

vision & language

CS 585, Fall 2018

Introduction to Natural Language Processing

<http://people.cs.umass.edu/~miyyer/cs585/>

Mohit Iyer

College of Information and Computer Sciences

University of Massachusetts Amherst

some slides adapted from Vicente Ordonez, Fei-Fei Li, and Jacob Andreas

poster specifications (due Dec 6!)

- poster template? **24" x 36"** in whatever software you want. info to include:
 - **title / authors**
 - **problem statement** (with examples): what is your problem and why is it interesting?
 - **your contributions:** did you accomplish what you set out to do in your proposal / progress report? if not, what remains and why is it so challenging?
 - **dataset properties** (with examples) and any **annotation** you did
 - **baseline & your approach:** explanation of how they work (ideally with pictures)
 - **experiments:** how does your approach compare to the baselines
 - **error analysis:** what are the strengths & weaknesses of the models you trained (ideally w/ examples from your own analysis of model outputs)

main criteria for grading

- does the poster effectively convey what you did and why it's interesting?
 - make it colorful, be creative!!!
 - if you just dump blocks of text from your writeups, we won't be happy :(
- does it look like a lot of work went into what you actually did?
 - if you haven't addressed comments made during the proposal or progress report, or you made no progress on things you said you were going to do, we won't be happy :(
- were your group members able to effectively answer questions that me / the TAs have about your project?

image captioning



a red truck is parked on
a street lined with trees

visual question answering



- Is this truck considered “vintage”?
- Does the road look new?
- What kind of tree is behind the truck?

we've seen how to compute
representations of words and
sentences. what about images?

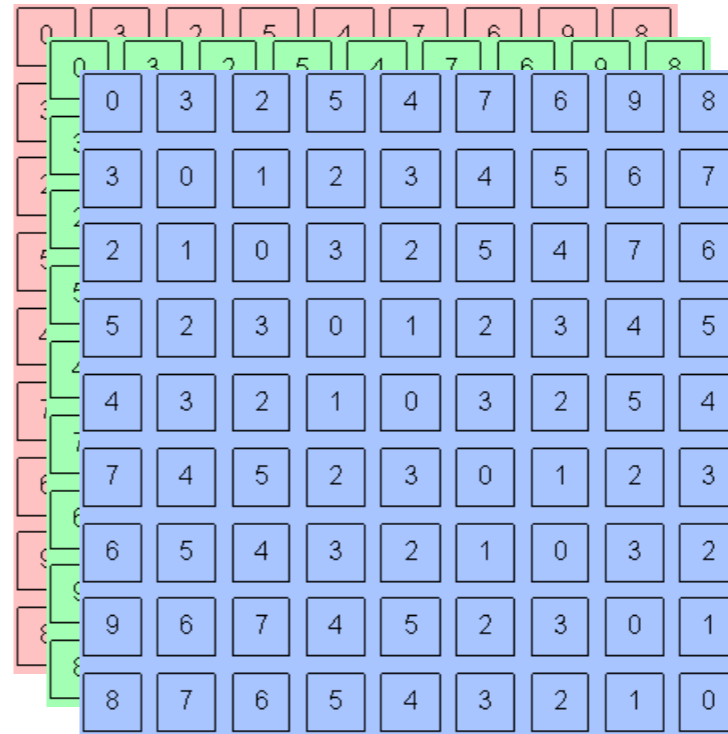
grayscale images are matrices



0	3	2	5	4	7	6	9	8
3	0	1	2	3	4	5	6	7
2	1	0	3	2	5	4	7	6
5	2	3	0	1	2	3	4	5
4	3	2	1	0	3	2	5	4
7	4	5	2	3	0	1	2	3
6	5	4	3	2	1	0	3	2
9	6	7	4	5	2	3	0	1
8	7	6	5	4	3	2	1	0

what range of values can each pixel take?

color images are tensors

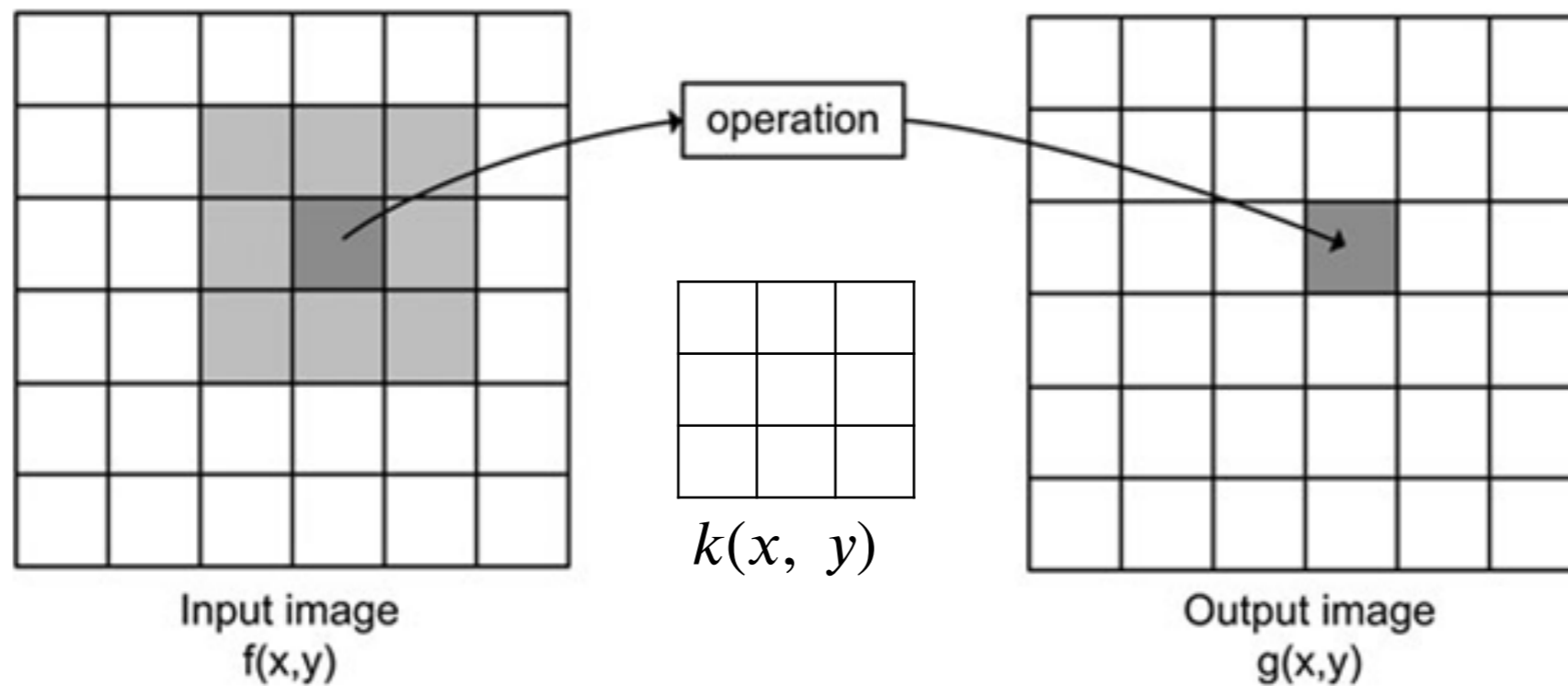


channel x height x width

Channels are usually RGB: Red, Green, and Blue

Other color spaces: HSV, HSL, LUV, XYZ, Lab, CMYK, etc

Convolution operator



$$g(x, y) = \sum_v \sum_u k(u, v) f(x - u, y - v)$$

(filter, kernel)

Input image

*

Weights



Output image

4	5	7	6	6
3	2	8	0	7
6	7	7	1	5
3	0	1	1	1
4	3	2	1	7

*

0	0	0
1	0	1
0	0	0

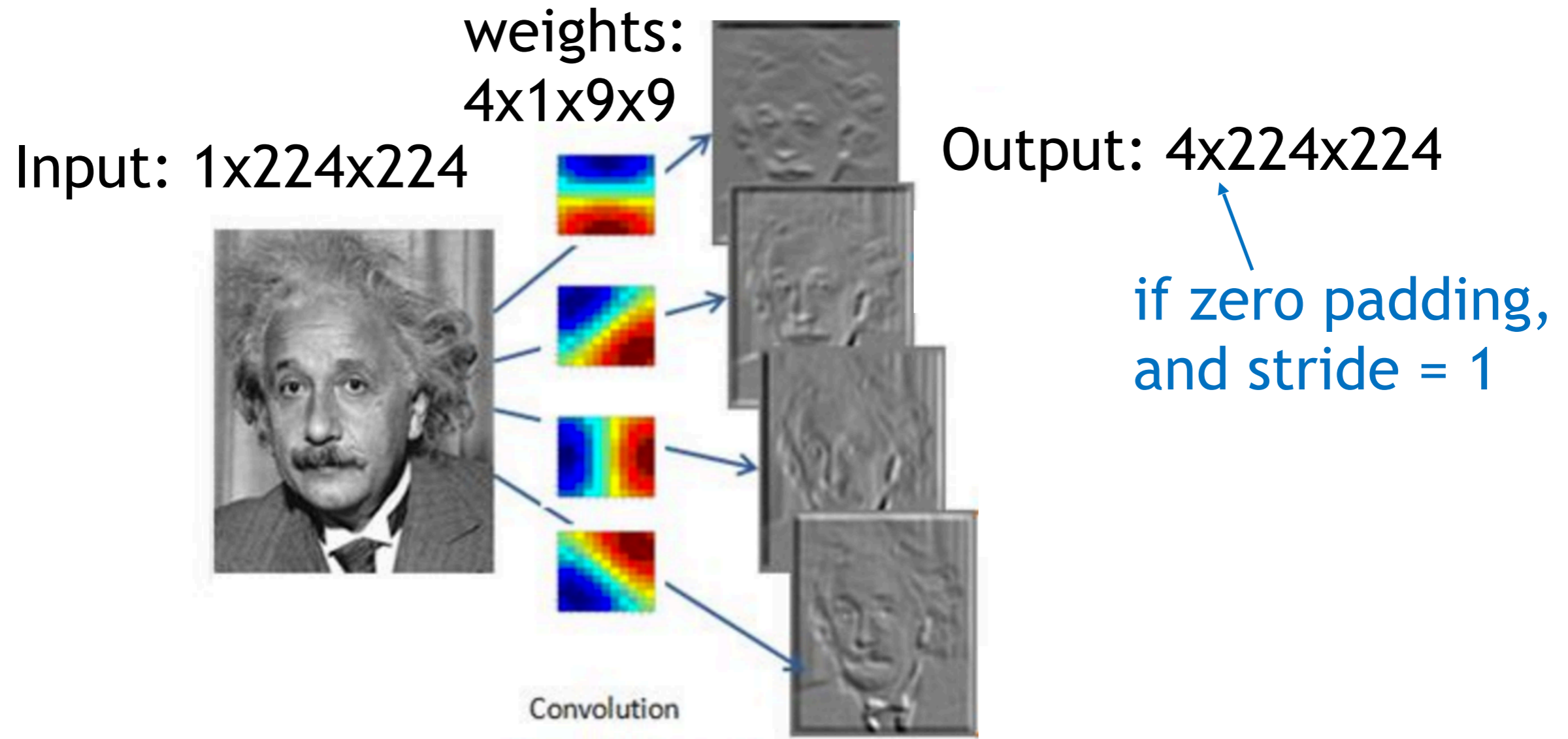


	11	2	15	
	13	8	12	
	?			

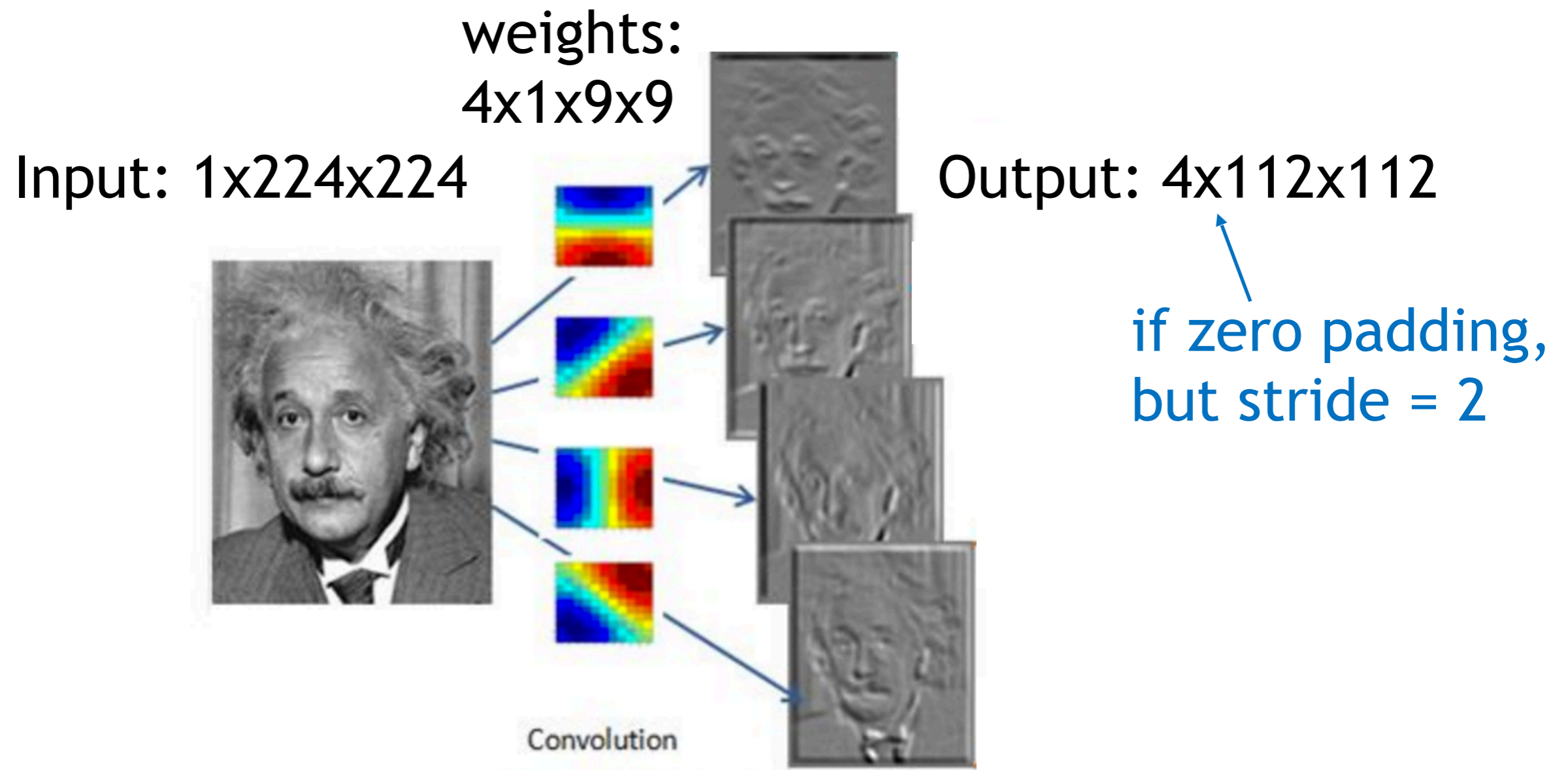
demo:

