: Noh et al

Deconvolution



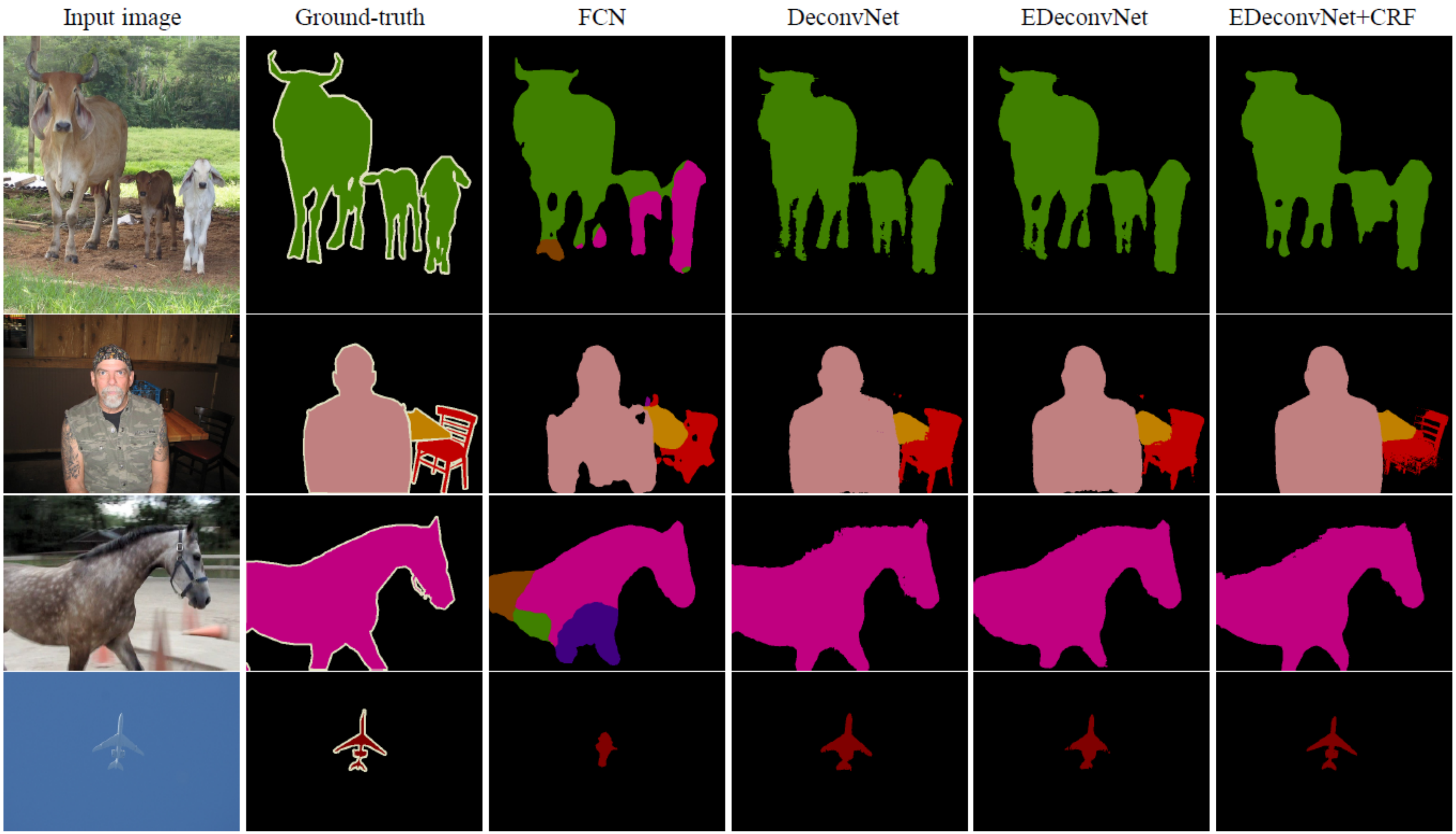
Credit: Noh et al

Unpooling and Deconv Effects



Credit: Noh et al

Results - segmentation



(a) Examples that our method produces better results than FCN [17].

Credit: Noh et al

Predicting Human Visual Memory

Memorability = The likelihood of remembering a particular image.

Welcome to the

Visual Memory Game

A stream of images will be presented on the screen for 1 second each.

Your task:

Clap anytime you see an image you saw before in this experiment.

Ready?

(Seriously, get ready to clap. The images go by fast…)

<clap!>

<clap!>

Measuring Memorability

Memorability is an intrinsic property of an image!

Memorability = Probability of correctly detecting a repeat after a single view of an image in a long sequence.

Understanding Image Memorability, Khosla et al, ICCV 2015

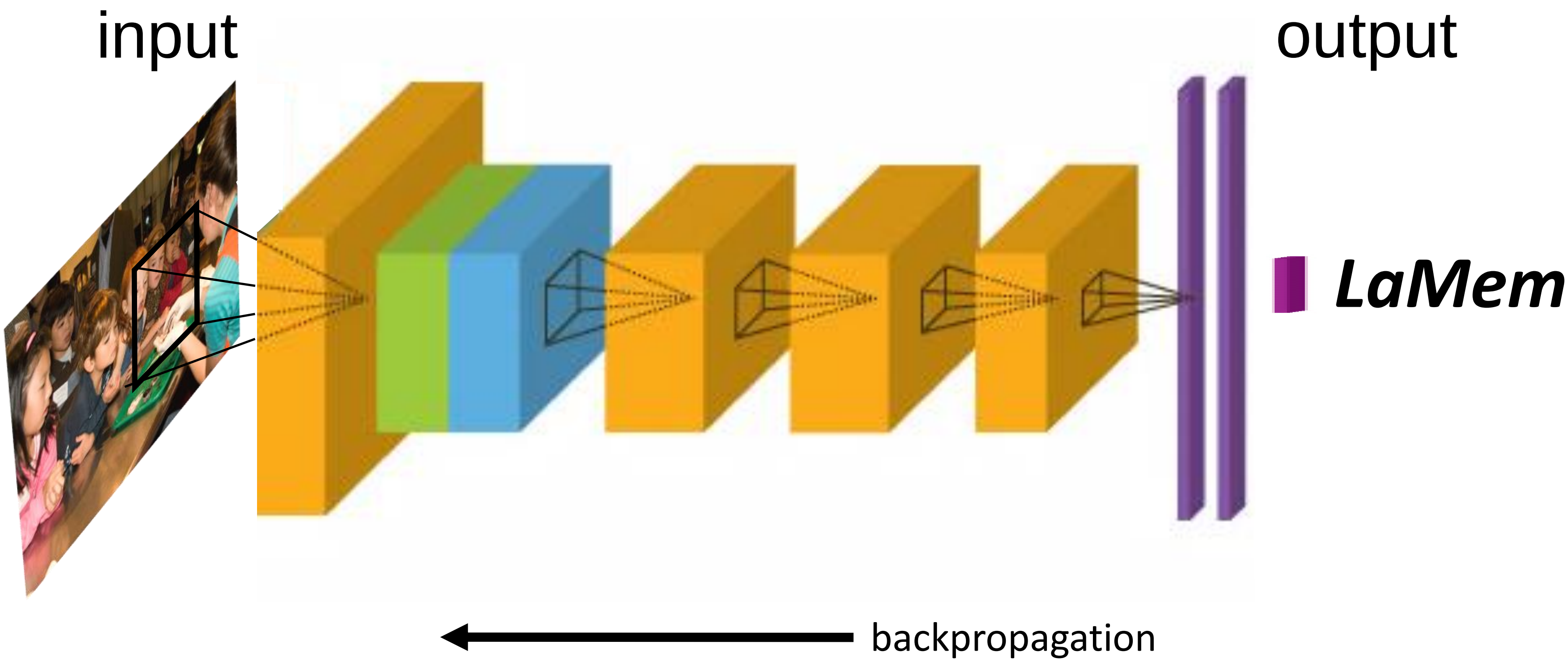
La Mem

- **No single focus**
- **Distant view**
- **Static**
- **Common**

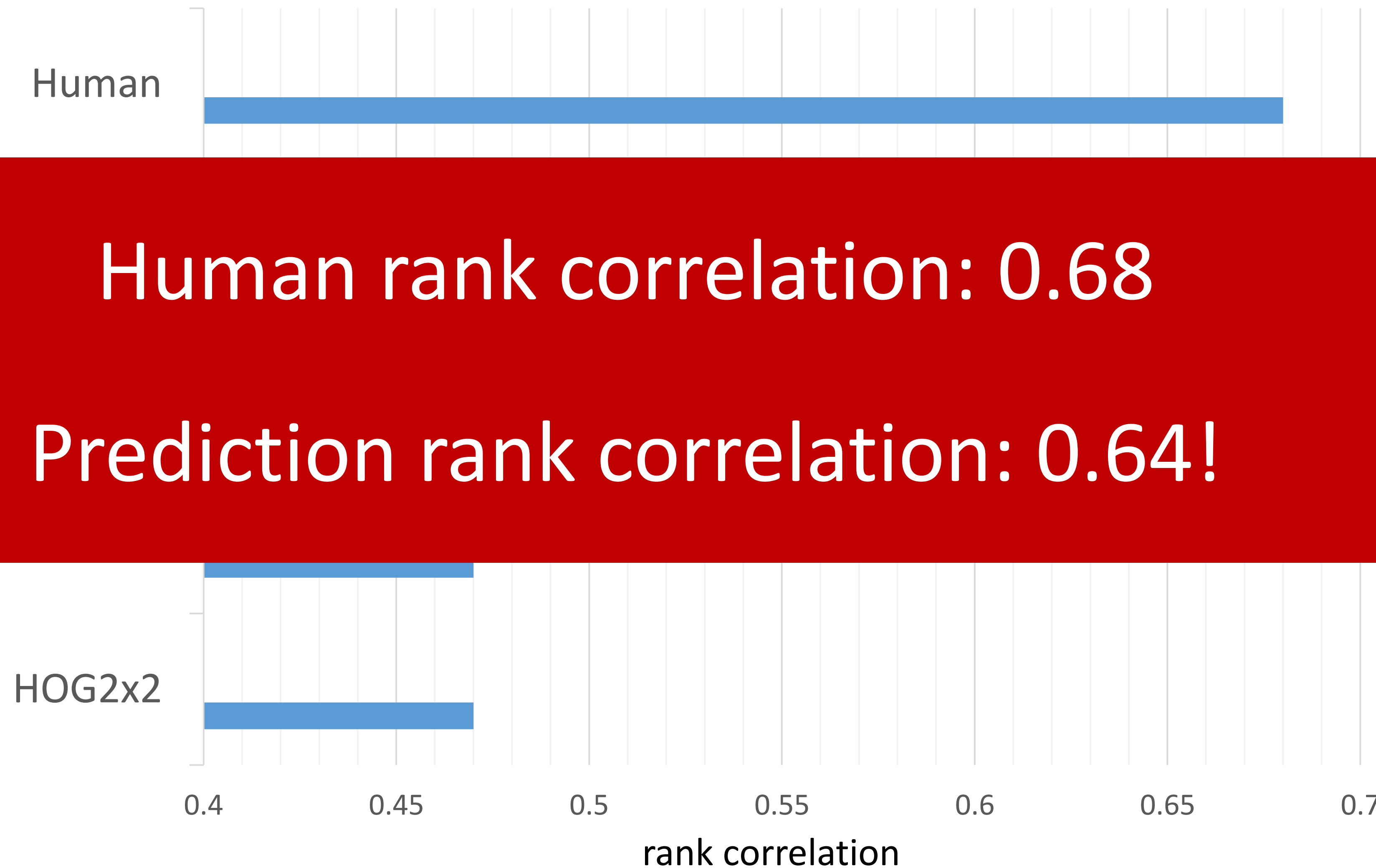
- Focused
- Enclosed Setting
- Dynamics
- Unusual

→ memorability

Training MemNet



MemNet Performance



Human rank correlation: 0.68

Prediction rank correlation: 0.64!

