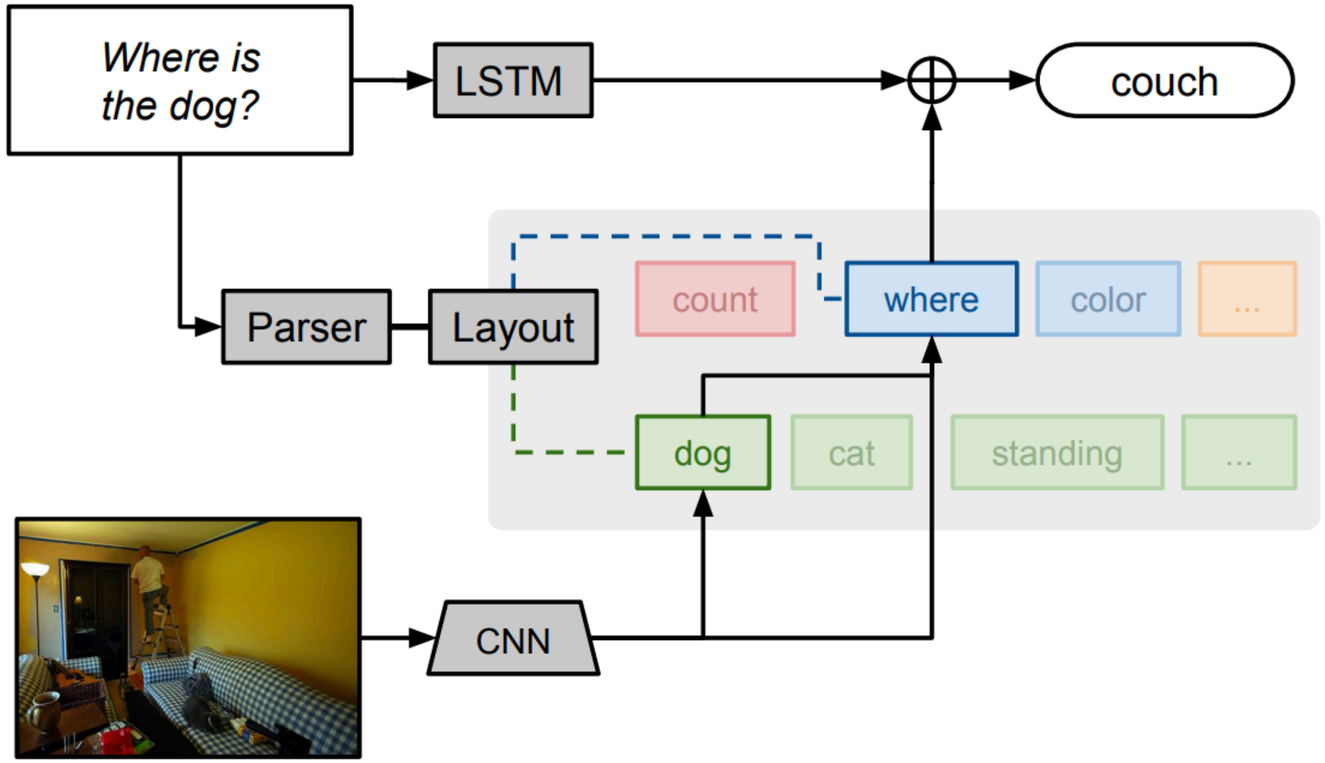
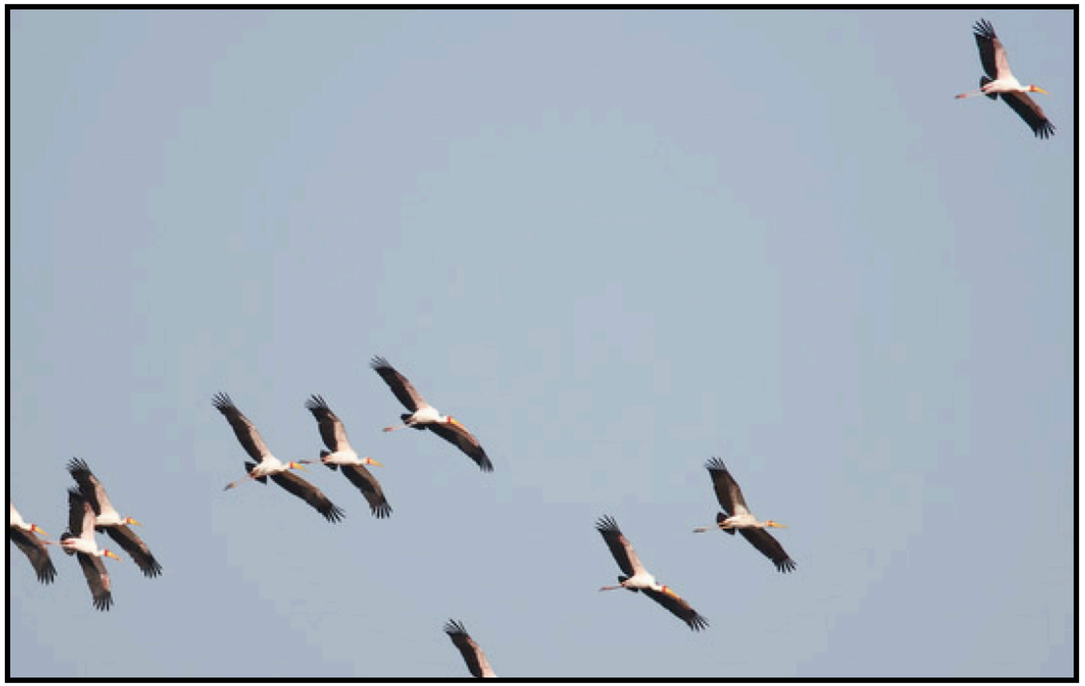


Language and Vision

Bill Freeman, Antonio Torralba, Phillip Isola

6.819 / 6.869

Image captioning



f

"A flock of birds against a gray sky"

Recipe for deep learning in a new domain

1. Transform your data into numbers (e.g., a vector)

2. Transform your goal into an equation (objective function)

3. #1 and #2 specify the “learning problem”

4. Use a generic optimizer (SGD) and an appropriate architecture (e.g., CNN or RNN) to solve the learning problem

How to represent words as numbers?

One-hot vector

Training data

$\mathbf{x}$	y
	"Fish"
	"Grizzly"
	"Chameleon"
⋮	

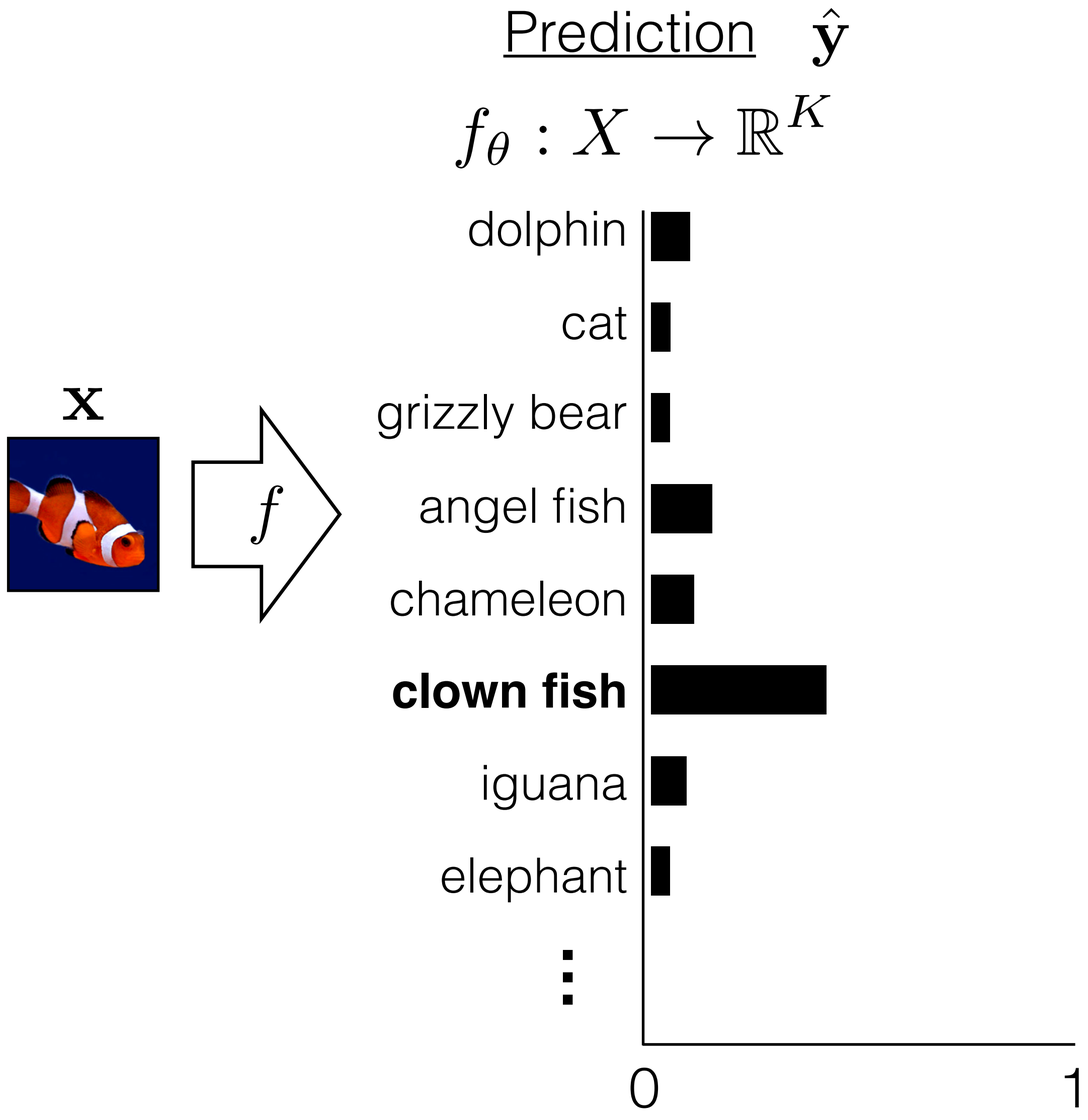
Training data

$\mathbf{x}$	y
	1
	2
	3
⋮	

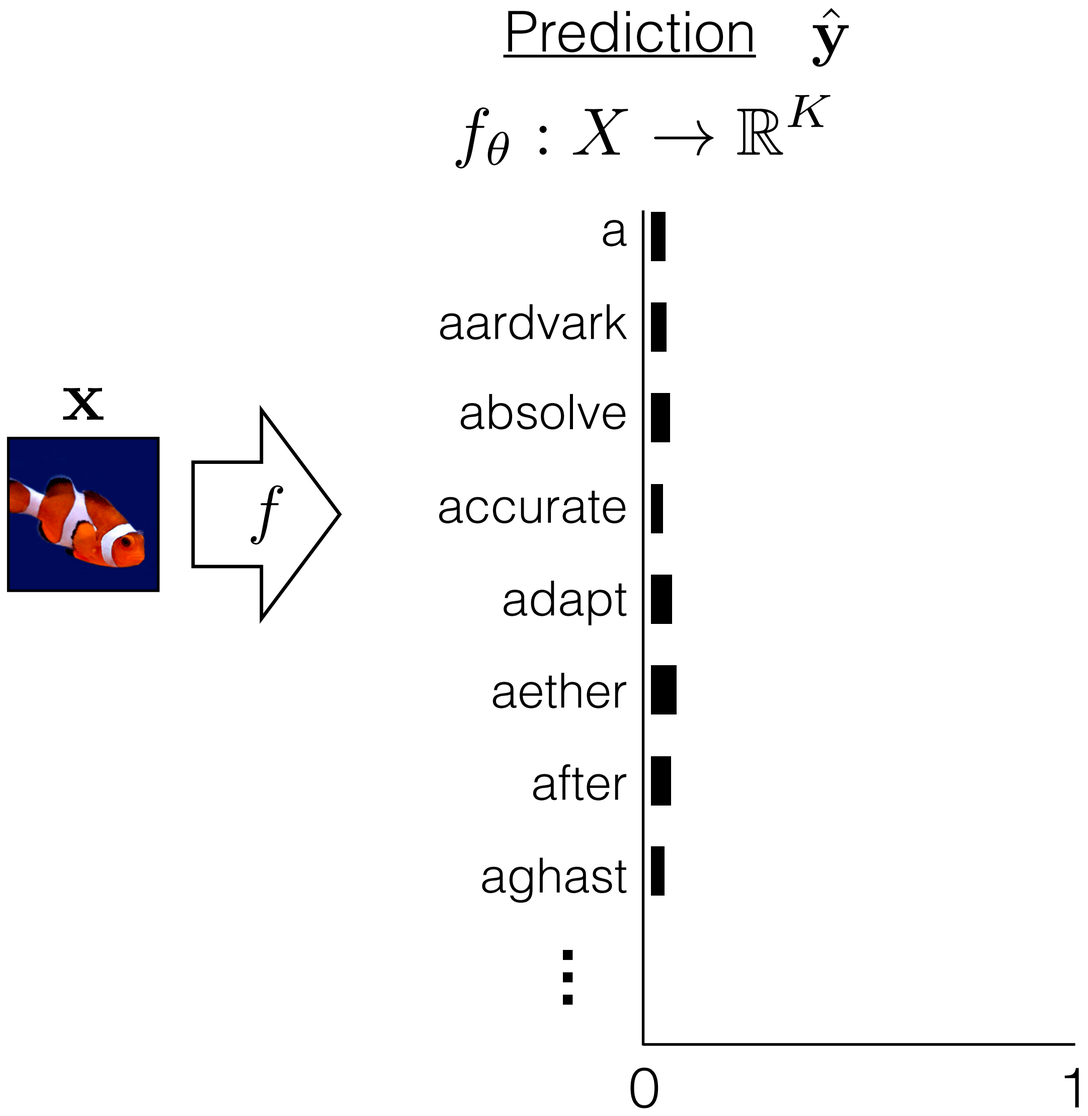
Training data

$\mathbf{x}$	$\mathbf{y}$
	[0,0,1]
	[0,1,0]
	[1,0,0]
⋮	

How to represent words as numbers?