<http://setosa.io/ev/image-kernels/>

Convolutional Layer (with 4 filters)

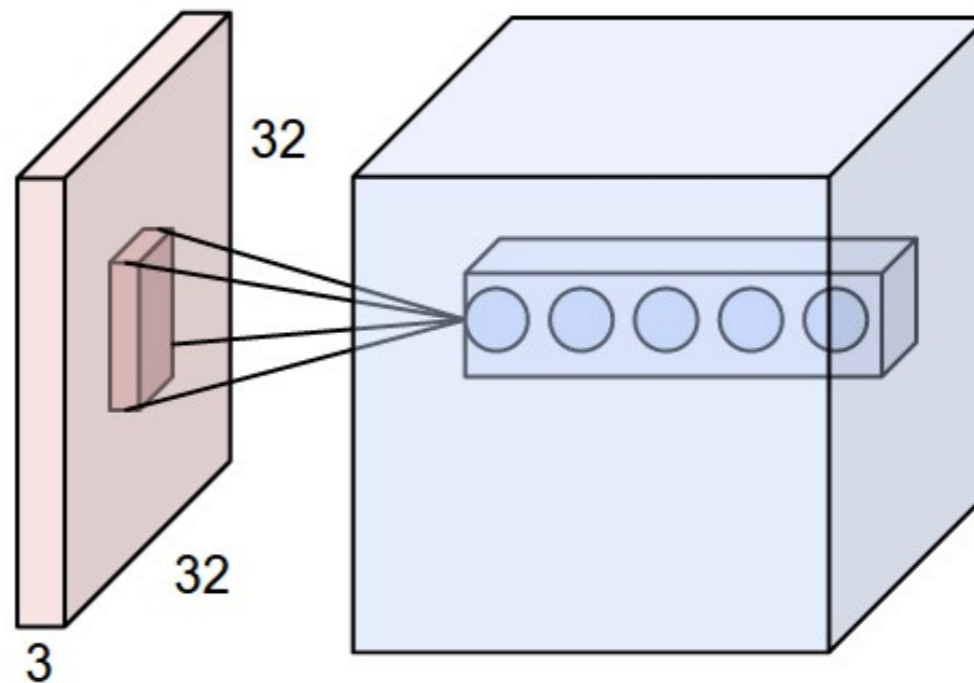


Convolutional Layer (with 4 filters)

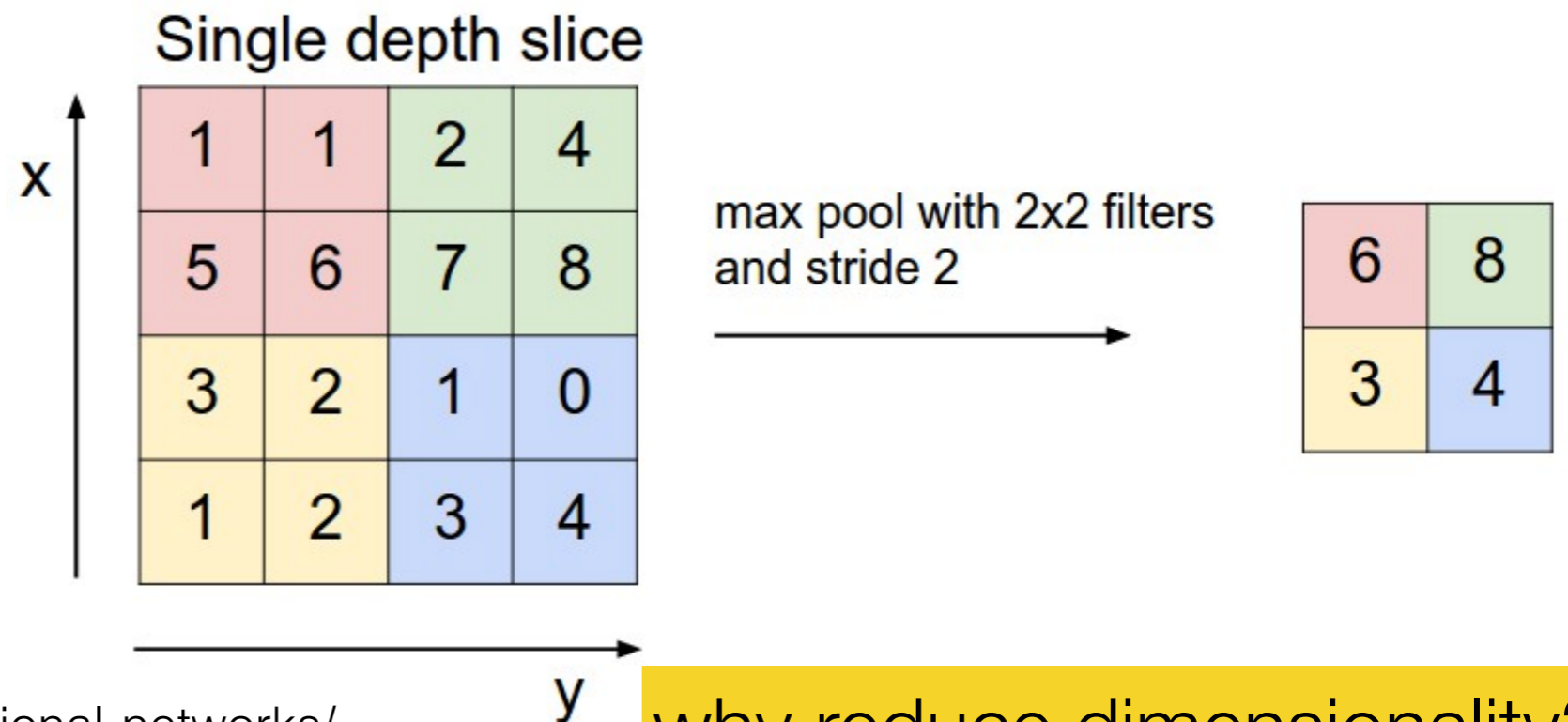


pooling layers also used to reduce dimensionality

Convolutional Layers:
slide a set of small filters over the image



Pooling Layers:
reduce dimensionality of representation



Alexnet

ImageNet Classification with Deep Convolutional Neural Networks

Alex Krizhevsky
University of Toronto
kriz@cs.utoronto.ca

Ilya Sutskever
University of Toronto
ilya@cs.utoronto.ca

Geoffrey E. Hinton
University of Toronto
hinton@cs.utoronto.ca

the paper that started the
deep learning revolution!

image classification

Classify an image into 1000 possible classes:

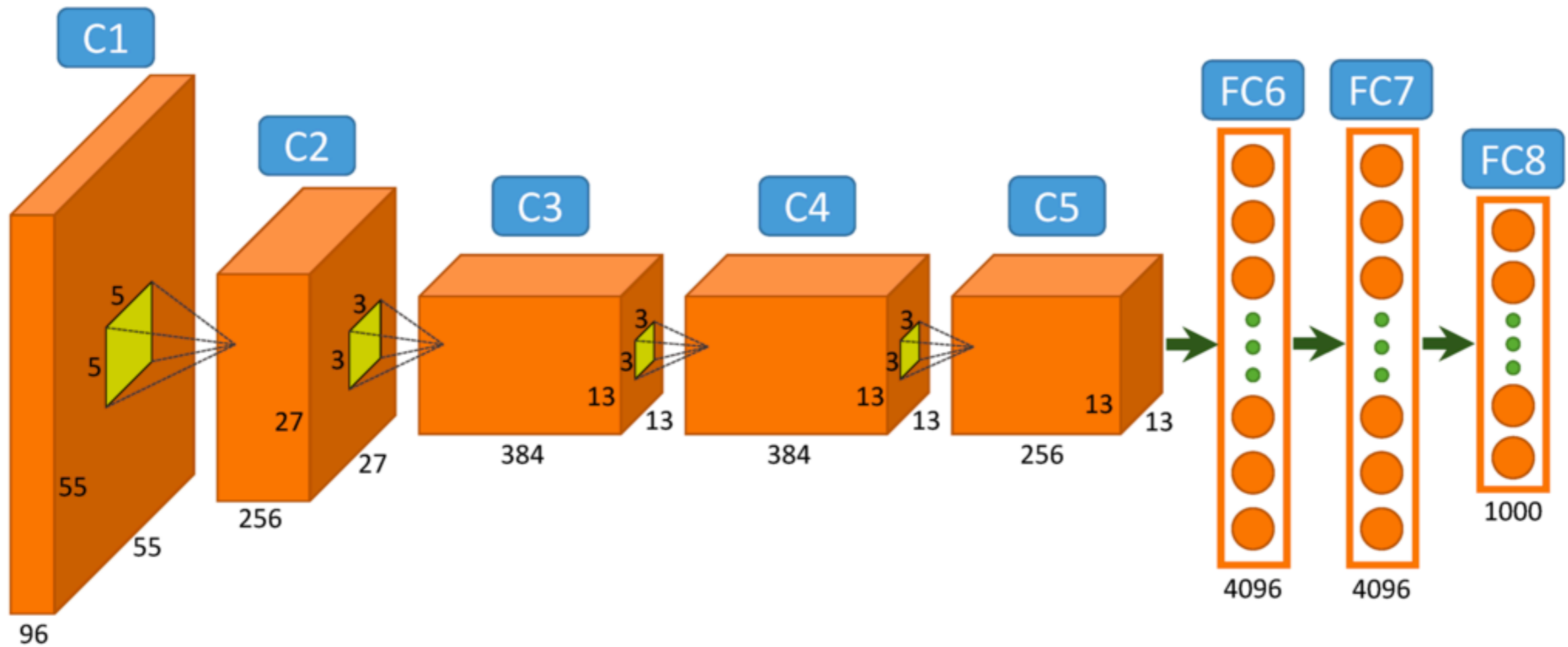
e.g. Abyssinian cat, Bulldog, French Terrier, Cormorant,
Chickadee,
red fox, banjo, barbell, hourglass, knot, maze, viaduct, etc.



cat, tabby cat (0.71)
Egyptian cat (0.22)
red fox (0.11)
.....