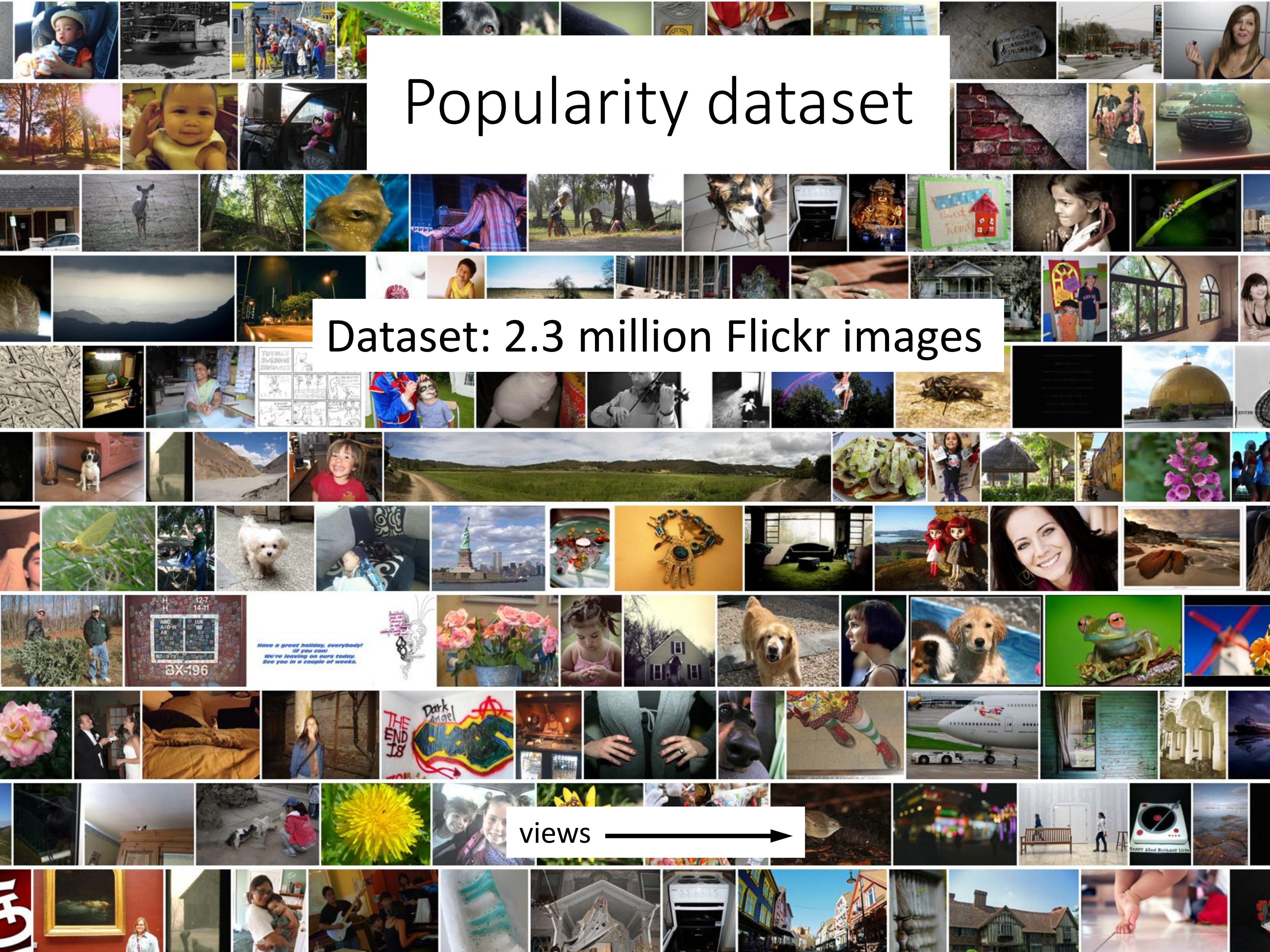
Visualizing Neurons

http://memorability.csail.mit.edu

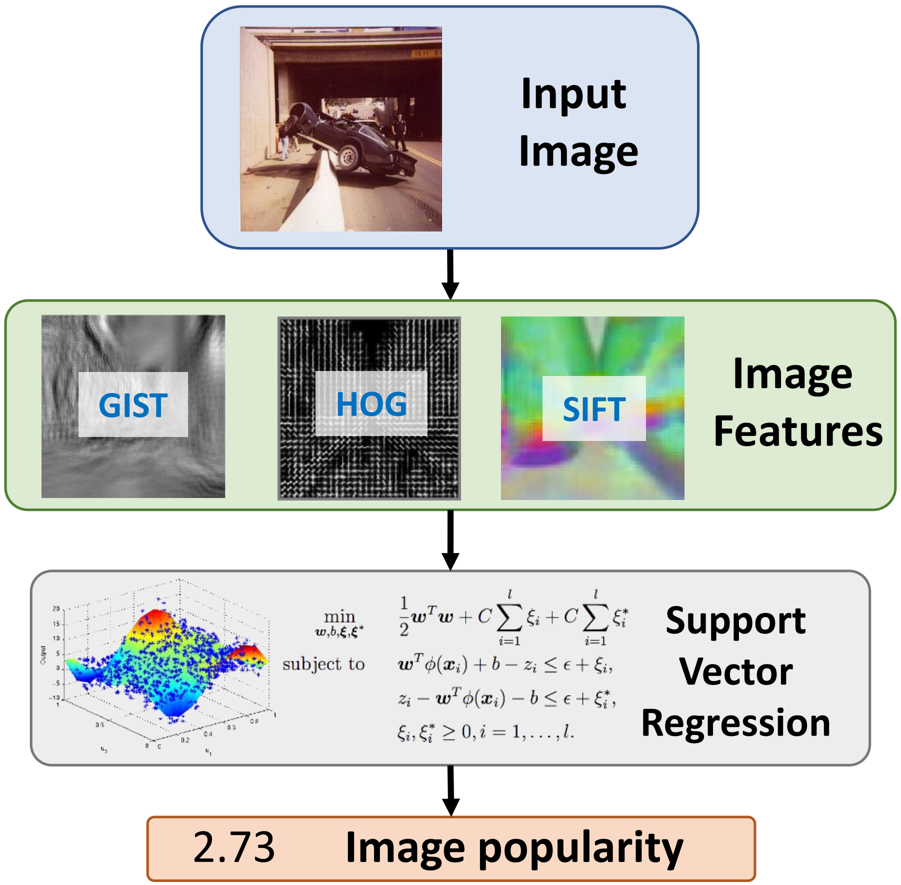
Predicting popularity

‘selfie selection’