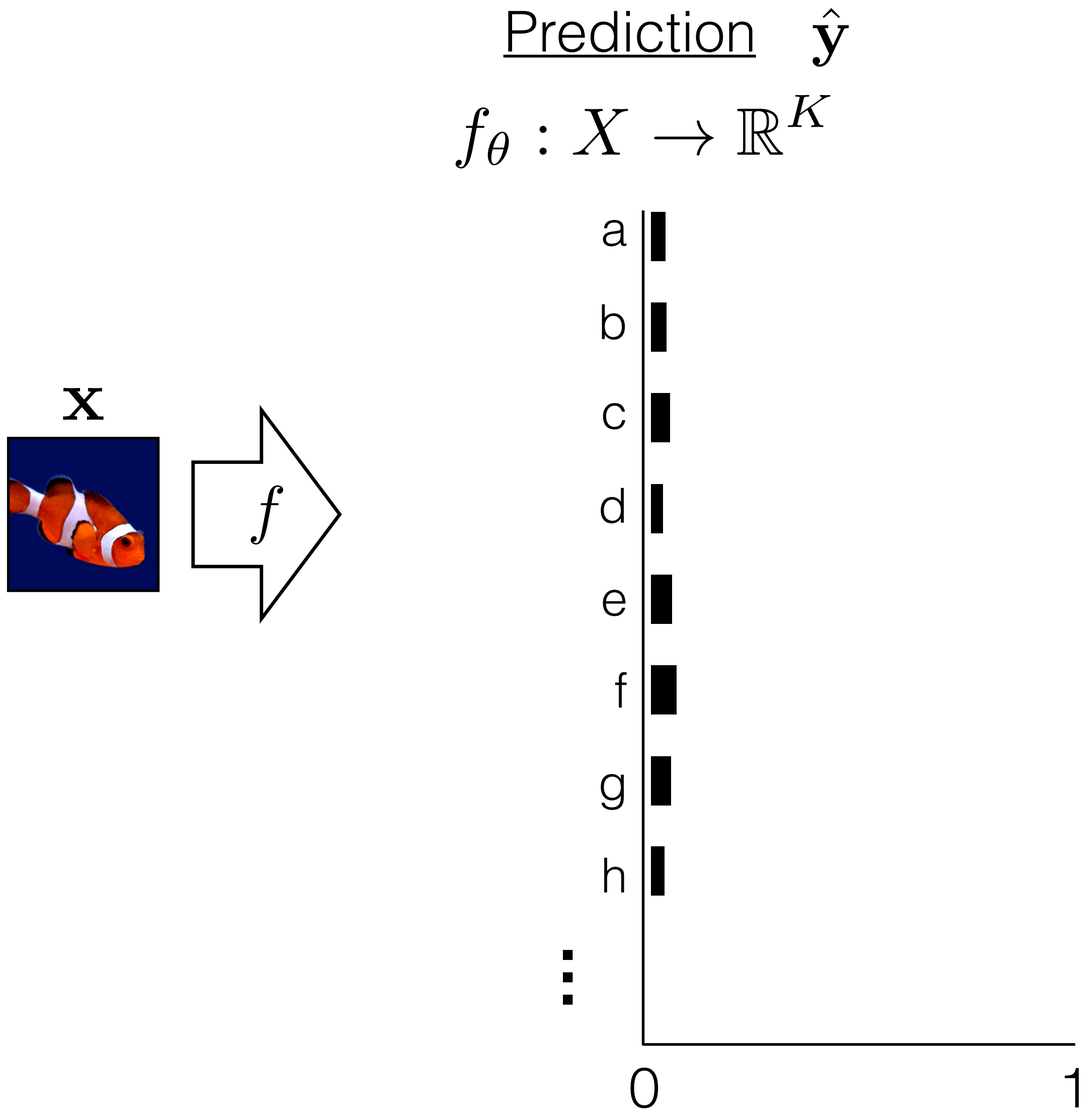


How to represent words as numbers?



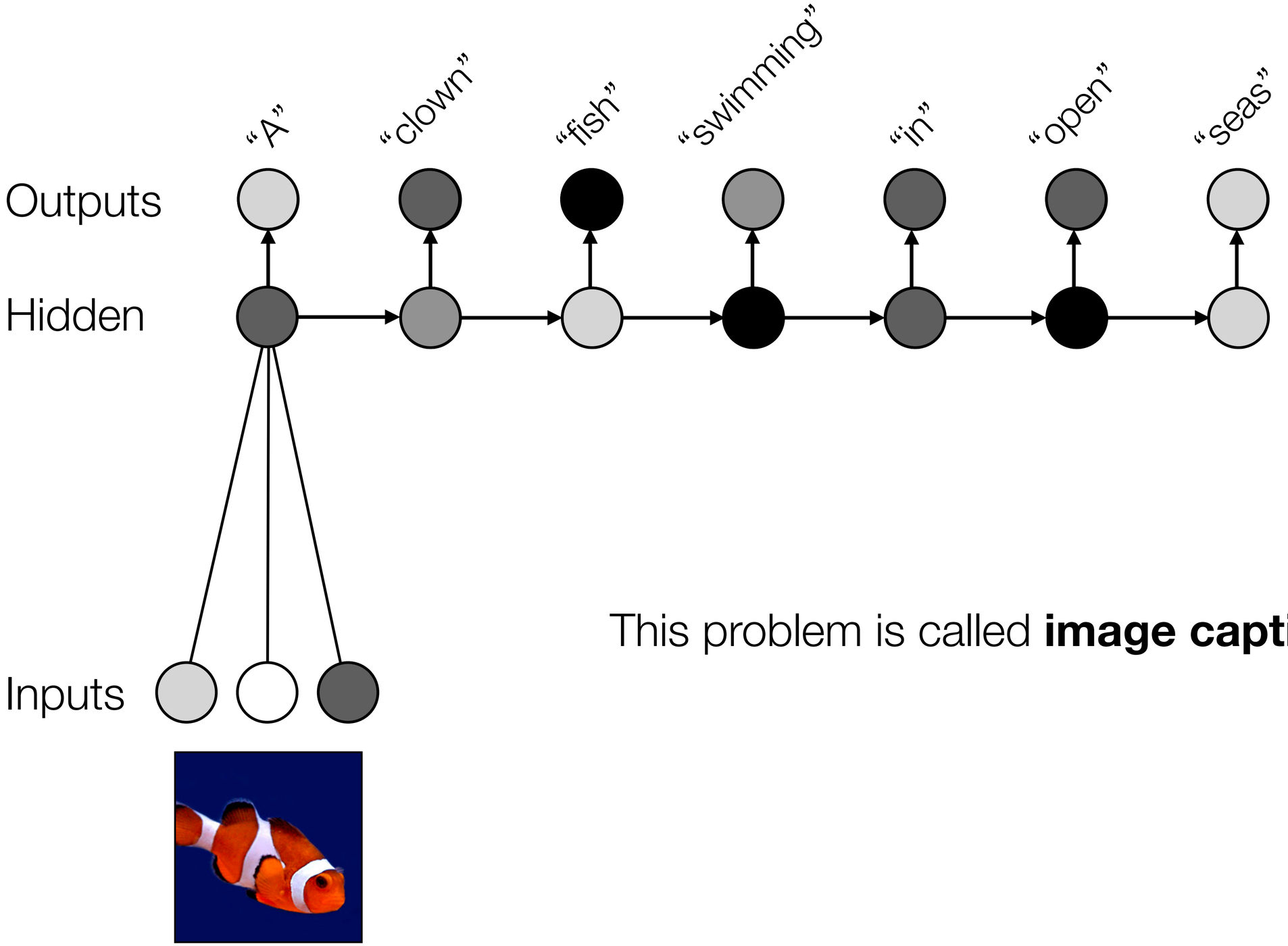
Rather than having just a handful of possible object classes, we can represent all words in a large vocabulary using a very large K (e.g., K=100,000).

How to represent words as numbers?

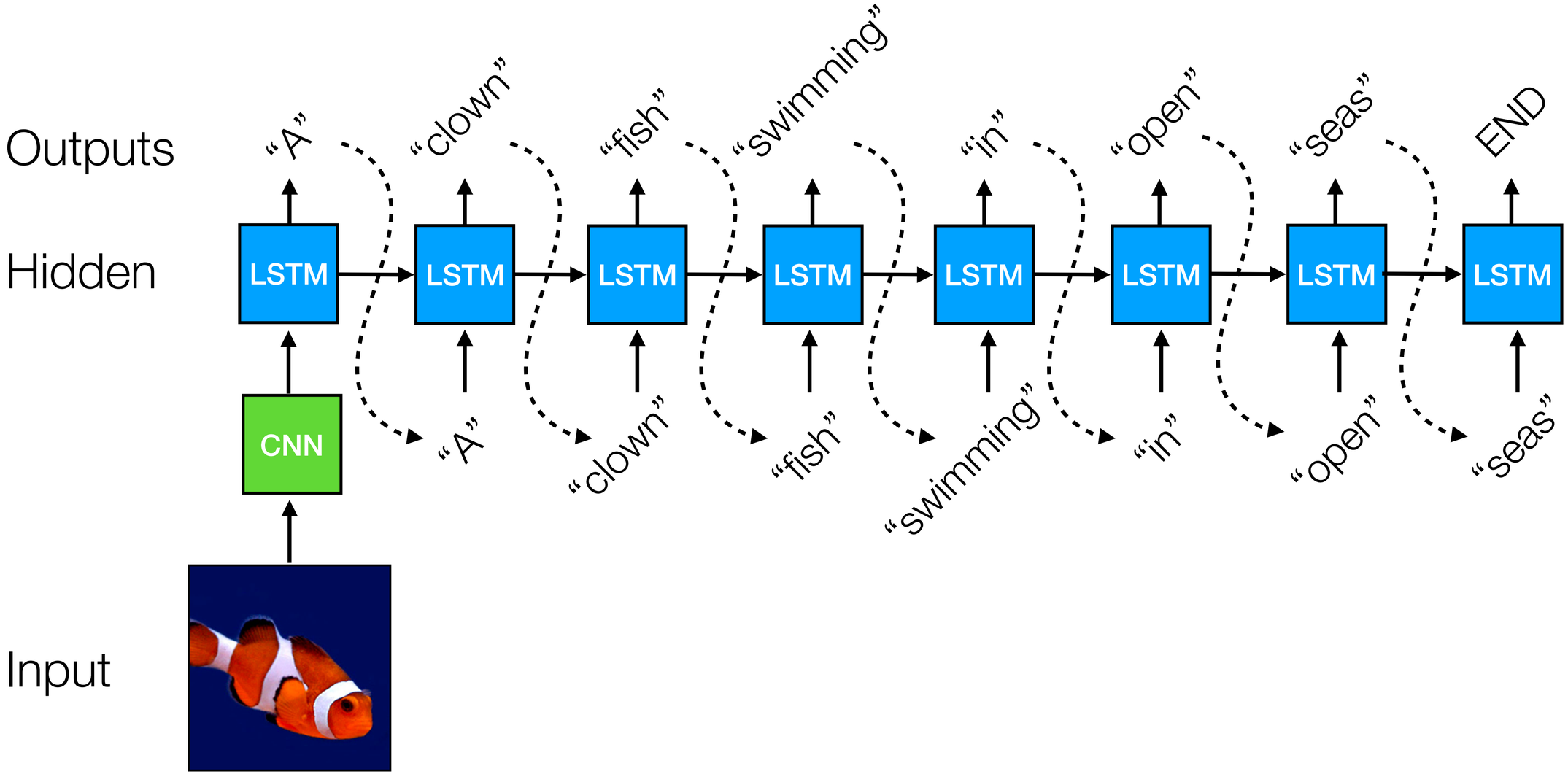


Or, represent each character as a class (e.g., K=26 for English letters),

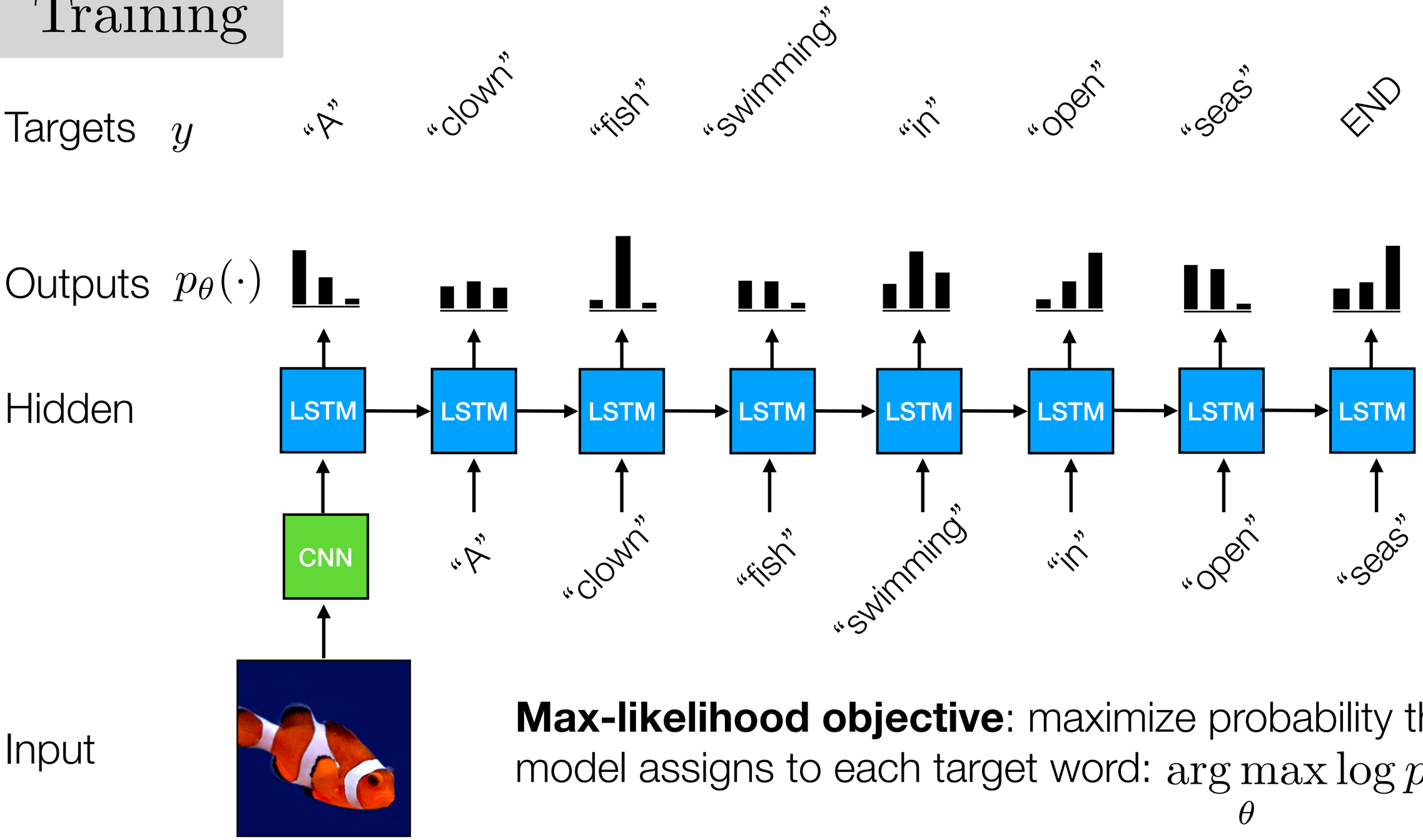
and represent words as a sequence of characters.