train on the ImageNet
challenge dataset,
~1.2 million images

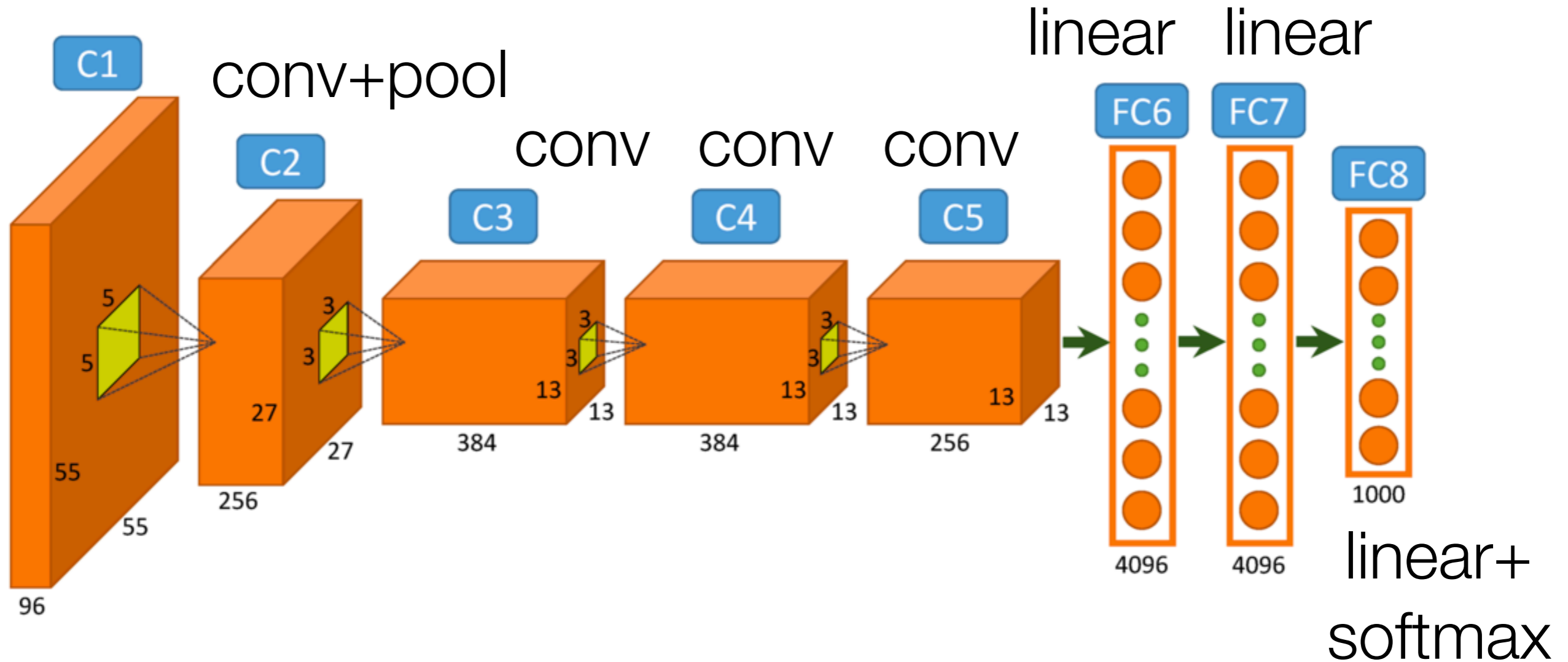
Alexnet



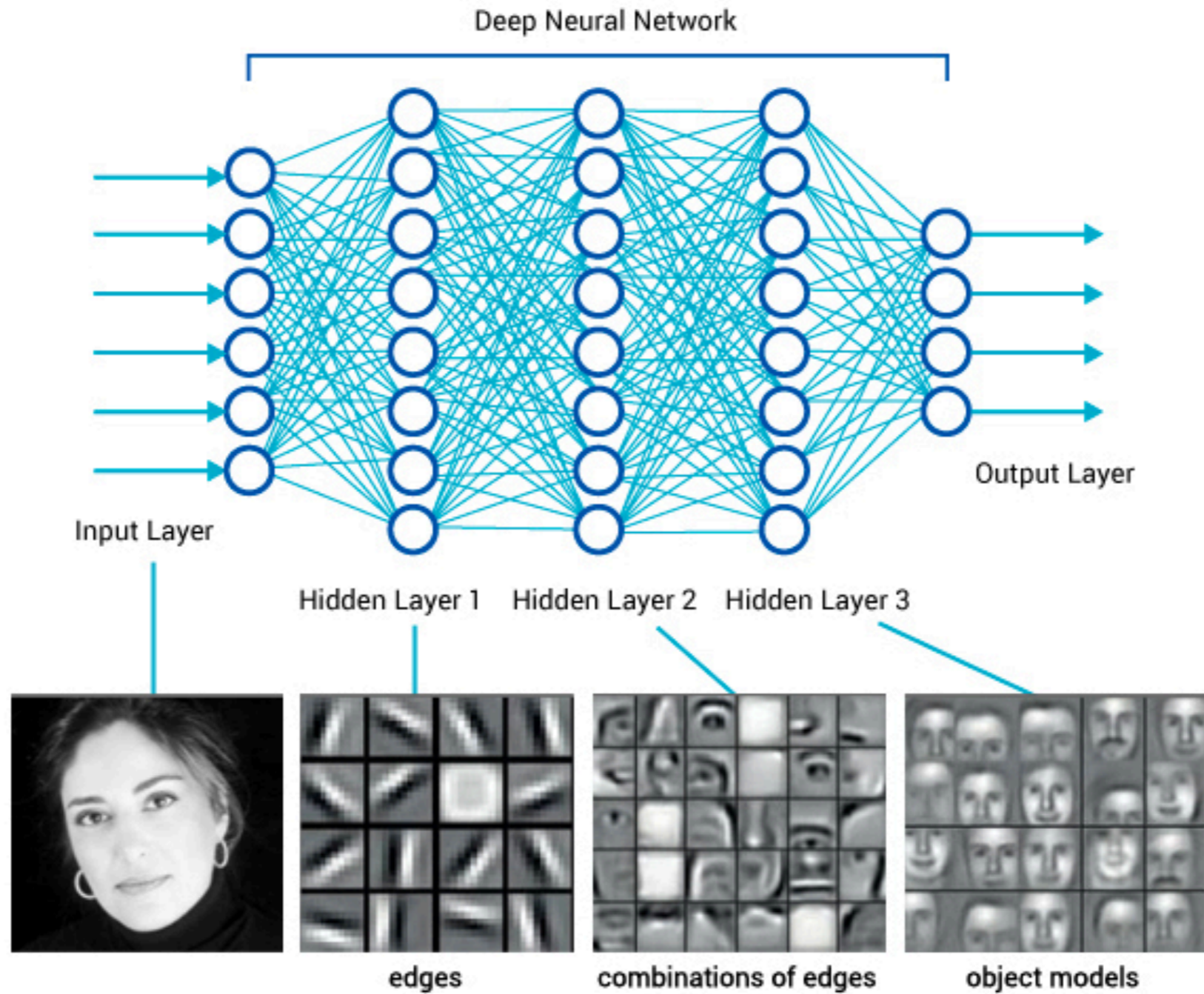
<https://www.saagie.com/fr/blog/object-detection-part1>

Alexnet

conv+pool



What is happening?



Revolution of Depth

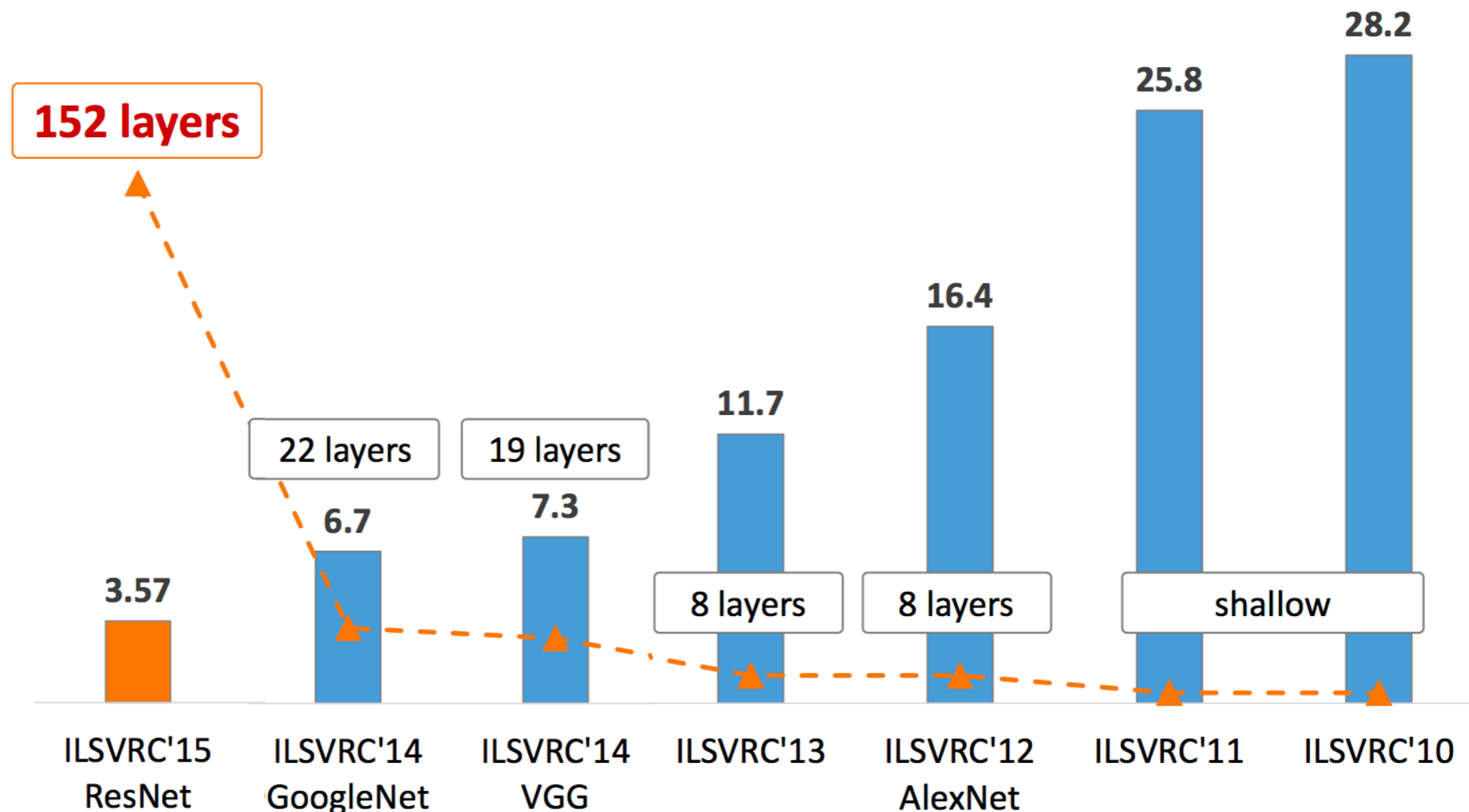
AlexNet, 8 layers
(ILSVRC 2012)



VGG, 19 layers
(ILSVRC 2014)

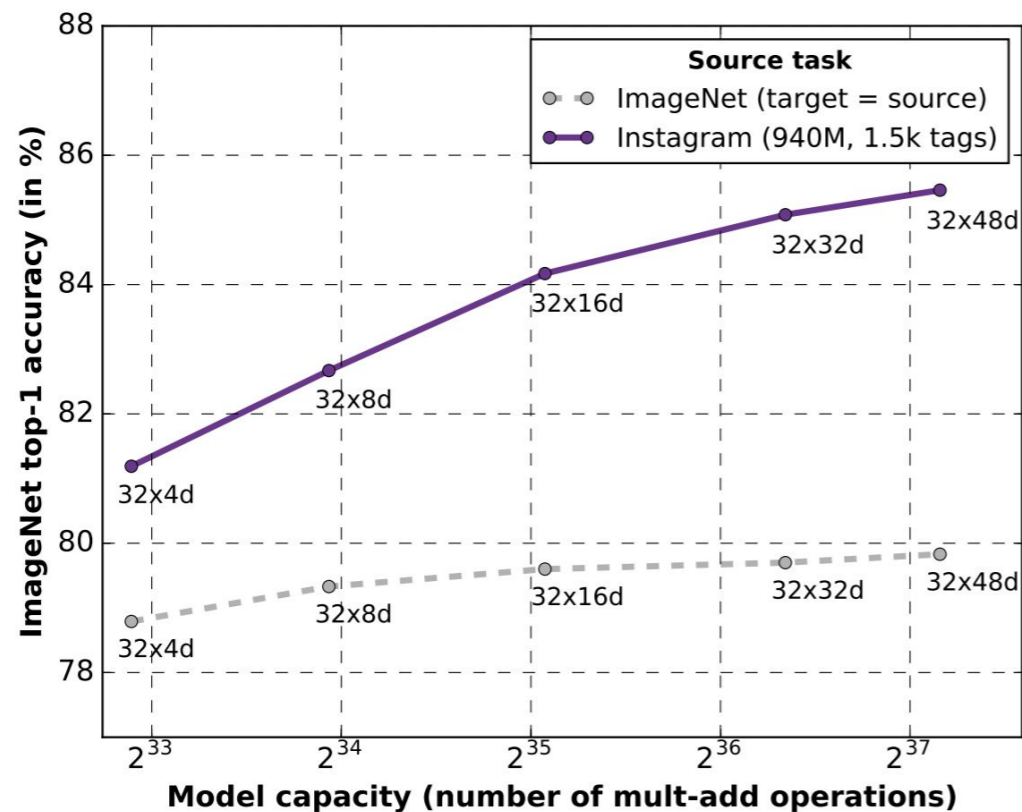


ResNet, **152 layers**
(ILSVRC 2015)