Popularity dataset

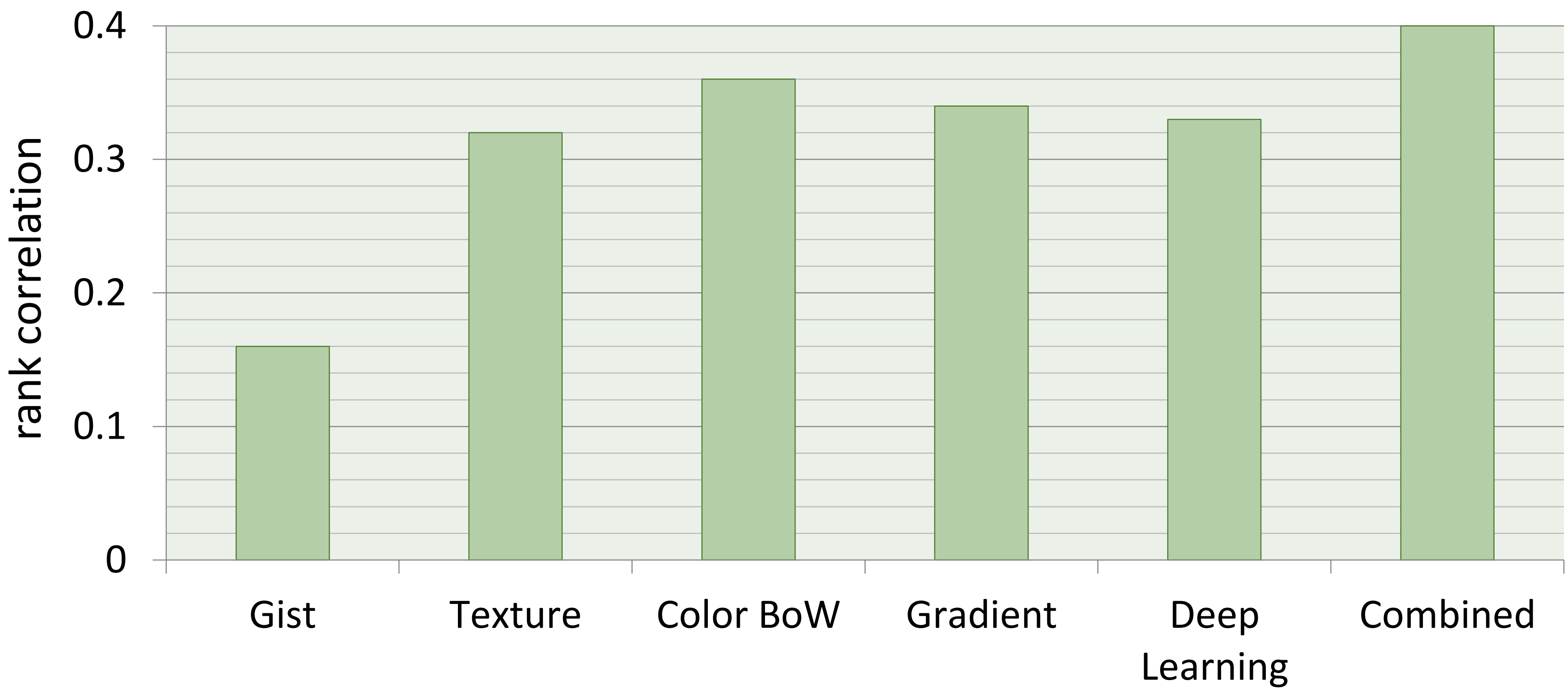


Predicting popularity



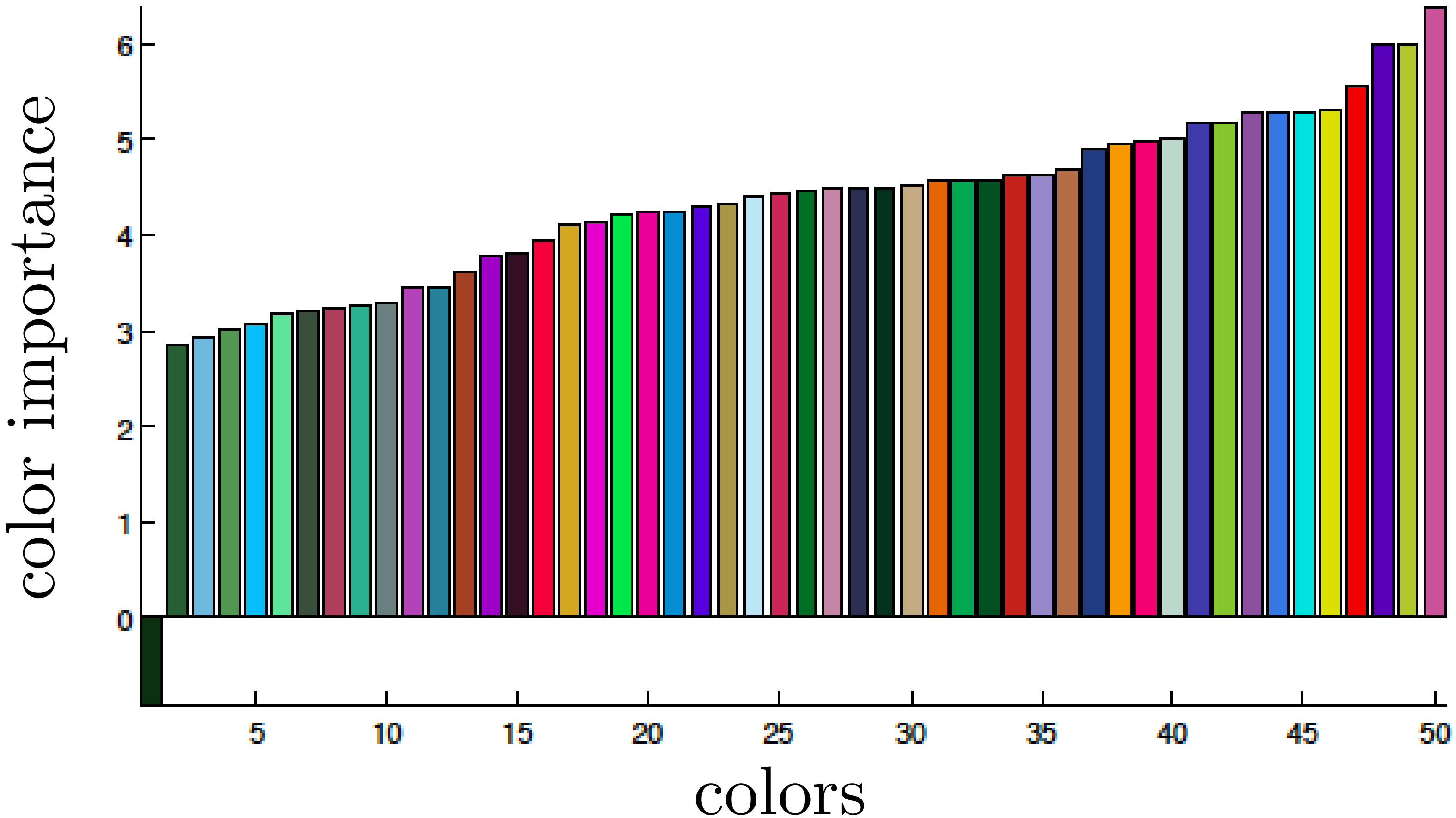
What makes an image popular? Khosla et al, WWW 2014

Predicting popularity



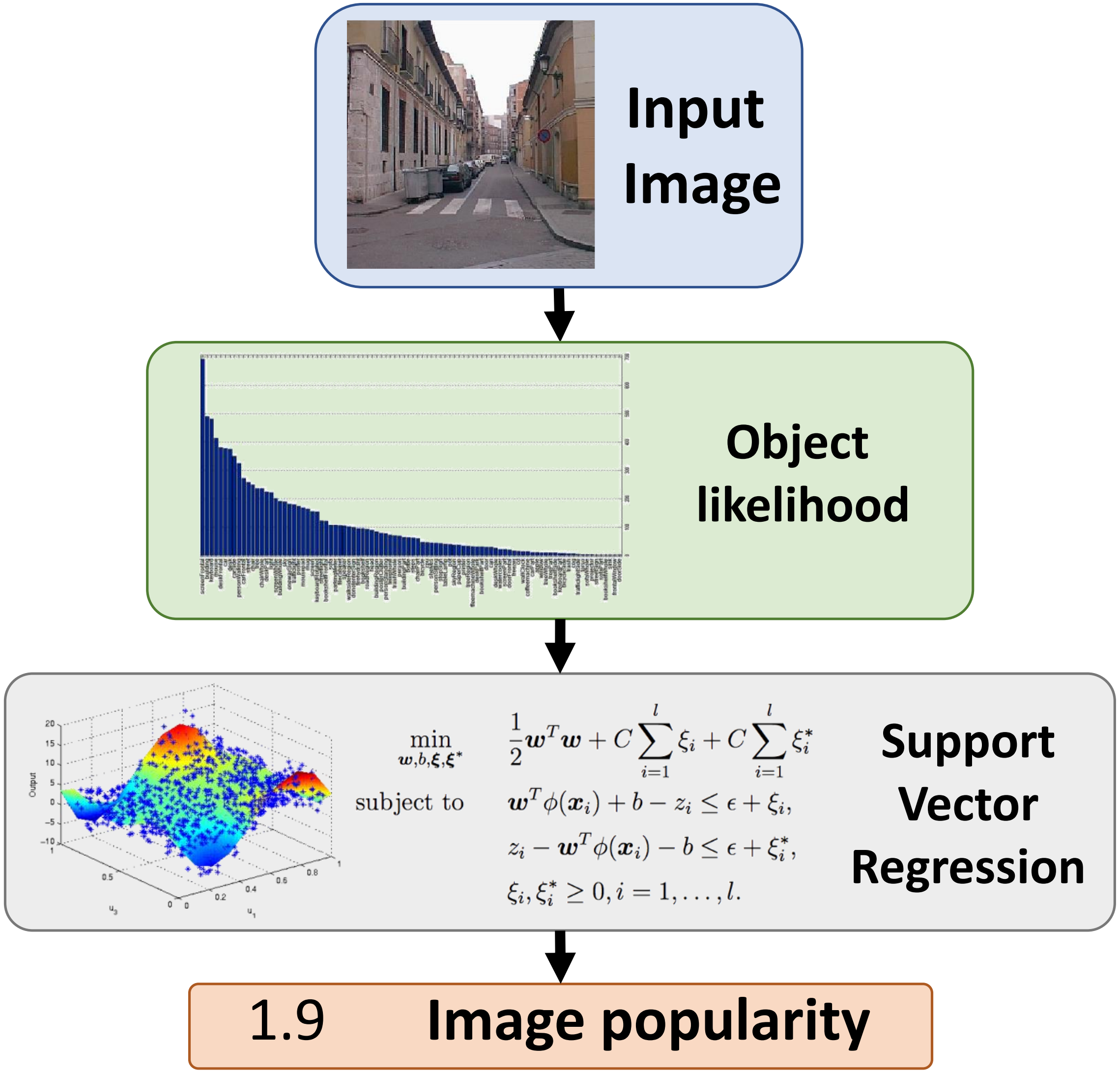
What makes an image popular? Khosla et al, WWW 2014

What makes an image popular?



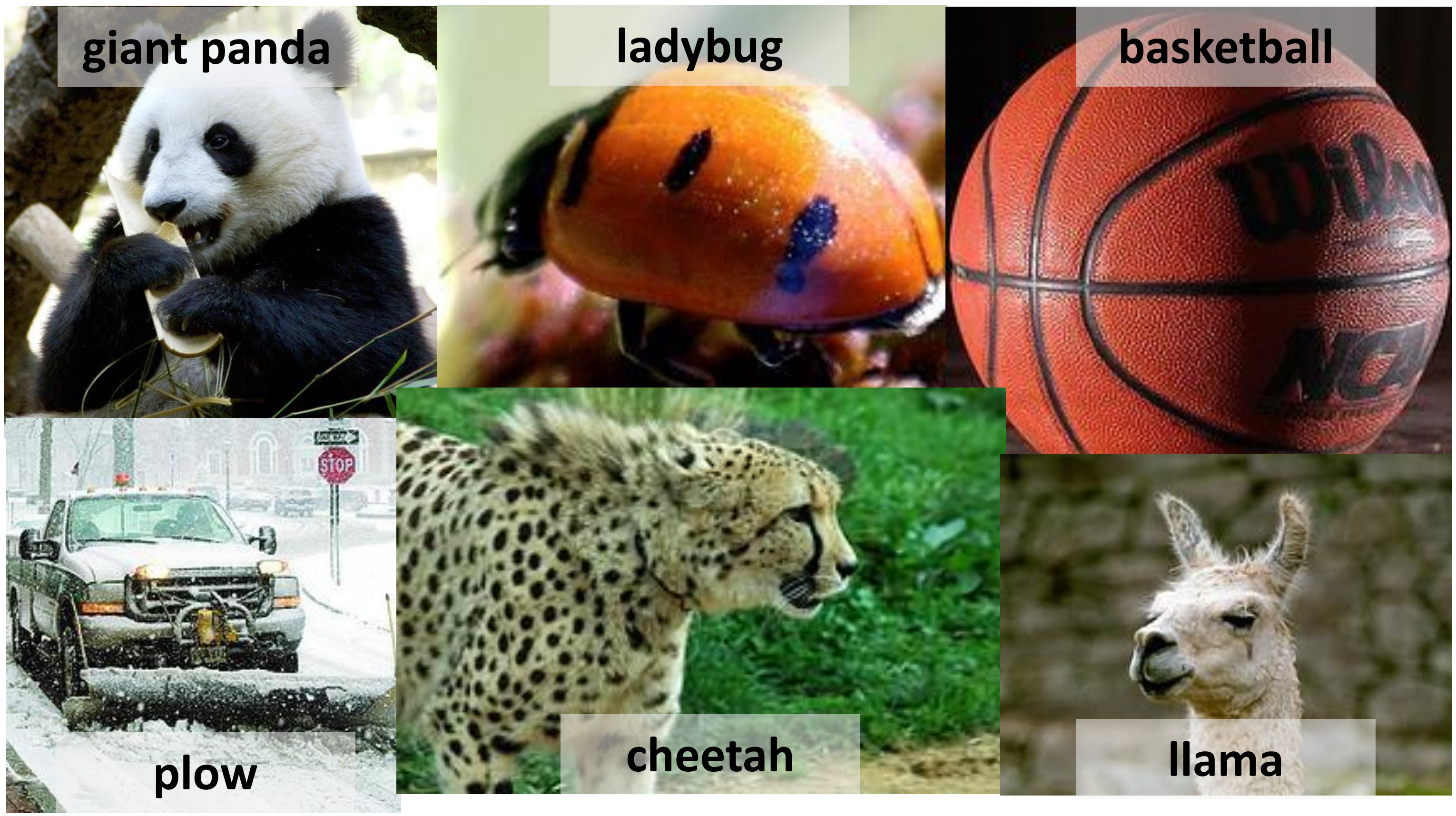
What makes an image popular? Khosla et al, WWW 2014

What makes an image popular?



