vision & language

CS 685, Spring 2022

Advanced Natural Language Processing

http://people.cs.umass.edu/~miyyer/cs685/

Mohit Iyyer

College of Information and Computer Sciences

University of Massachusetts Amherst

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

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?

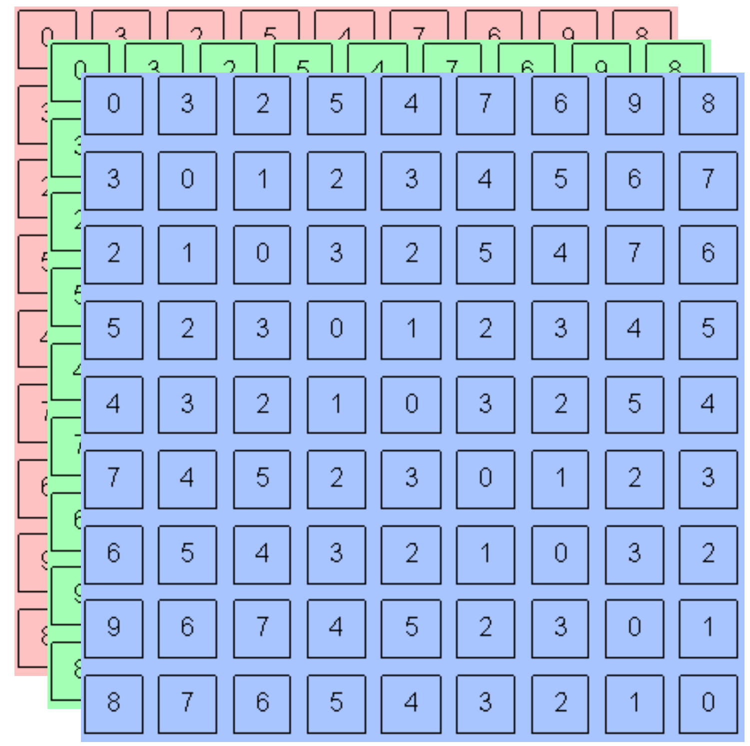
we’ve seen how to compute representations of words and sentences. what about images?

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

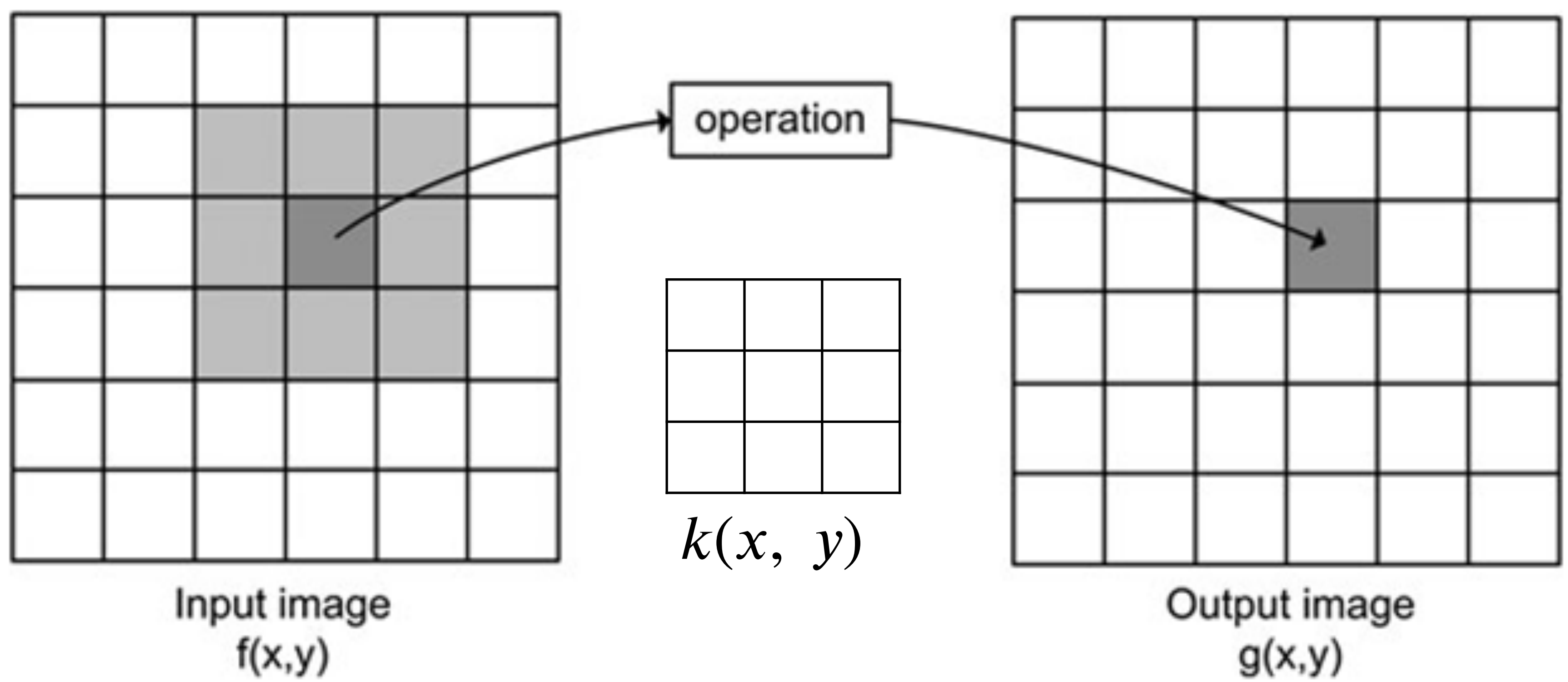


channel x height x width

Channels are usually RGB: Red, Green, and Blue

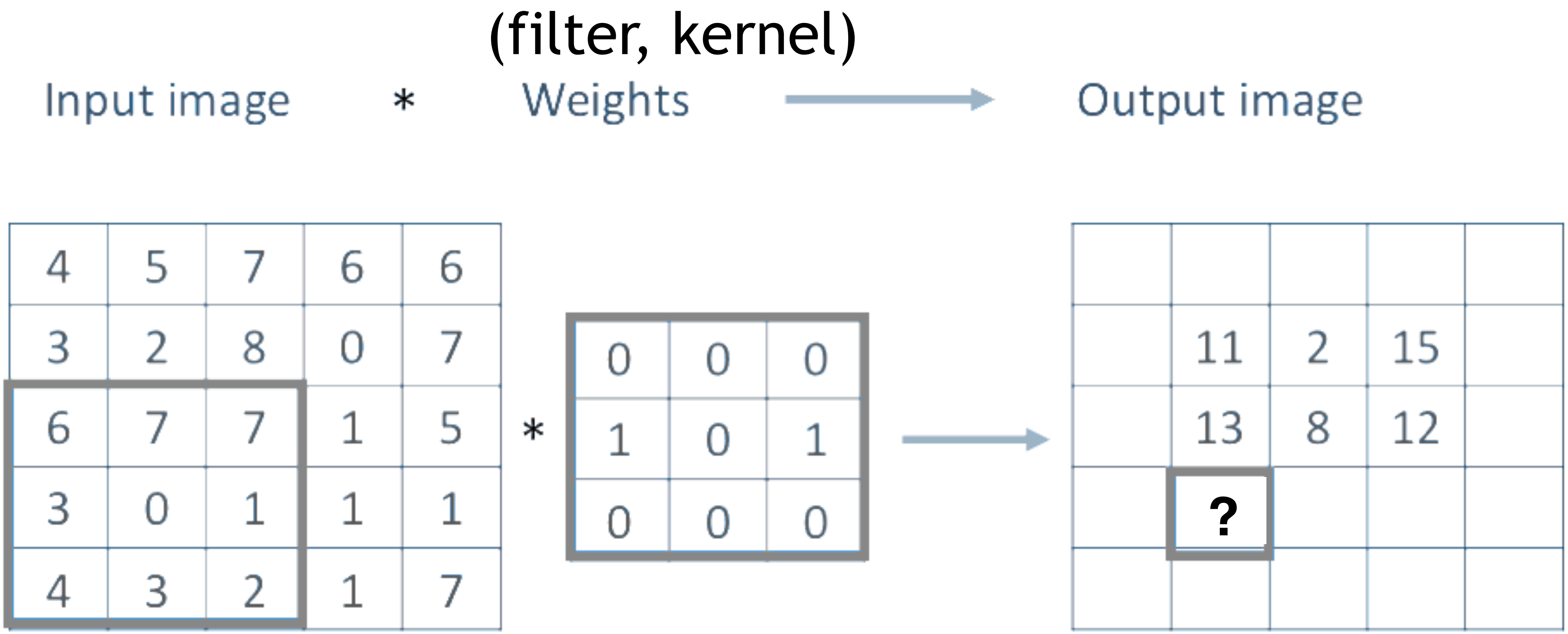
Other color spaces: HSV, HSL, LUV, XYZ, Lab, CMYK, etc

Convolution operator



$$g\left(x,\ y\right) = \sum_{v} \sum_{u} k(u,v) f(x\ - u,\ y - v)$$

Image Credit: http://what-when-how.com/introduction-to-video-and-image-processing/neighborhood-processing-introduction-to-video-and-image-processing-part-1/



(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/

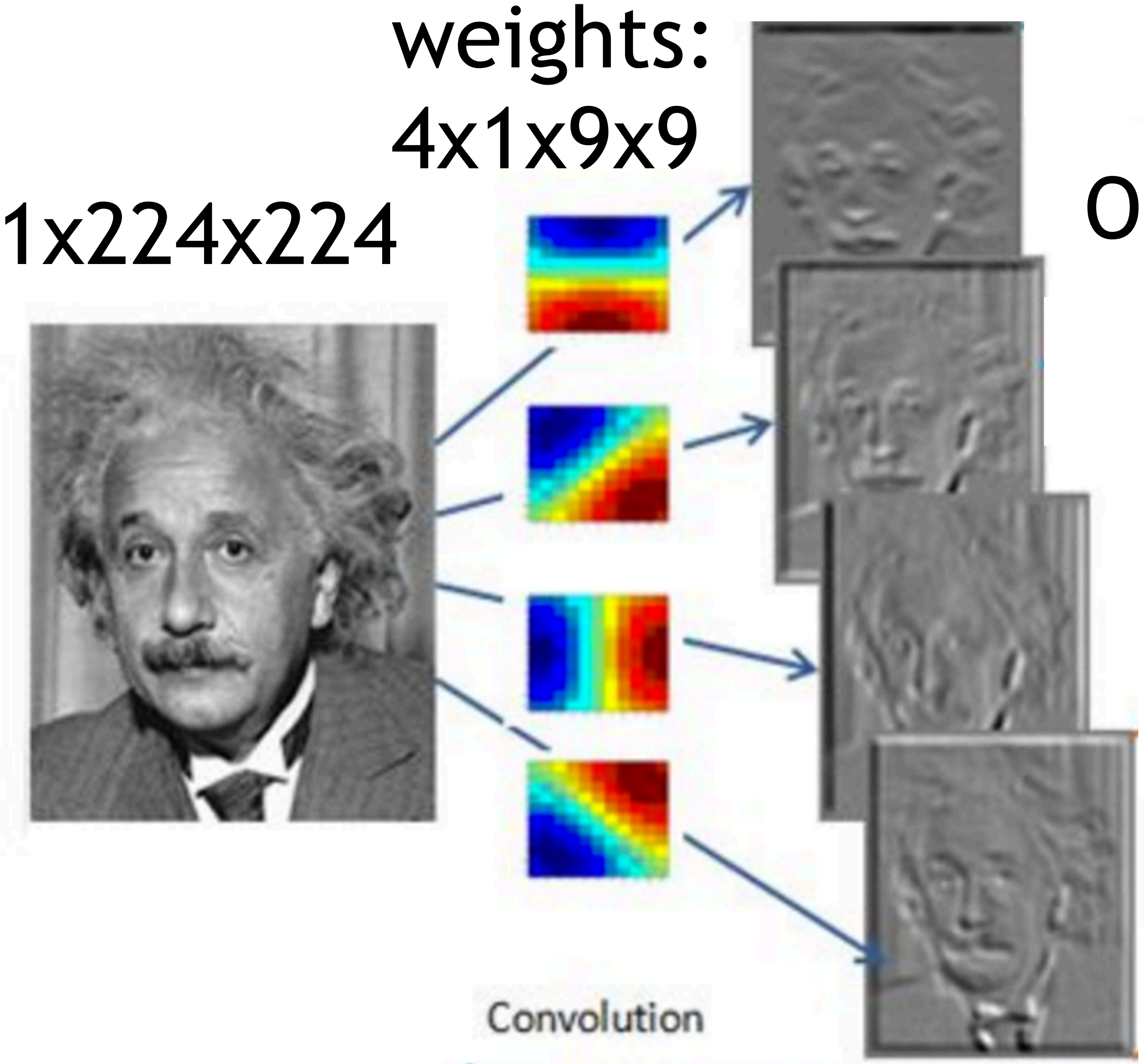
Convolutional Layer (with 4 filters)

weights:
4x1x9x9

Input: 1x224x224

Output: 4x224x224

if zero padding,
and stride = 1



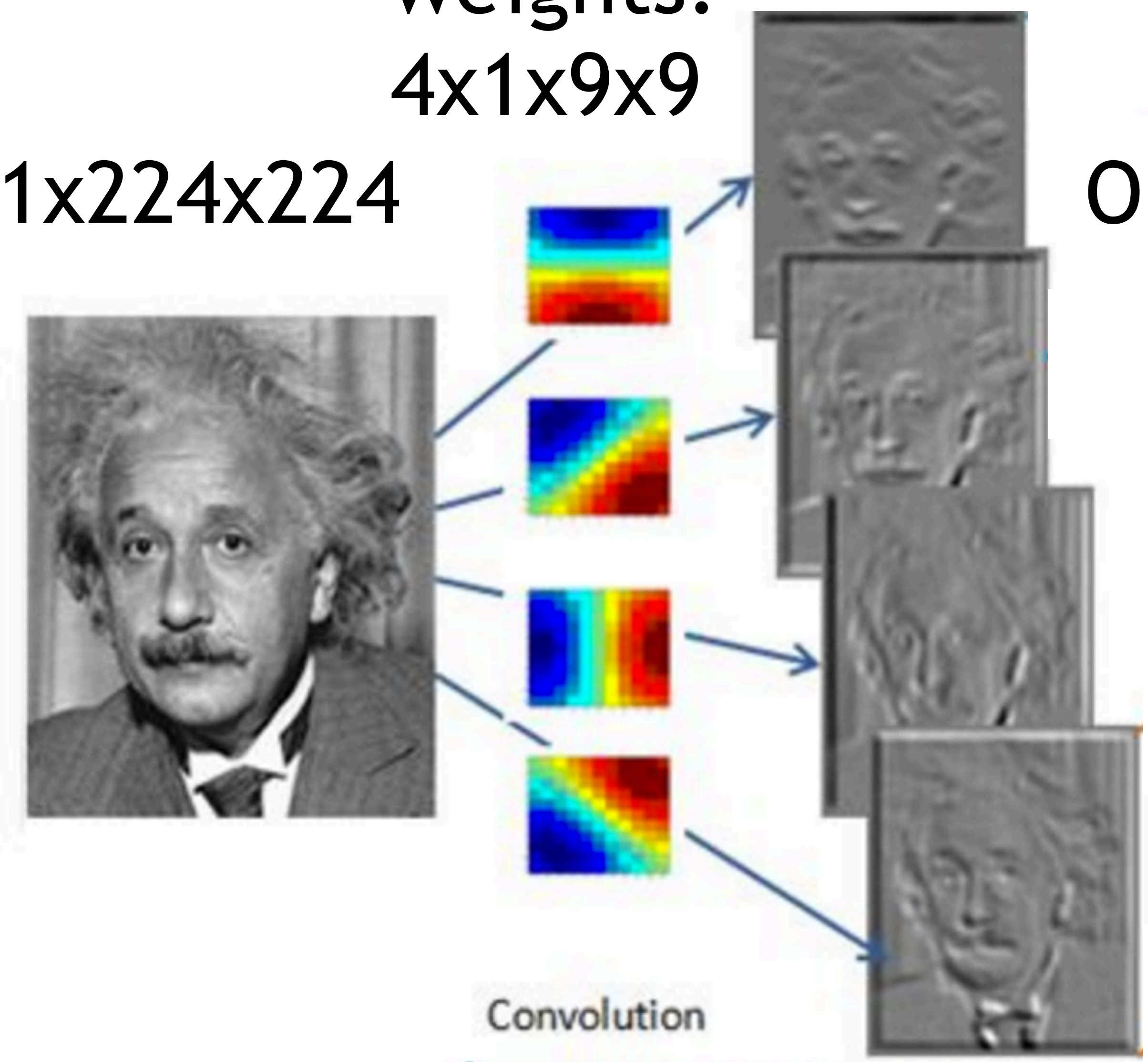
Convolutional Layer (with 4 filters)