ImageNet pretraining -> Instagram pretraining

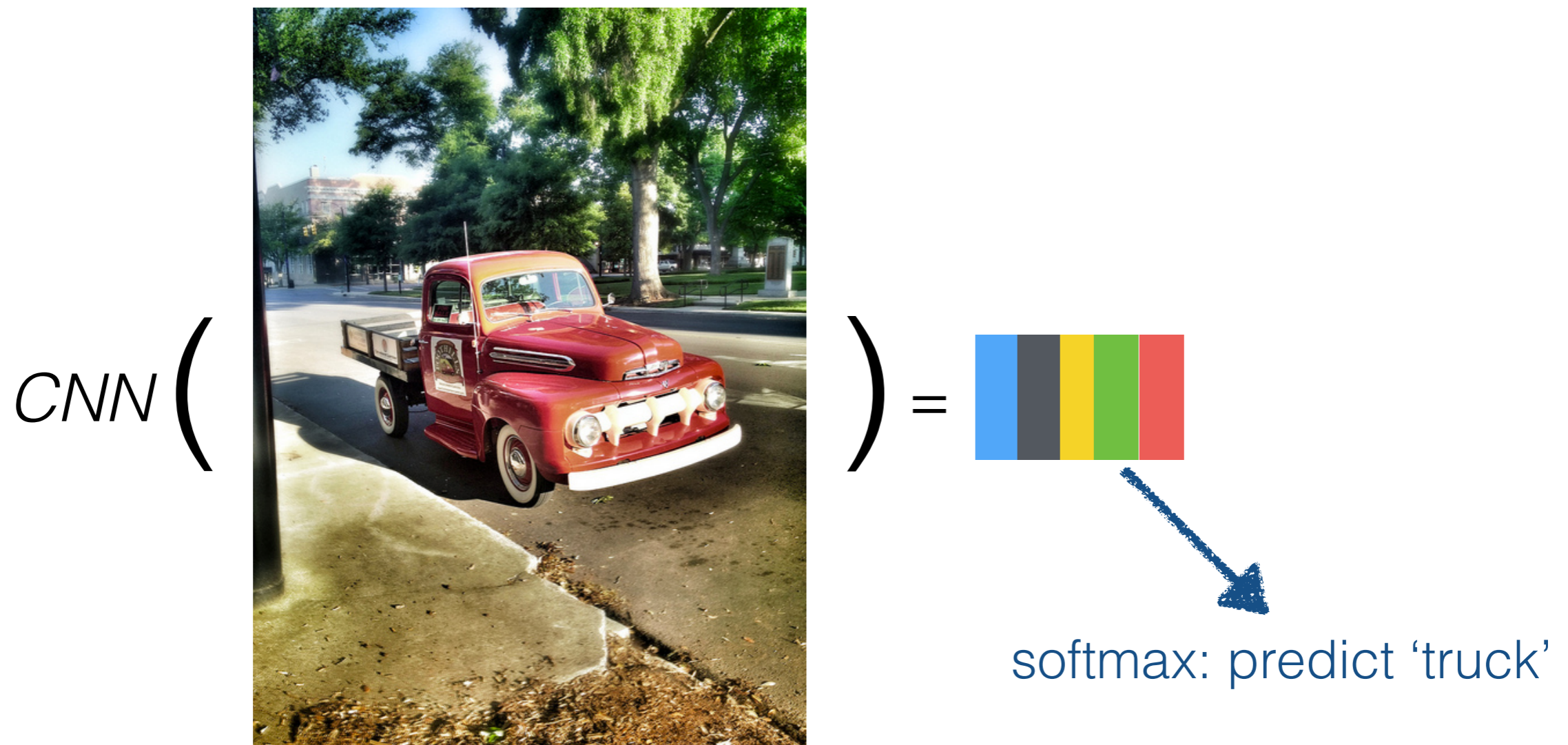
Bigger models are saturated on ImageNet, but with more data bigger models do better



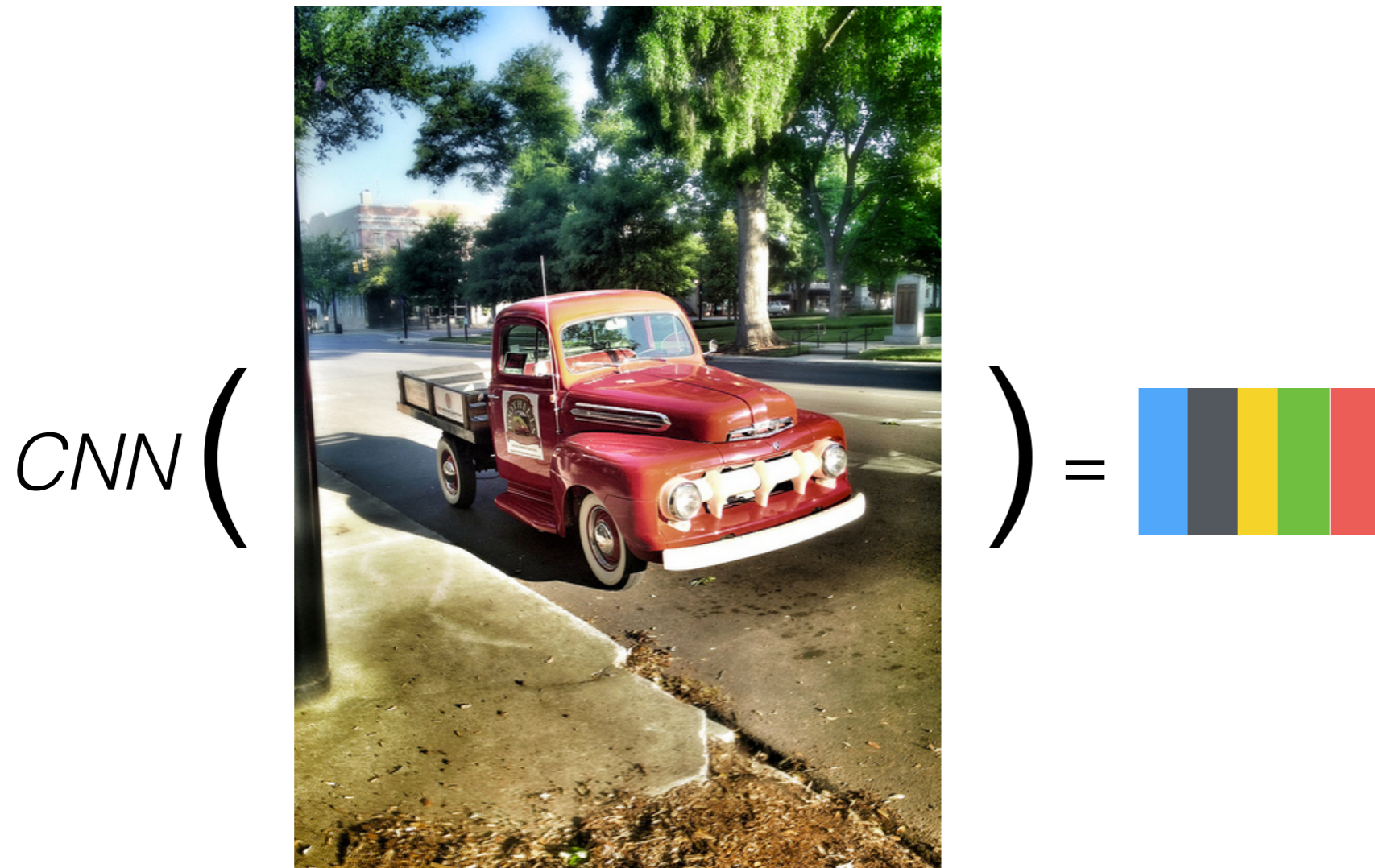
Biggest network was pretrained on 3.5B Instagram images

Trained on 336 GPUs for 22 days

at the end of the day, we generate a fixed size vector from an image and run a classifier over it



key insight: this vector is useful for many more tasks than just image classification!
we can use it for *transfer learning*



simple visual QA

- $i = \text{CNN}(\text{image})$ > use an existing network trained for image classification and freeze weights
- $q = \text{RNN}(\text{question})$ > learn weights
- $\text{answer} = \text{softmax}(\text{linear}([i; q]))$

why isn't this a good way of doing visual QA?



How many benches are shown?

visual attention

- Use the question representation q to determine where in the image to look

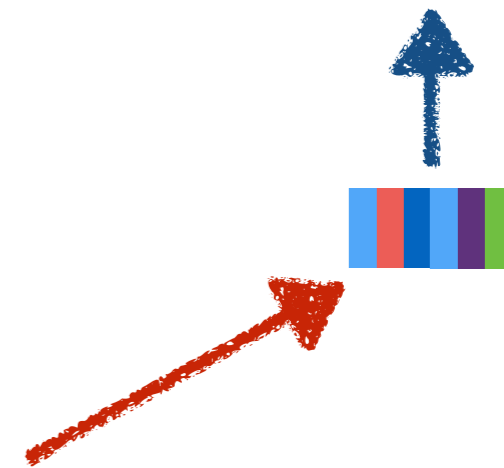
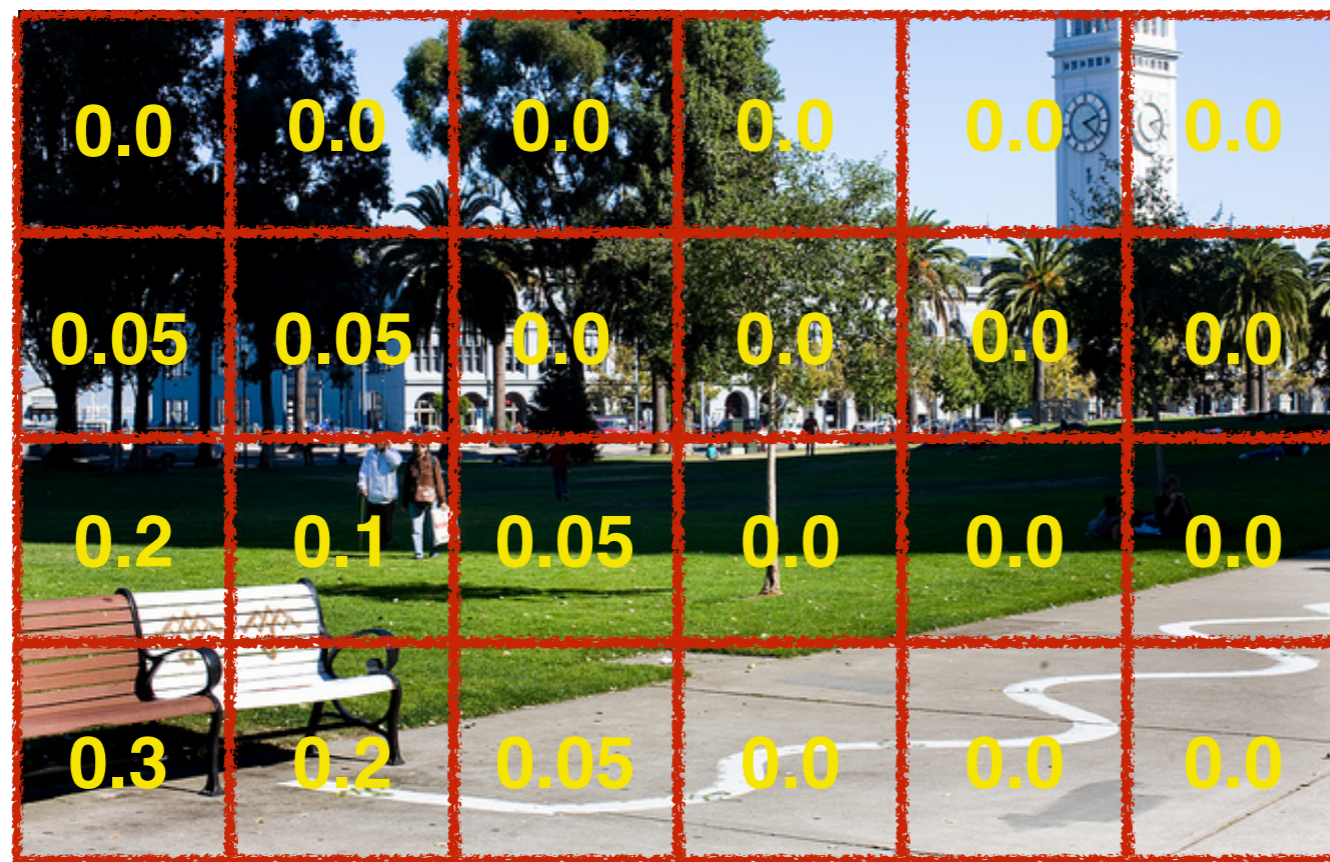


How many benches are shown?

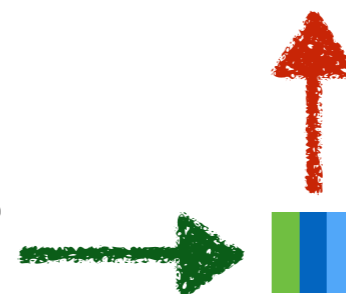


softmax:
predict answer

attention over final convolutional
layer in network: 196 boxes, captures
color and positional information