What makes an image popular?

Medium positive impact



What makes an image popular?

Strong positive impact

What makes an image popular?

Negative impact

http://popularity.csail.mit.edu

Popularity Demo

How likely is your image to become popular? Upload it to find out!

Upload: Choose File No file chosen Run

or

URL: http:// Run

or

Click One:

Popularity API

Usage: http://popularity.csail.mit.edu/cgi-bin/image.py?url=IMG_URL

Example: http://popularity.csail.mit.edu/cgi-bin/image.py?url=http://popularity.csail.mit.edu/demo/1.jpg

Notice: Please do not overload our server by querying repeatedly in a short period of time. This is a free service for academic research and education purposes only. It has no guarantee of any kind. For any questions or comments regarding this API or potential commercial applications, please contact Aditya Khosla.

Media coverage

THE VERGE THE HUFFINGTON POST TechCrunch TIME

Entrepreneur YAHOO! NEWS THE TIMES OF INDIA The Washington Post

http://popularity.csail.mit.edu

Popularity Demo

How likely is your image to become popular? Upload it to

Popularity API

Usage: http://popularity.csail.mit.edu/cgi-

To: khosla@csail.mit.edu

popularity

Dear Aditya Khosla,

This popularity calculator is a nice initiative, but i think these girls deserve much better scores than 5.

Best regards,

THE VERGE

HUFFINGTON POST

TechCrunch

TIME

Entrepreneur

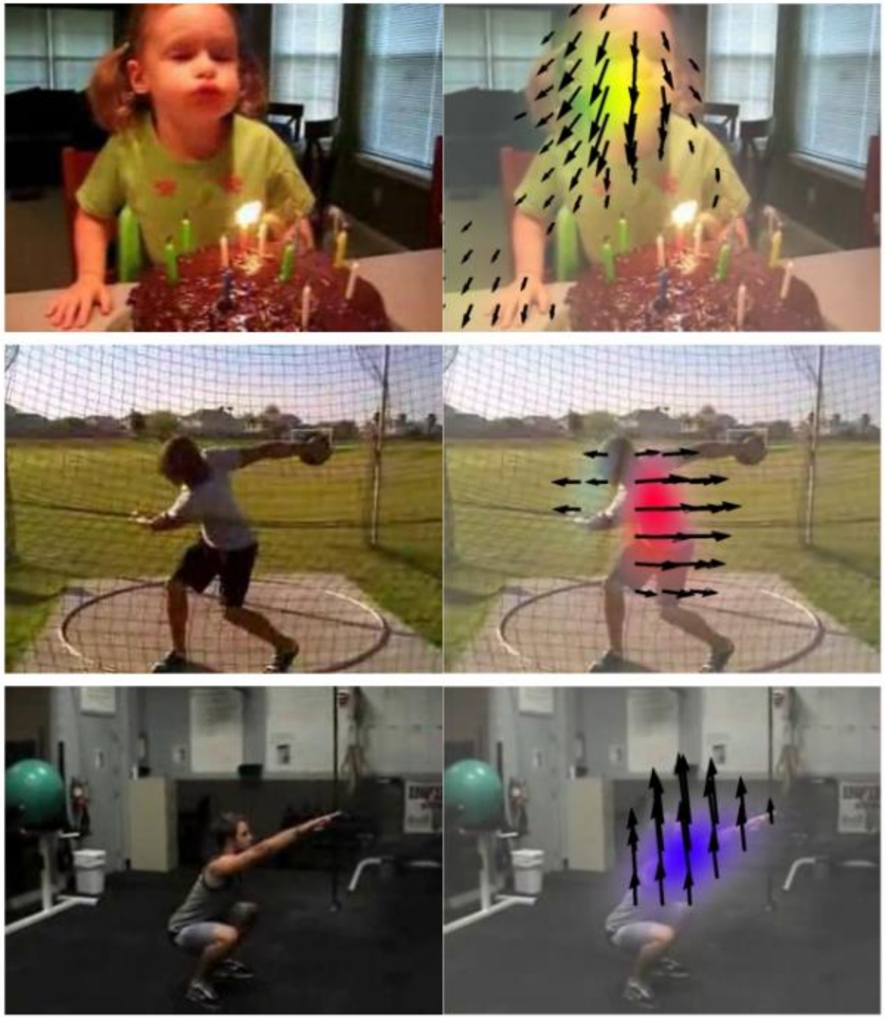
YAHOO! NEWS

THE TIMES OF INDIA

The Washington Post

Predicting the future

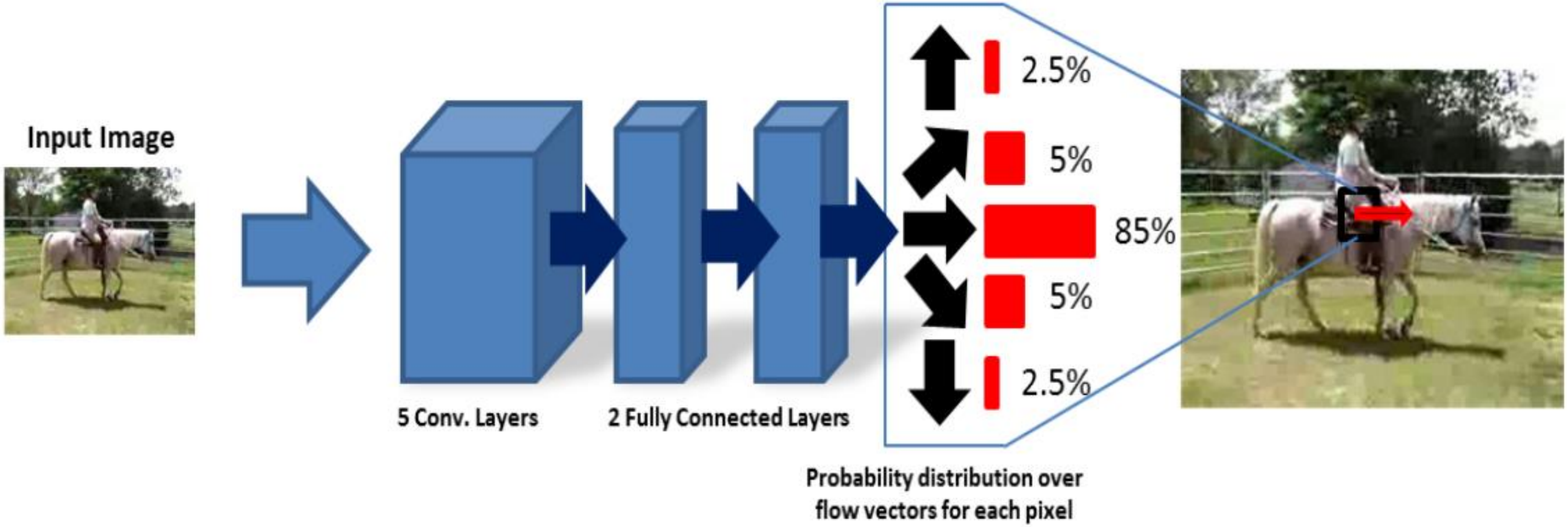
Goal:
 - predict Lucas-Kanade optical flow given just one image!