weights:
4x1x9x9

Input: 1x224x224

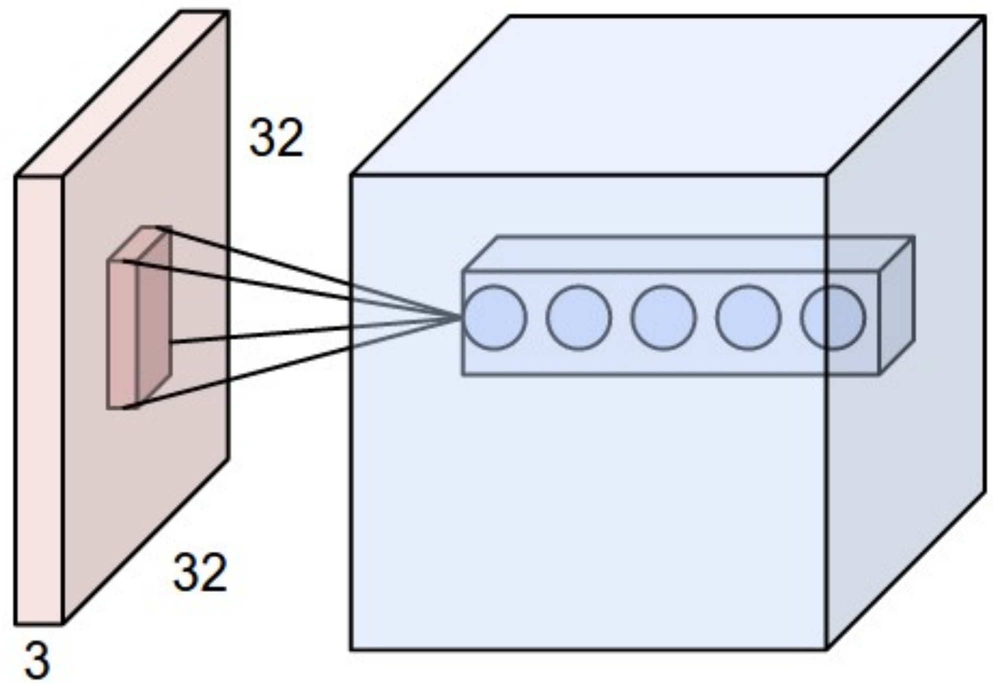
Output: 4x112x112

if zero padding,
but stride = 2



pooling layers also used to reduce dimensionality

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



Pooling Layers:
reduce dimensionality of representation

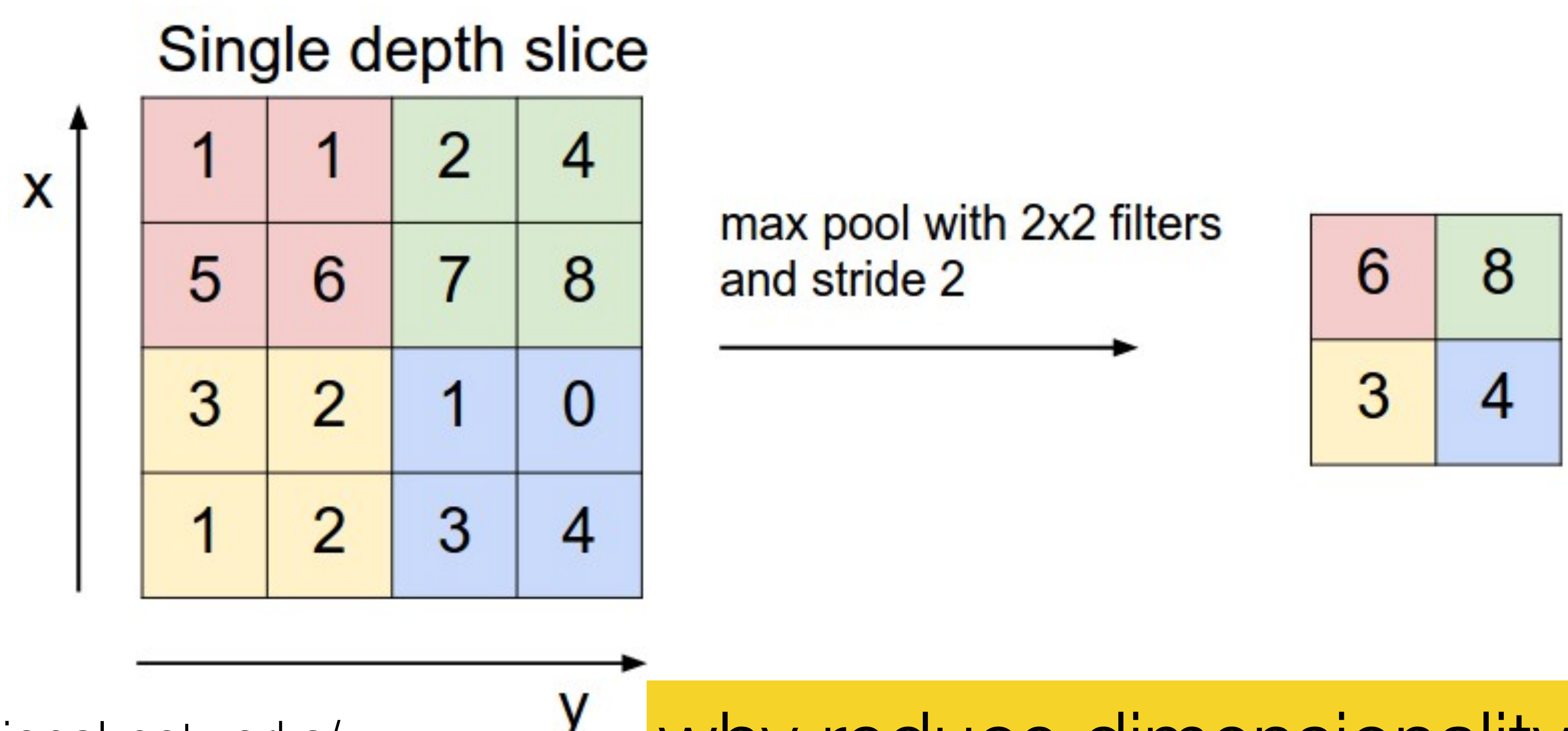


image: https://cs231n.github.io/convolutional-networks/

why reduce dimensionality?

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!

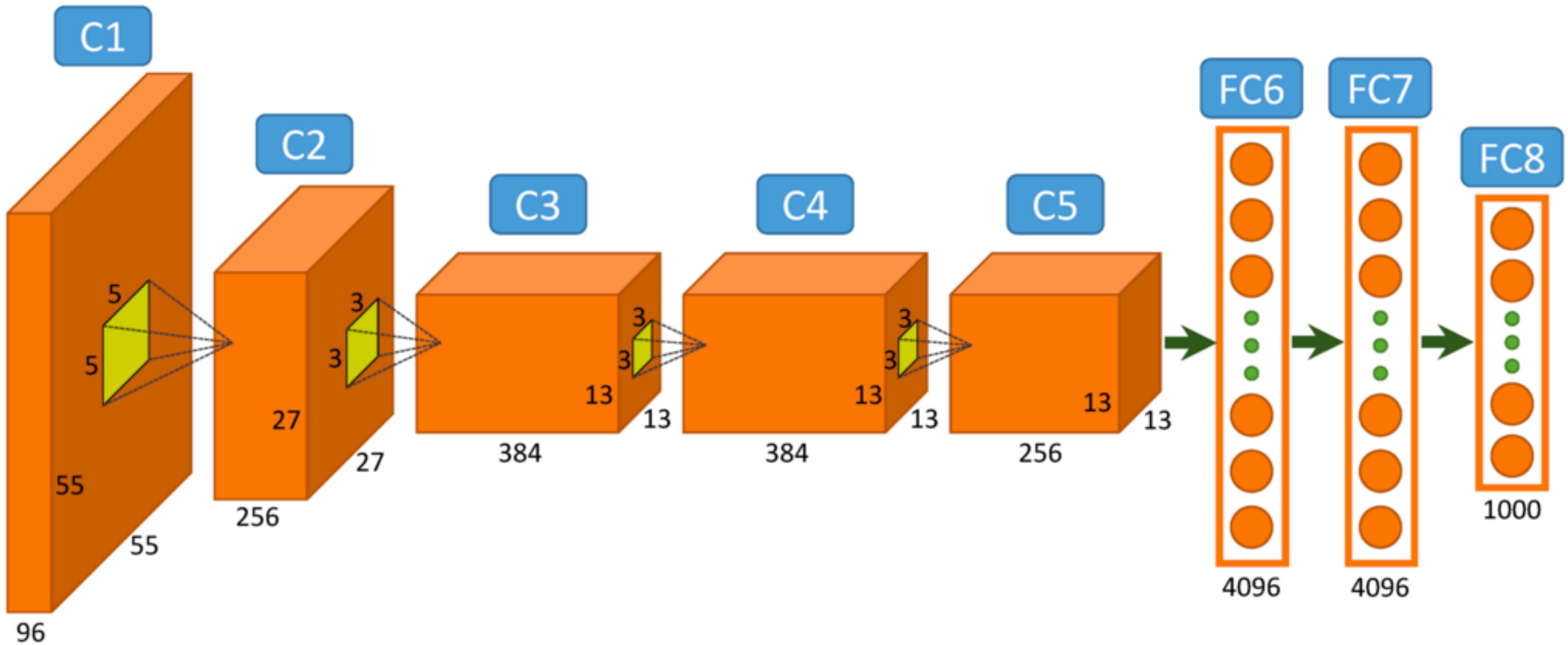
image classification

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)
.....

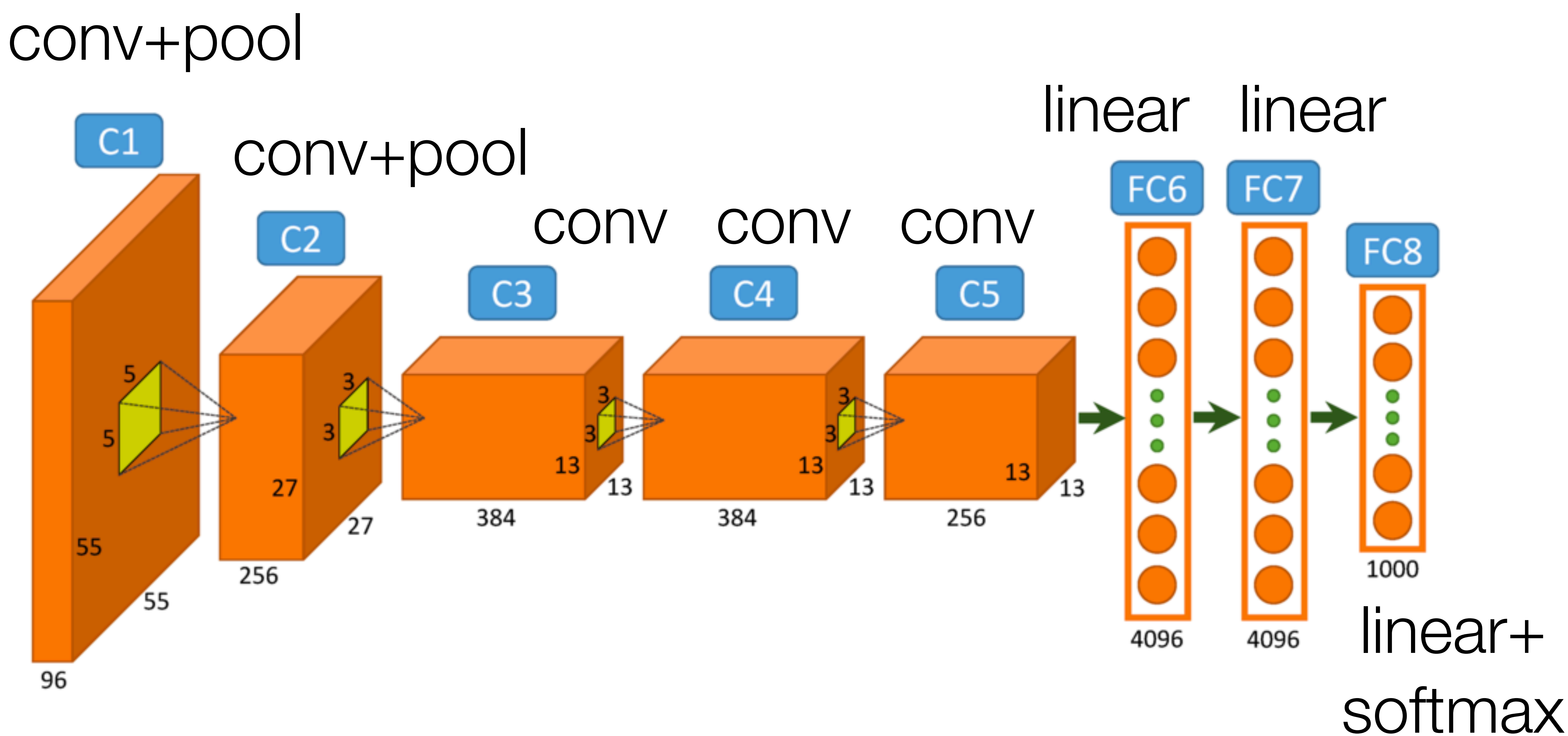
train on the ImageNet challenge dataset, ~1.2 million images

Alexnet



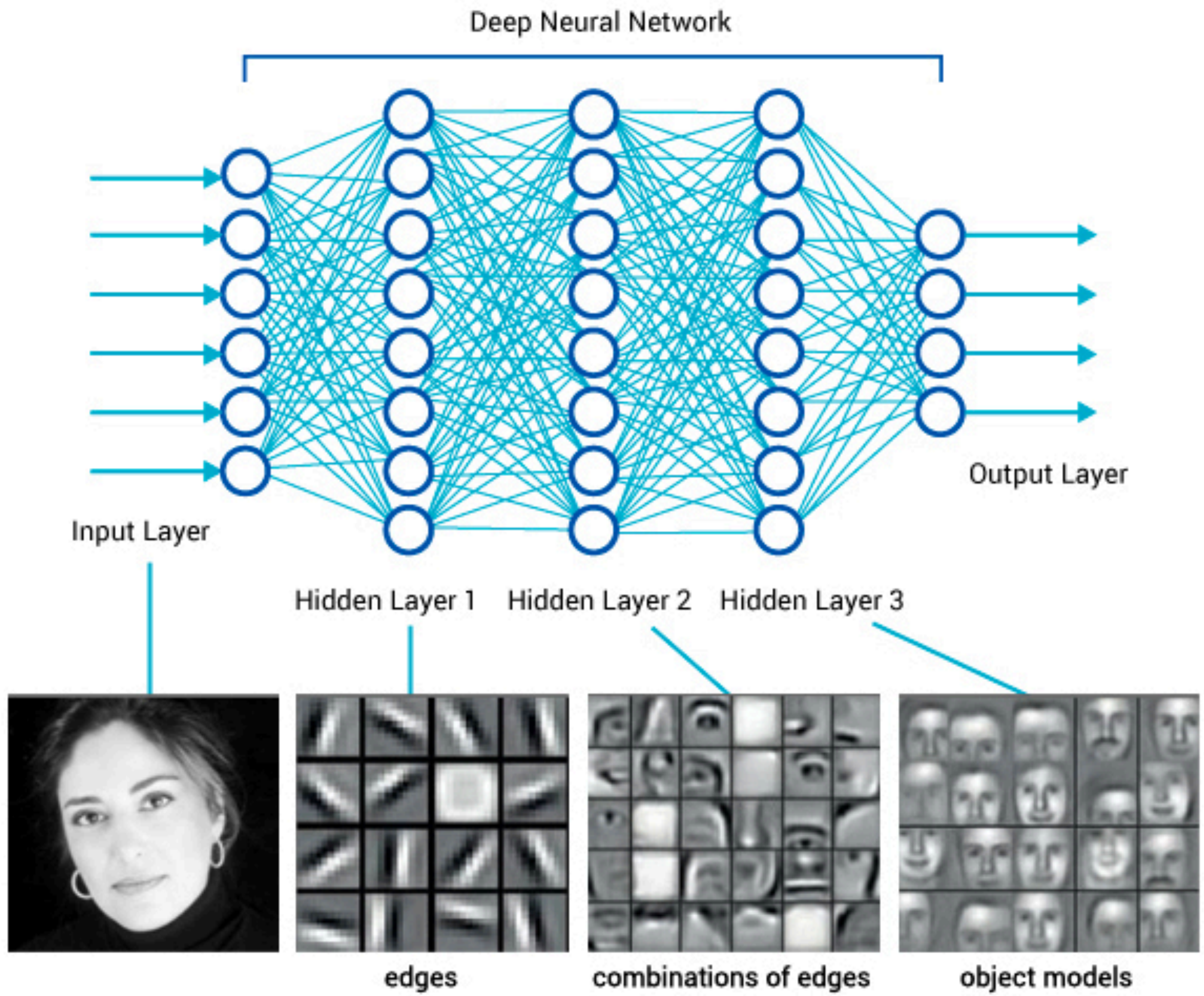
https://www.saagie.com/fr/blog/object-detection-part1

Alexnet



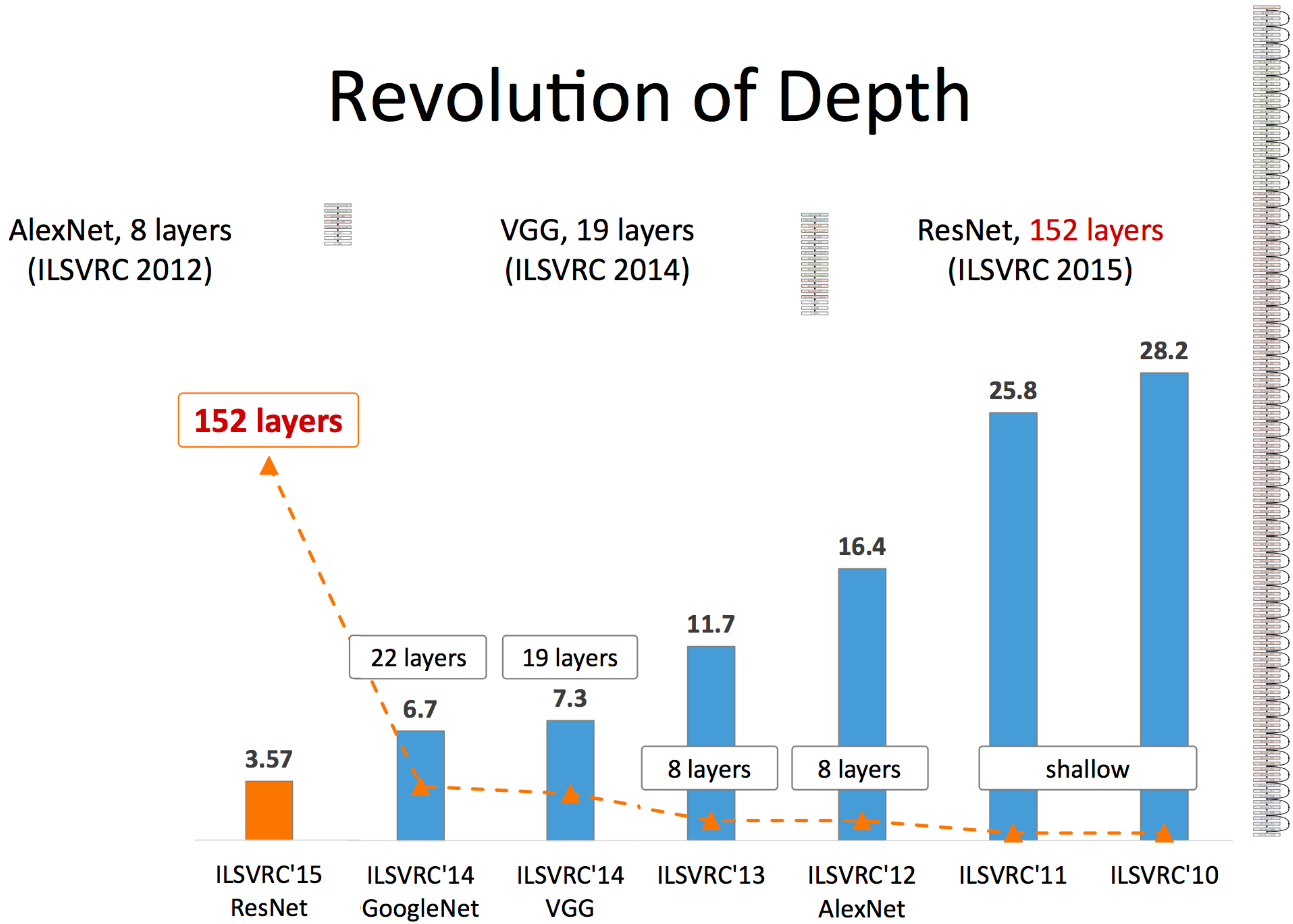
https://www.saagie.com/fr/blog/object-detection-part1

What is happening?