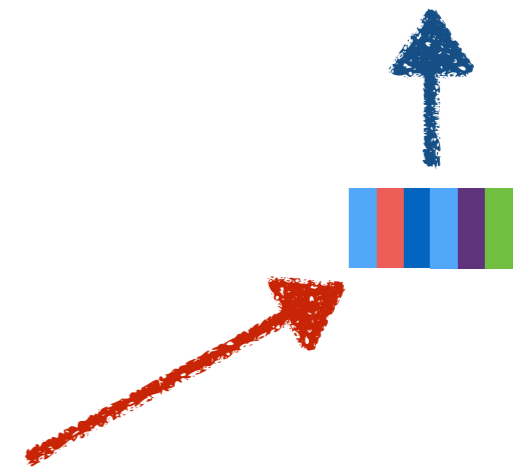
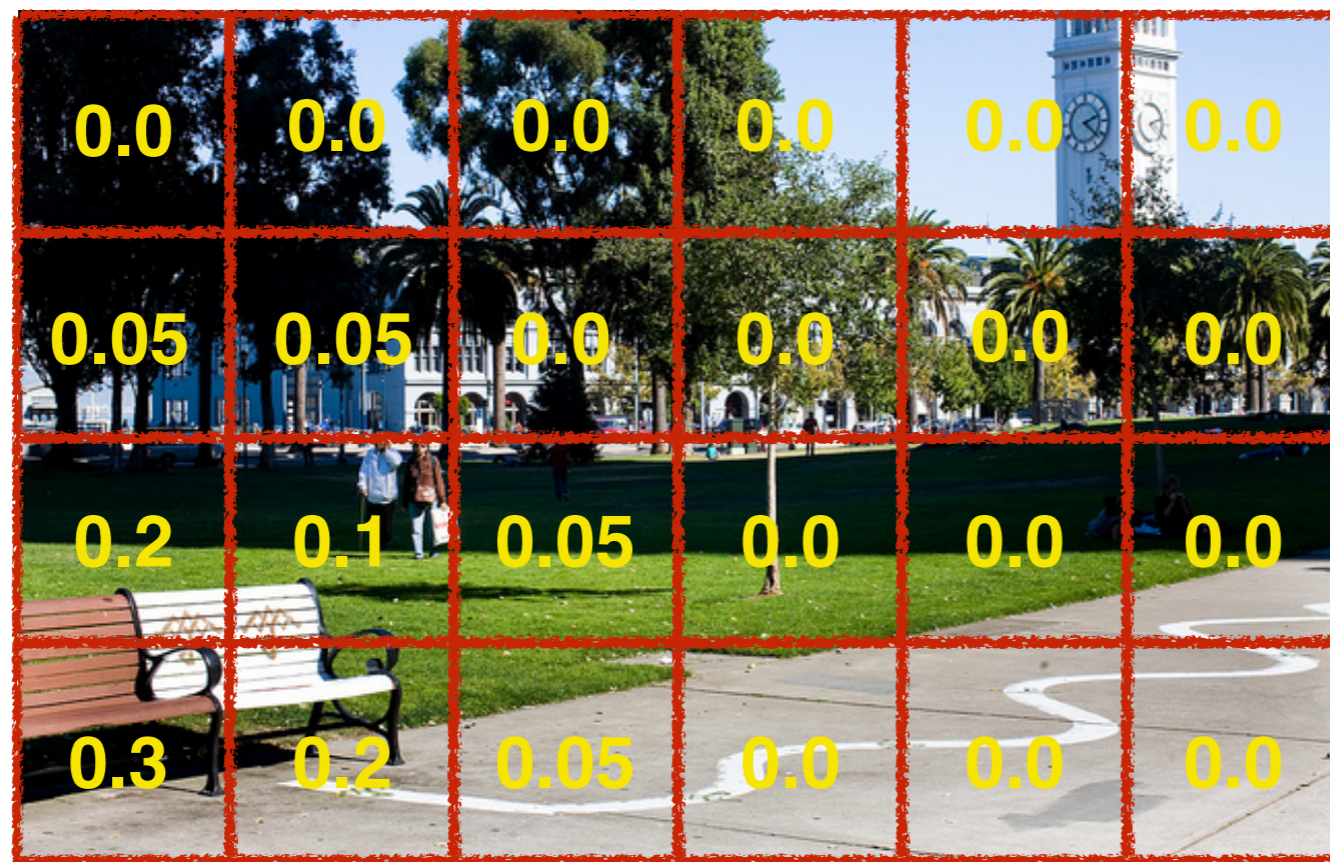


How many benches are shown?



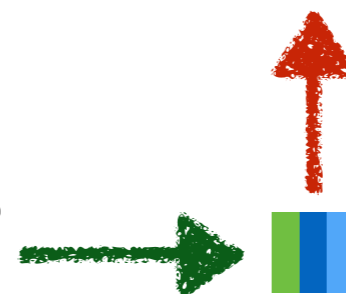
softmax:
predict answer

attention over final convolutional
layer in network: 196 boxes, captures
color and positional information



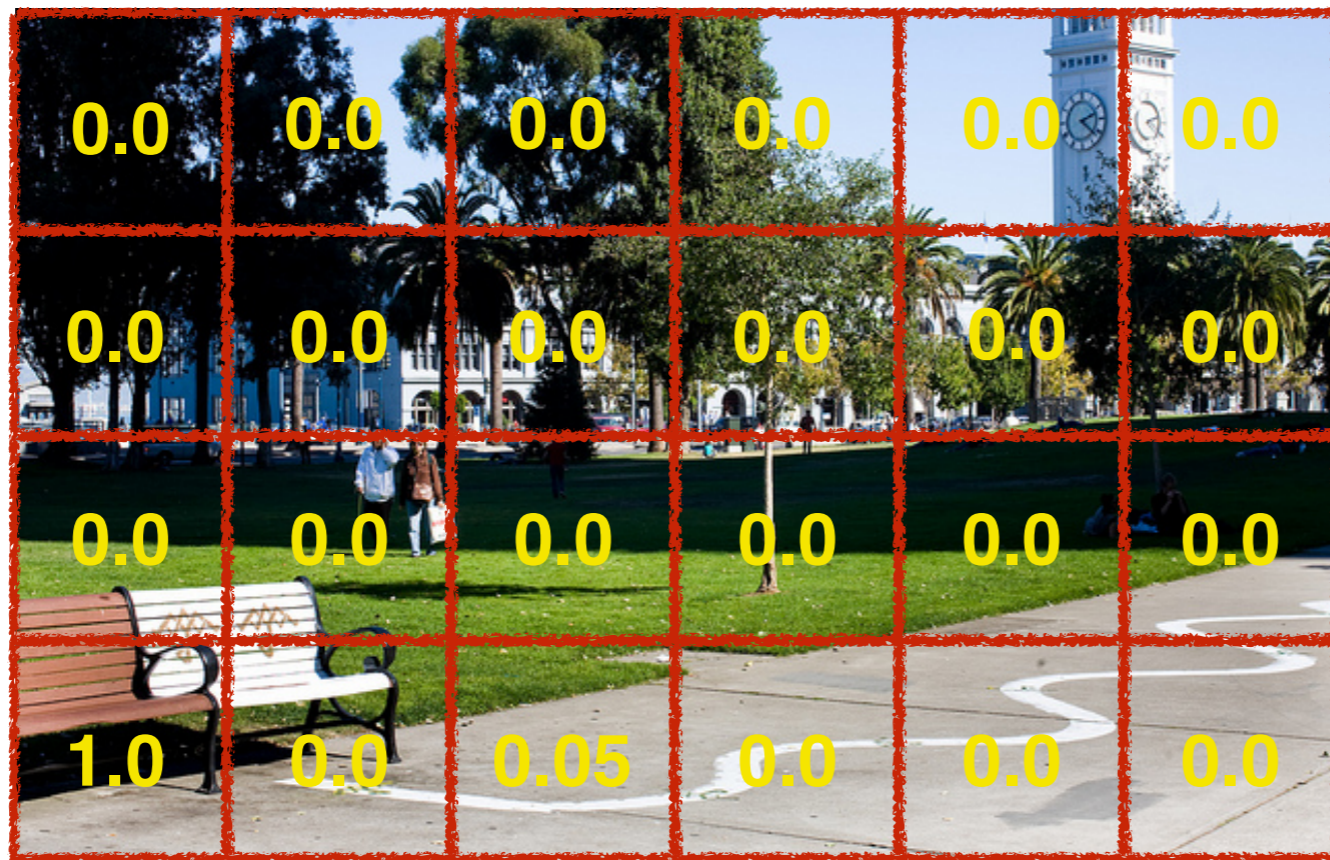
how can we
compute these
attention
scores?

How many benches are shown?

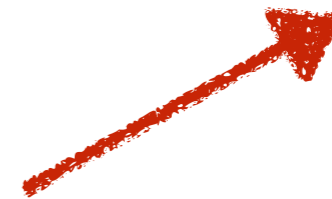
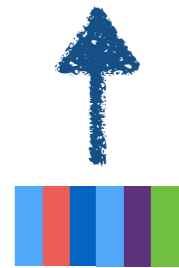


hard attention

attention over final convolutional layer in network: 196 boxes, captures color and positional information

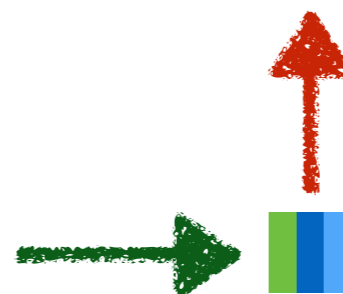


softmax:
predict answer



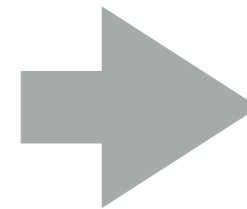
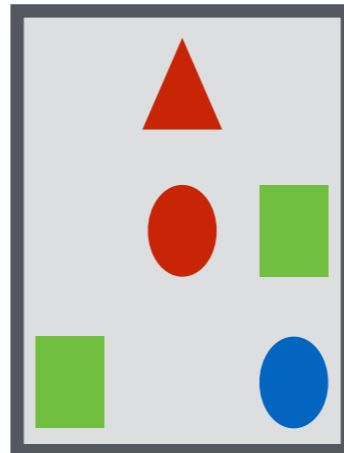
we can use *reinforcement learning* to focus on just one box

How many benches are shown?



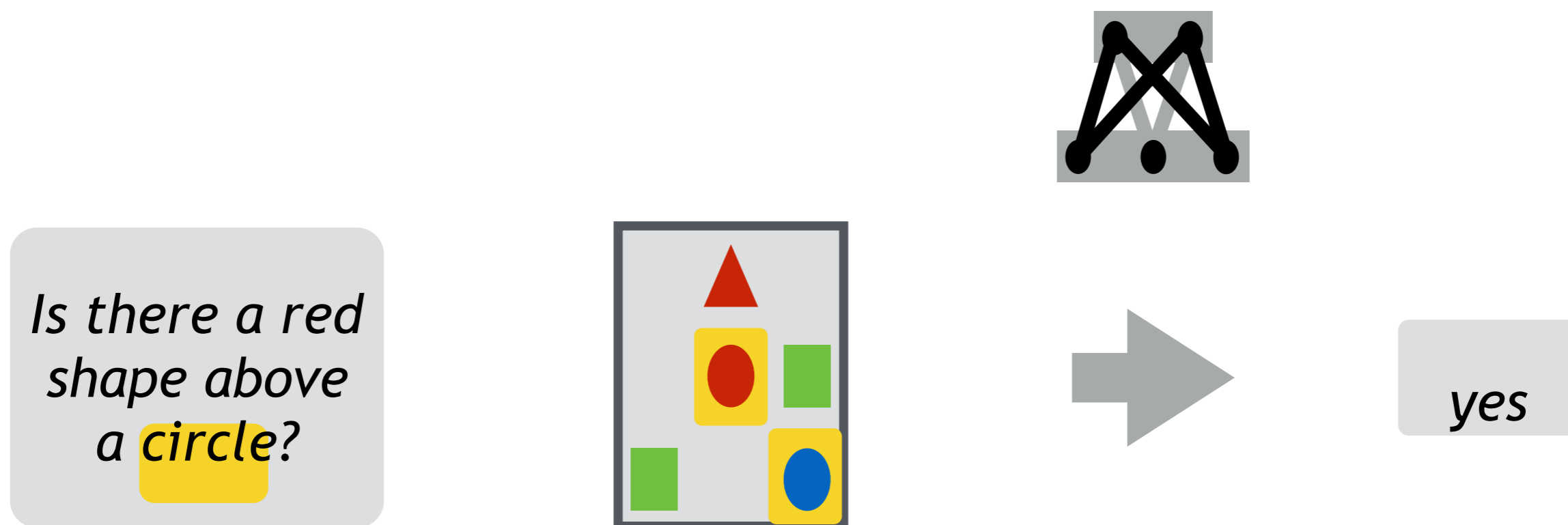
Grounded question answering

*Is there a red
shape above
a circle?*



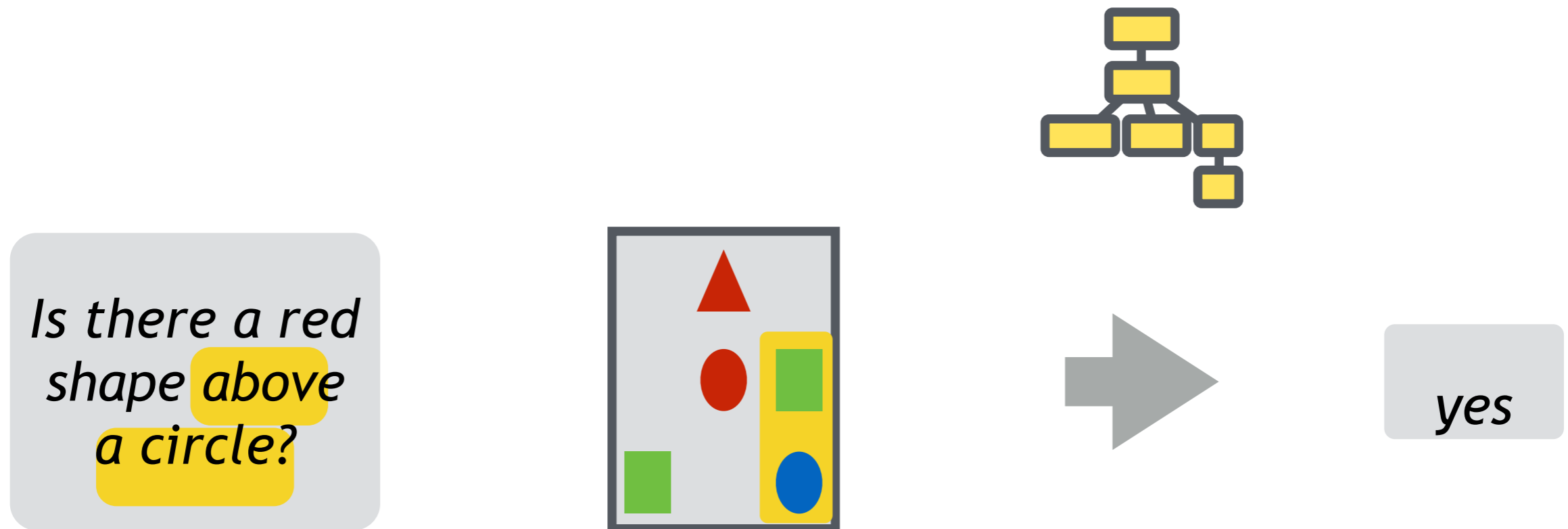
yes

Neural nets learn lexical groundings



[Iyyer et al. 2014, Bordes et al. 2014,
Yang et al. 2015, Malinowski et al., 2015]

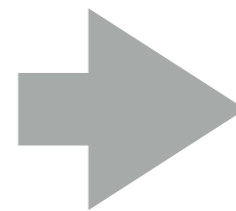
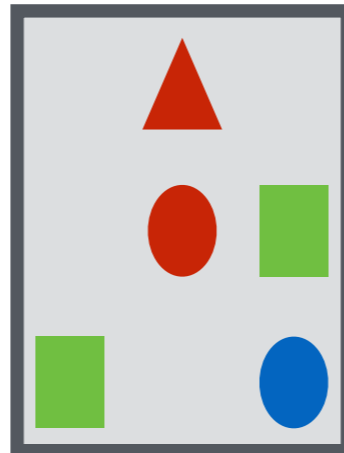
Semantic parsers learn composition



[Wong & Mooney 2007, Kwiatkowski et al. 2010, Liang et al. 2011, A et al. 2013]

Neural module networks learn both!