This problem is called **image captioning**

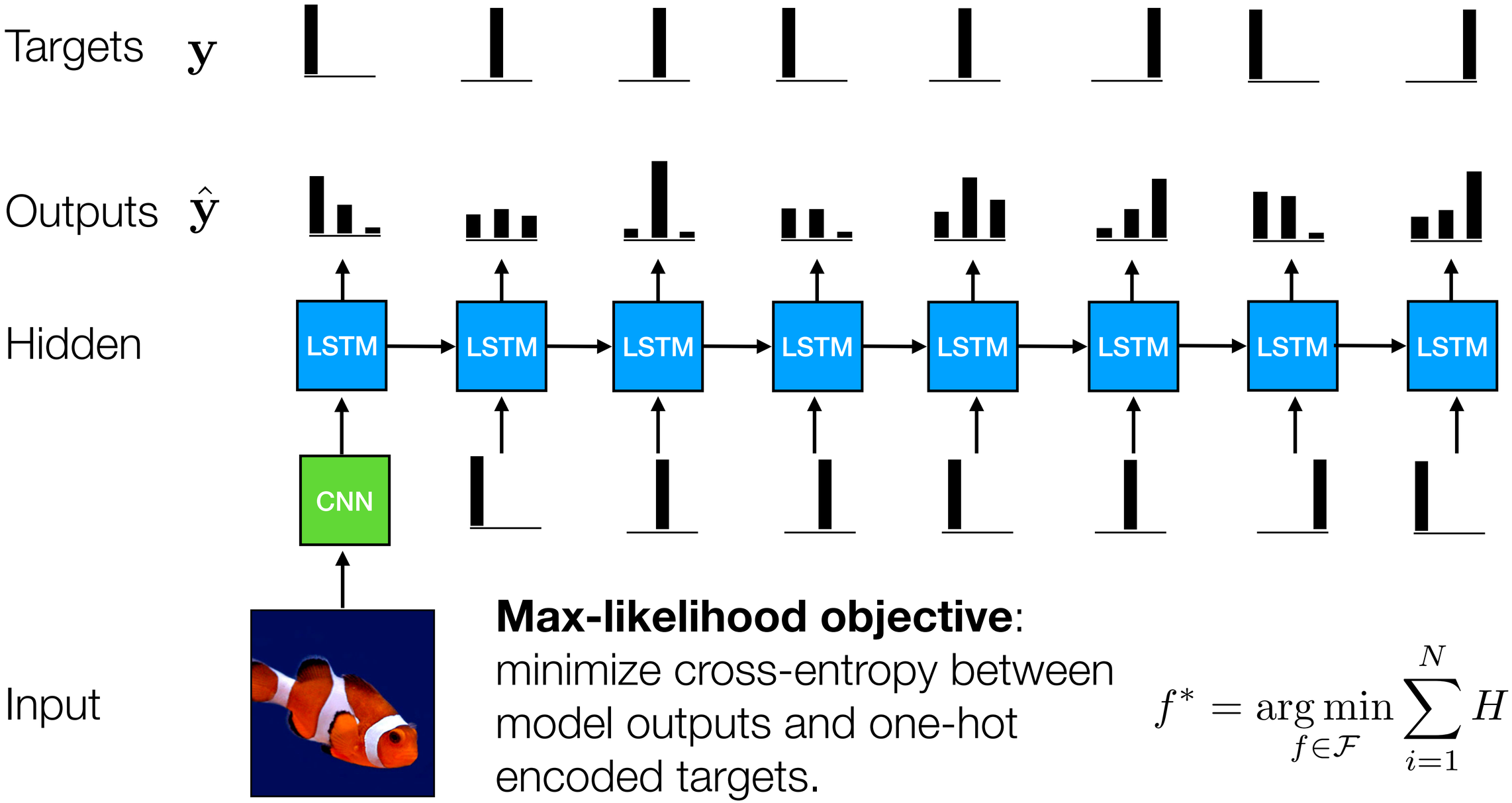

Outputs
Hidden
Input
CNN
LSTM
LSTM
LSTM
LSTM
LSTM
LSTM
LSTM
LSTM
"A"
"clown"
"fish"
"swimming"
"in"
"open"
"seas"
END
"A"
"clown"
"fish"
"swimming"
"in"
"open"
"seas"

Training



Max-likelihood objective: maximize probability the model assigns to each target word: $\arg\max_\theta \log p_\theta(y)$

Training



Max-likelihood objective: minimize cross-entropy between model outputs and one-hot encoded targets.

$$f^* = \arg\min_{f \in \mathcal{F}} \sum_{i=1}^{N} H(\mathbf{y}_i, \hat{\mathbf{y}}_i)$$

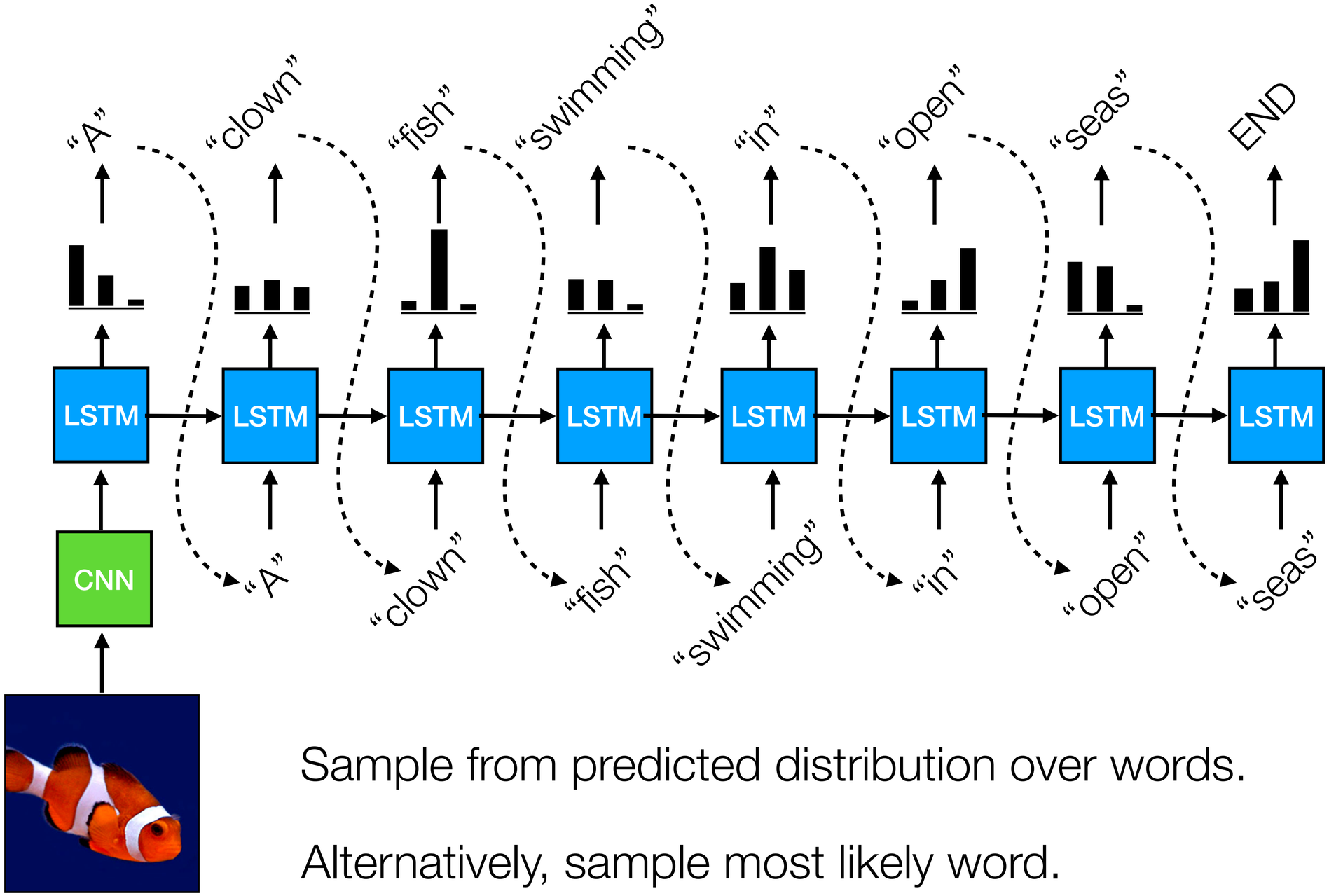
Testing

Samples

Outputs $p_\theta(\cdot)$

Hidden

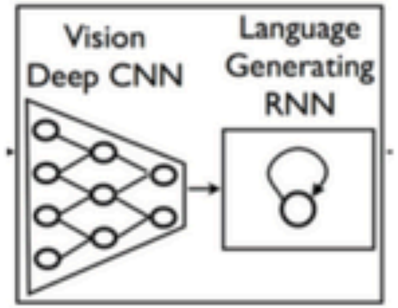
Input



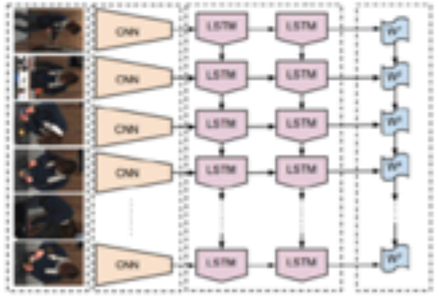
Sample from predicted distribution over words.

Alternatively, sample most likely word.

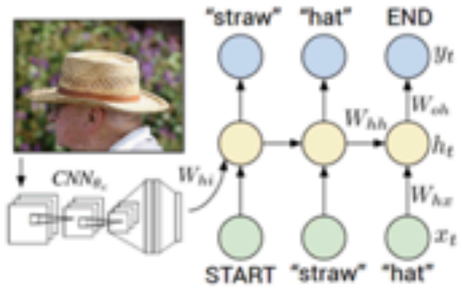
It was very popular a few years ago



Vinyals et al., 2015



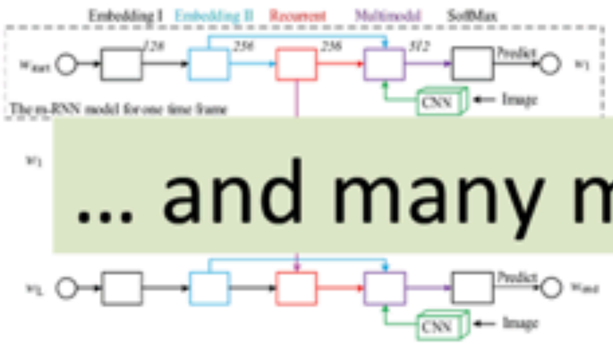
Donahue et al., 2015



Karpathy and Fei-Fei, 2015

Hodosh et al., 2013

Fang et al., 2015

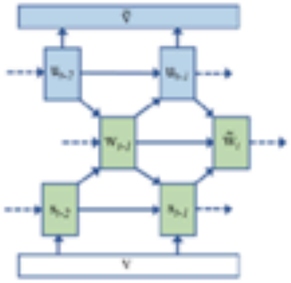


Mao et al., 2015

Ordonez et al., 2011

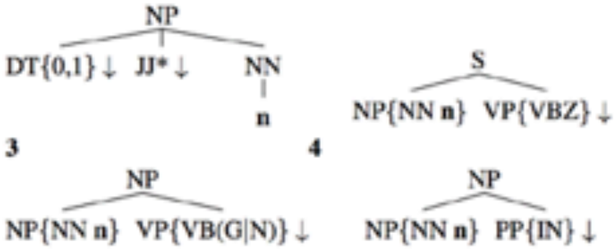
Kulkarni et al., 2011

... and many more

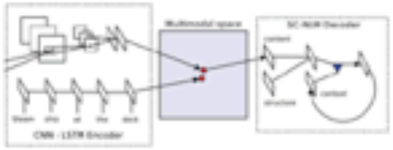


Chen and Zitnick, 2015

Farhadi et al., 2010



Mitchell et al., 2012

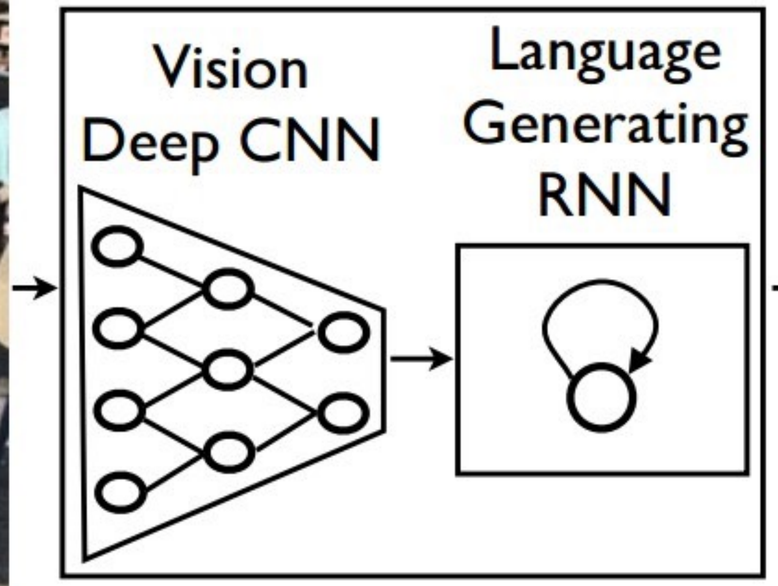


Kiros et al., 2015

Slide credit: Devi Parikh

Show and Tell: A Neural Image Caption Generator

[Vinyals et. al., CVPR 2015]

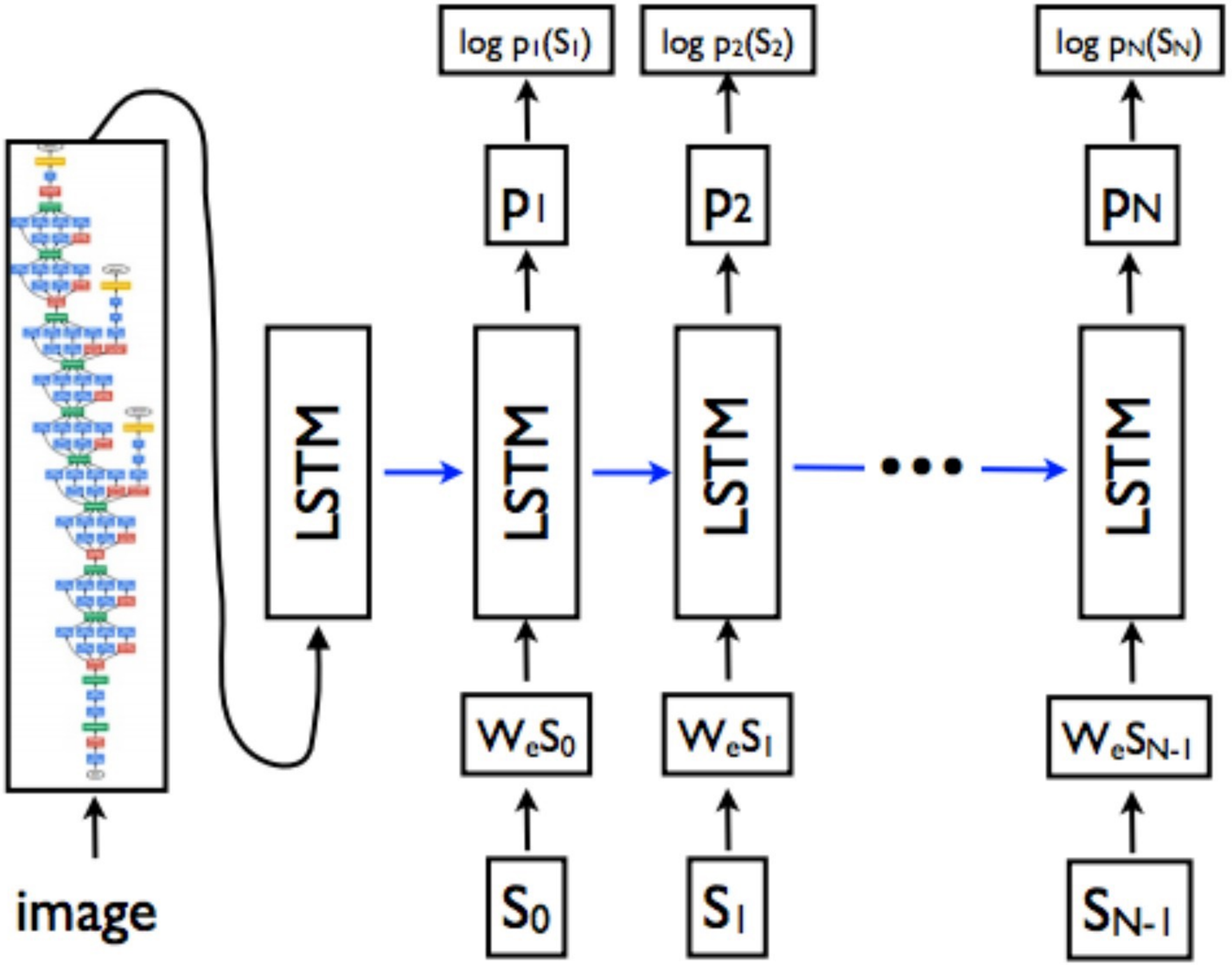


A group of people shopping at an outdoor market.