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

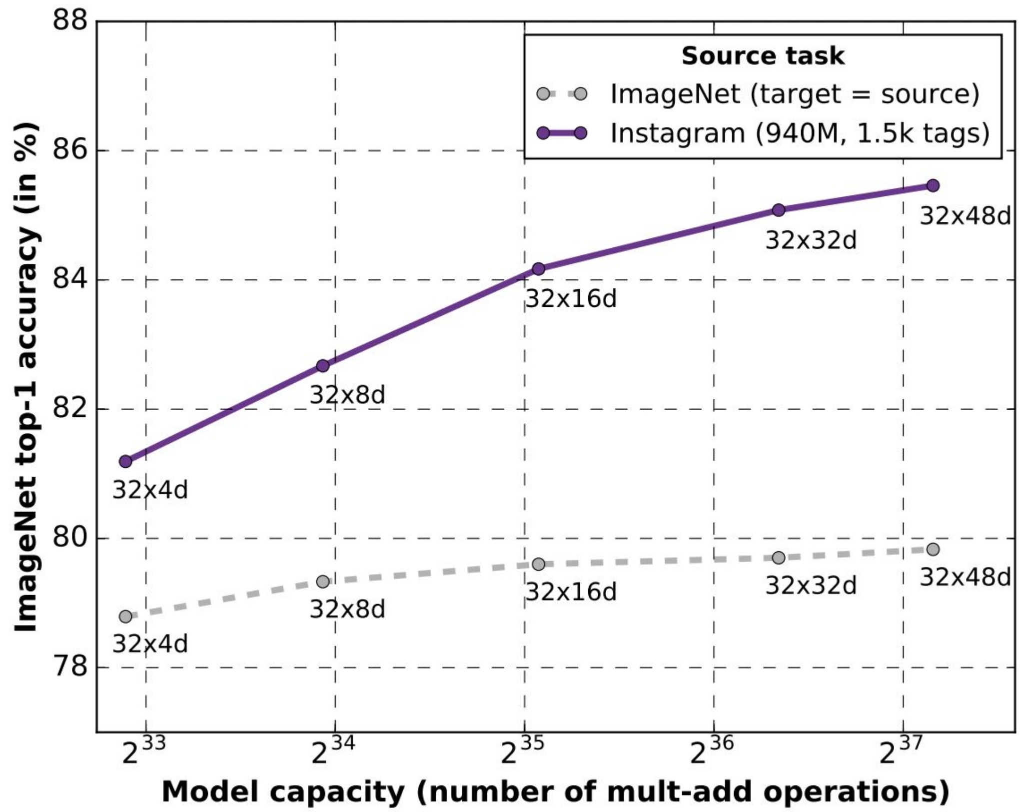
Revolution of Depth



Slide by Mohammad Rastegari

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

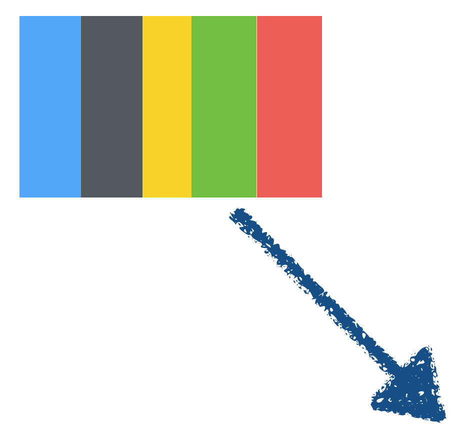
Trained on 336 GPUs for 22 days

Mahajan et al, “Exploring the Limits of Weakly Supervised Pretraining”, arXiv 2018

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

CNN (

) =

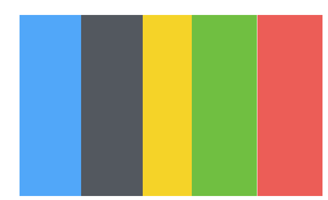


softmax: predict ‘truck’

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

CNN (

) =



simple visual QA

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

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

How many benches are shown?

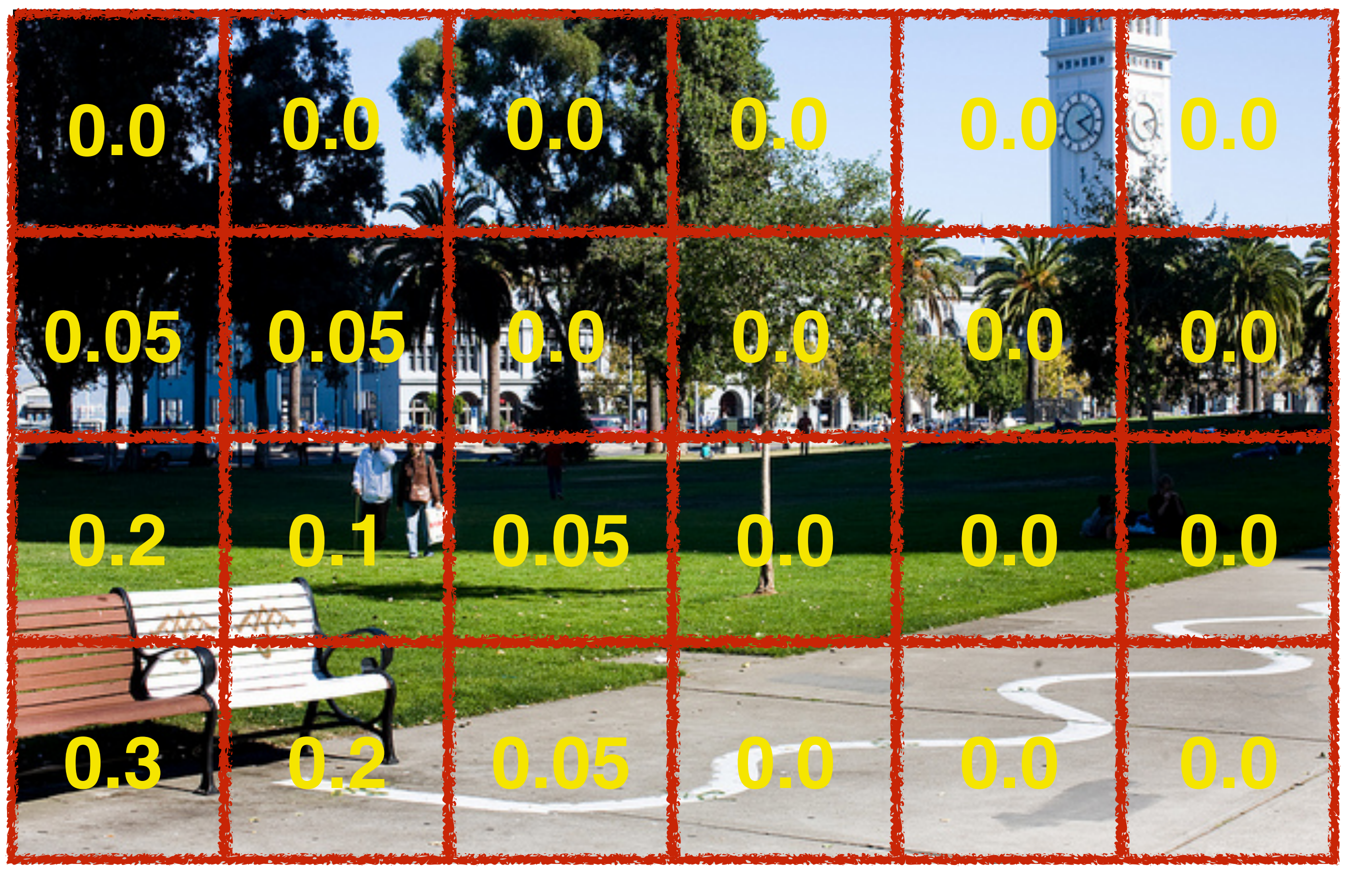
visual attention

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

How many benches are shown?

softmax:
predict answer

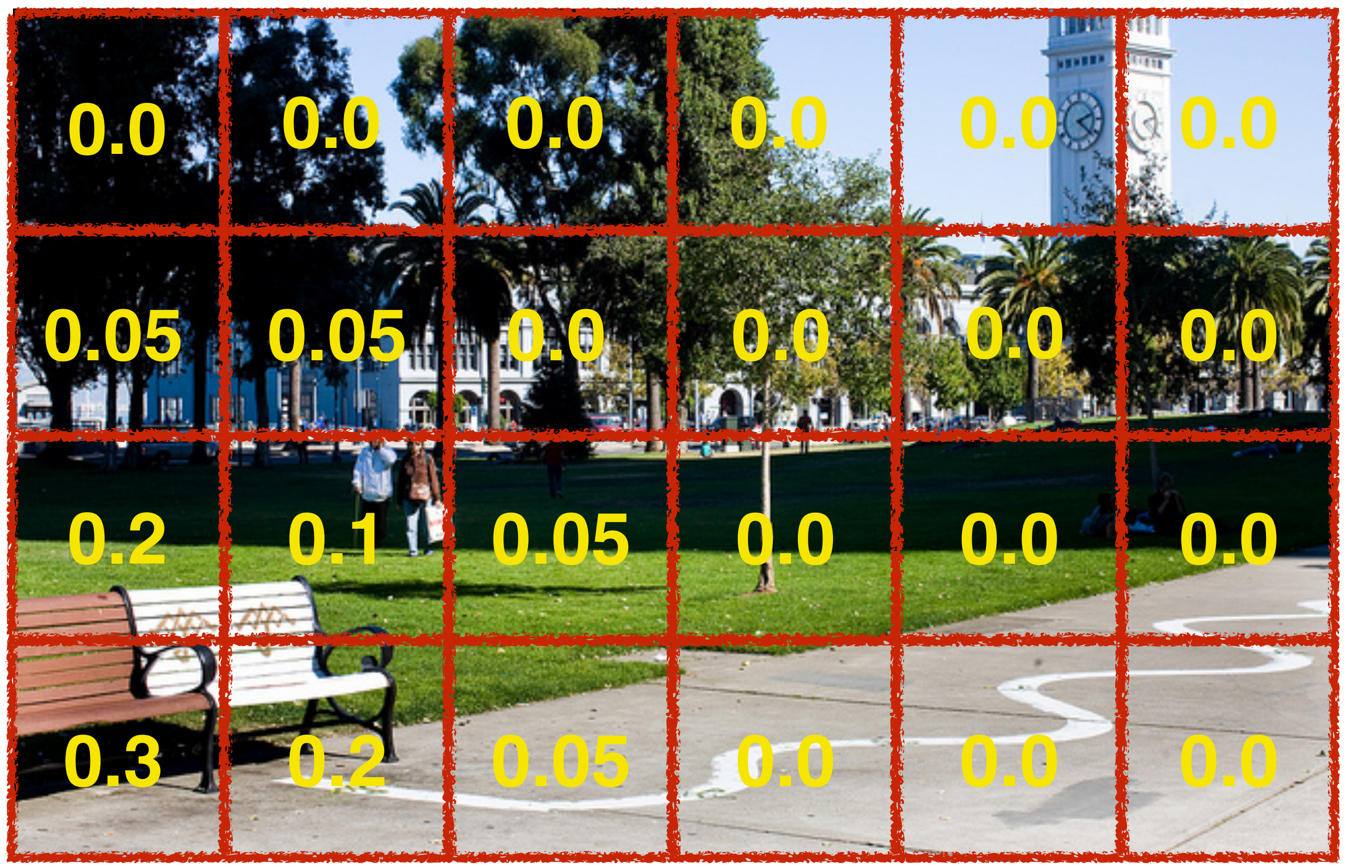
attention over final convolutional layer in network: 196 boxes, captures color and positional information



How many benches are shown?

softmax:
predict answer

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



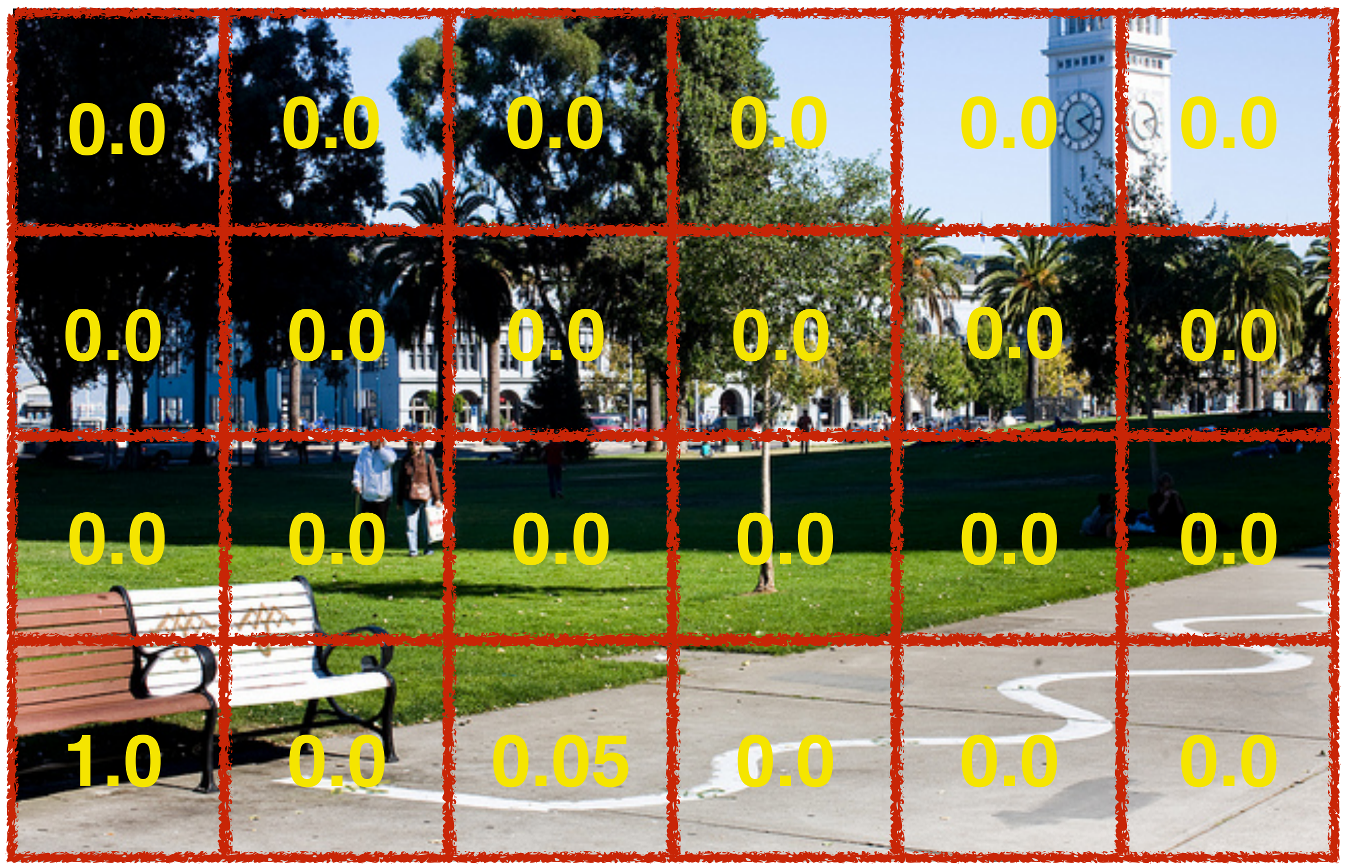
how can we compute these attention scores?