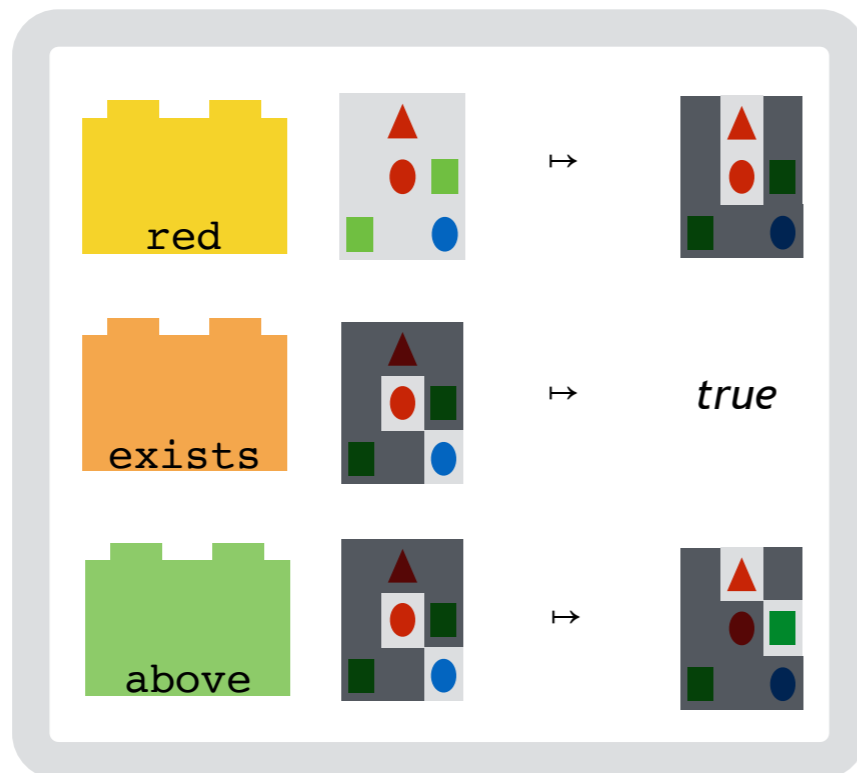
Is there a red shape above a circle?



yes

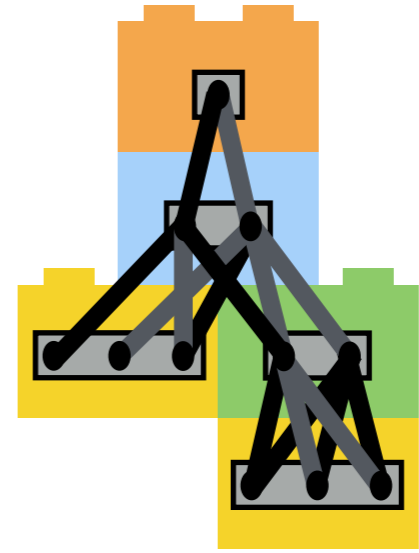
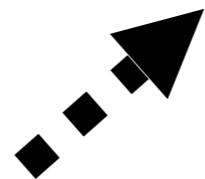
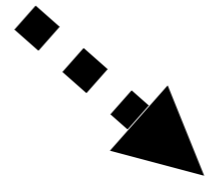
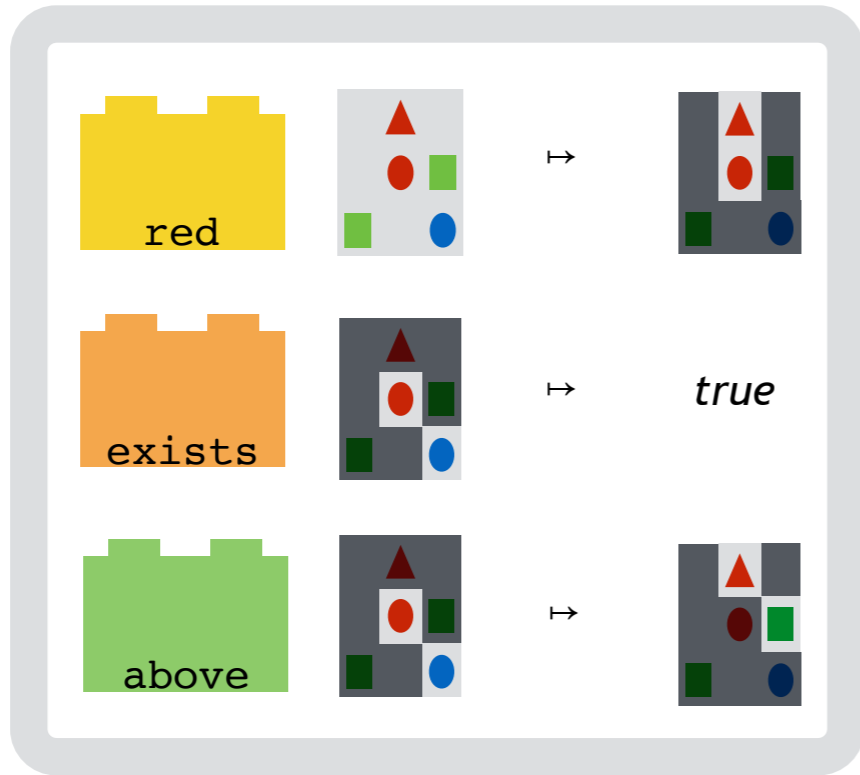
Neural module networks

*Is there a red shape
above a circle?*

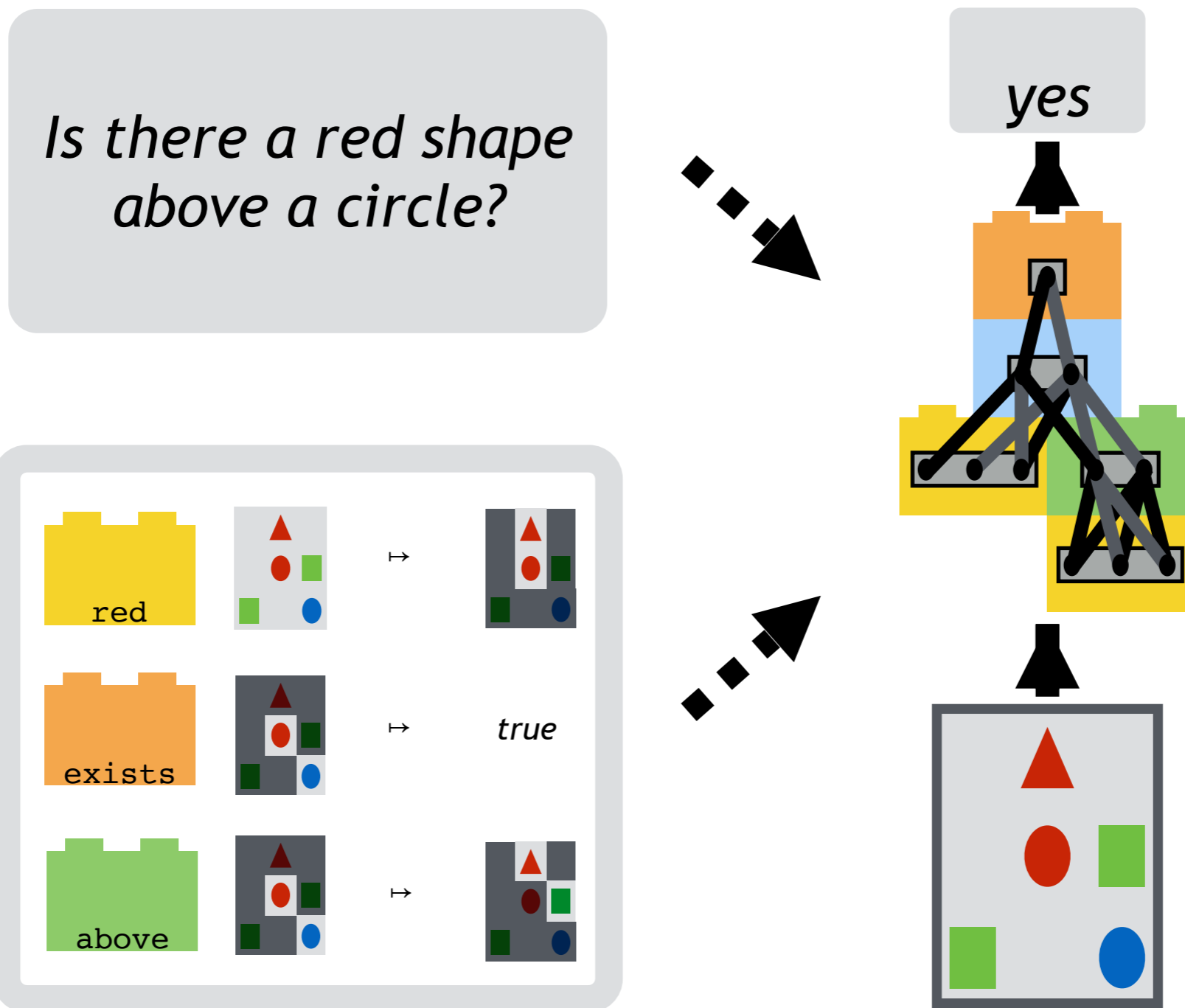


Neural module networks

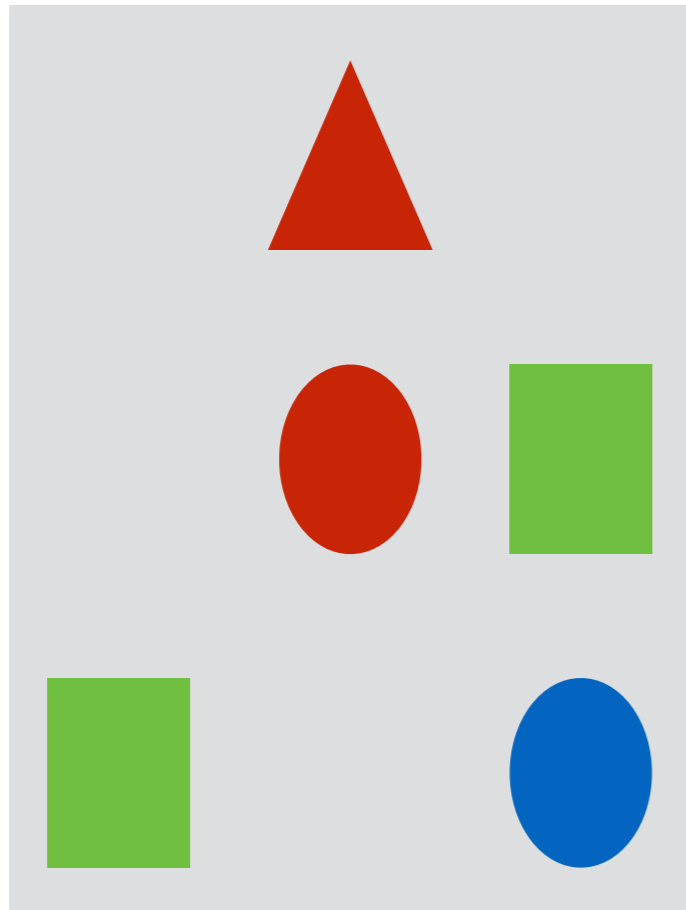
*Is there a red shape
above a circle?*



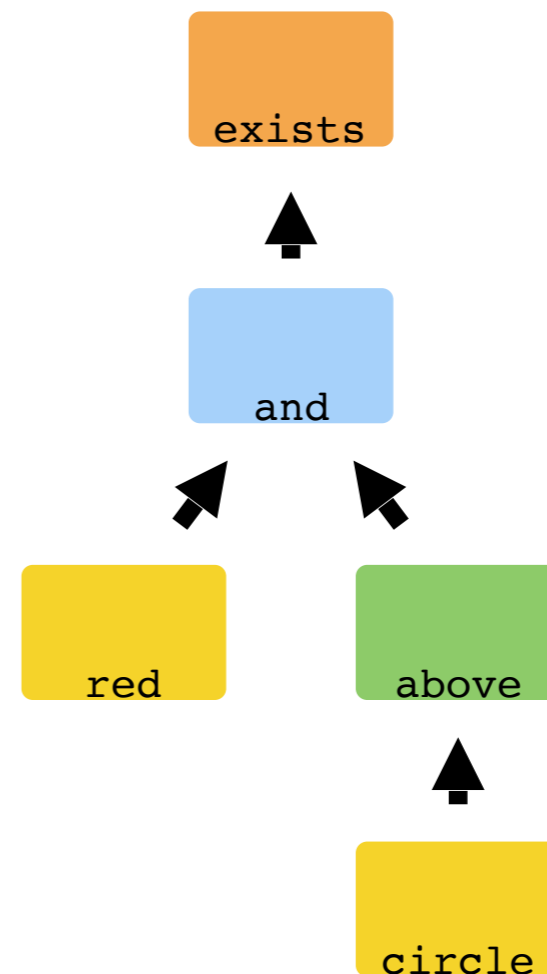
Neural module networks



Sentence meanings are computations



Is there a red shape above a circle?