There are many vegetables at the fruit stand.

Show and Tell: A Neural Image Caption Generator

[Vinyals et. al., CVPR 2015]

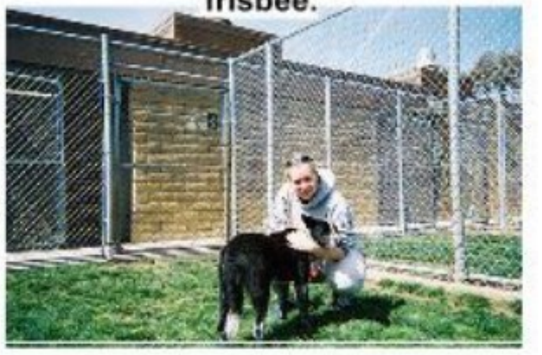


A person riding a motorcycle on a dirt road.

Two dogs play in the grass.

A skateboarder does a trick on a ramp.

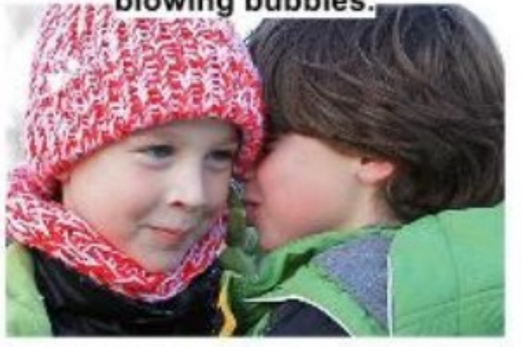
A dog is jumping to catch a frisbee.



A group of young people playing a game of frisbee.

Two hockey players are fighting over the puck.

A little girl in a pink hat is blowing bubbles.



A refrigerator filled with lots of food and drinks.

A herd of elephants walking across a dry grass field.

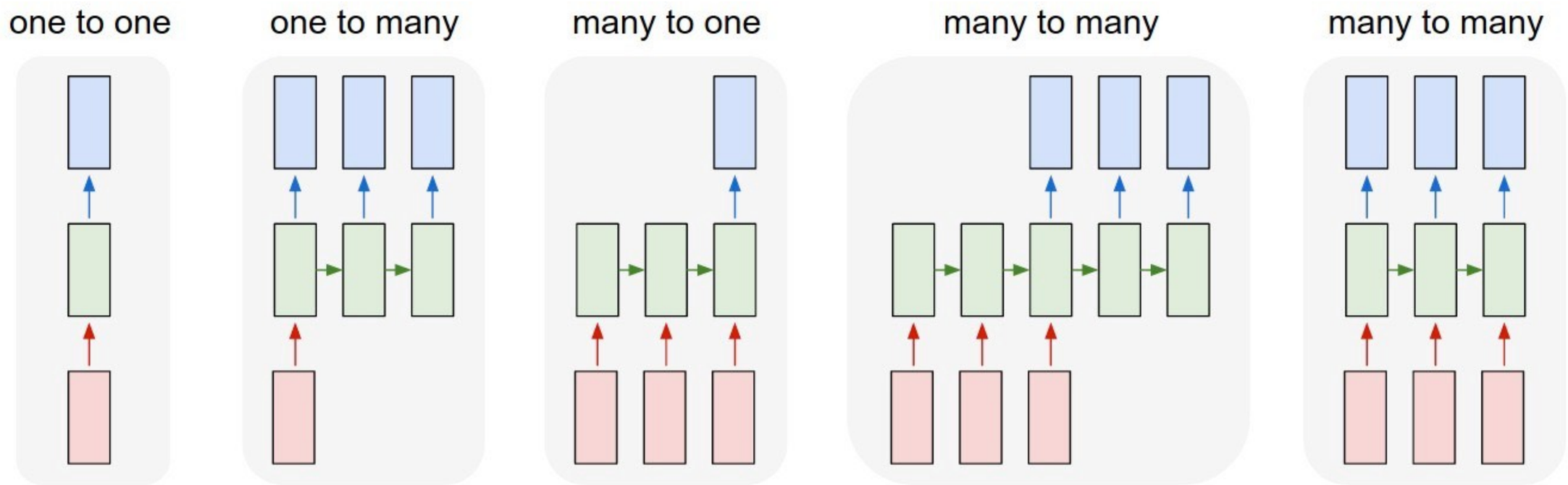
A close up of a cat laying on a couch.

A red motorcycle parked on the side of the road.

A yellow school bus parked in a parking lot.

Describes without errors	Describes with minor errors	Somewhat related to the image	Unrelated to the image

How do we model sequences?



Input: No sequence

Output: No sequence

Example: “standard” classification / regression problems

Input: No sequence

Output: Sequence

Example: Im2Caption

Input: Sequence

Output: No sequence

Example: sentence classification, multiple-choice question answering

Input: Sequence

Output: Sequence

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

[Slide credit: Dhruv Batra, Andrej Karpathy]

http://karpathy.github.io/2015/05/21/rnn-effectiveness/

VQA: Visual Question Answering

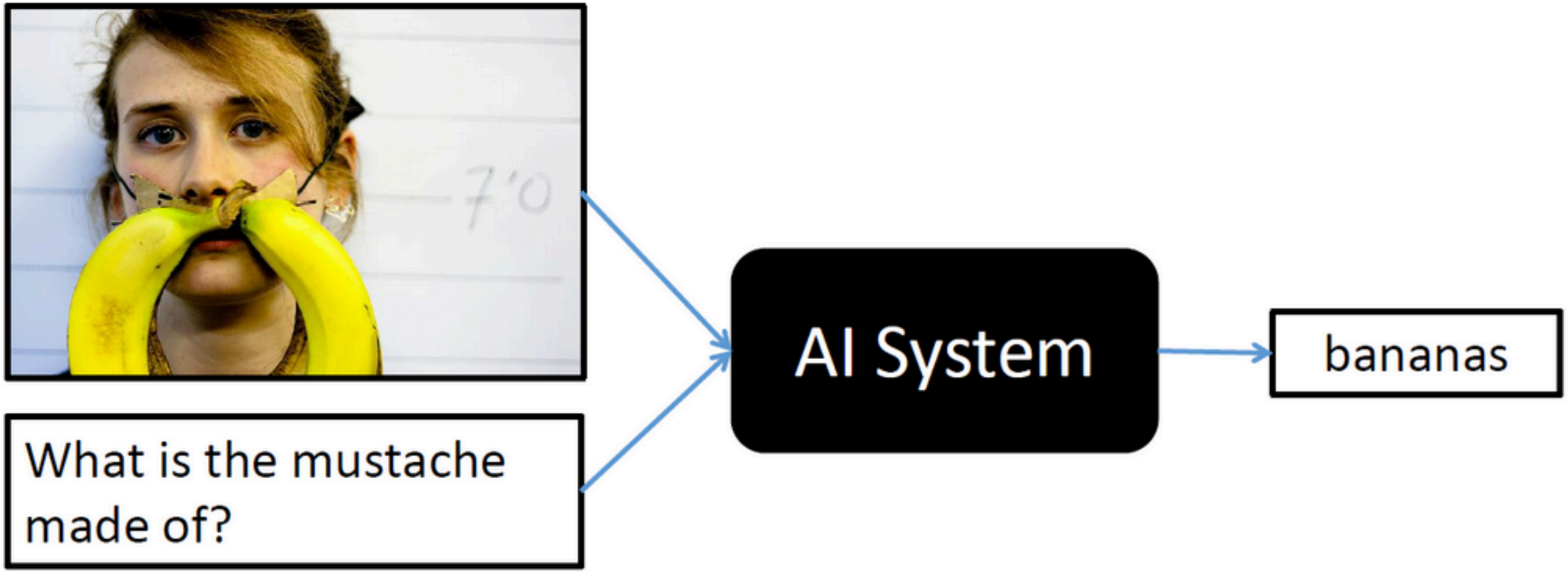
www.visualqa.org

Aishwarya Agrawal*, Jiasen Lu*, Stanislaw Antol*,
Margaret Mitchell, C. Lawrence Zitnick, Dhruv Batra, Devi Parikh

Abstract—We propose the task of *free-form* and *open-ended* Visual Question Answering (VQA). Given an image and a natural language question about the image, the task is to provide an accurate natural language answer. Mirroring real-world scenarios, such as helping the visually impaired, both the questions and answers are open-ended. Visual questions selectively target different areas of an image, including background details and underlying context. As a result, a system that succeeds at VQA typically needs a more detailed understanding of the image and complex reasoning than a system producing generic image captions. Moreover, VQA is amenable to automatic evaluation, since many open-ended answers contain only a few words or a closed set of answers that can be provided in a multiple-choice format. We provide a dataset containing ~0.25M images, ~0.76M questions, and ~10M answers (www.visualqa.org), and discuss the information it provides. Numerous baselines and methods for VQA are provided and compared with human performance.

2016

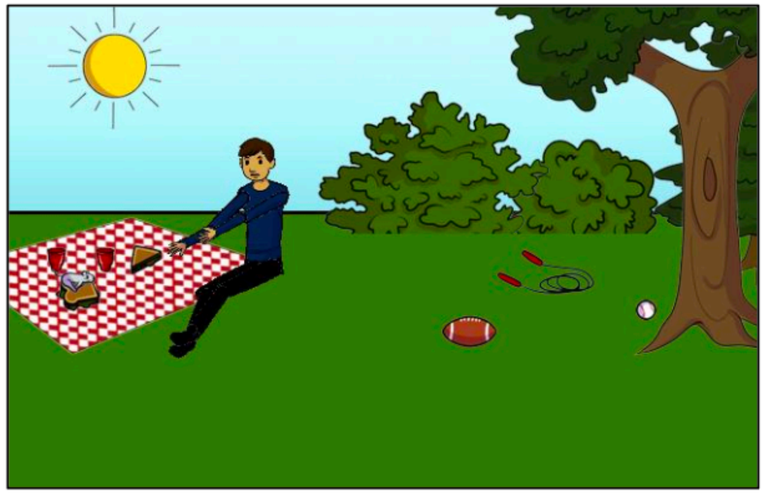
https://arxiv.org/pdf/1505.00468v6.pdf



http://www.visualqa.org/challenge.html

What color are her eyes?
What is the mustache made of?

How many slices of pizza are there?
Is this a vegetarian pizza?

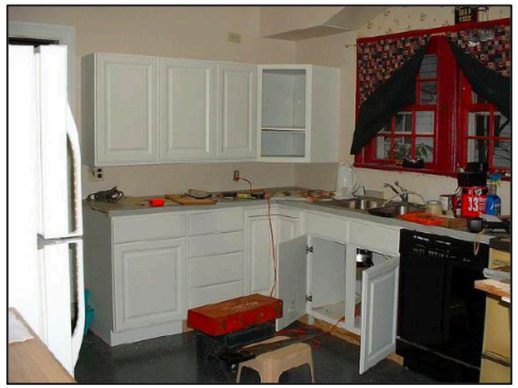


Is this person expecting company?
What is just under the tree?

Does it appear to be rainy?
Does this person have 20/20 vision?

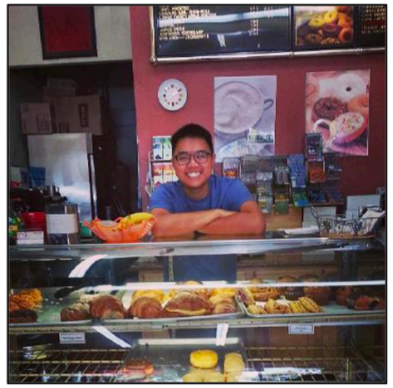
Fig. 1: Examples of free-form, open-ended questions collected for images via Amazon Mechanical Turk. Note that commonsense knowledge is needed along with a visual understanding of the scene to answer many questions.

Questions and answers collected with Amazon Mechanical Turk

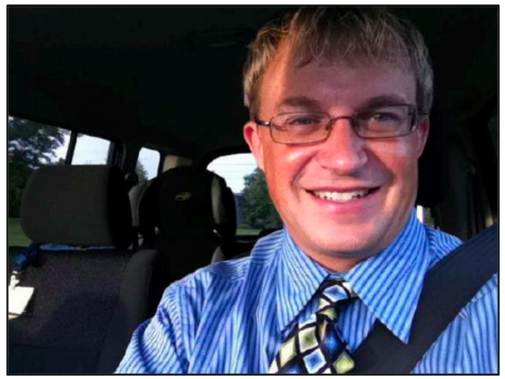


Is something under the sink broken?	yes yes yes	no no no
What number do you see?	33 33 33	5 6 7

Can you park here?	no no no	no no yes
What color is the hydrant?	white and orange white and orange white and orange	red red yellow



What kind of store is this?	bakery bakery pastry	art supplies grocery grocery
Is the display case as full as it could be?	no no no	no yes yes



Does this man have children?	yes yes yes	yes yes yes
Is this man crying?	no no no	no yes yes

Has the pizza been baked?	yes yes yes	yes yes yes
What kind of cheese is topped on this pizza?	feta feta ricotta	mozzarella mozzarella mozzarella

How many pickles are on the plate?	1 1 1	1 1 1
What is the shape of the plate?	circle round round	circle round round

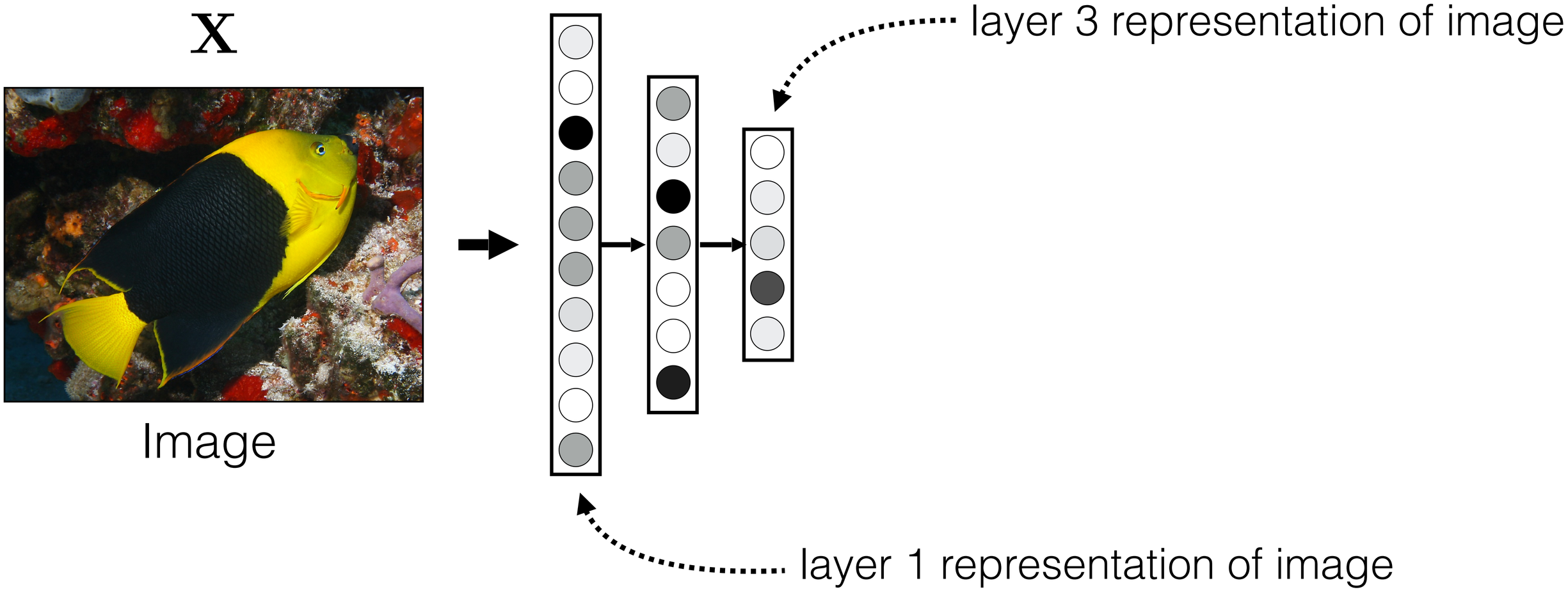
Fig. 2: Examples of questions (black), (a subset of the) answers given when looking at the image (green), and answers given when not looking at the image (blue) for numerous representative examples of the dataset. See the appendix for more examples.