How many benches are shown?

hard attention

softmax:
predict answer

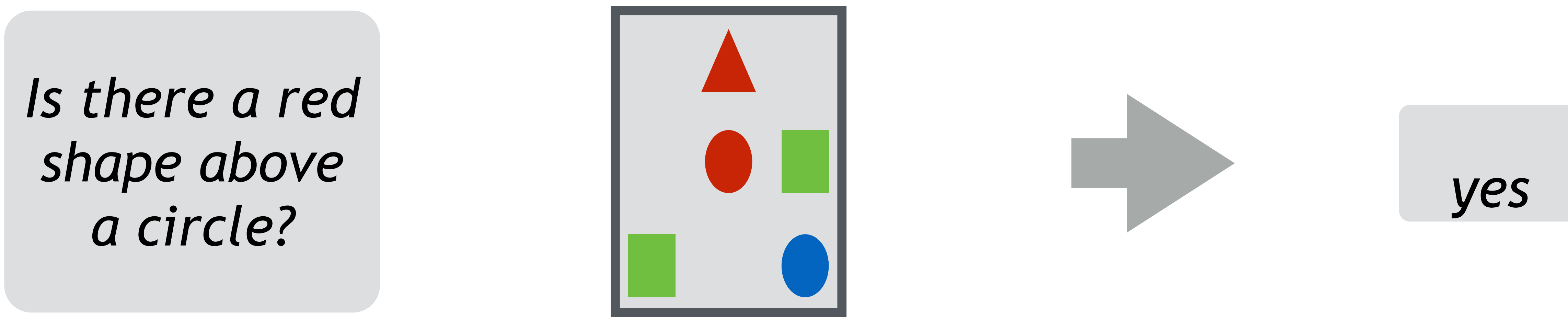
attention over final convolutional layer in network: 196 boxes, captures color and positional information



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

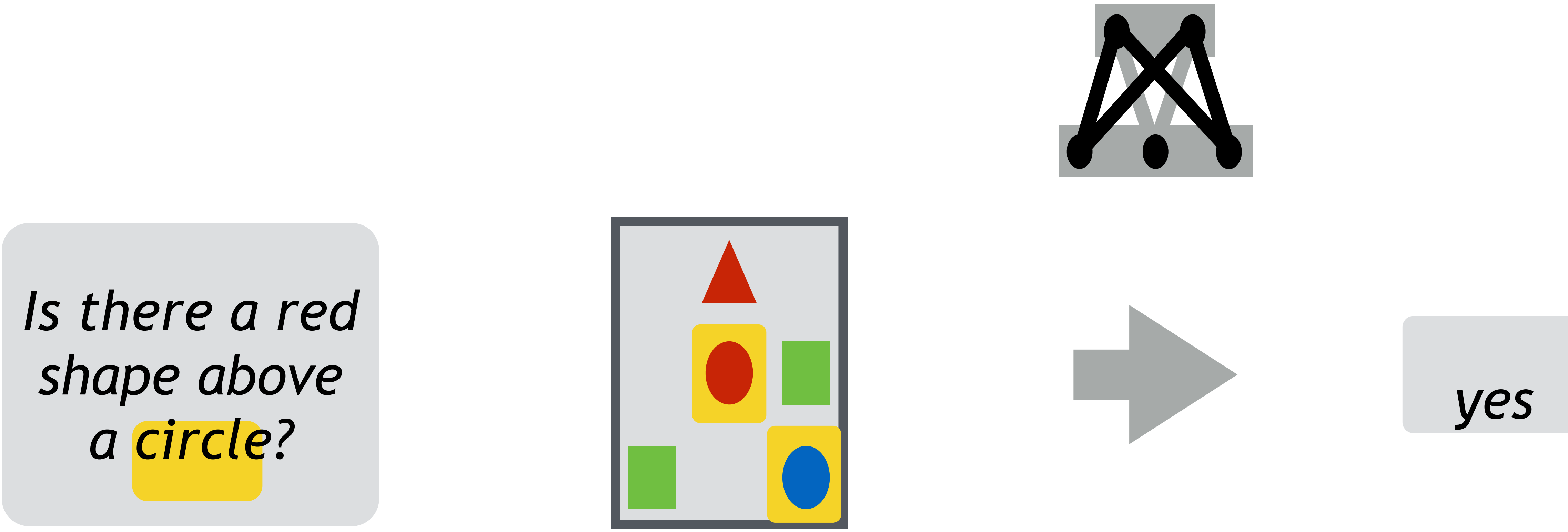
How many benches are shown?

Grounded question answering



Slide credit: Jacob Andreas

Neural nets learn lexical groundings

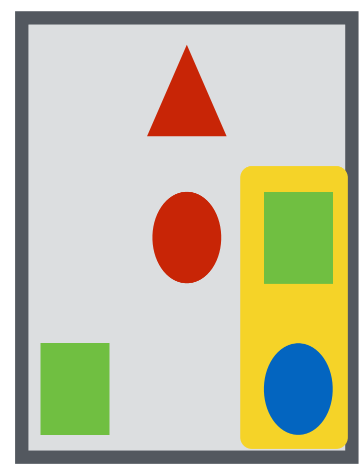


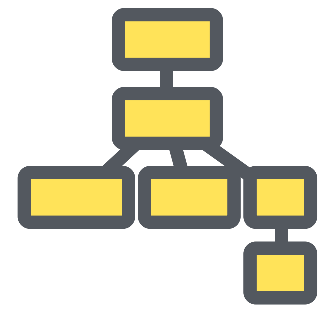
[Iyyer et al. 2014, Bordes et al. 2014,
Yang et al. 2015, Malinowski et al., 2015]

Slide credit: Jacob Andreas

Semantic parsers learn composition

Is there a red shape above a circle?



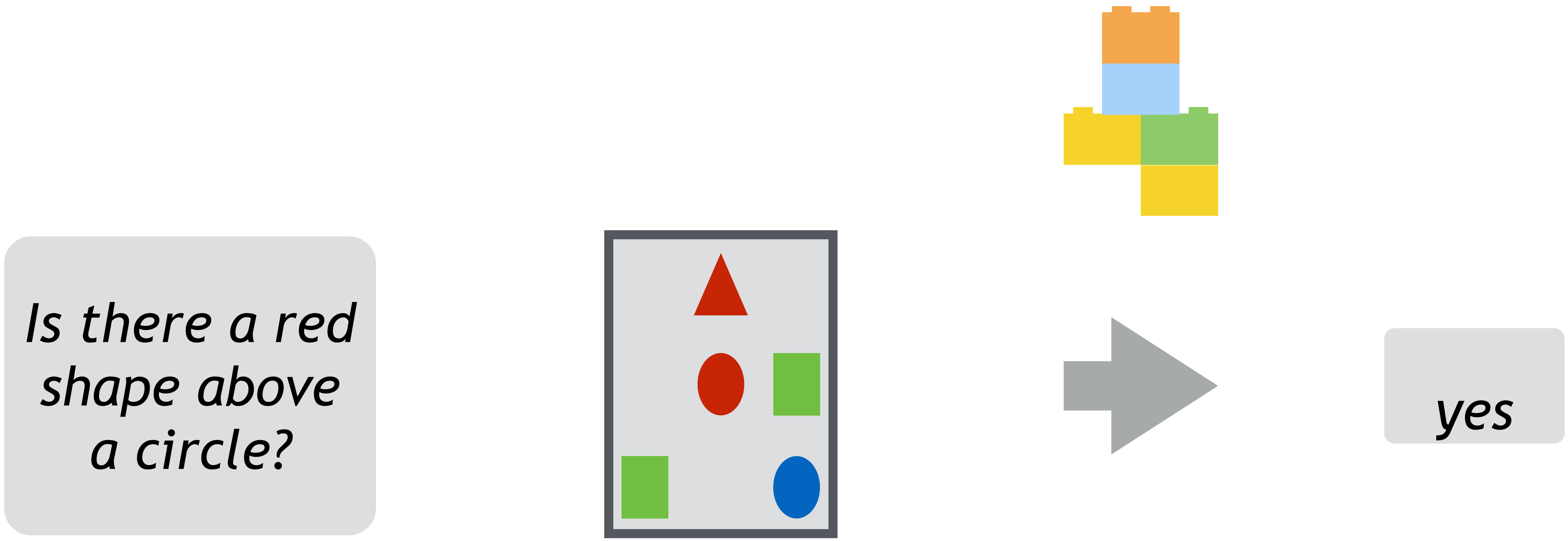


yes

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

Slide credit: Jacob Andreas

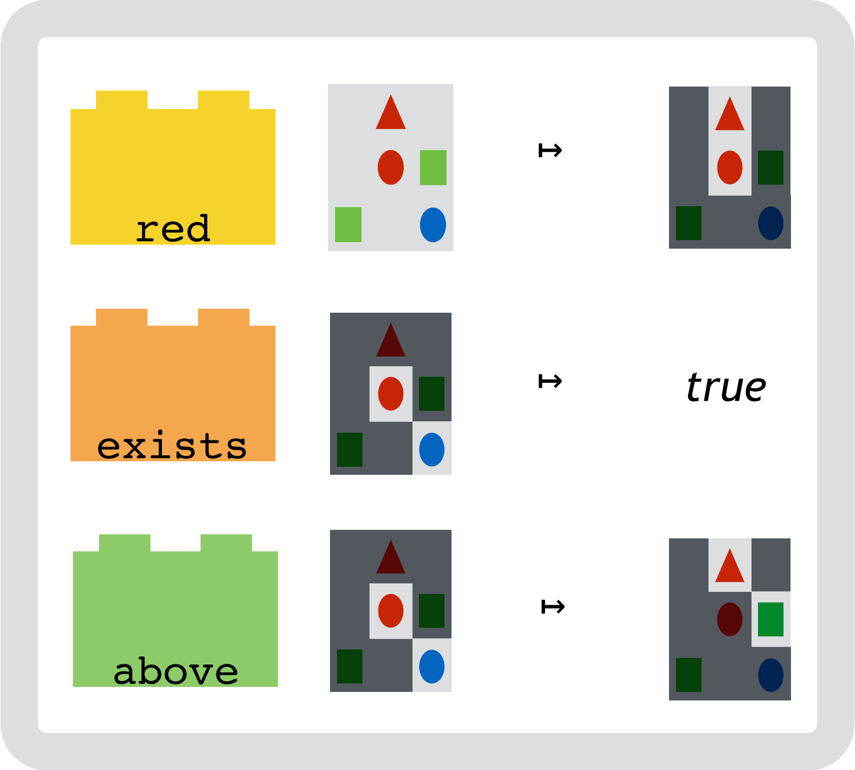
Neural module networks learn both!



Slide credit: Jacob Andreas

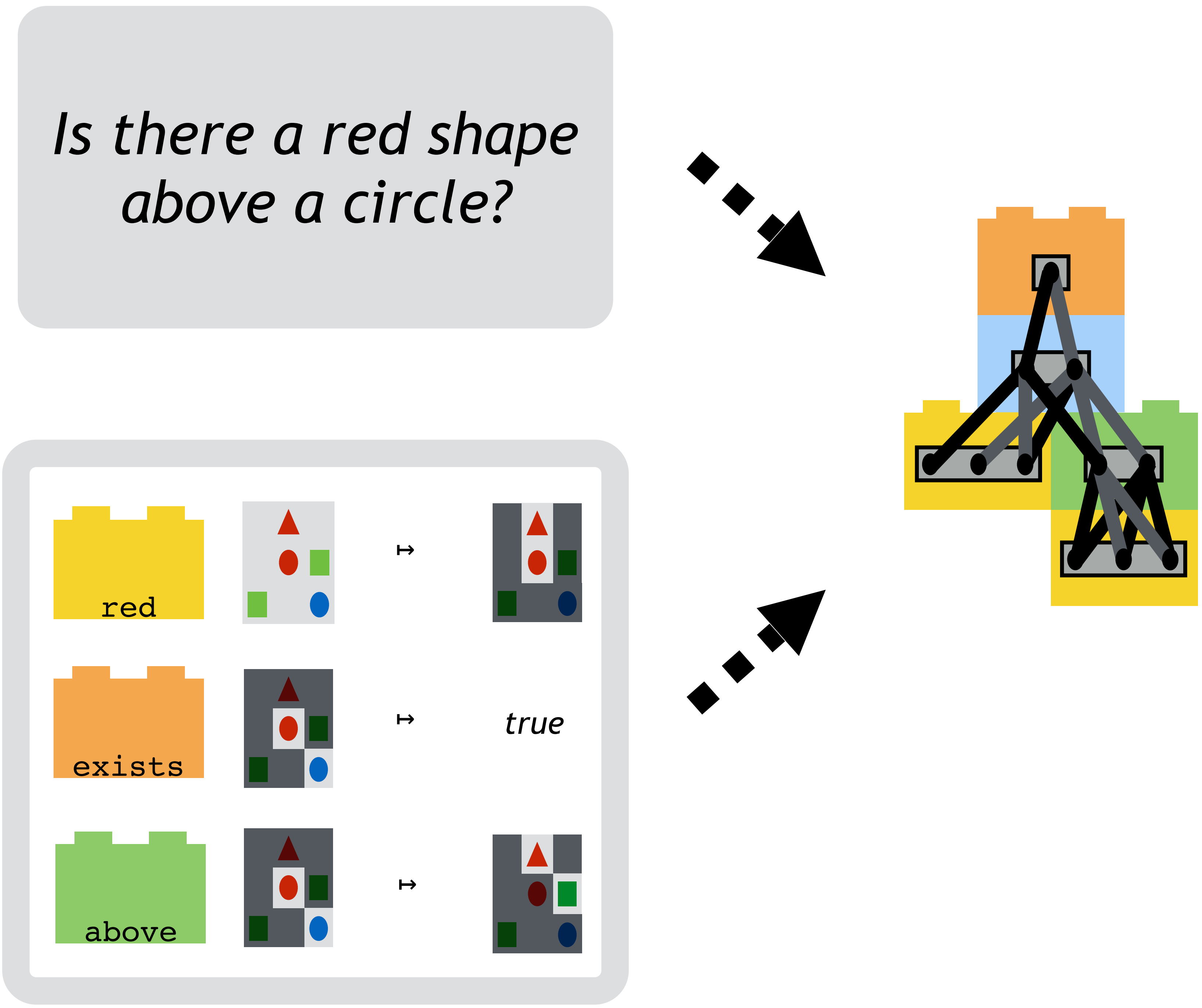
Neural module networks

Is there a red shape above a circle?