(a) Input Image (b) Prediction

Credit: Walker et al

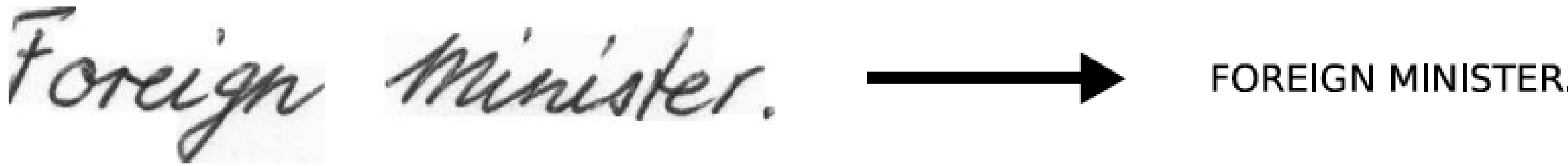
Predicting the future

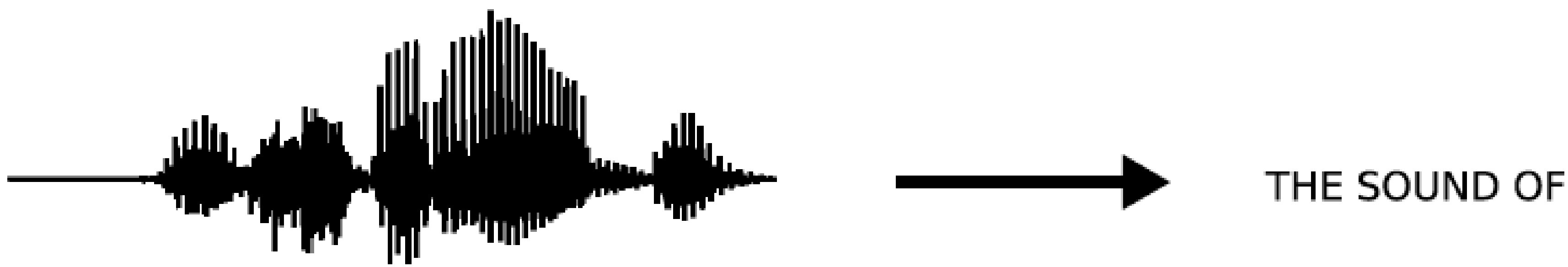


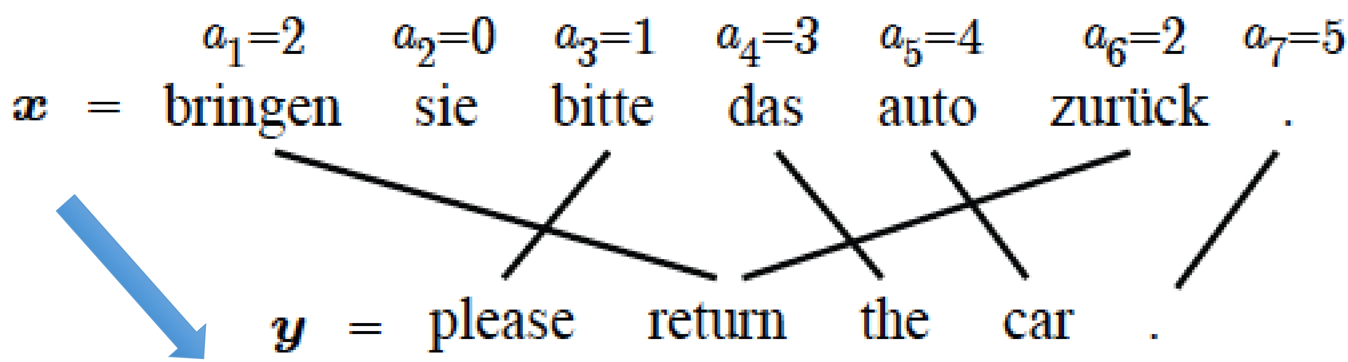
Credit: Walker et al

Learning sequences

Sequences are everywhere…

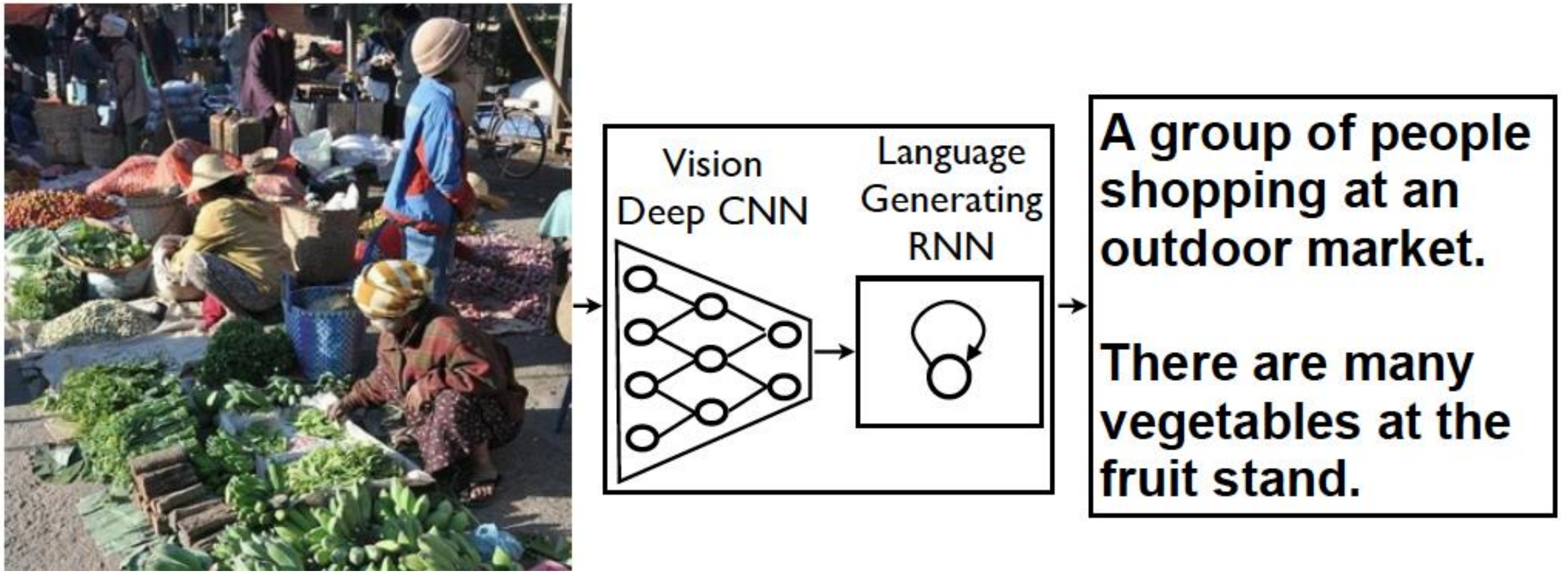




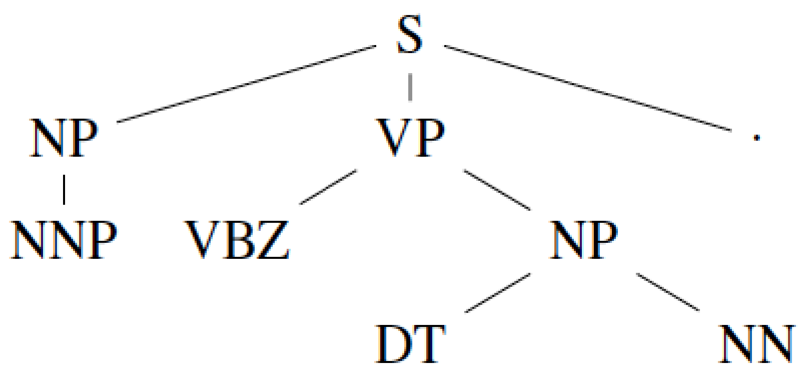


Credit: Alex Graves, Kevin Gimpel, Dhruv Batra

Even where you might not expect a sequence…



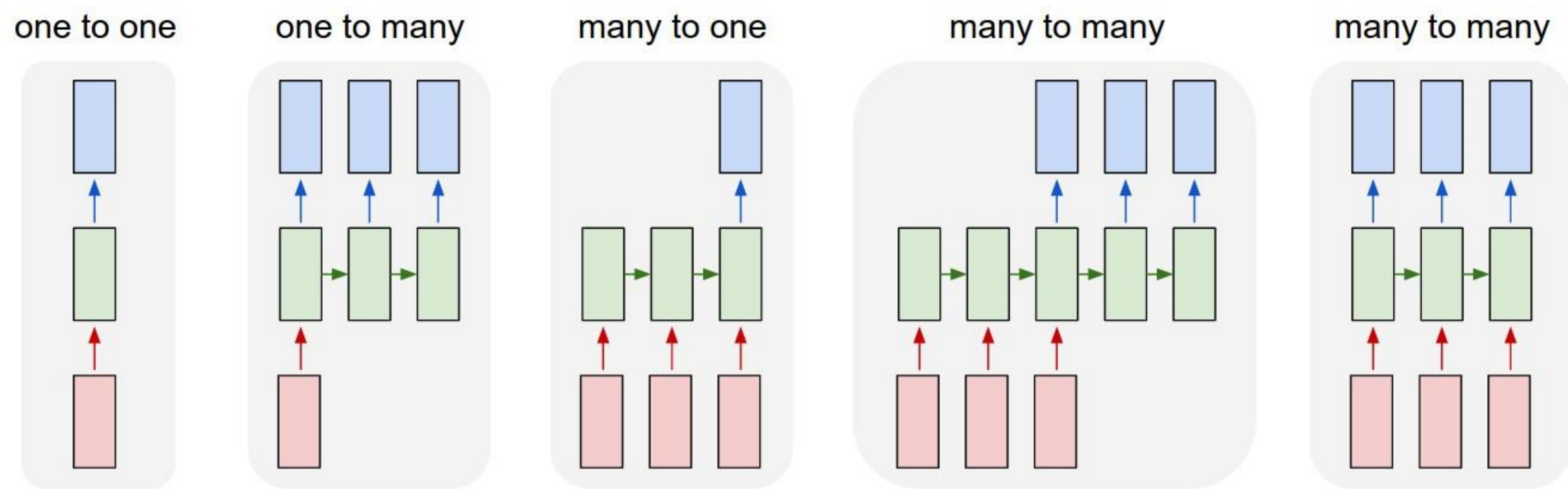
John has a dog . →



John has a dog . → (S (NP NNP $)_{NP}$ (VP VBZ (NP DT NN $)_{NP}$ $)_{VP}$. $)_{S}$

Credit: Dhruv Batra, Vinyals et al.

How do we model sequences?



one to one

Input: No sequence

Output: No sequence

Example: “standard” classification / regression problems

one to many

Input: No sequence

Output: Sequence

Example: Im2Caption

many to one

Input: Sequence

Output: No sequence

Example: sentence classification, multiple-choice question answering

many to many

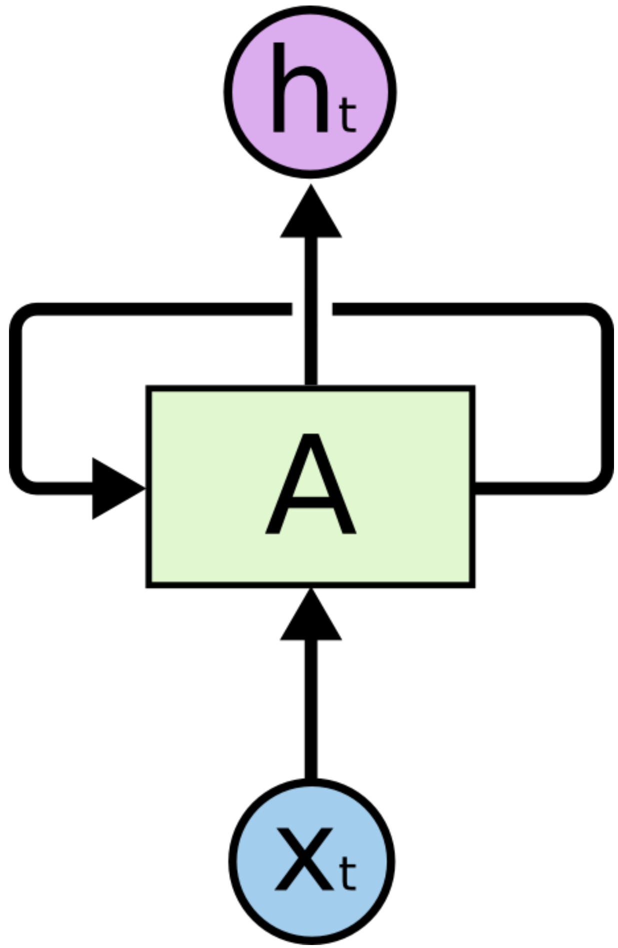
Input: Sequence

Output: Sequence

Example: machine translation, video captioning, open-ended question answering, video question answering

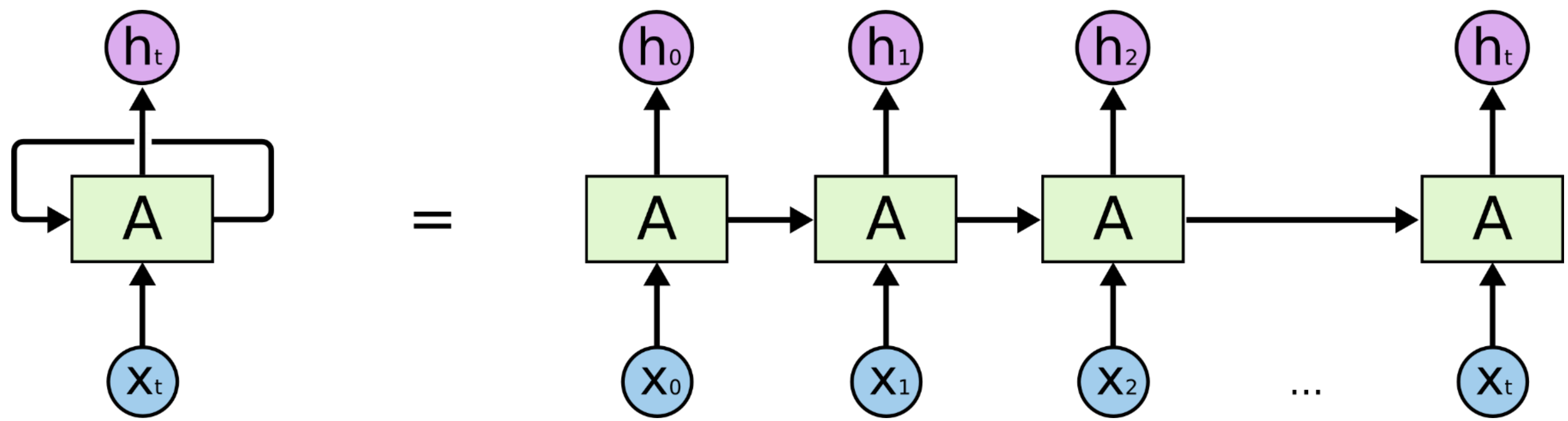
Credit: Dhruv Batra, Andrej Karpathy

Recurrent Neural Networks (RNNs)