Architecture

How to represent words as numbers

“Fish”	→	[**1**,0,0,0,0,0,0,…]
“Shark”	→	[0,**1**,0,0,0,0,0,…]
“Whale”	→	[0,0,**1**,0,0,0,0,…]
“Water”	→	[0,0,0,**1**,0,0,0,…]
“Cat”	→	[0,0,0,0,**1**,0,0,…]
“Couch”	→	[0,0,0,0,0,**1**,0,…]
“Sun”	→	[0,0,0,0,0,0,**1**,…]

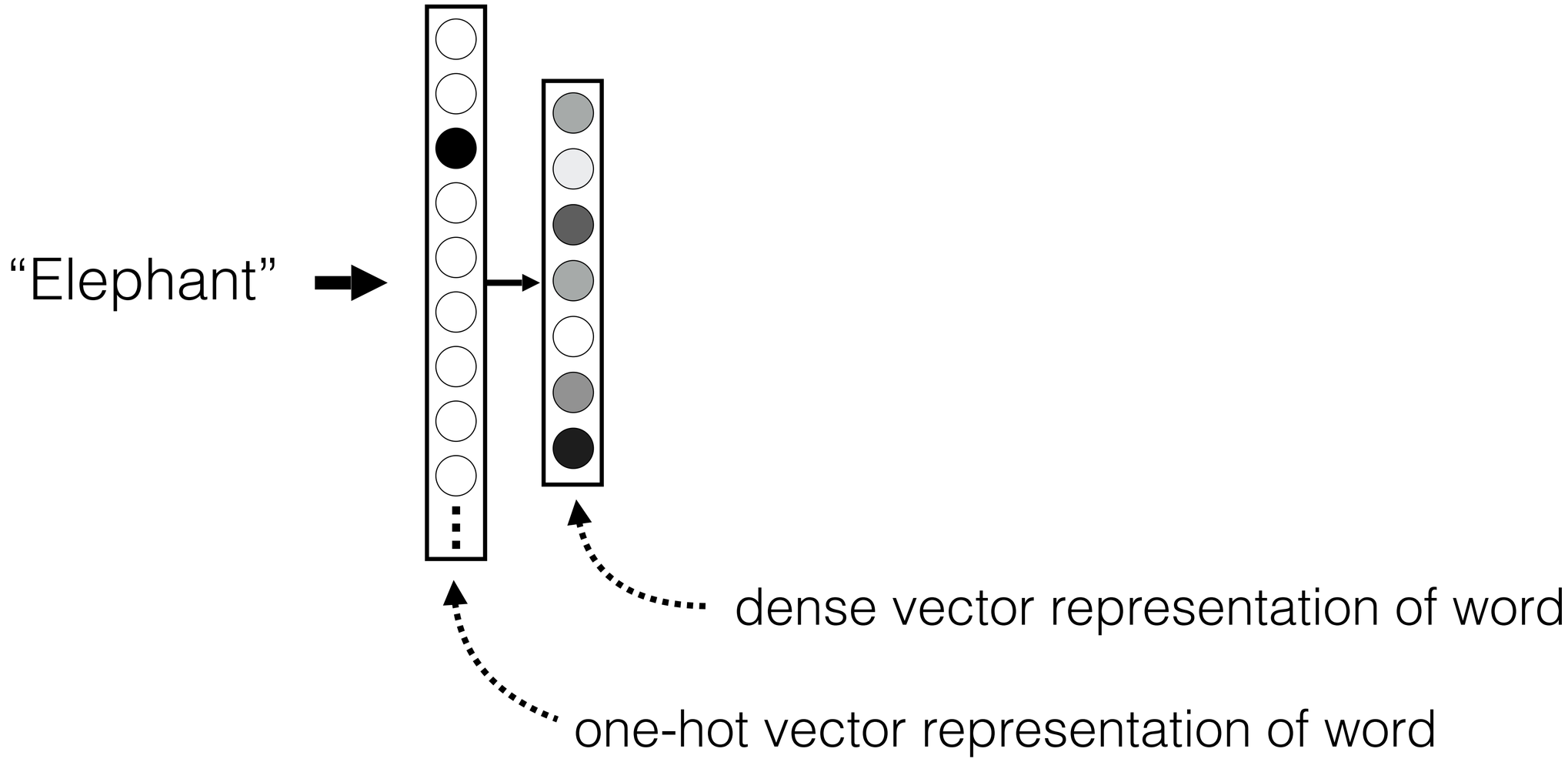
im2vec



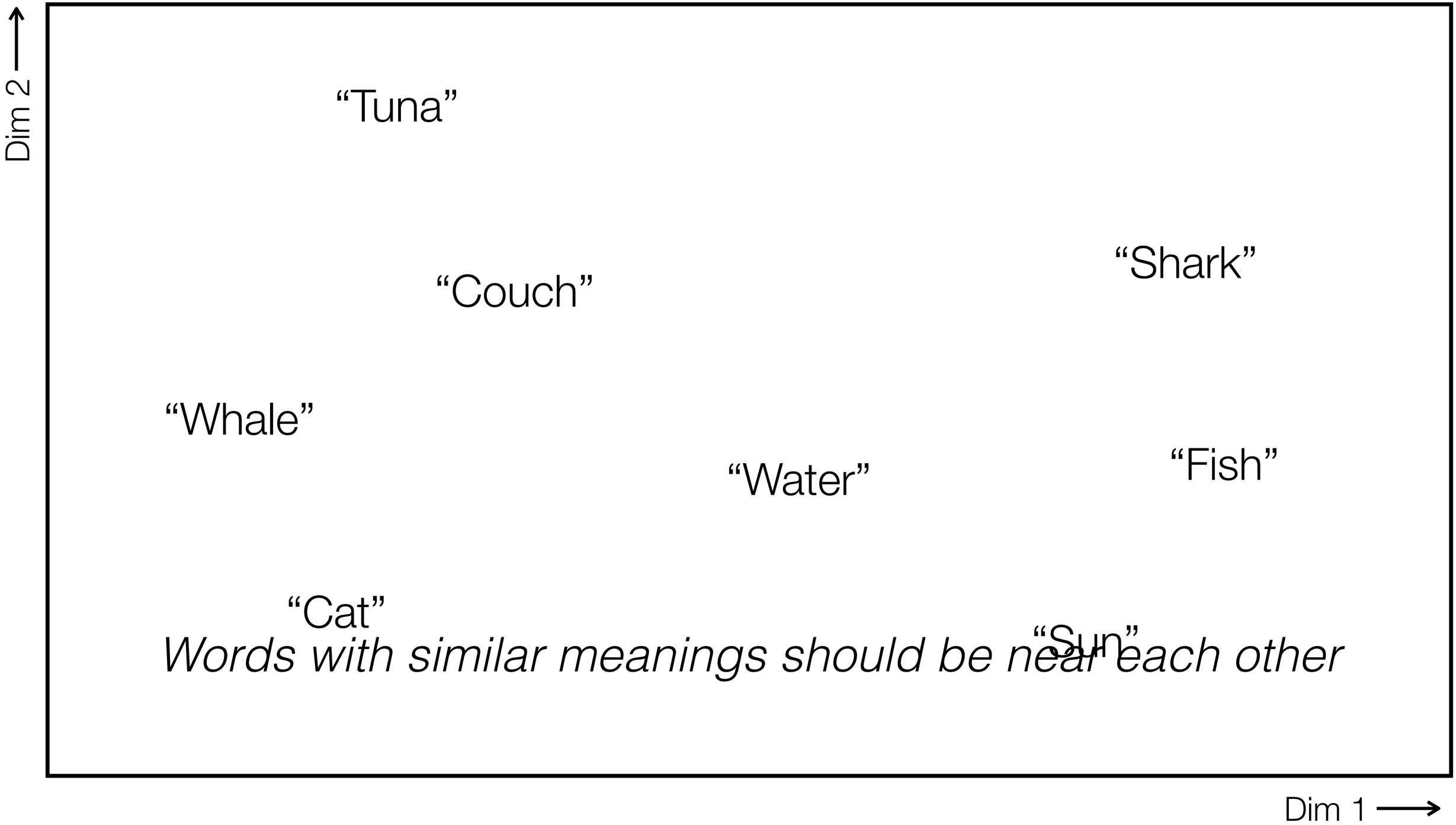
Represent image as a vector of neural activations

(perhaps representing a vector of detected texture patterns or object parts)

word2vec



X2vec methods are also called embeddings of X, e.g., a **word embedding**



Words with similar meanings should be near each other

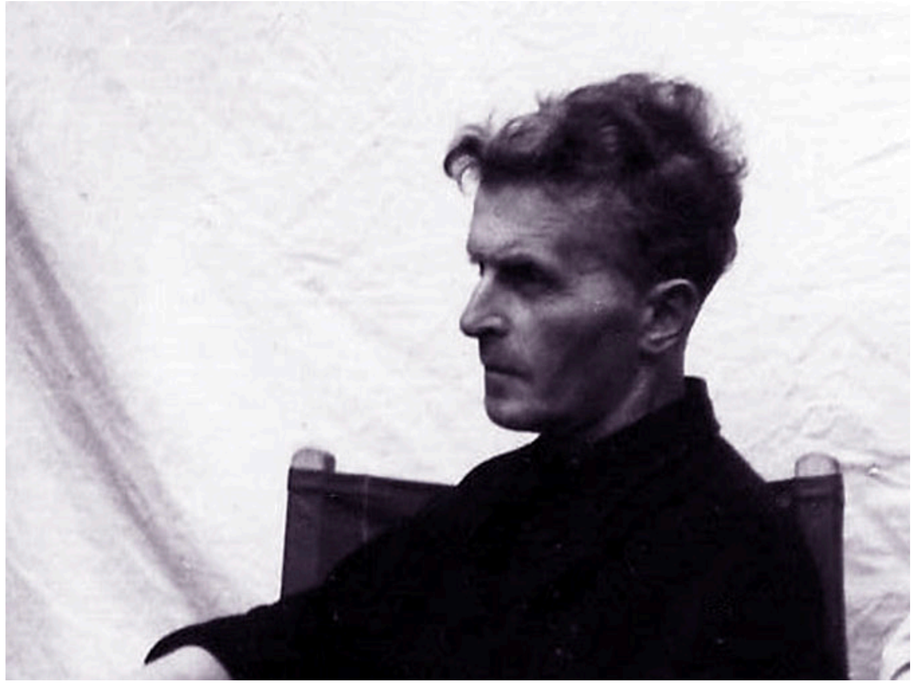
word2vec

Words with similar meanings should be near each other

Proxy: words that are used in the same context tend to have similar meanings

words with similar contexts should be near each other

“Meaning is use” — Wittgenstein



Next to the 'sofa' is a desk, and a 'person' is sitting behind it.

'armchair'	'man'
'bench'	'woman'
'chair'	'child'
'deck chair'	'teenager'
'ottoman'	'girl'
'seat'	'boy'
'stool'	'baby'
'swivel chair'	'daughter'
'loveseat'	'son'
...	...

word2vec

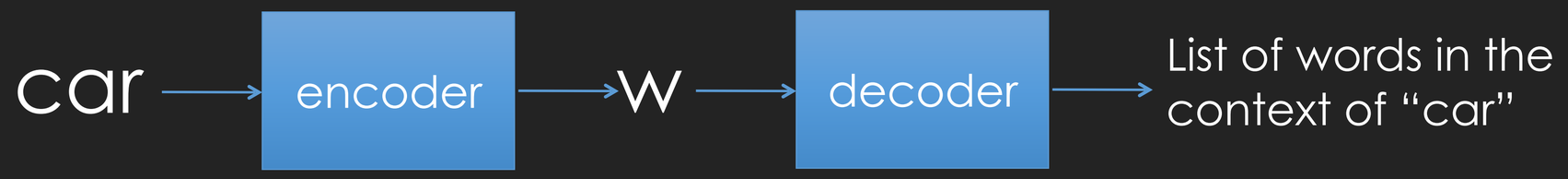
I parked the car in a nearby street. It is a red car with two doors, …

I parked the vehicle in a nearby street…

T. Mikolov, K. Chen, G. Corrado, J. Dean. Efficient Estimation of Word Representations in Vector Space. arXiv:1301.3781, 2013