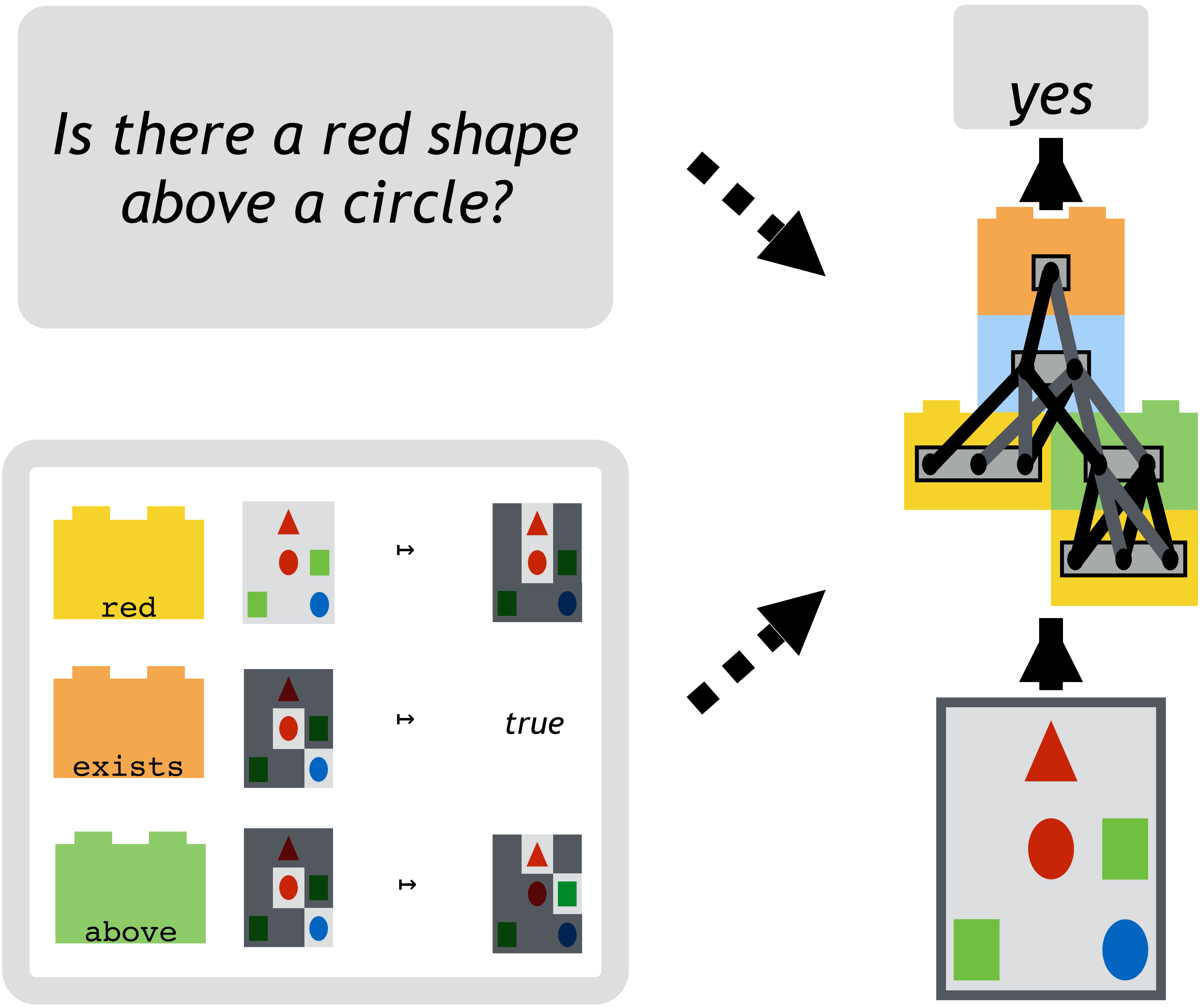
Slide credit: Jacob Andreas

Neural module networks



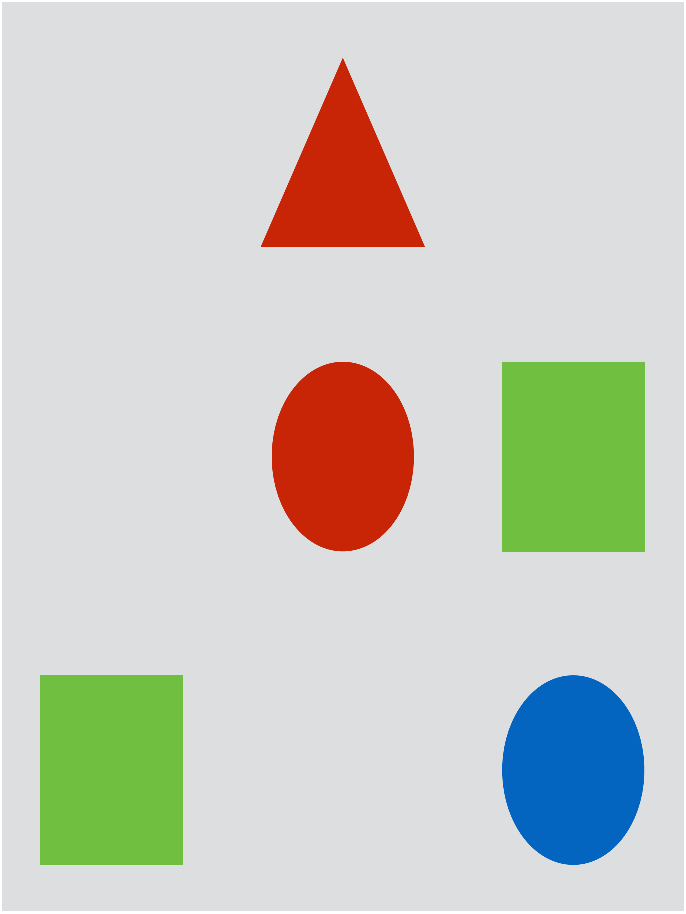
Slide credit: Jacob Andreas

Neural module networks

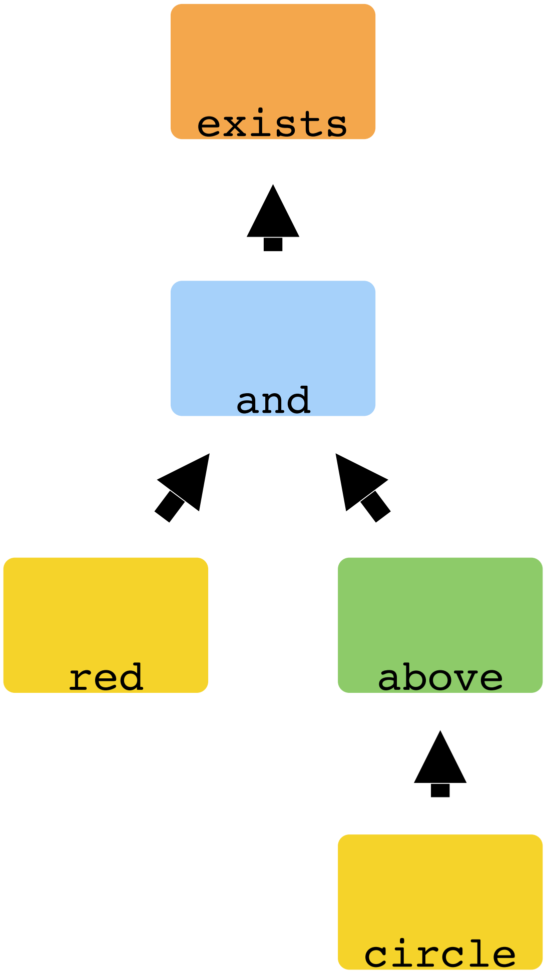


Slide credit: Jacob Andreas

Sentence meanings are computations



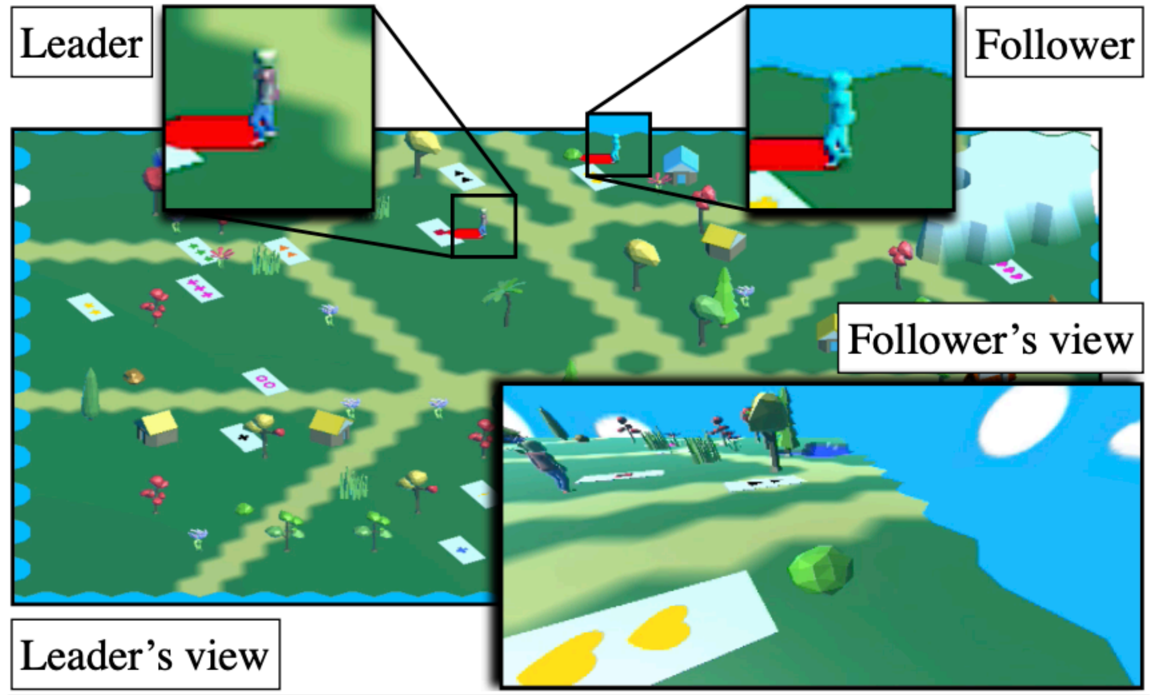
Is there a red shape above a circle?



Slide credit: Jacob Andreas

NLVR^2: 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.



...

$\bar{x}_3$: *turn left and head toward the yellow hearts, but don't pick them up yet. I'll get the next card first.*

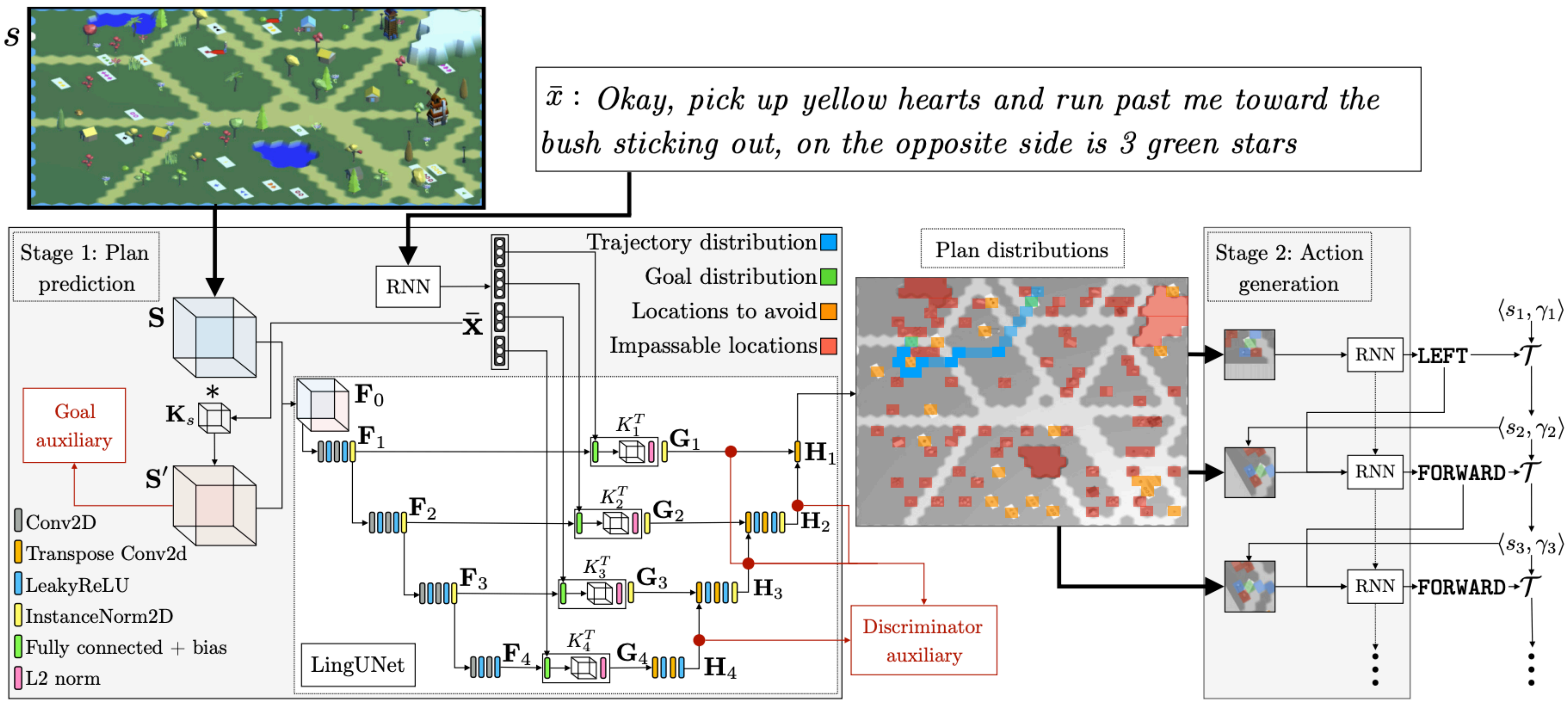
$\bar{x}_4$: ***Okay, pick up yellow hearts and run past me toward the bush sticking out, on the opposite side is 3 green stars***

`[Set made. New score: 4]`

...

Figure 1: A snapshot from an interaction in CEREALBAR. The current instruction is in bold. The large image shows the entire environment. This overhead view is available only to the leader. The follower sees a first-person view only (bottom right). The zoom boxes (top) show the leader and follower.

Suhr et al., 2019 ("CEREALBAR")



Suhr et al., 2019 (“CEREALBAR”)