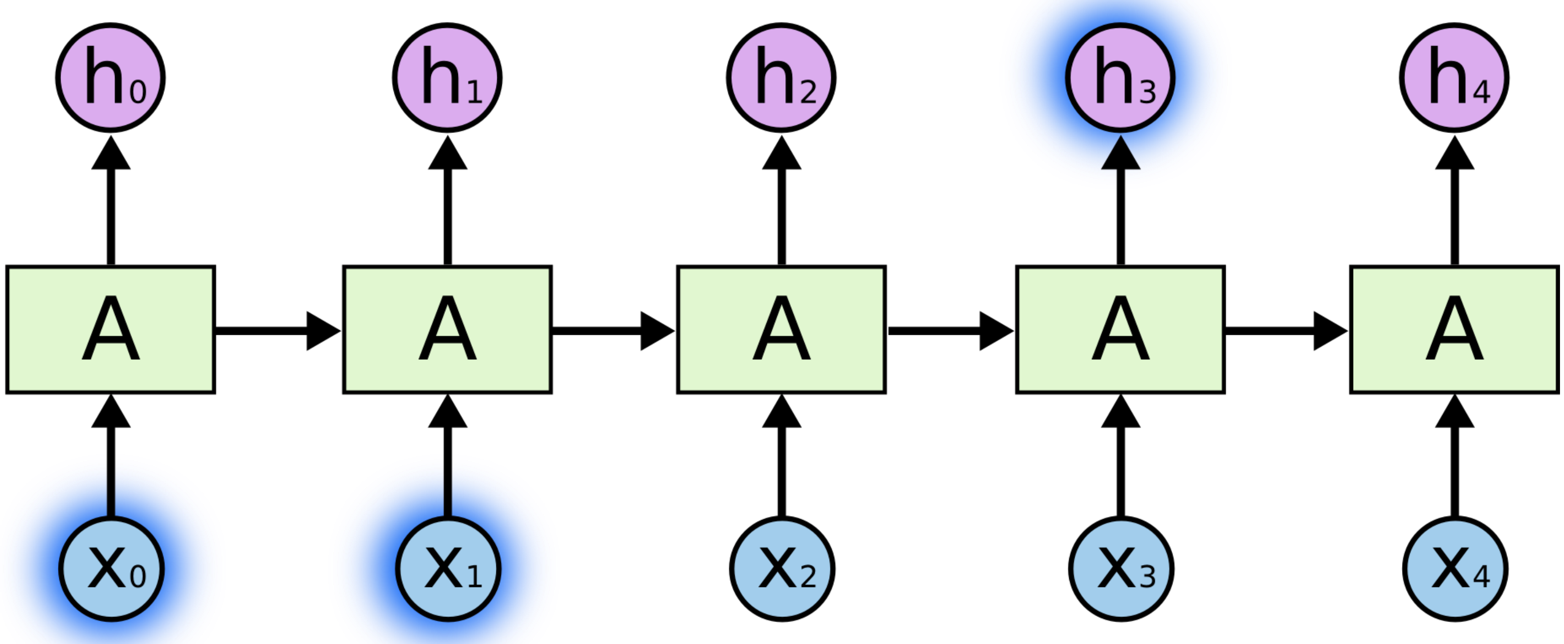
Credit: Christopher Olah

Recurrent Neural Networks (RNNs)



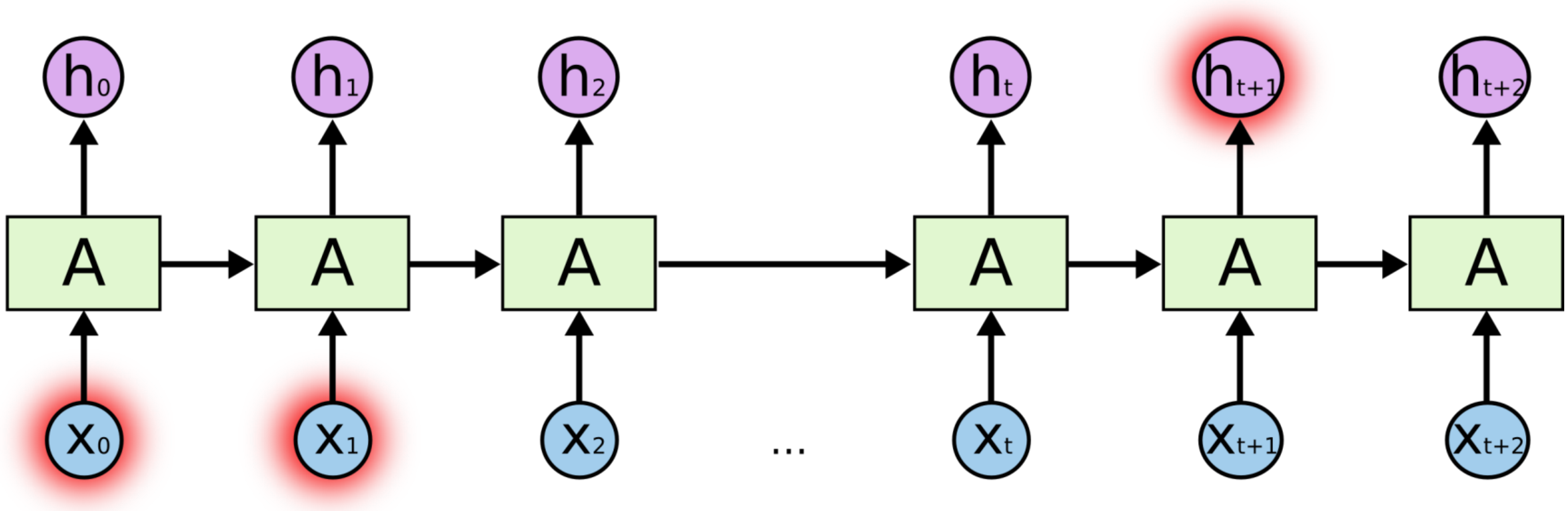
Credit: Christopher Olah

Recurrent Neural Networks (RNNs)



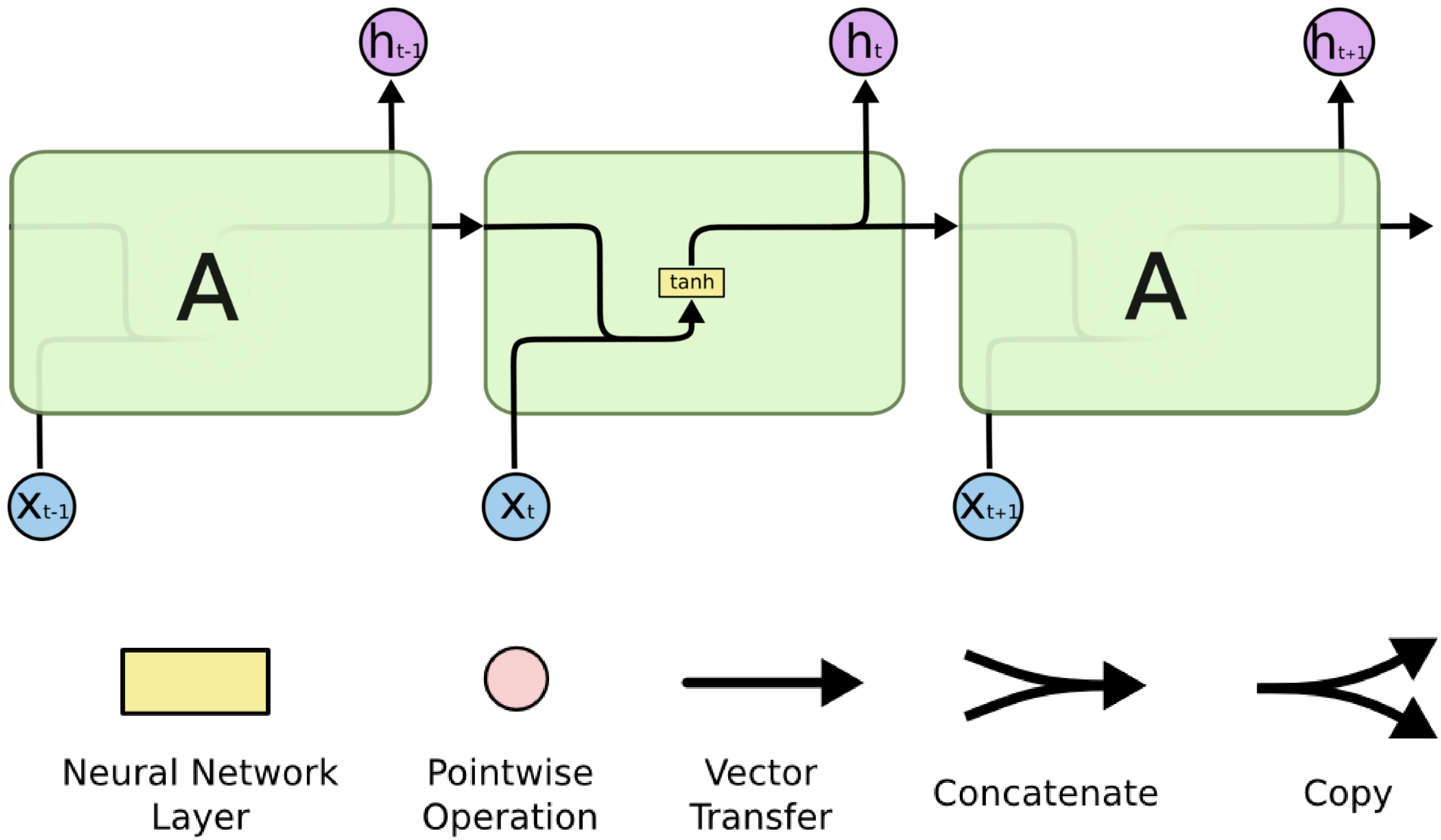
Credit: Christopher Olah

Long-term dependencies–hard to model!