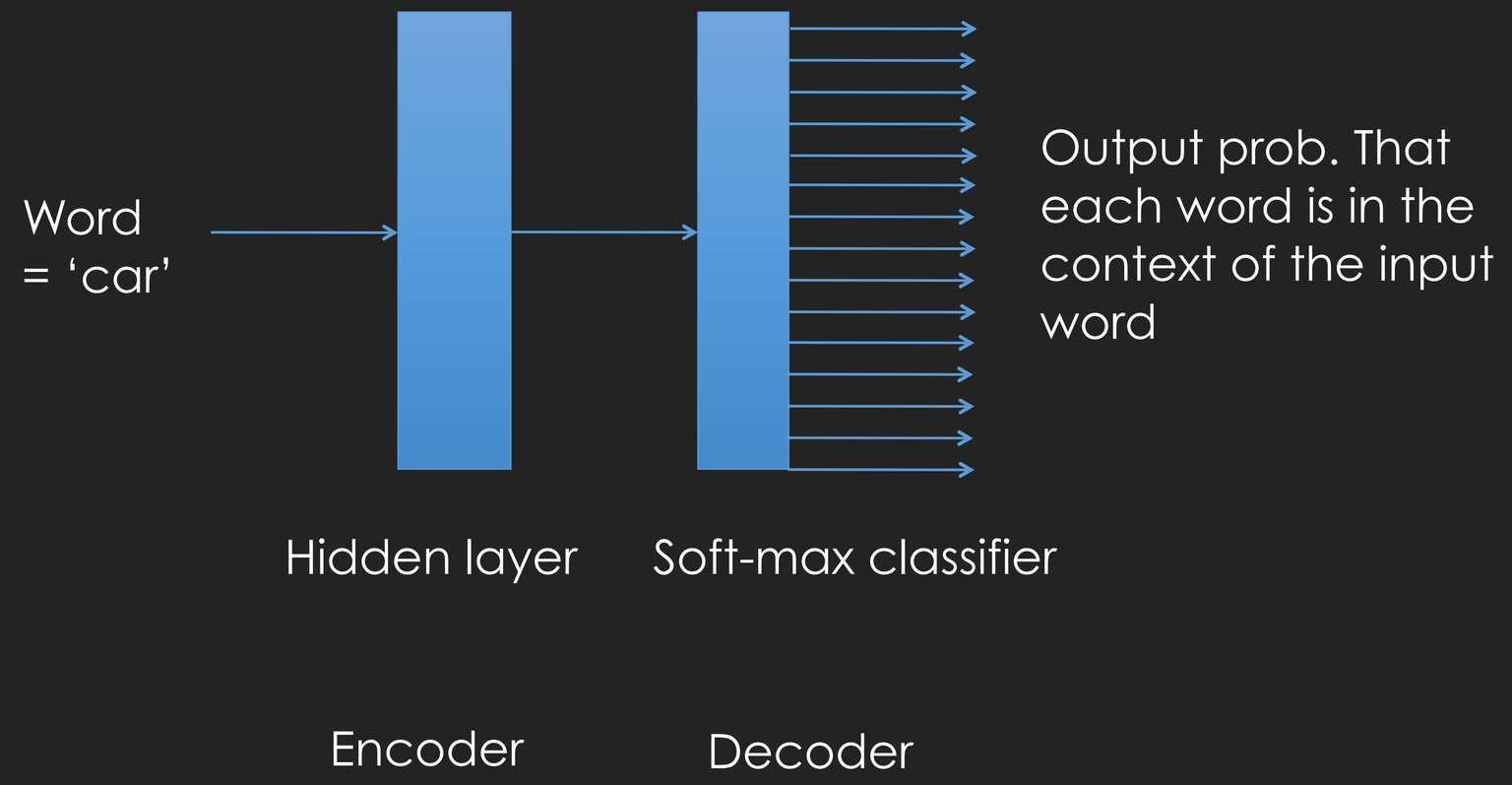
word2vec

I parked the car in a nearby street. It is a red car with two doors, …



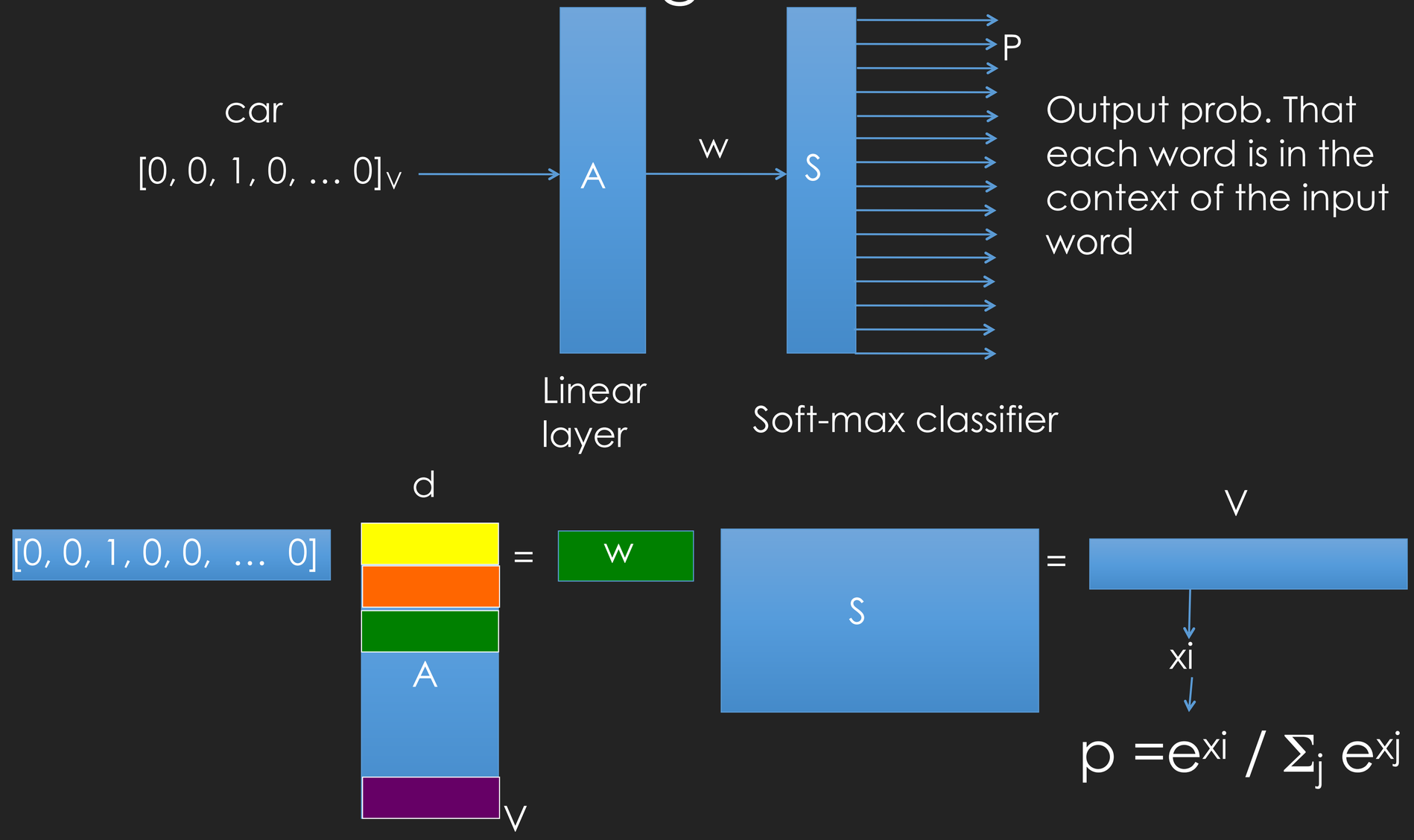
T. Mikolov, K. Chen, G. Corrado, J. Dean. Efficient Estimation of Word Representations in Vector Space. arXiv:1301.3781, 2013

word2vec



T. Mikolov, K. Chen, G. Corrado, J. Dean. Efficient Estimation of Word Representations in Vector Space. arXiv:1301.3781, 2013

word2vec, training



T. Mikolov, K. Chen, G. Corrado, J. Dean. Efficient Estimation of Word Representations in Vector Space. arXiv:1301.3781, 2013

word2vec, training

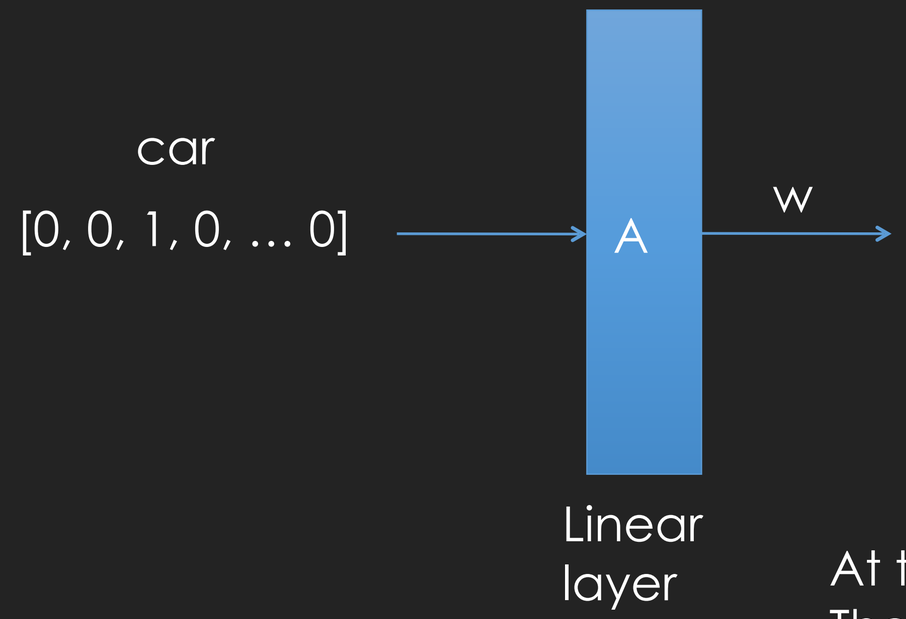
- In training maximize log-likelihood over the training set:

$$\sum_{t=1}^{T} \sum_{i=-c}^{c} \log p(w_{t+i} | w_t)$$

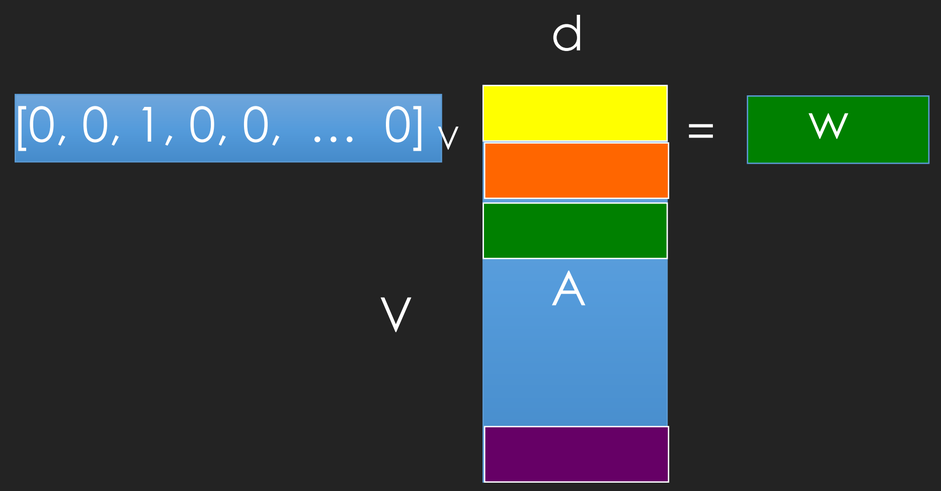
T … training set size
c … context window size

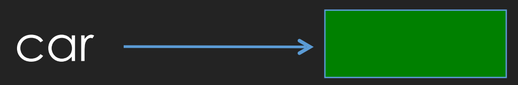
T. Mikolov, K. Chen, G. Corrado, J. Dean. Efficient Estimation of Word Representations in Vector Space. arXiv:1301.3781, 2013

word2vec, test time



At test time, w is our word embedding.
The encoding is just a look up table.



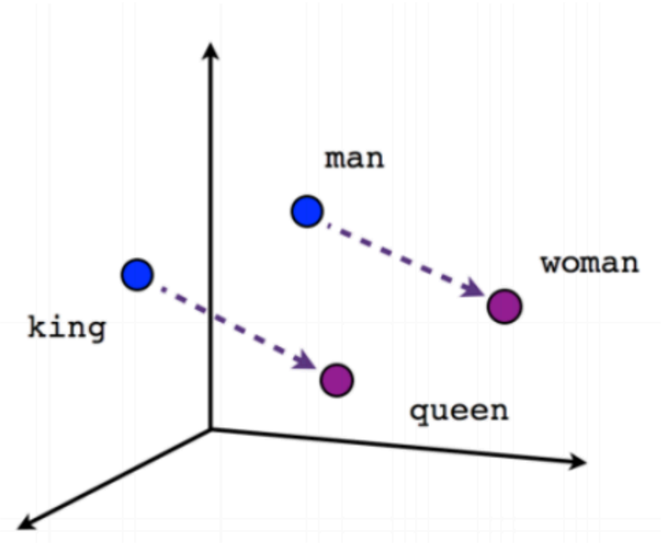


T. Mikolov, K. Chen, G. Corrado, J. Dean. Efficient Estimation of Word Representations in Vector Space. arXiv:1301.3781, 2013

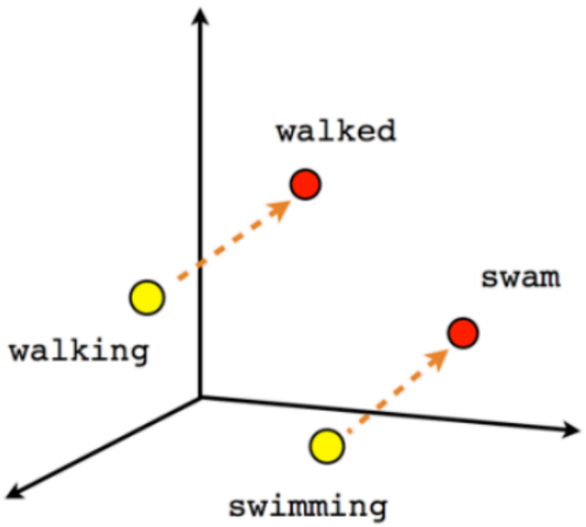
Algebraic operations with the vector representation of words

X = Vector(“Paris”) – vector(“France”) + vector(“Italy”)

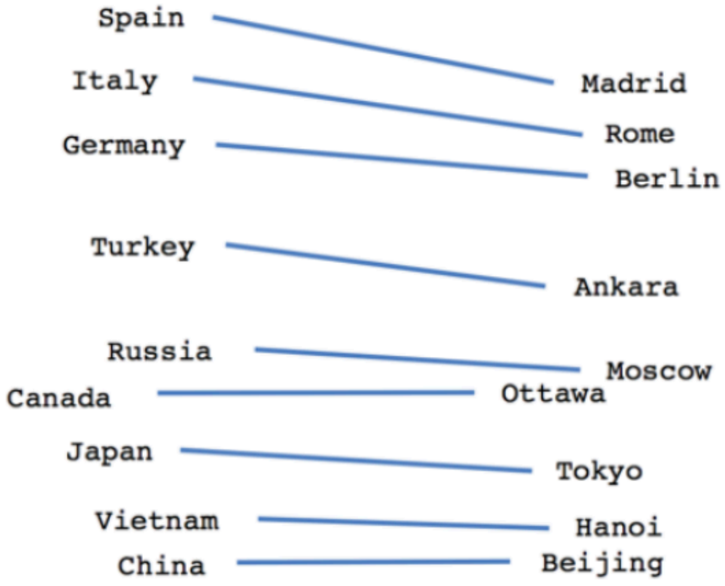
Closest nearest neighbor to X is vector(“Rome”)