NLVR²: natural language for visual reasoning! (Suhr et al., 2018)



TRUE OR FALSE: the left image contains twice the number of dogs as the right image, and at least two dogs in total are standing.

image captioning

Image Captioning

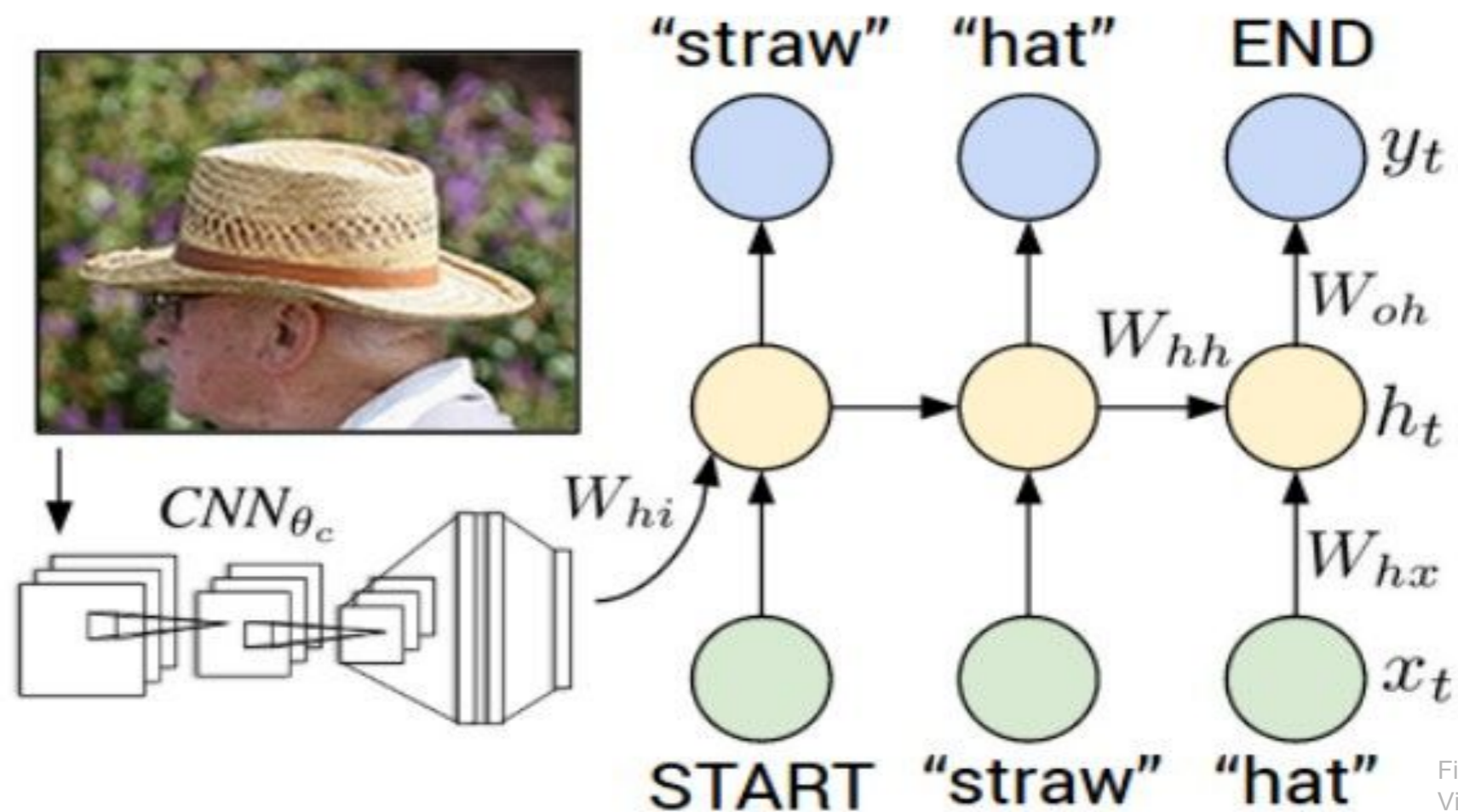


Figure from
Visual-Semantic
Image Description
copyright
Reproduced

Explain Images with Multimodal Recurrent Neural Networks, Mao et al.

Deep Visual-Semantic Alignments for Generating Image Descriptions, Karpathy and Fei-Fei

Show and Tell: A Neural Image Caption Generator, Vinyals et al.

Long-term Recurrent Convolutional Networks for Visual Recognition and Description, Donahue et al.

Learning a Recurrent Visual Representation for Image Caption Generation, Chen and Zitnick

test image



This image is CC0 public domain



test image

this is our ImageNet CNN, now used as a feature extractor

image

conv-64

conv-64

maxpool

conv-128

conv-128

maxpool

conv-256

conv-256

maxpool

conv-512

conv-512

maxpool

conv-512

conv-512

maxpool

FC-4096

FC-4096

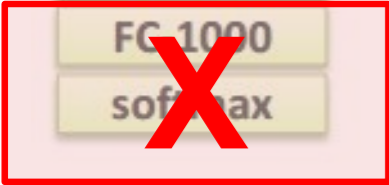
FC-1000

softmax



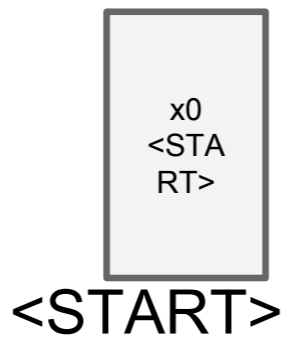
test image

this is our ImageNet CNN, now used as a feature extractor





test image



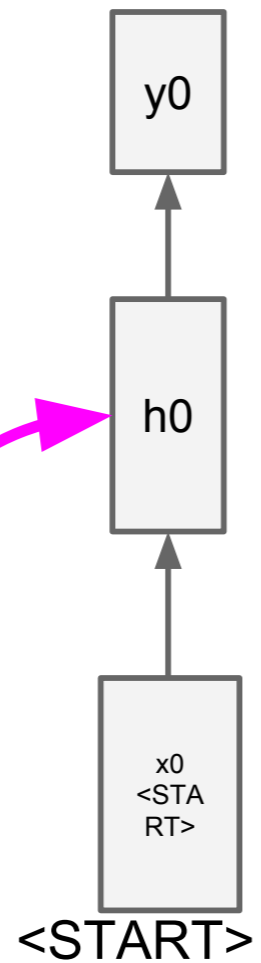


test image



v

Wih



before:

$$h = \tanh(W_{xh} * x + W_{hh} * h)$$

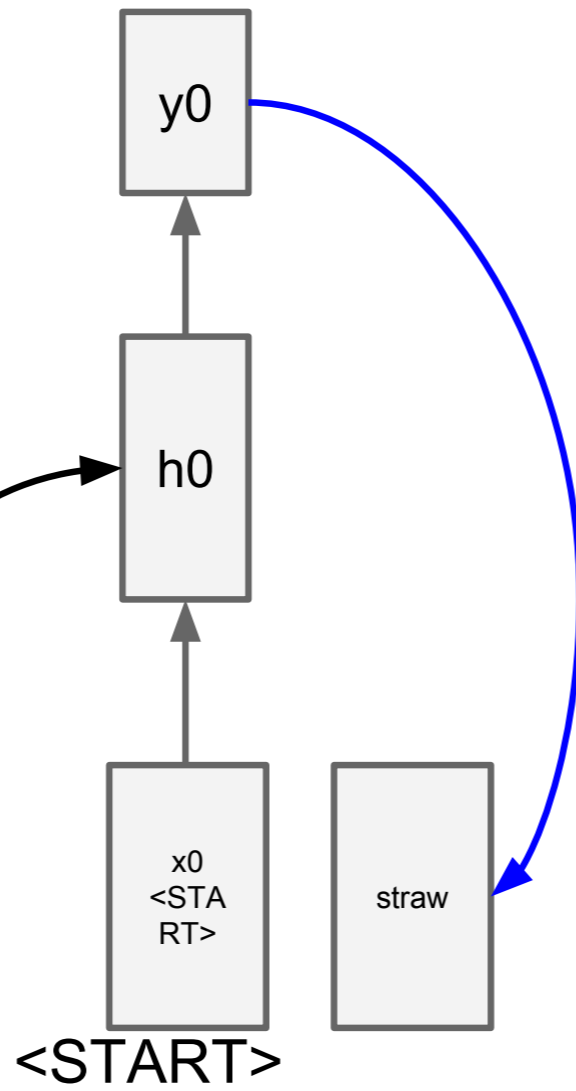
now:

$$h = \tanh(W_{xh} * x + W_{hh} * h + W_{ih} * v)$$

let's use the image features to create a conditional LM

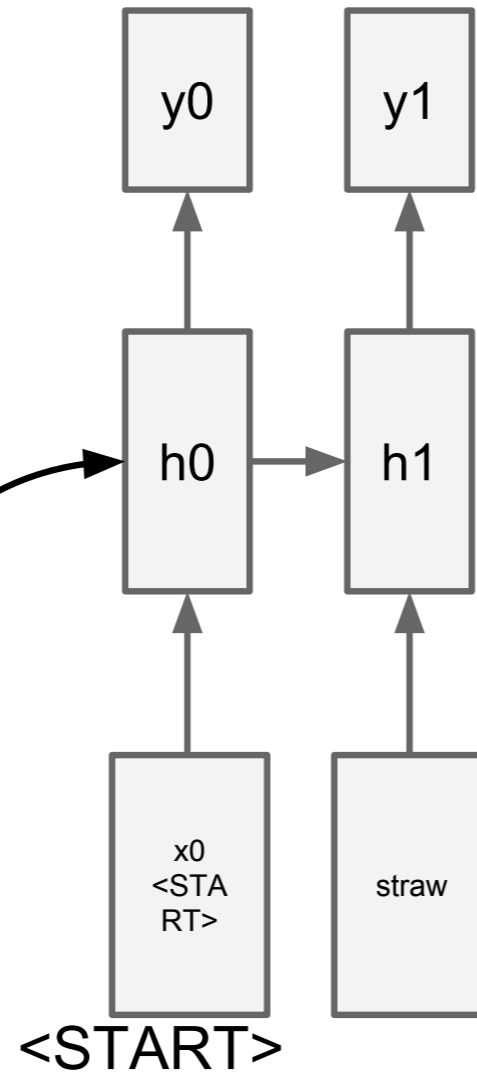


test image



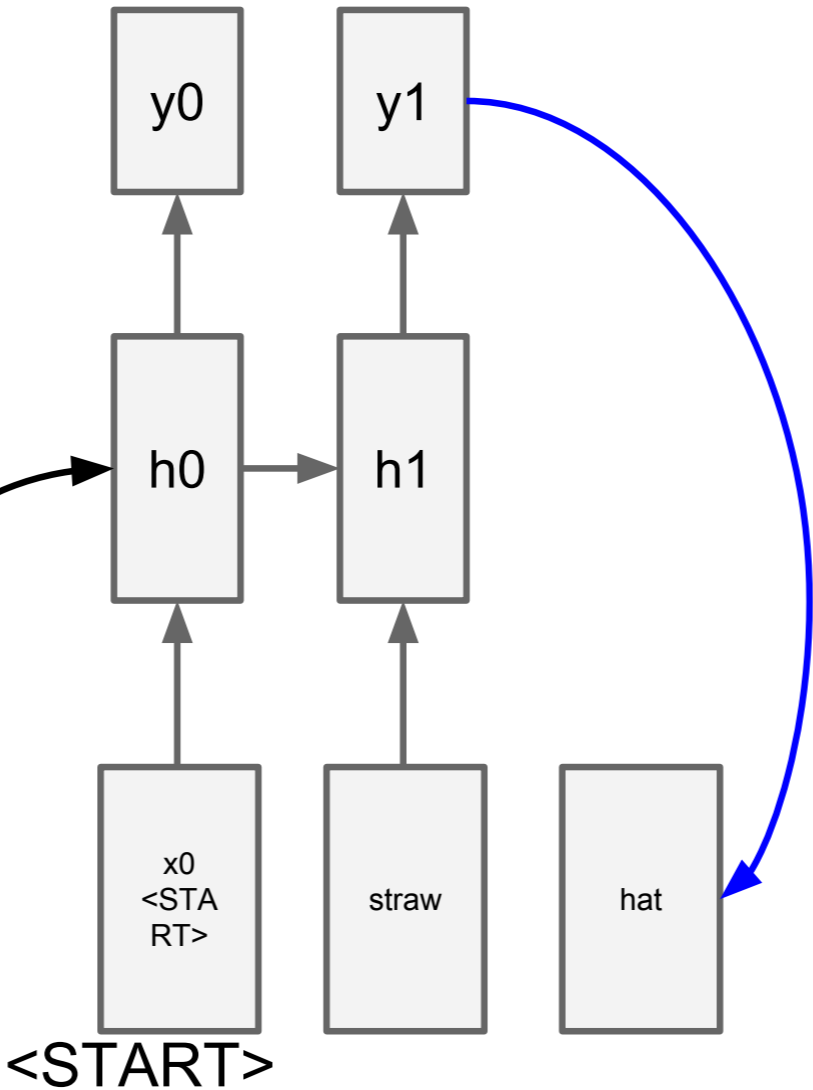
sample!