Credit: Christopher Olah

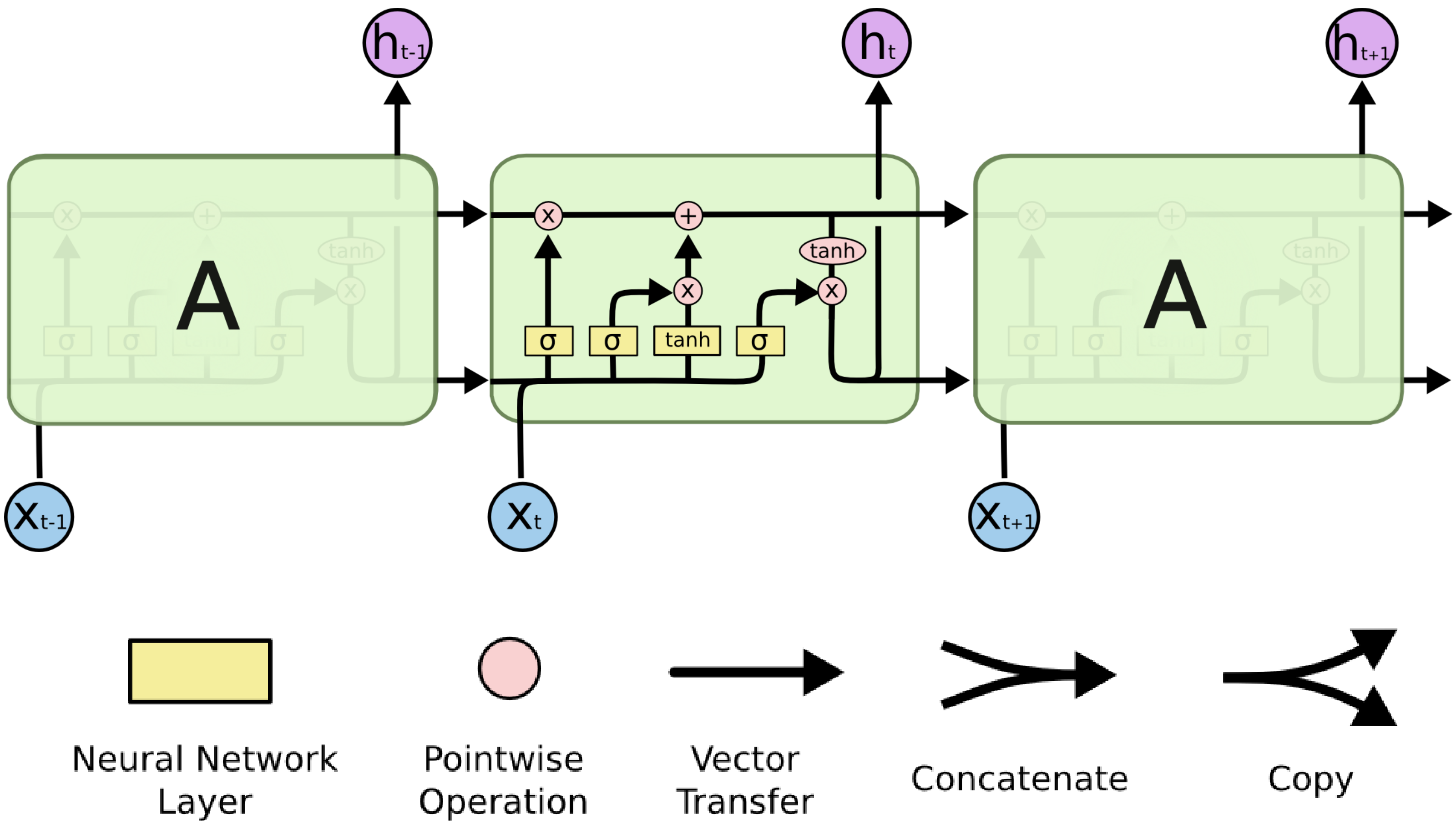
From plain RNNs to LSTMs



(LSTM: Long Short Term Memory Networks)

Credit: Christopher Olah

From plain RNNs to LSTMs

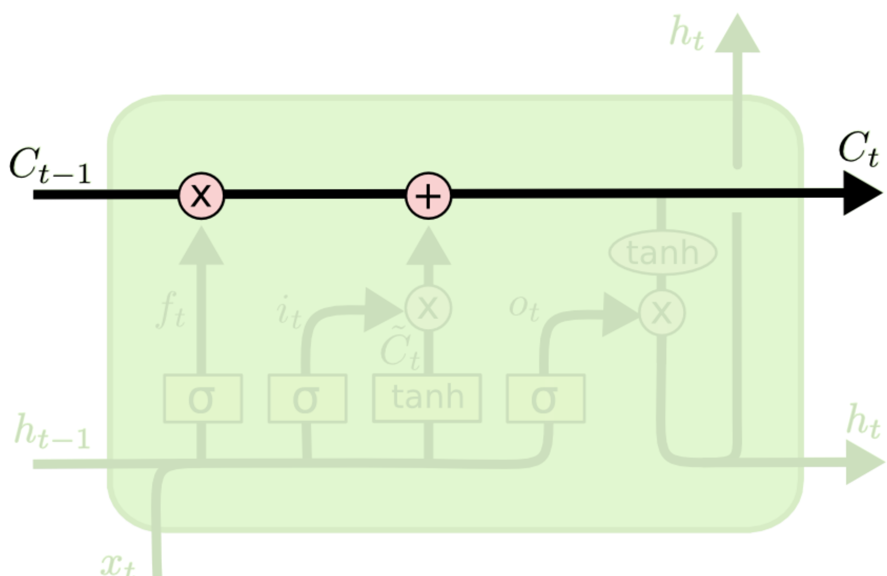


(LSTM: Long Short Term Memory Networks)

Credit: Christopher Olah

LSTMs Intuition: Memory

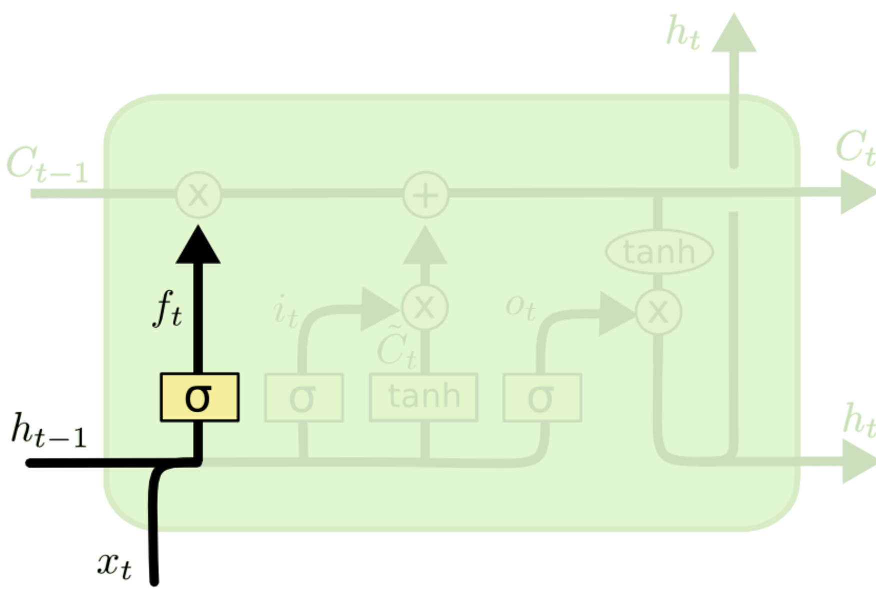
- Cell State / Memory



Credit: Christopher Olah

LSTMs Intuition: Forget Gate

- Should we continue to remember this "bit" of information or not?

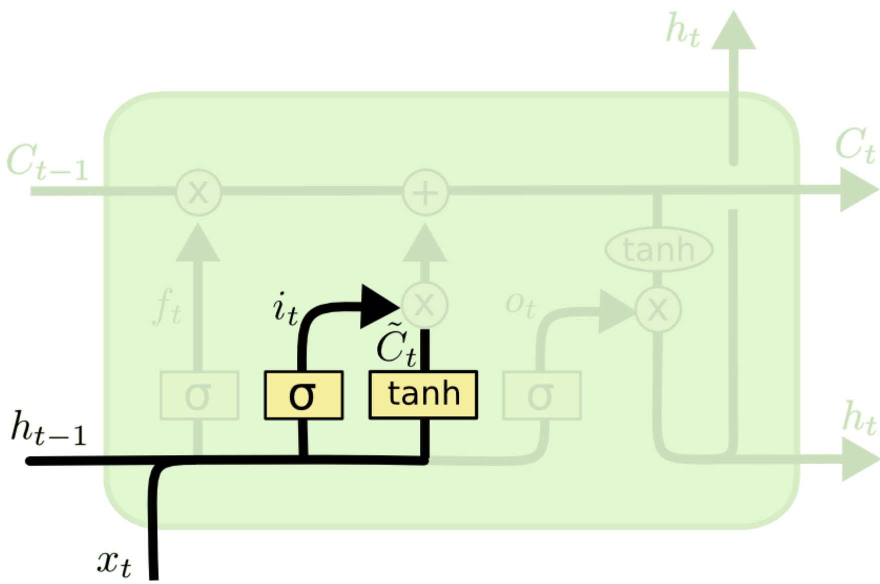


$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] \ + \ b_f\right)$$

Credit: Christopher Olah

LSTMs Intuition: Input Gate

- Should we update this “bit” of information or not?
  - If so, with what?



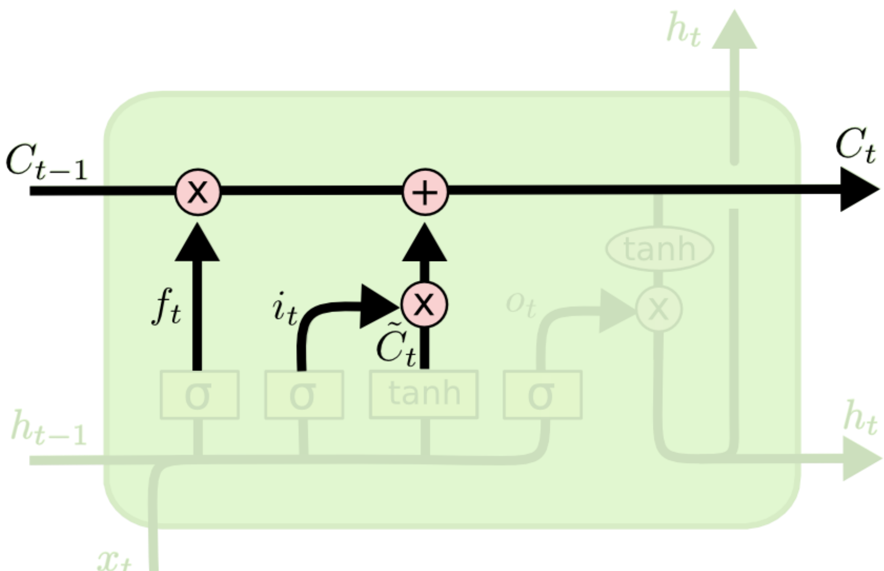
$$i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] \ + \ b_i\right)$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] \ + \ b_C)$$

Credit: Christopher Olah

LSTMs Intuition: Memory Update

- Forget that + memorize this

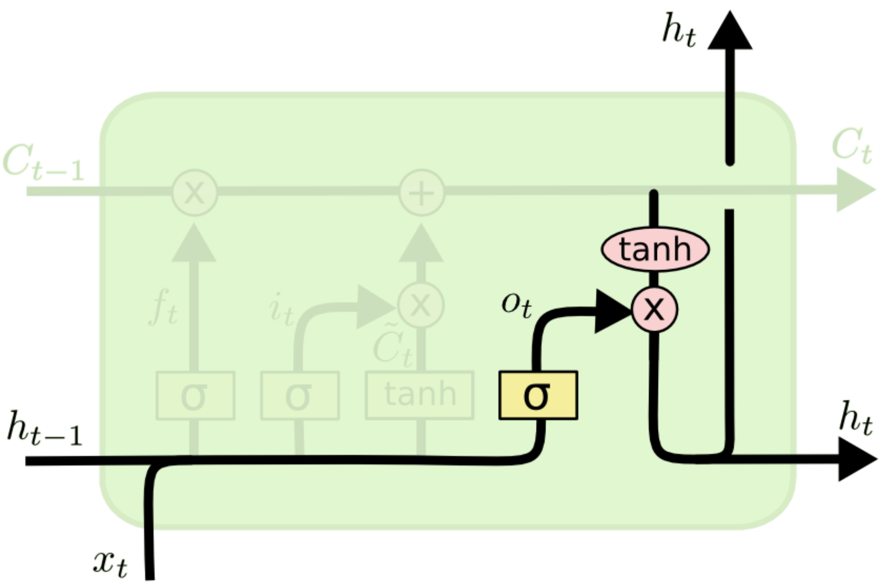


$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$$

Credit: Christopher Olah

LSTMs Intuition: Output Gate

- Should we output this “bit” of information to “deeper” layers?