Male-Female



Verb tense



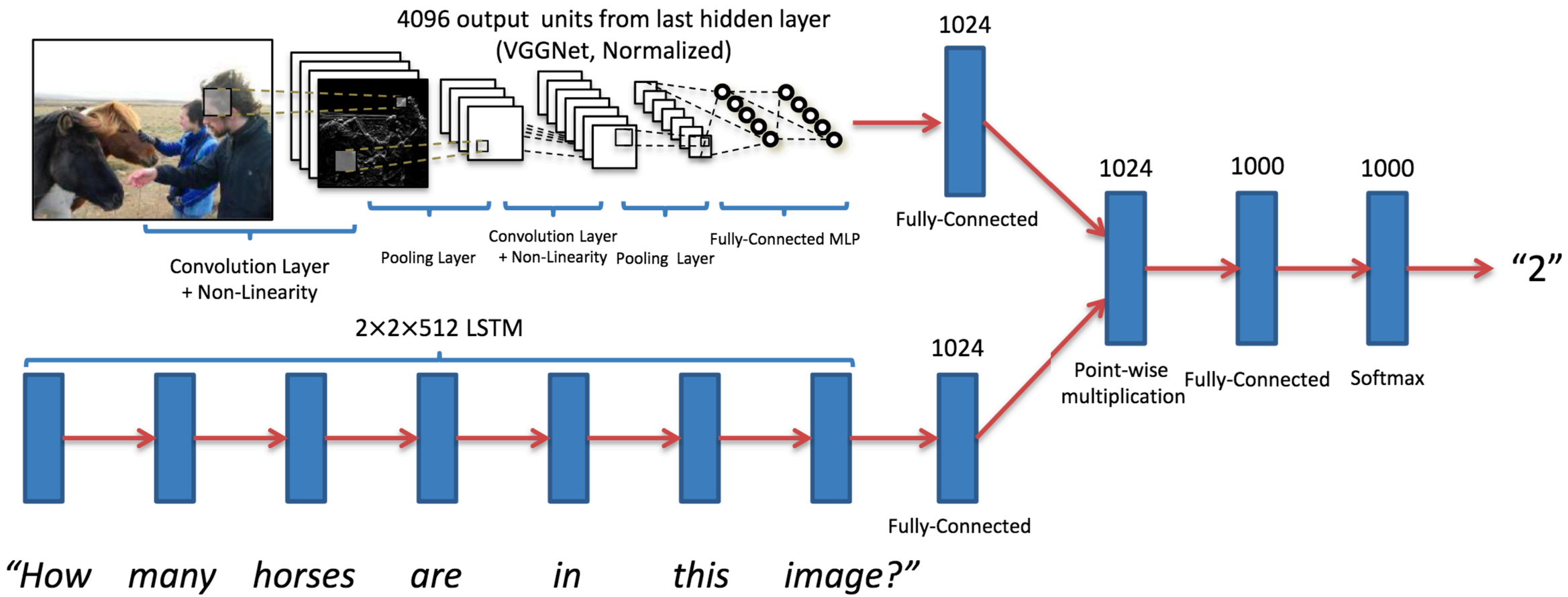
Country-Capital

Examples from https://www.tensorflow.org/tutorials/representation/word2vec

Architecture

often, we work with word embeddings, rather than one-hot representations of words

Architecture



There are 1000 possible answers in this system.
Questions are unlimited.

what is on the ground?

Submit

Predicted top-5 answers with confidence:

sand

90.748%

snow

2.858%

beach

1.418%

surfboards

0.677%

water

0.528%

what color is the umbrella?

Submit

Predicted top-5 answers with confidence:

yellow

95.090%

white

1.811%

black

0.663%

blue

0.541%

gray

0.362%

are we alone in the universe?

Submit

Predicted top-5 answers with confidence:

no

78.234%

yes

21.763%

people

0.001%

birds

0.000%

out

0.000%

what is the meaning of life?

Submit

Predicted top-5 answers with confidence:

beach

15.262%

sand

8.537%

seagull

4.708%

tower

2.393%

rocks

1.746%

what is the yellow thing?

Submit

Predicted top-5 answers with confidence:

frisbee

79.844%

surfboard

7.319%

banana

2.844%

lemon

2.438%

surfboards

1.252%

how many trains are in the picture?

Submit

Predicted top-5 answers with confidence:

3

30.233%

5

18.270%

4

17.000%

2

11.343%

6

7.806%

Neural Module Networks

Jacob Andreas

(with Dan Klein, Marcus Rohrbach and Trevor Darrell)

[Slides credit: Jacob Andreas]

Grounded question answering

What rivers are in South Carolina?

name	type	coastal
Columbia	**city**	**no**
Cooper	**river**	**yes**
Charleston	**city**	**yes**

Cooper

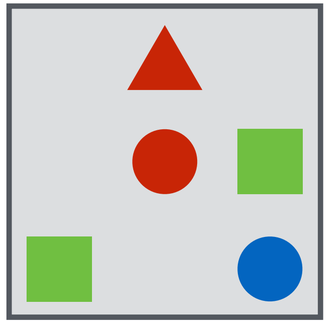
Grounded question answering

What color is the necktie?

yellow

Grounded question answering

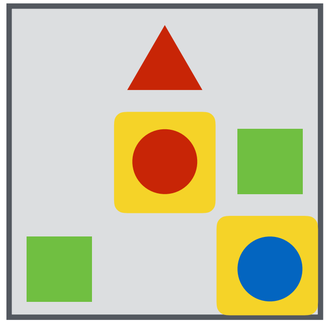
Is there a red shape above a circle?

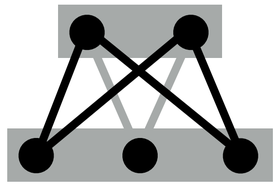


yes

Neural nets learn lexical groundings

Is there a red shape above a circle?



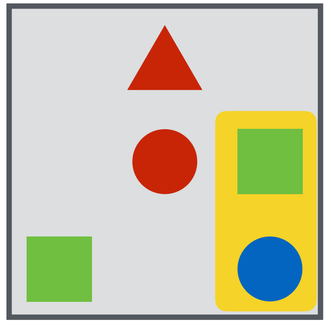


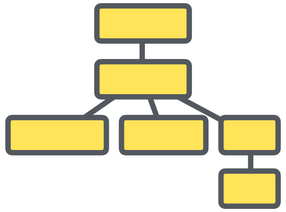
yes

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

Semantic parsers learn composition

Is there a red shape above a circle?



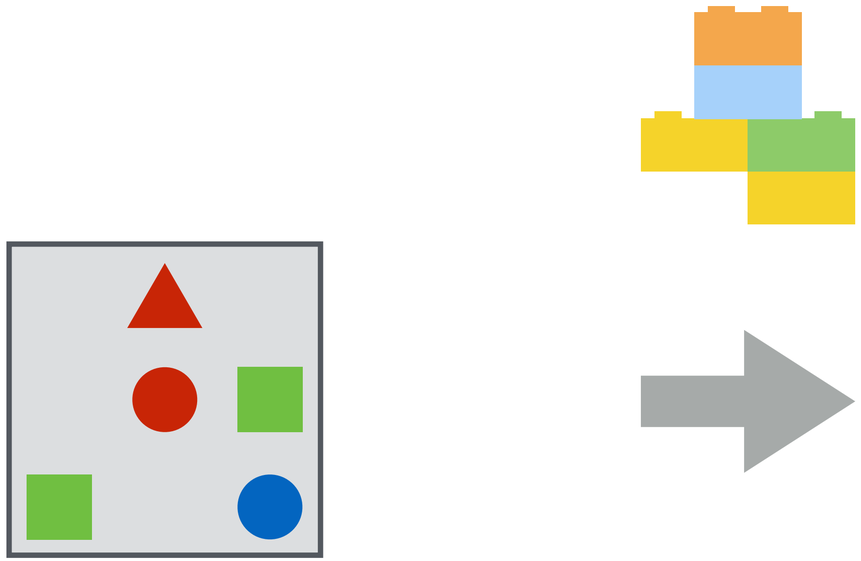


yes

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

Neural module networks learn both!

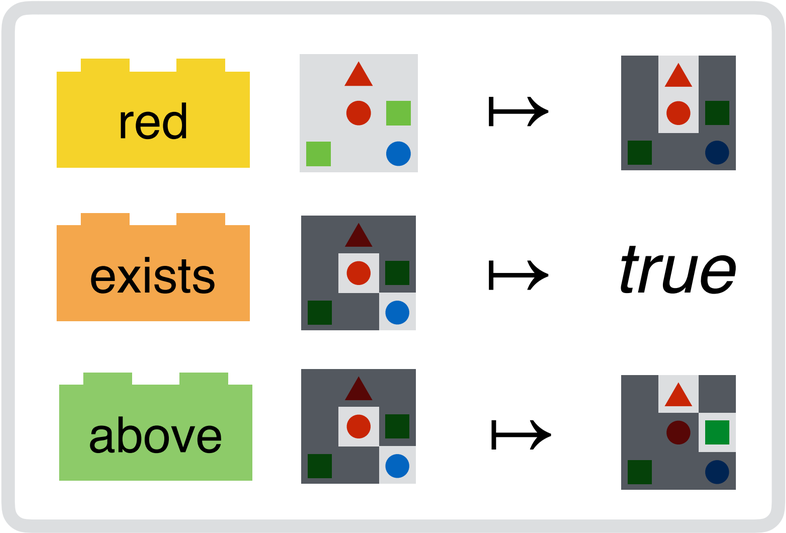
Is there a red shape above a circle?



yes

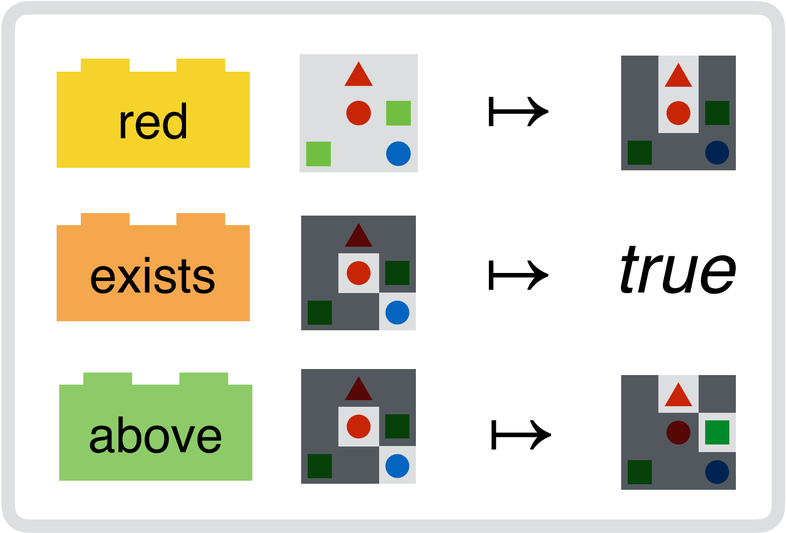
Neural module networks

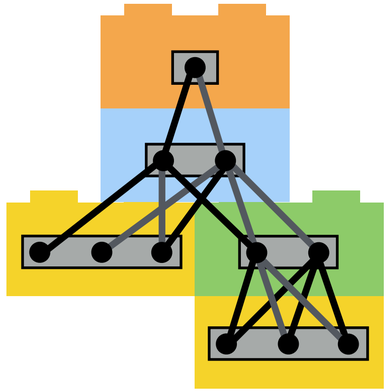
Is there a red shape above a circle?



Neural module networks

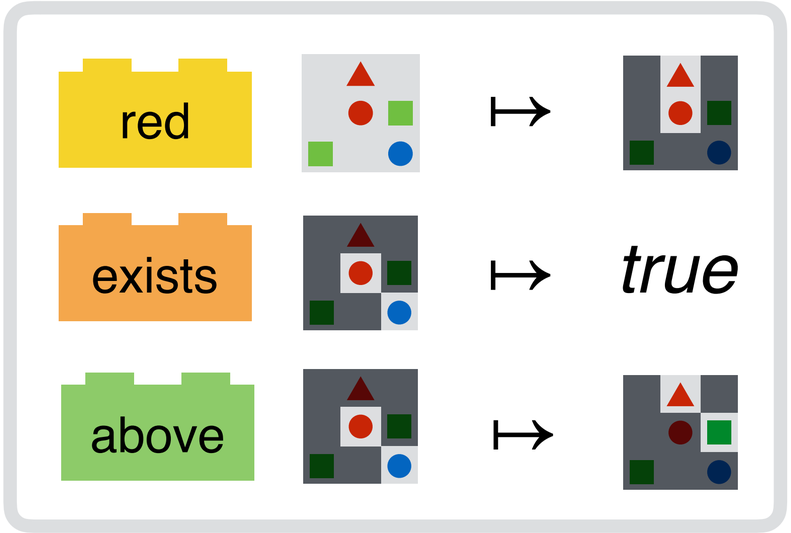
Is there a red shape above a circle?

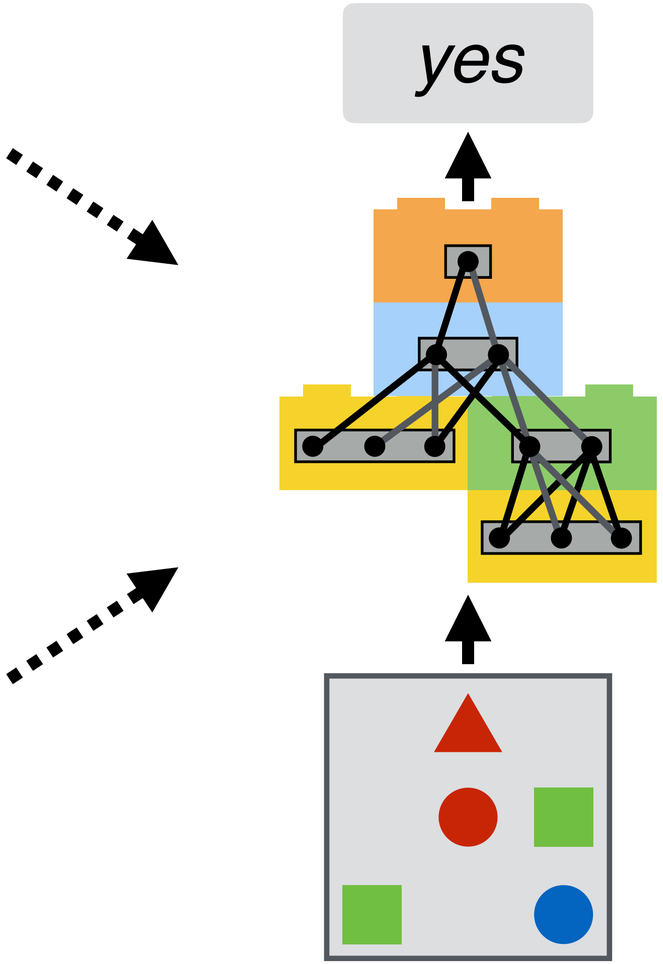




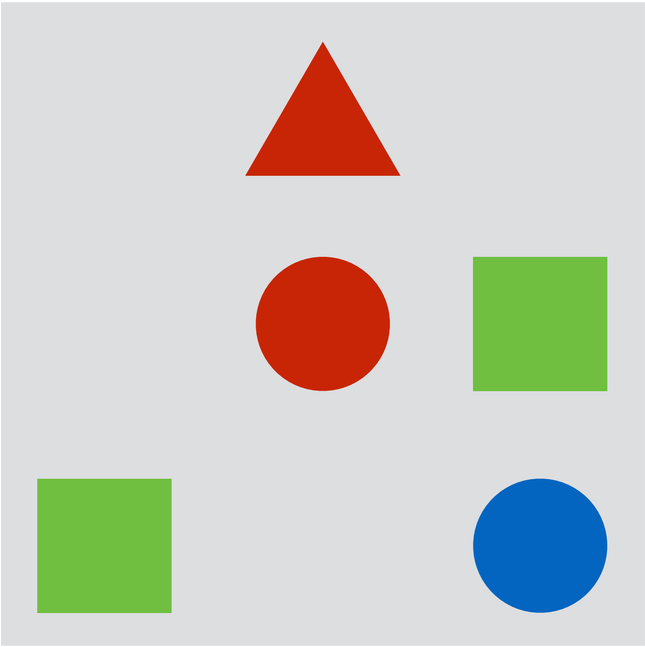
Neural module networks

Is there a red shape above a circle?