Image Captioning

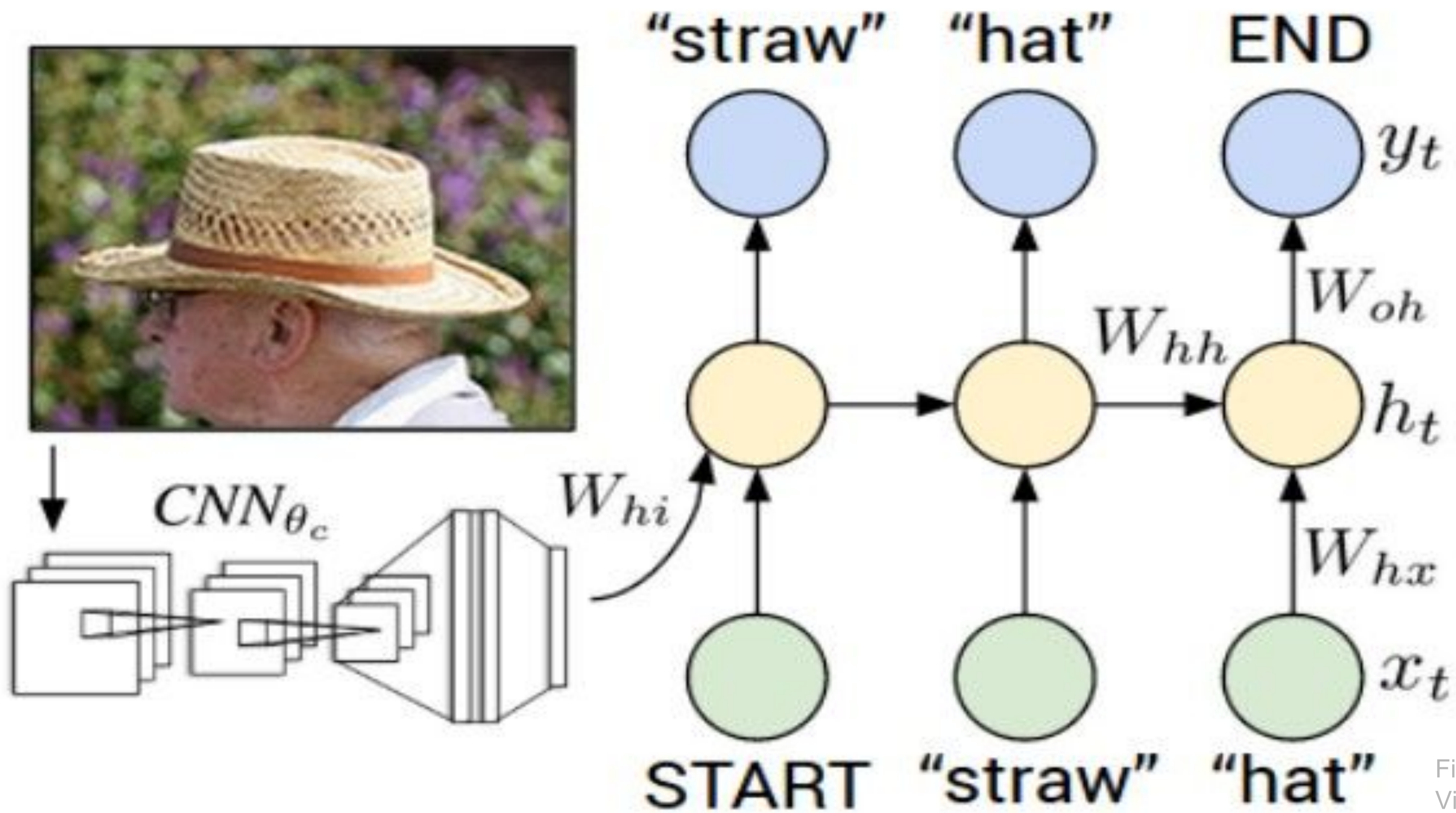


Figure fr
Visual-S
Image D
copyrigh
Reprodu

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

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

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

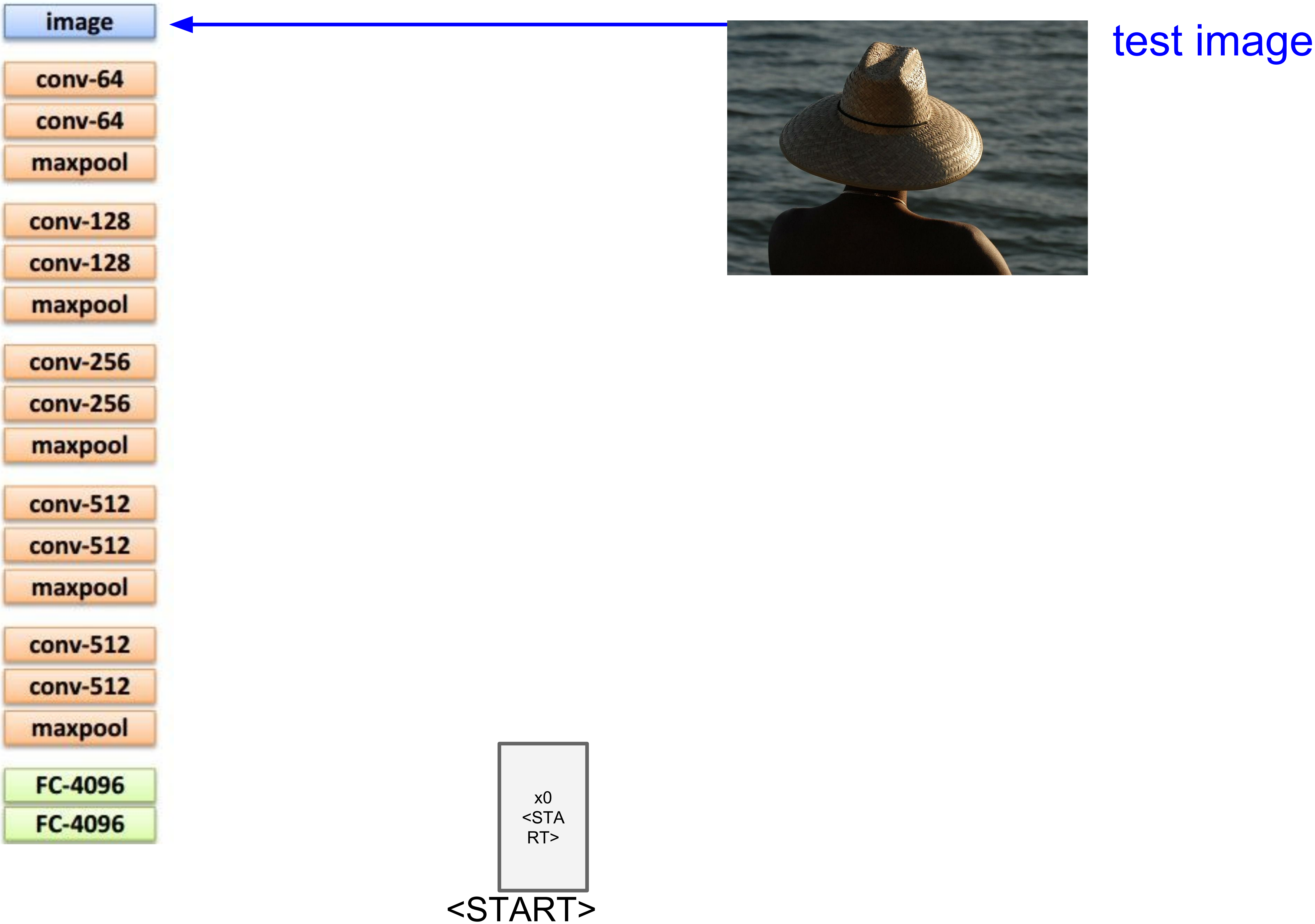
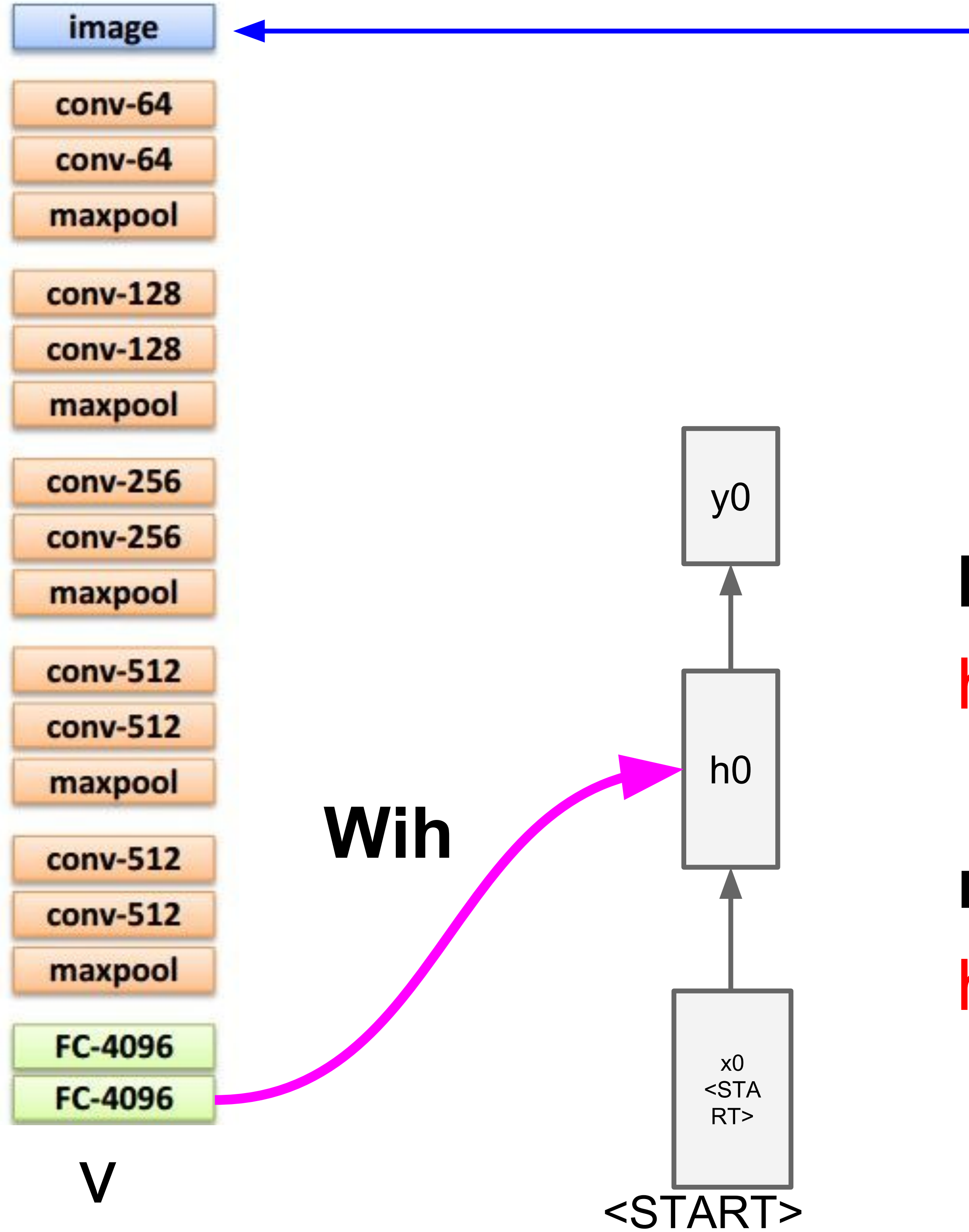


image
conv-64
conv-64
maxpool
conv-128
conv-128
maxpool
conv-256
conv-256
maxpool
conv-512
conv-512
maxpool
conv-512
conv-512
maxpool
FC-4096
FC-4096
test image
x0
<STA
RT>
<START>



test image

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

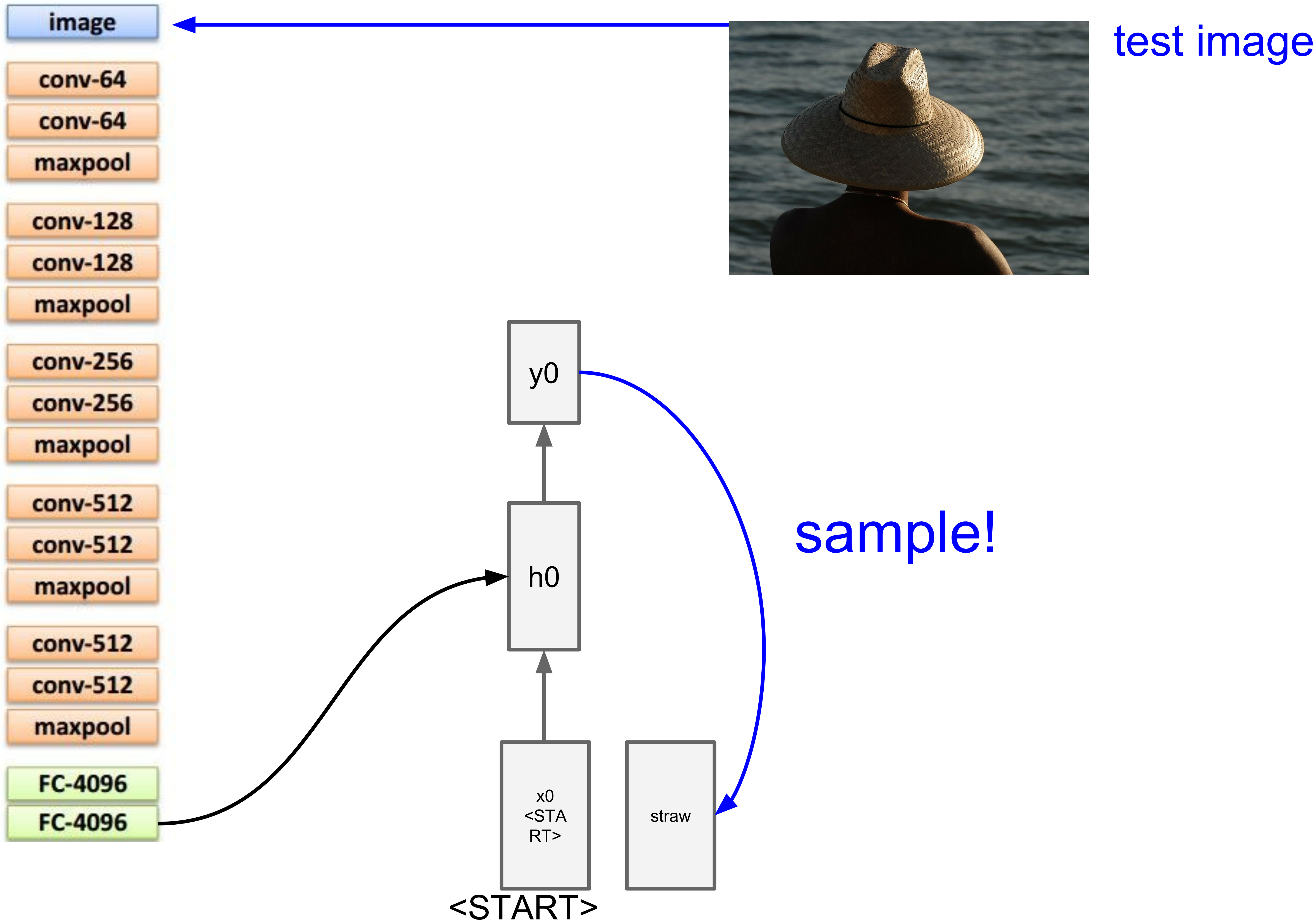


image
conv-64
conv-64
maxpool
conv-128
conv-128
maxpool
conv-256
conv-256
maxpool
conv-512
conv-512
maxpool
conv-512
conv-512
maxpool
FC-4096
FC-4096
test image
y0
sample!
h0
x0
<STA
RT>
straw
<START>

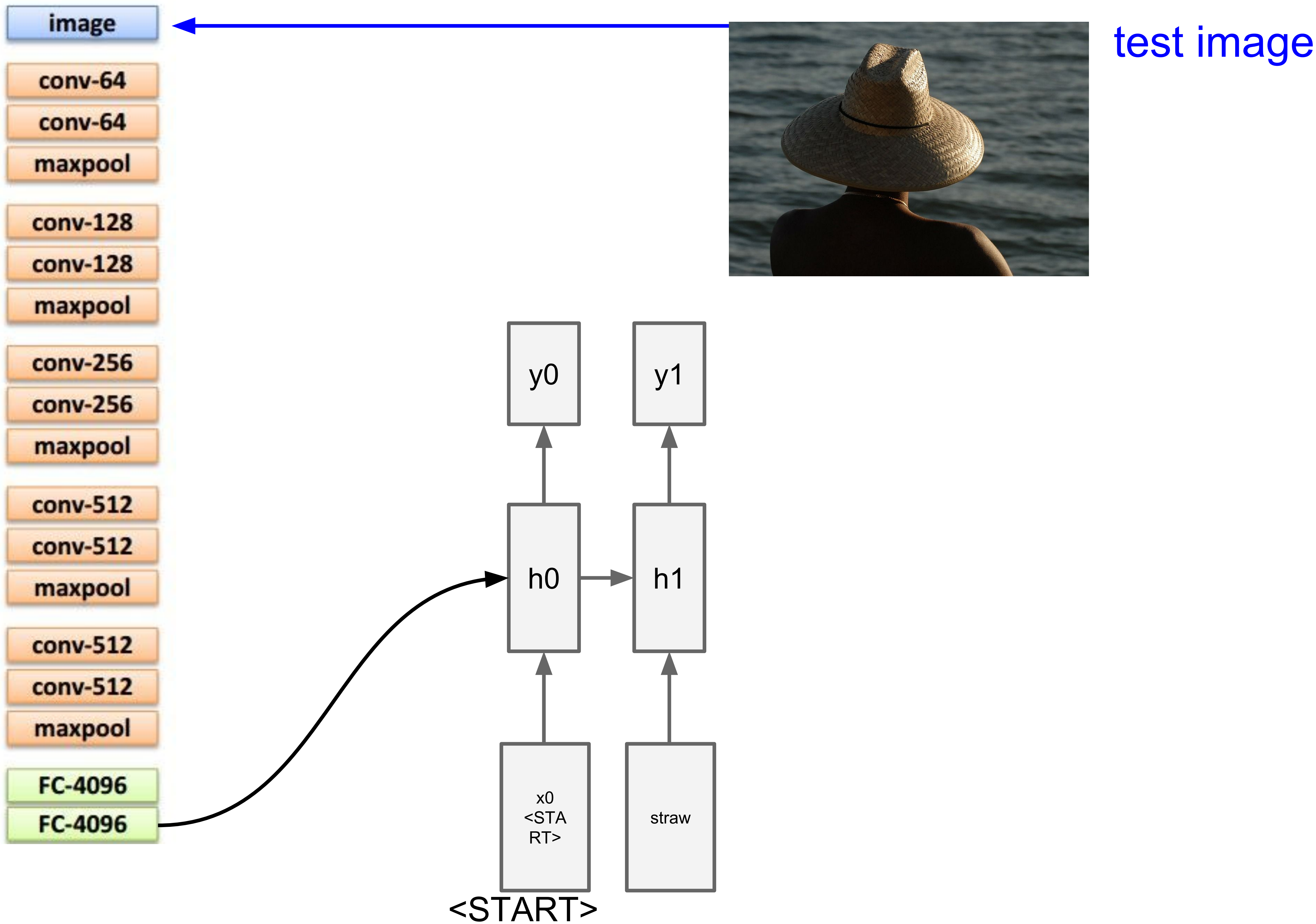


image
conv-64
conv-64
maxpool
conv-128
conv-128
maxpool
conv-256
conv-256
maxpool
conv-512
conv-512
maxpool
conv-512
conv-512
maxpool
FC-4096
FC-4096
test image
y0
y1
h0
h1
x0
<STA
RT>
straw
<START>

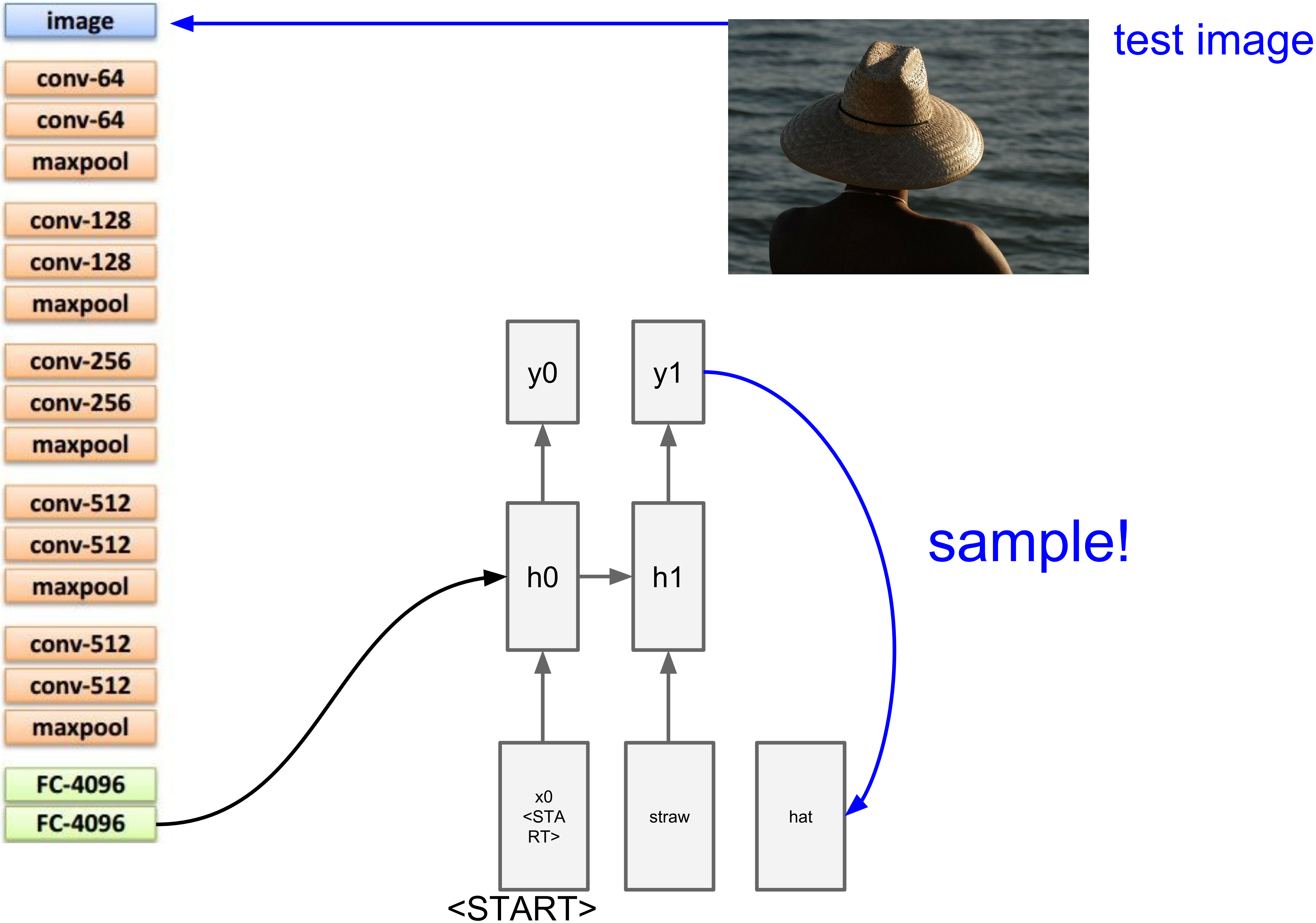


image
conv-64
conv-64
maxpool
conv-128
conv-128
maxpool
conv-256
conv-256
maxpool
conv-512
conv-512
maxpool
conv-512
conv-512
maxpool
FC-4096
FC-4096
test image
y0
y1
h0
h1
sample!
x0 <STA RT>
straw
hat
<START>