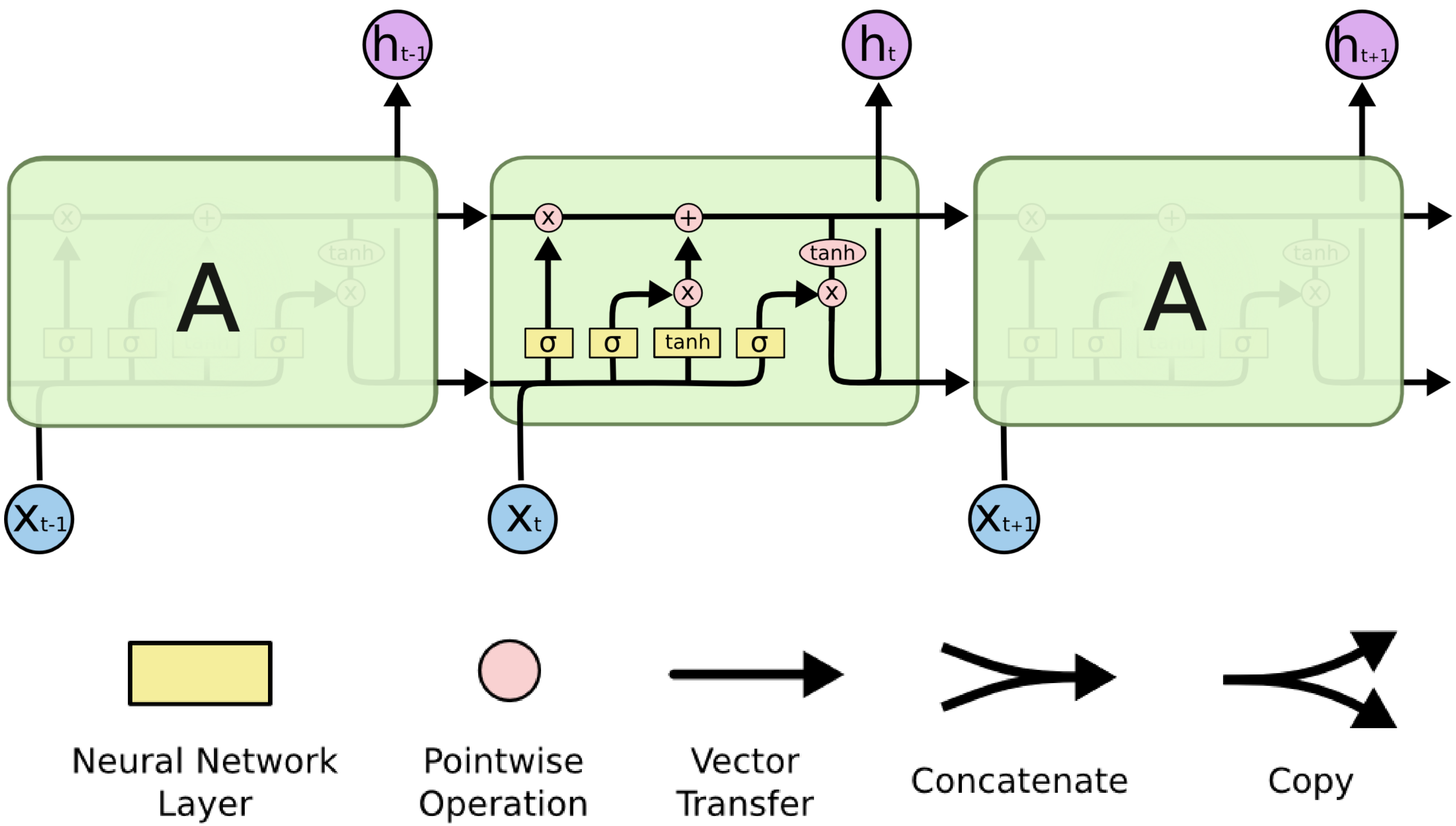


$$o_t = \sigma\left(W_o \; [h_{t-1}, x_t] \; + \; b_o\right)$$

$$h_t = o_t * \tanh\left(C_t\right)$$

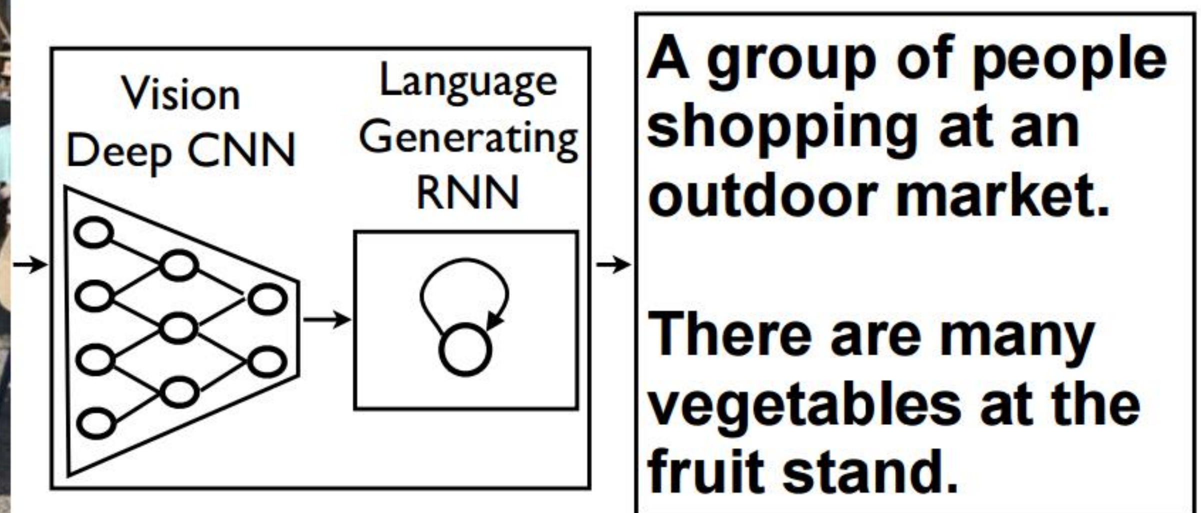
Credit: Christopher Olah

LSTM: A pretty sophisticated cell



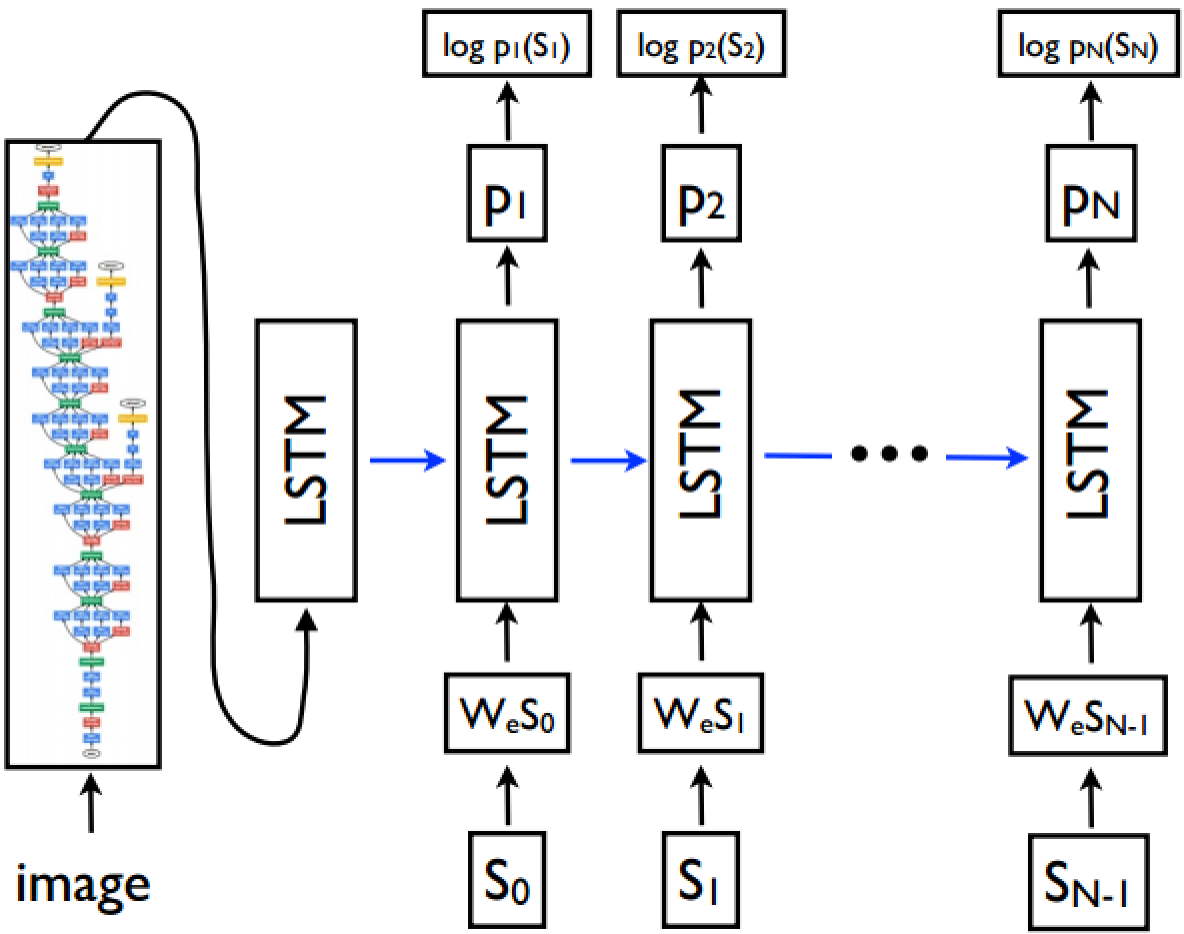
Credit: Christopher Olah

Generating image captions



Credit: Vinyals et al

Generating image captions



Credit: Vinyals et al

Credit: Vinyals et al