test image





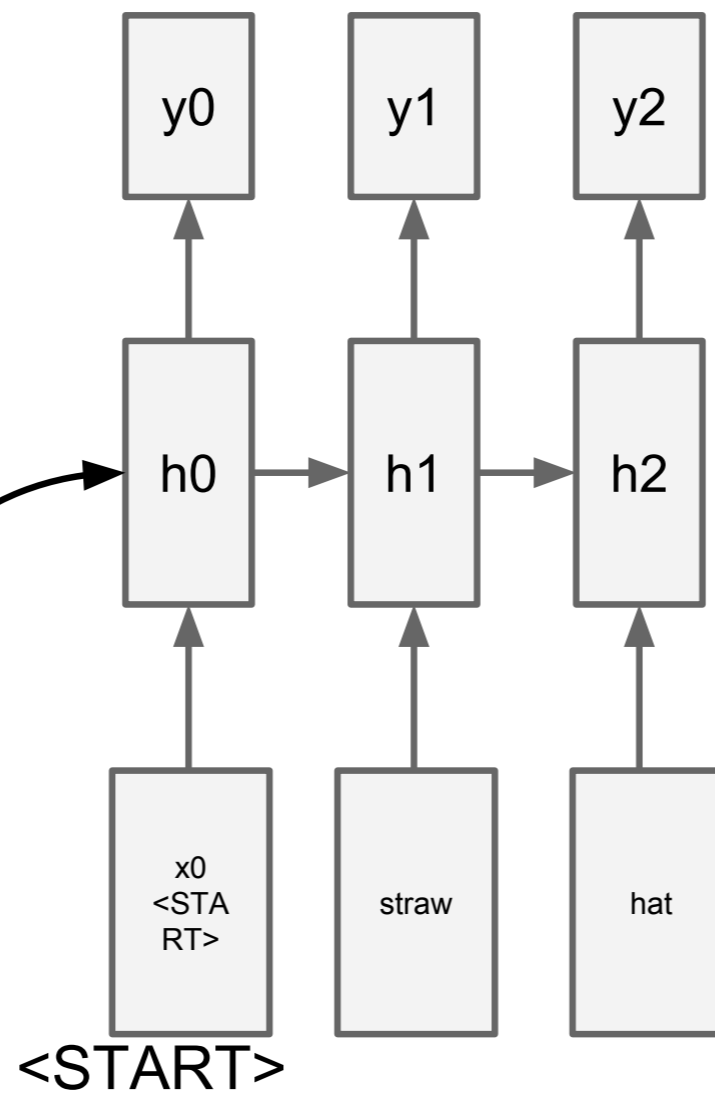
test image



sample!

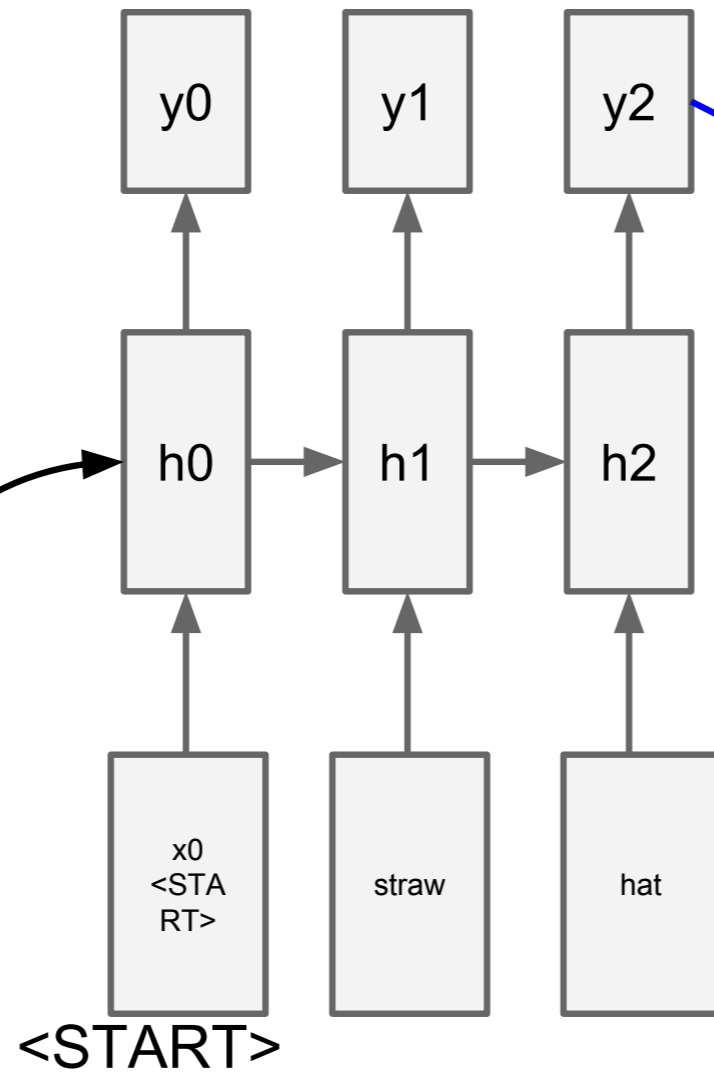


test image





test image



sample
<END> token
=> finish.

Image Captioning: Example Results

Captions generated using [neuraltalk2](#)
All images are [CC0 Public domain](#):
[cat suitcase](#), [cat tree](#), [dog](#), [bear](#),
[surfers](#), [tennis](#), [giraffe](#), [motorcycle](#)



A cat sitting on a suitcase on the floor



A cat is sitting on a tree branch



A dog is running in the grass with a frisbee



A white teddy bear sitting in the grass



Two people walking on the beach with surfboards



A tennis player in action on the court



Two giraffes standing in a grassy field



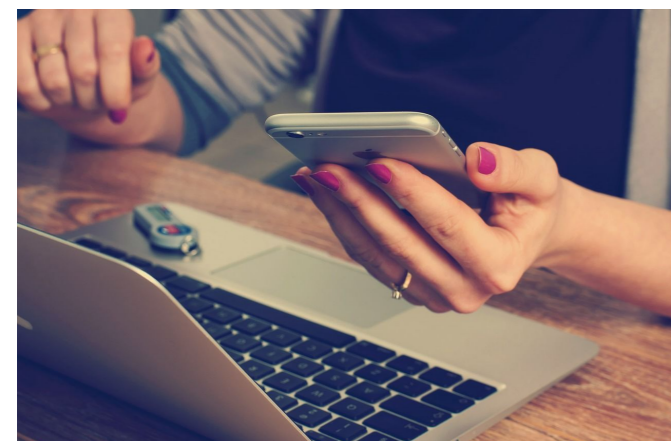
A man riding a dirt bike on a dirt track

Image Captioning: Failure Cases

Captions generated using [neuraltalk2](#)
All images are [CC0 Public domain](#): [fur coat](#), [handstand](#), [spider web](#), [baseball](#)



A woman is holding a cat in her hand



A person holding a computer mouse on a desk



A woman standing on a beach holding a surfboard



A bird is perched on a tree branch



A man in a baseball uniform throwing a ball

Image Captioning with Attention

RNN focuses its attention at a different spatial location when generating each word

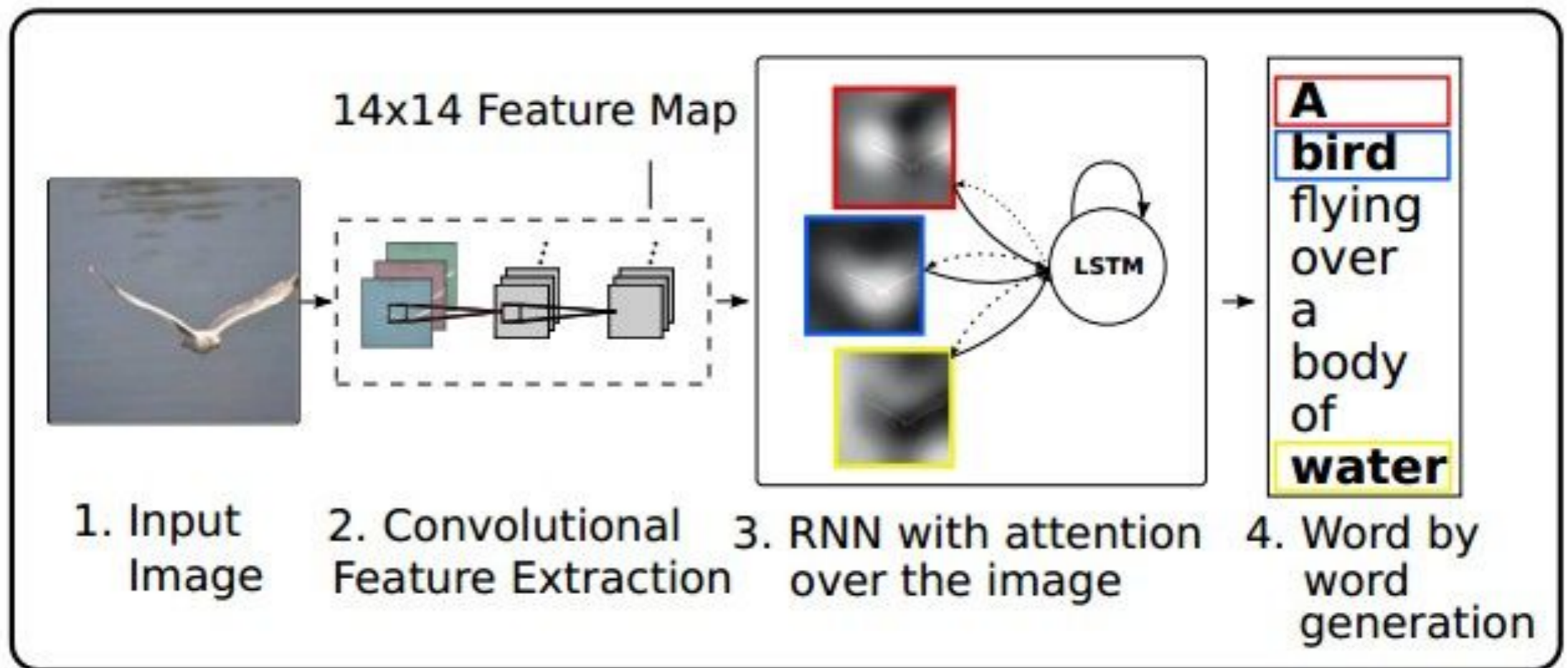


Image Captioning with Attention

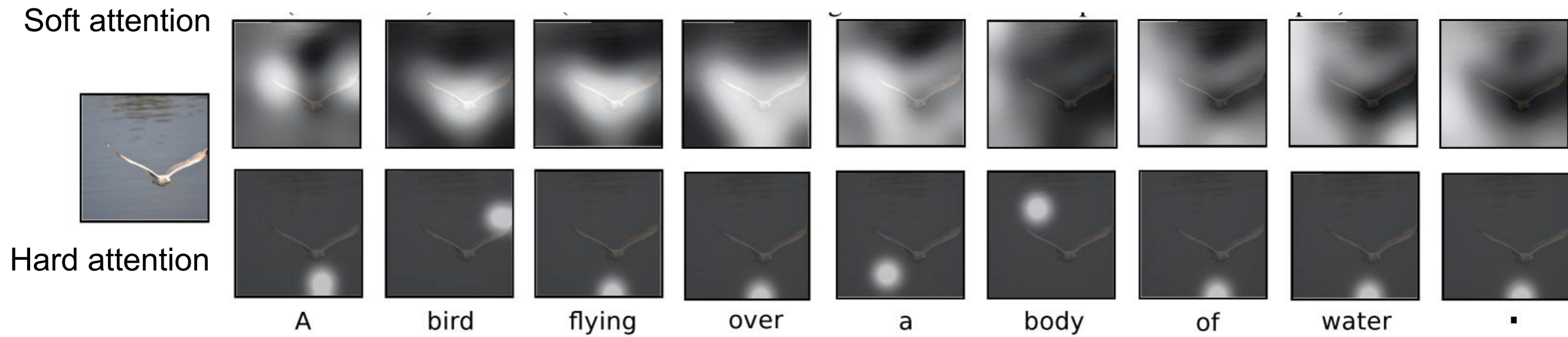


Image Captioning with Attention



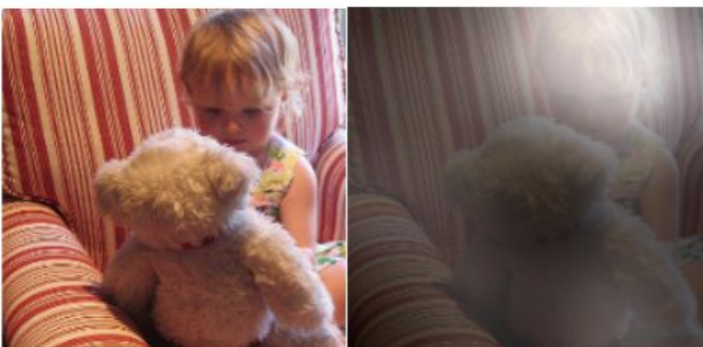
A woman is throwing a frisbee in a park.



A dog is standing on a hardwood floor.



A stop sign is on a road with a mountain in the background.



A little girl sitting on a bed with a teddy bear.



A group of people sitting on a boat in the water.



A giraffe standing in a forest with trees in the background.