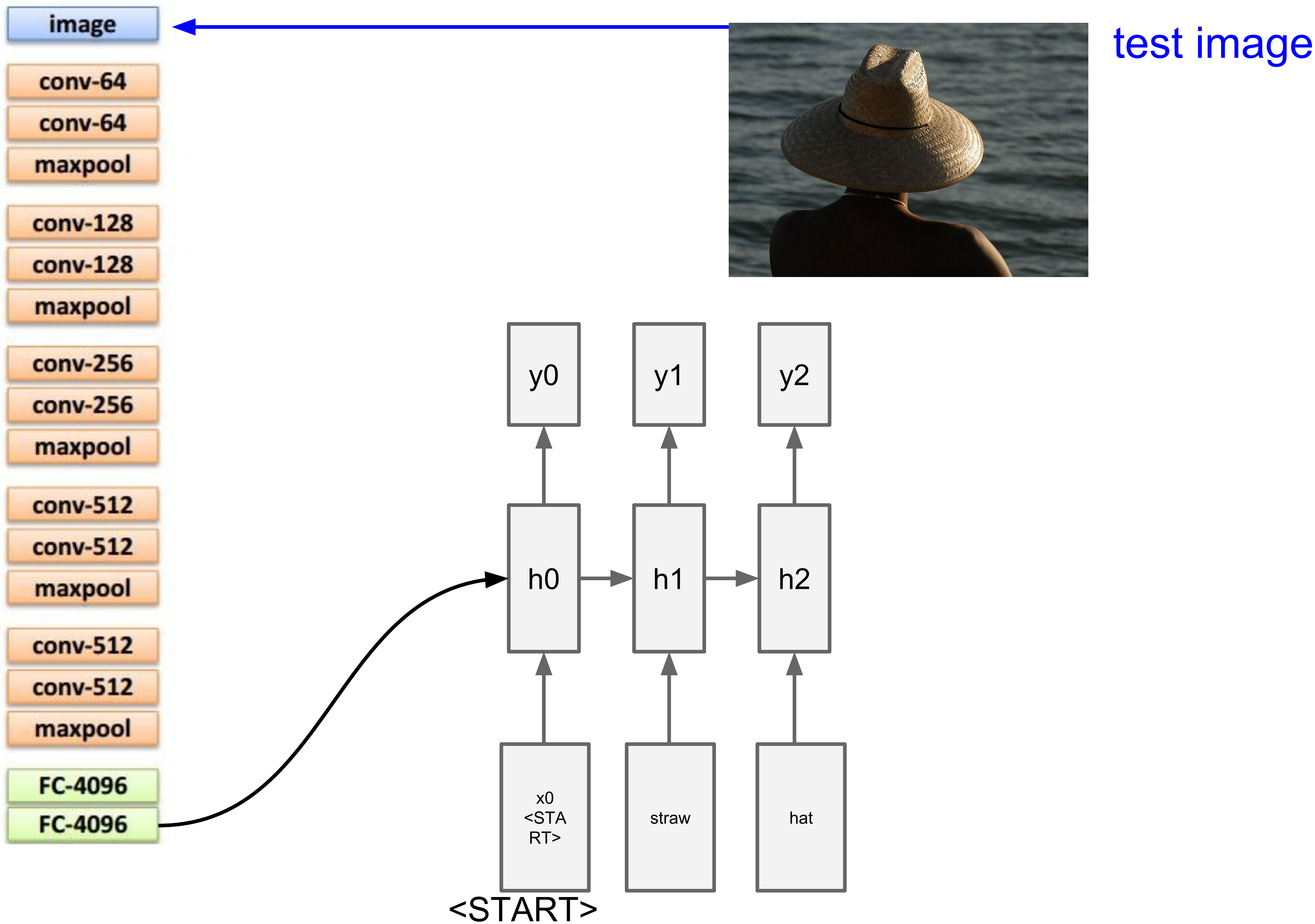


image
conv-64
conv-64
maxpool
conv-128
conv-128
maxpool
conv-256
conv-256
maxpool
conv-512
conv-512
maxpool
conv-512
conv-512
maxpool
FC-4096
FC-4096
test image
y0
y1
y2
h0
h1
h2
x0 <STA RT>
straw
hat
<START>

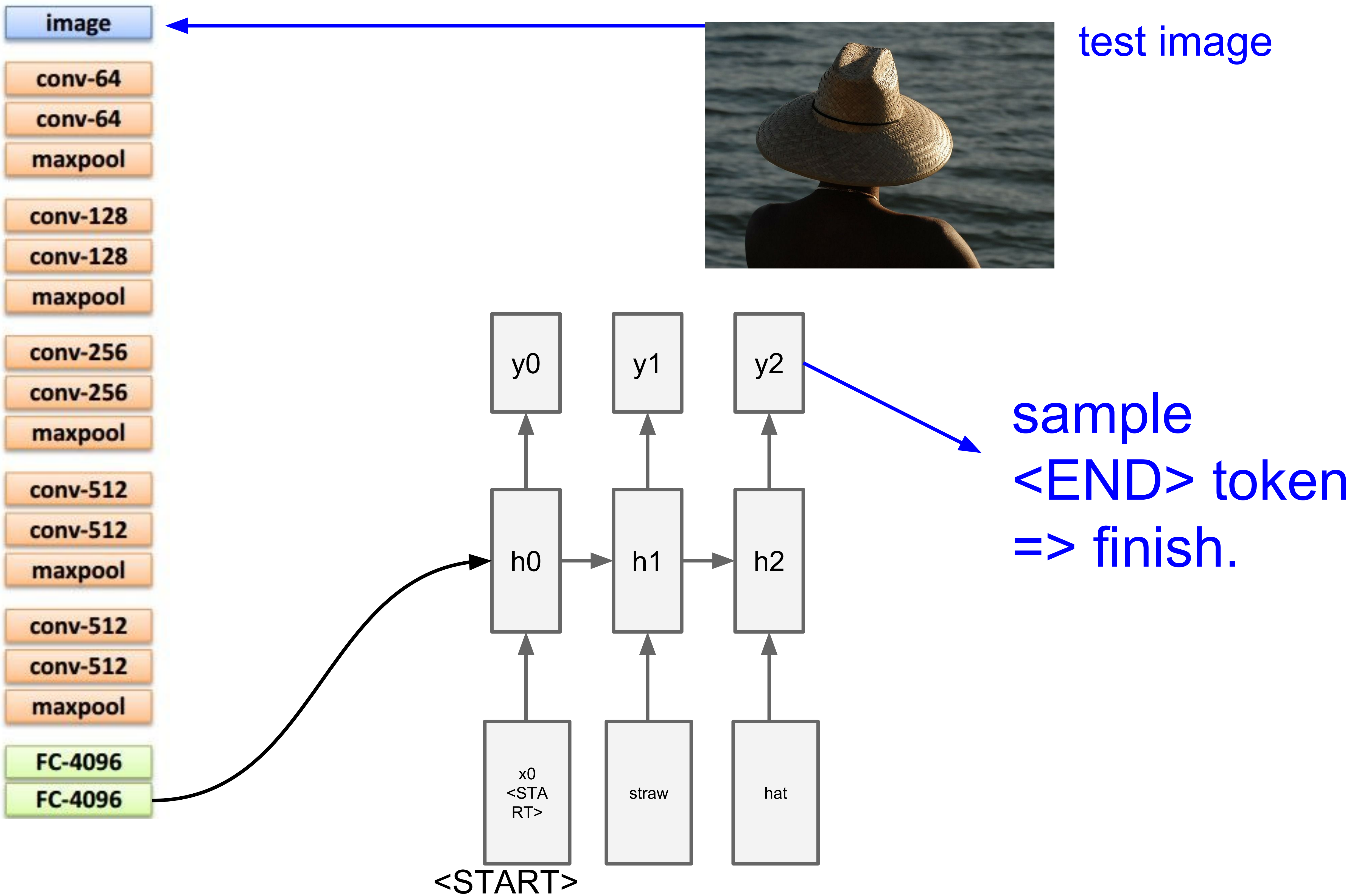
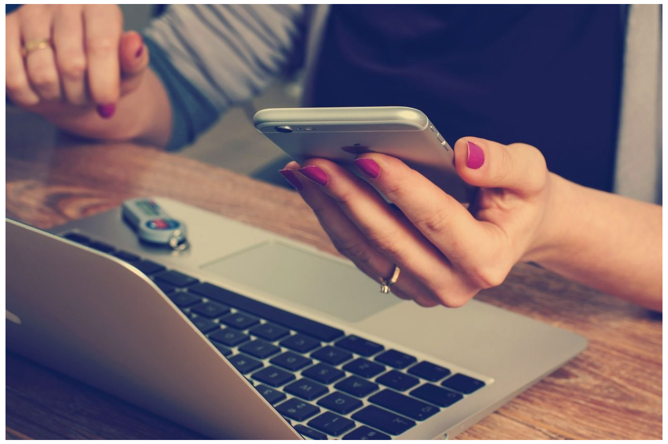


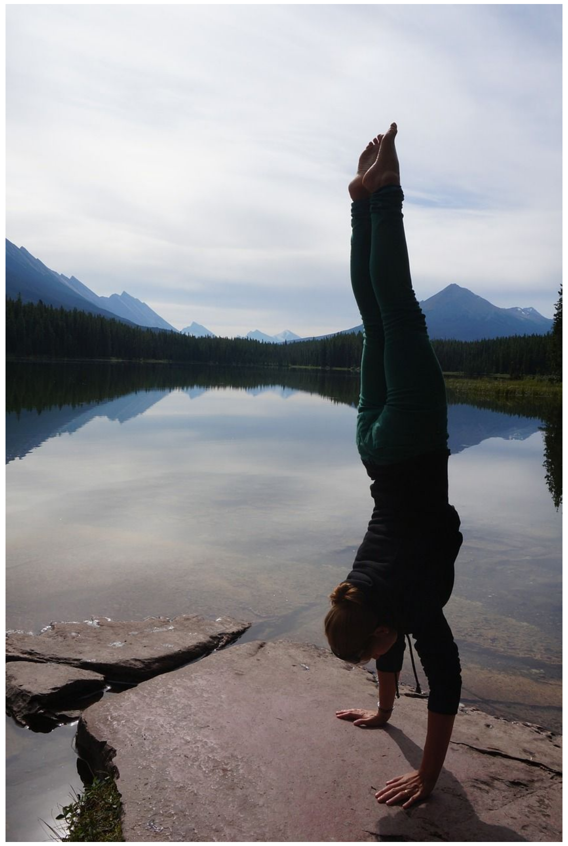
image
conv-64
conv-64
maxpool
conv-128
conv-128
maxpool
conv-256
conv-256
maxpool
conv-512
conv-512
maxpool
conv-512
conv-512
maxpool
FC-4096
FC-4096
test image
y0
y1
y2
h0
h1
h2
x0 <START>
straw
hat
<START>
sample <END> token
=> finish.

Image Captioning: Failure Cases

A woman is holding a cat in her hand



A person holding a computer mouse on a desk



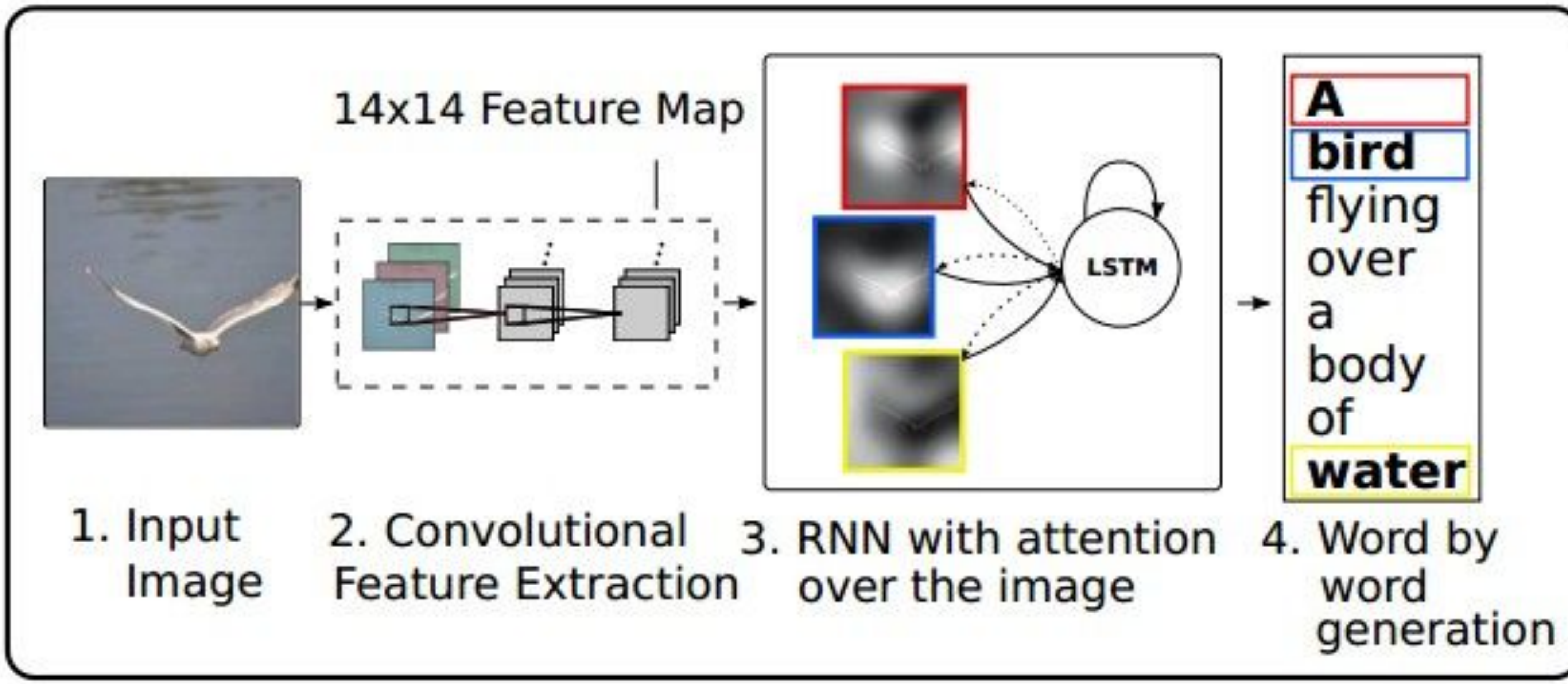
A woman standing on a beach holding a surfboard

A bird is perched on a tree branch

A man in a baseball uniform throwing a ball

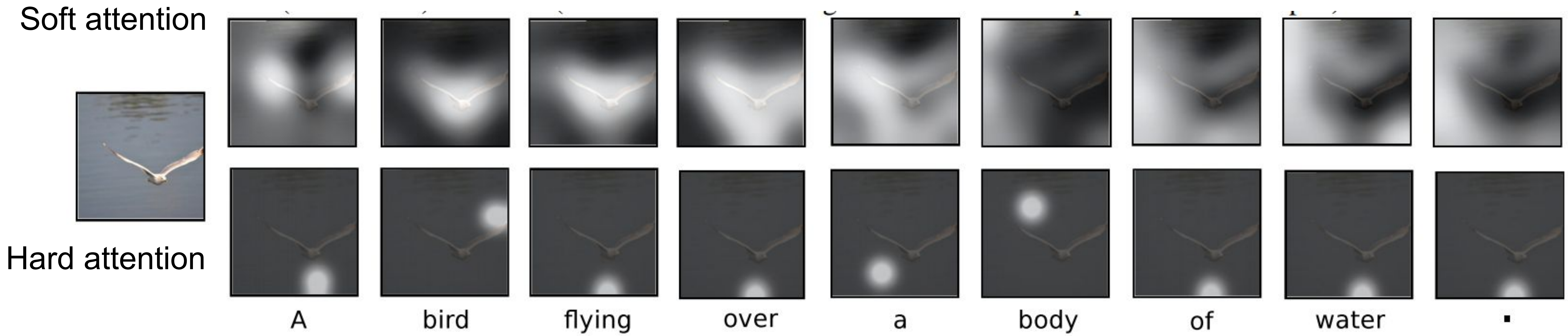
Image Captioning with Attention

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



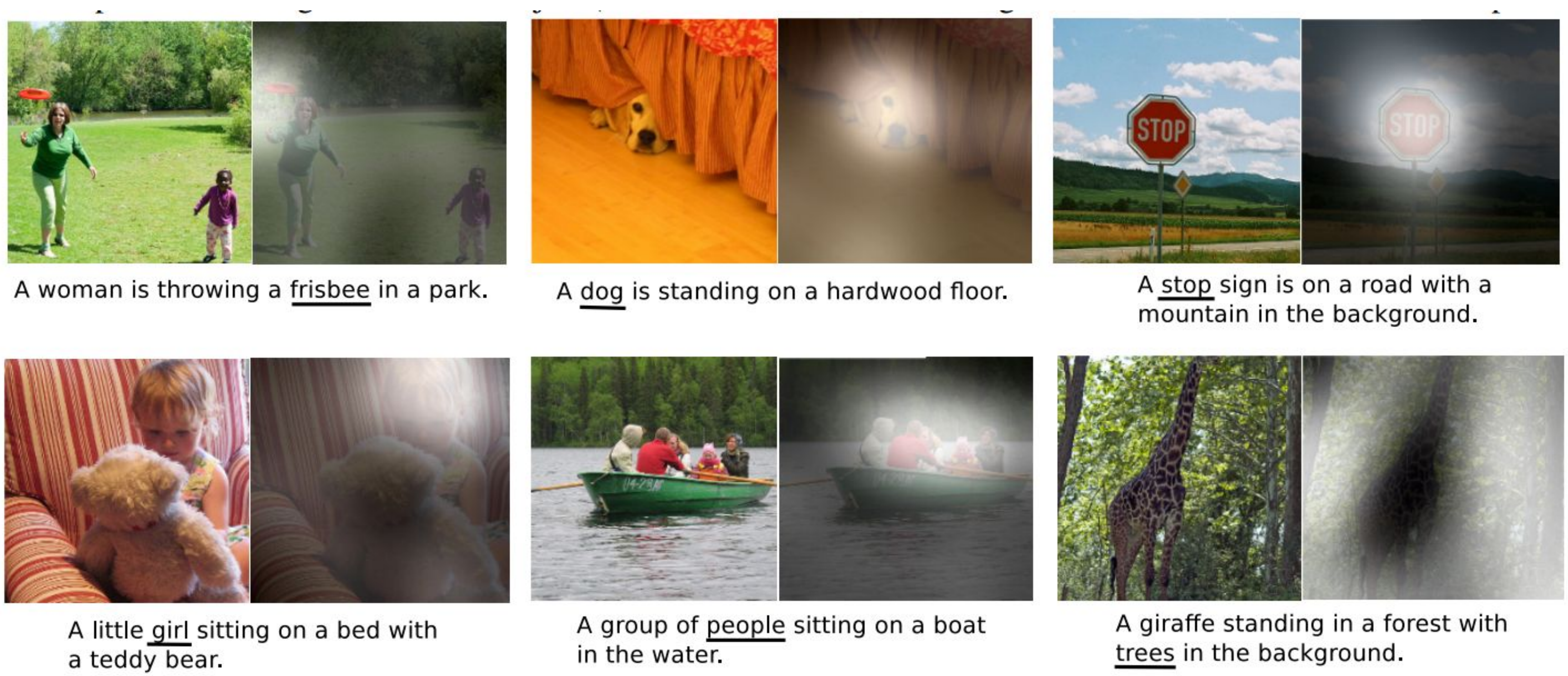
Xu et al, “Show, Attend, and Tell: Neural Image Caption Generation with Visual Attention”, ICML 2015

Image Captioning with Attention



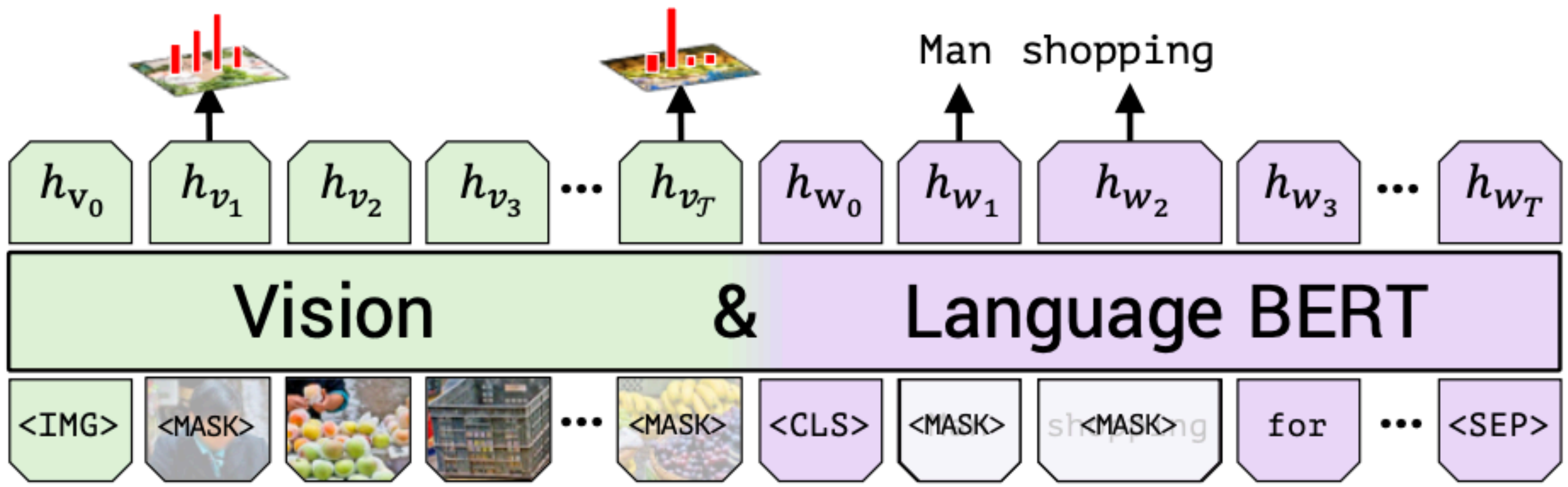
Xu et al, “Show, Attend, and Tell: Neural Image Caption Generation with Visual Attention”, ICML 2015

Image Captioning with Attention

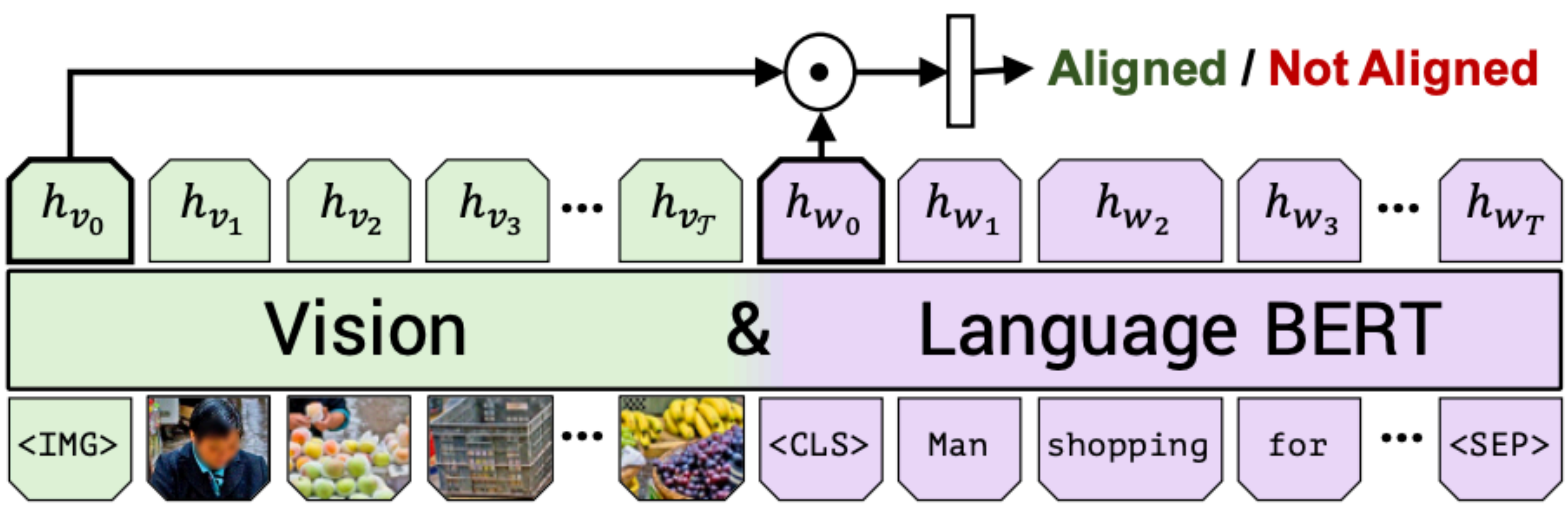


Xu et al, “Show, Attend, and Tell: Neural Image Caption Generation with Visual Attention”, ICML 2015

ViLBERT (vision and language BERT)



(a) Masked multi-modal learning



(b) Multi-modal alignment prediction

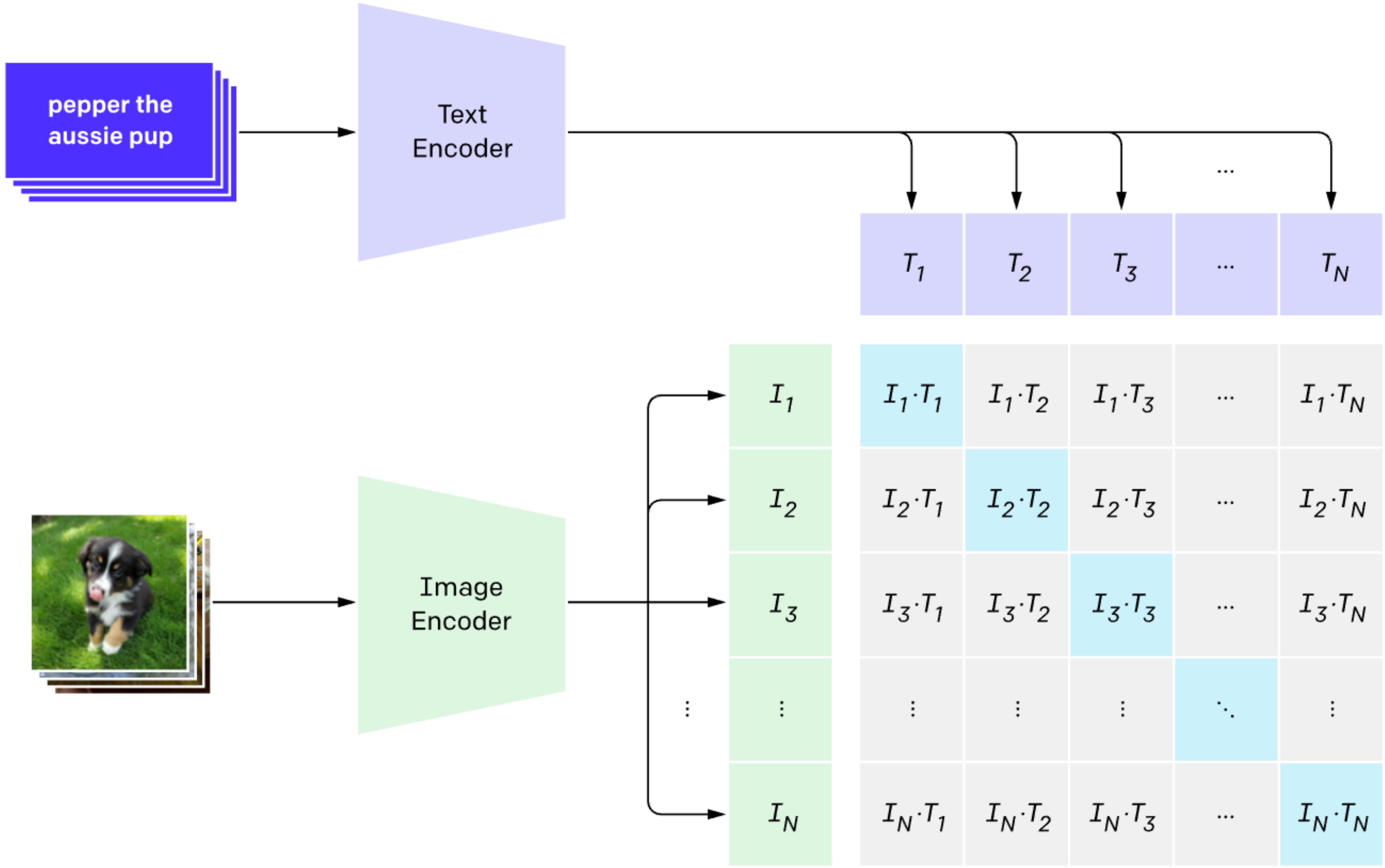
Lu et al., 2019 (“ViLBERT”)

OpenAI's CLIP: Contrastive language-image pretraining

- VilBERT and similar methods (e.g. LXMERT) rely on small labeled datasets like MSCOCO and Visual Genome (~100K images each)
- OpenAI collect 400 million (image, text) pairs from the web
- Then, they train an image encoder and a text encoder with a simple contrastive loss: given a collection of images and text, predict which (image, text) pairs actually occurred in the dataset

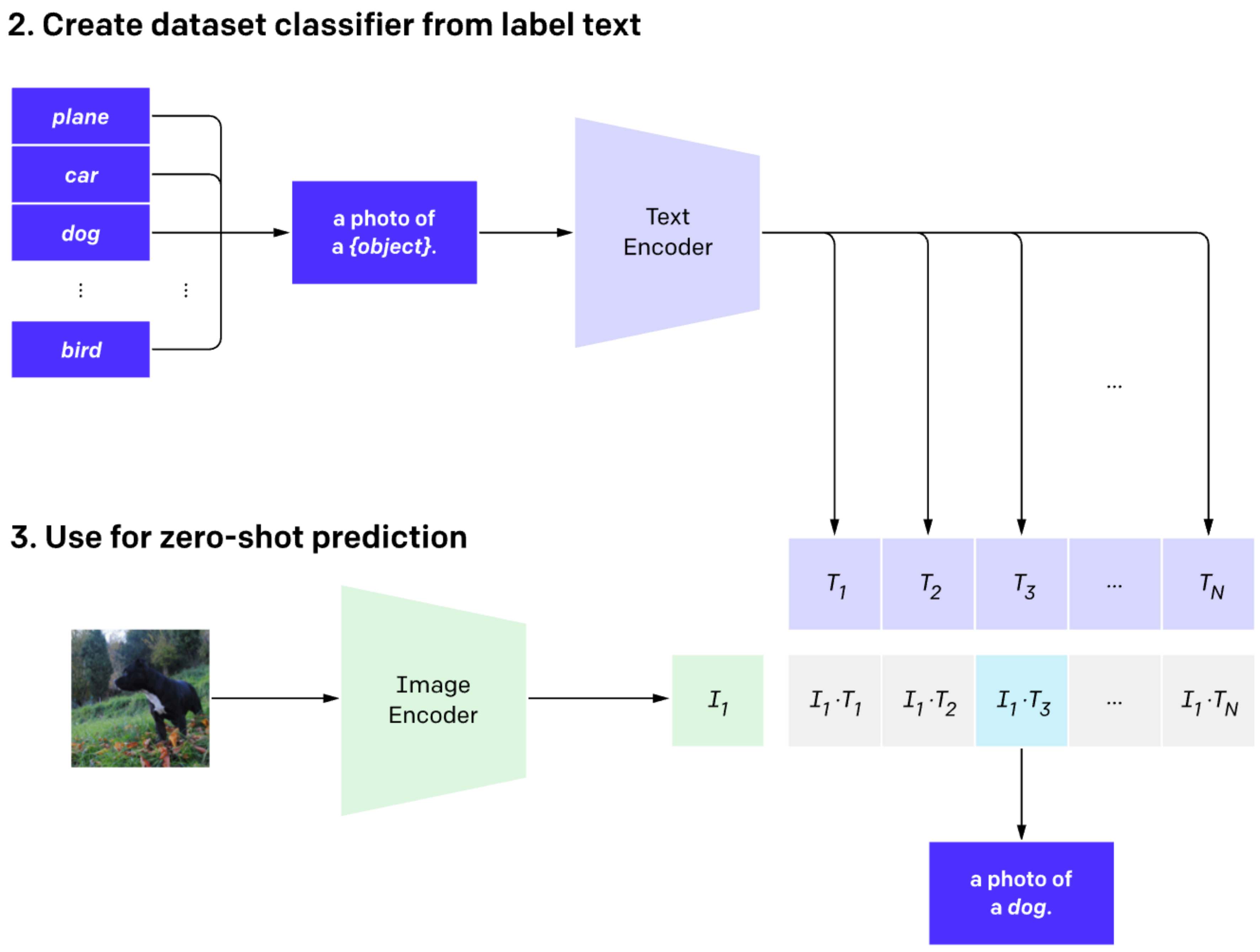
Radford et al., 2021 ("CLIP")

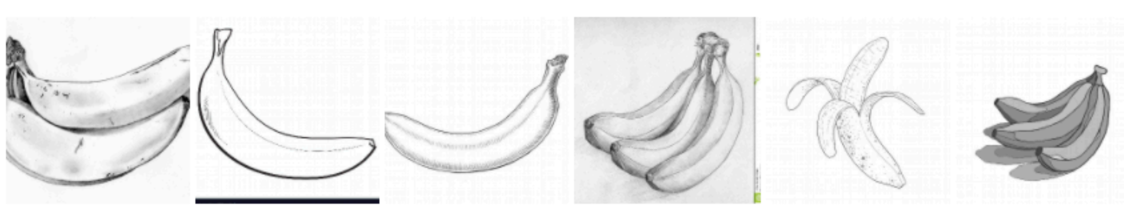
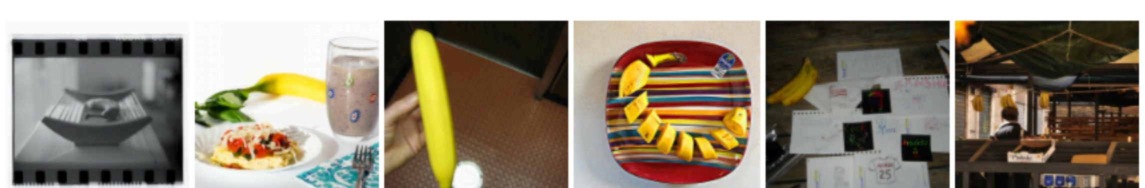
1. Contrastive pre-training



https://openai.com/blog/clip/

Similar to GPT-3, you can use CLIP for zero-shot learning



DATASET	IMAGENET RESNET101	CLIP VIT-L
 ImageNet	76.2%	76.2%
 ImageNet V2	64.3%	70.1%
 ImageNet Rendition	37.7%	88.9%
 ObjectNet	32.6%	72.3%
 ImageNet Sketch	25.2%	60.2%
 ImageNet Adversarial	2.7%	77.1%