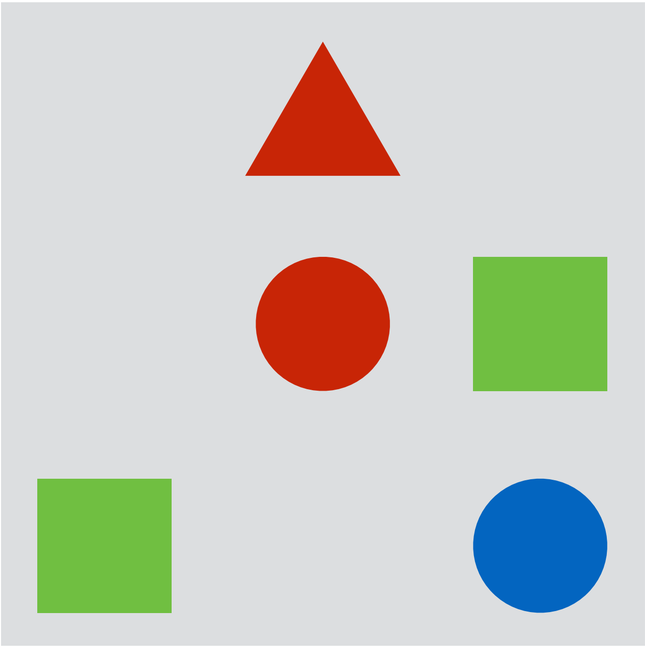


Representing meaning



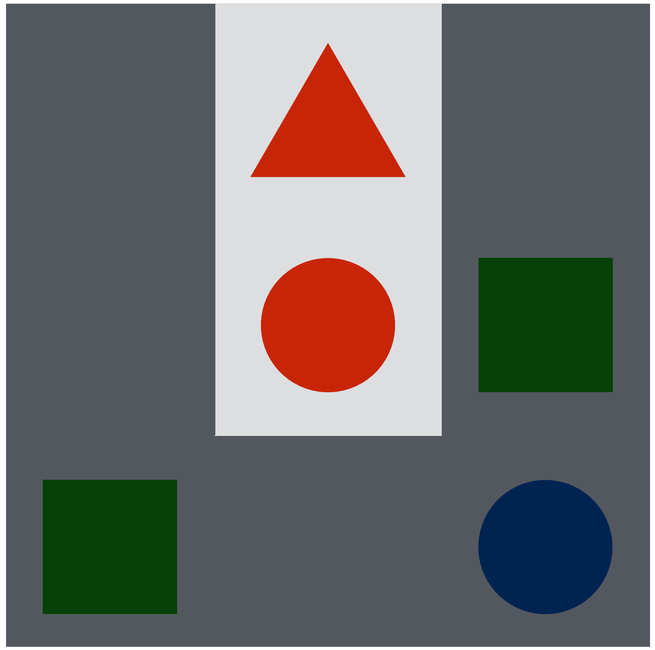
Is there a red shape above a circle?

Representing meaning



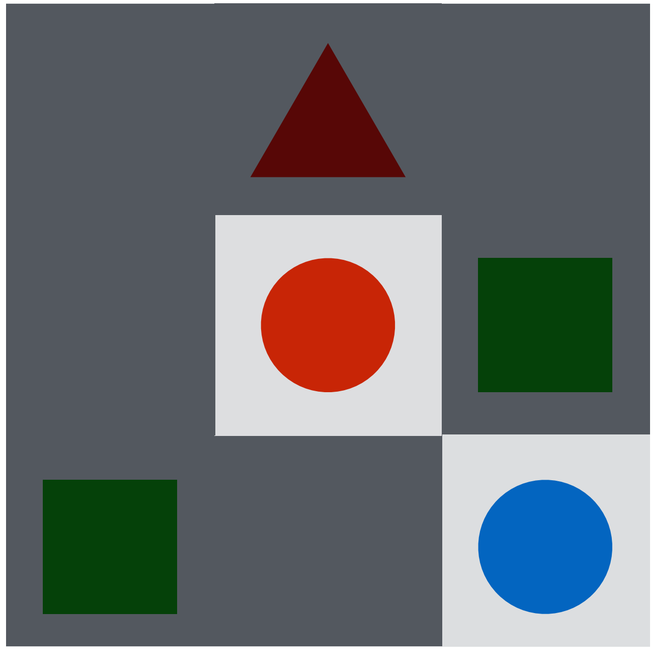
Is there a red shape above a circle?

Sets encode meaning



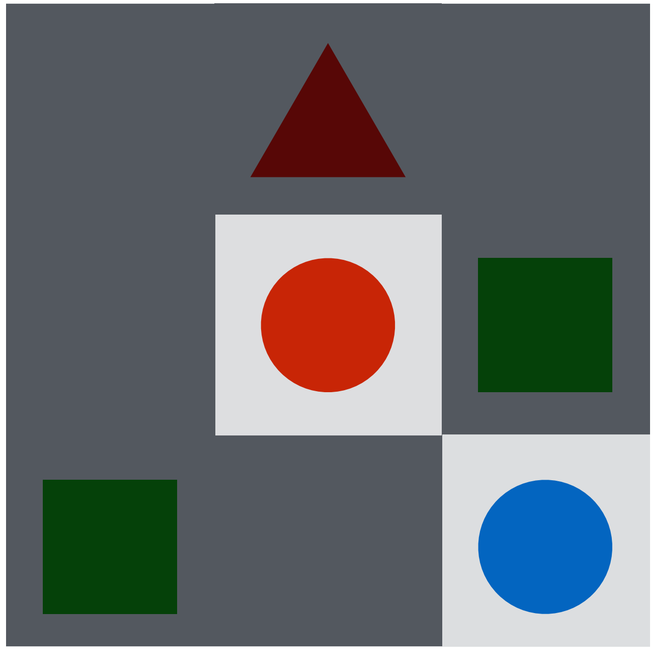
Is there a red shape above a circle?

Sets encode meaning



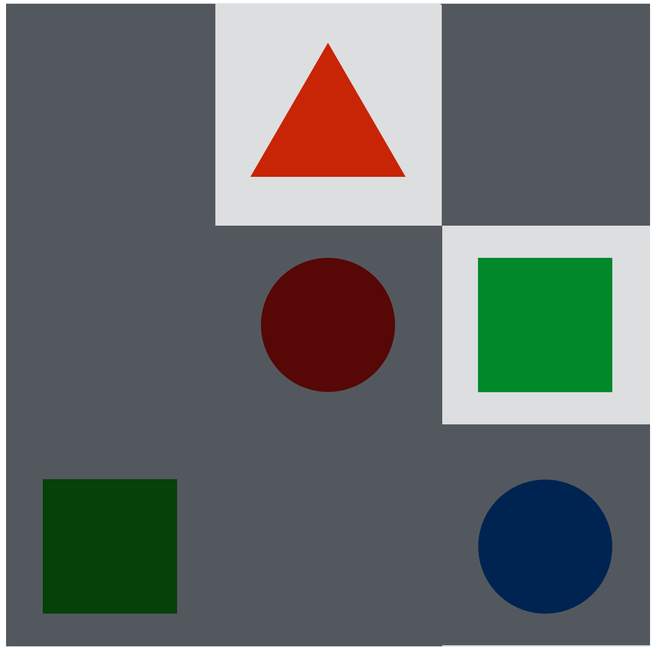
Is there a red shape above a circle?

Set transformations encode meaning



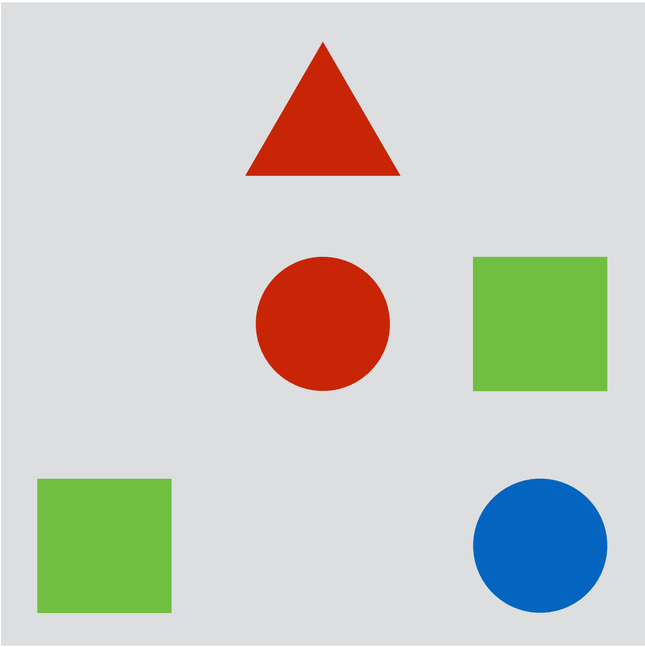
Is there a red shape above a circle?

Set transformations encode meaning

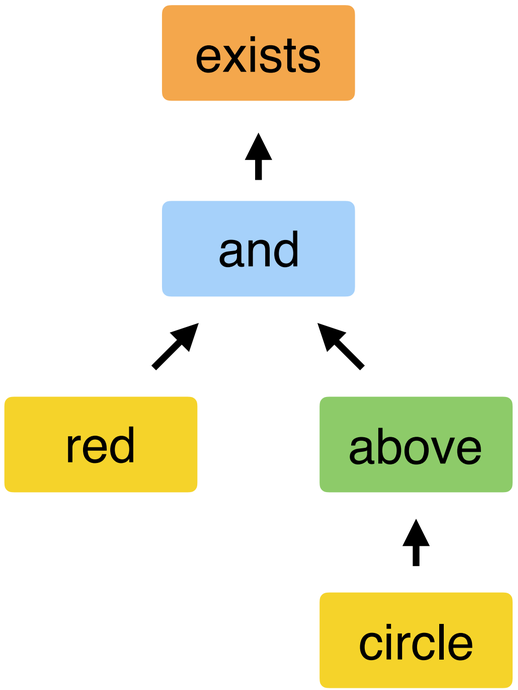


Is there a red shape above a circle?

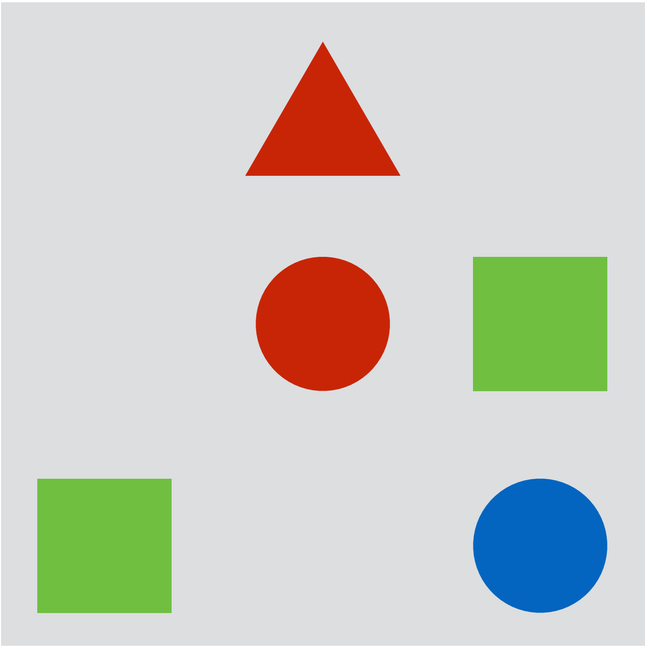
Sentence meanings are computations



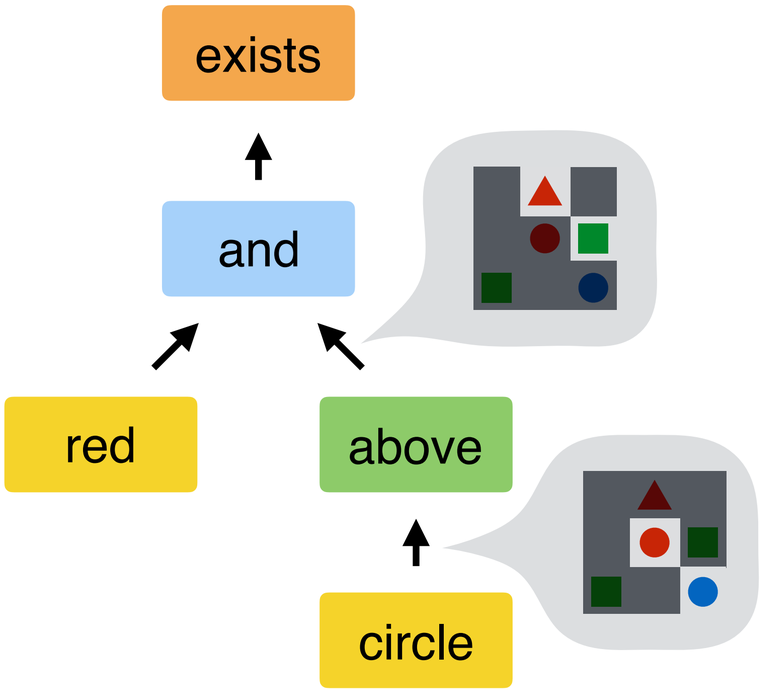
Is there a red shape above a circle?



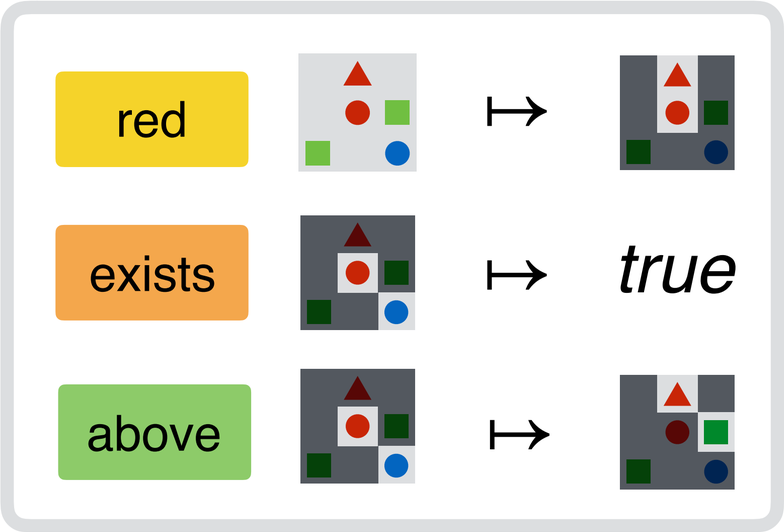
Sentence meanings are computations

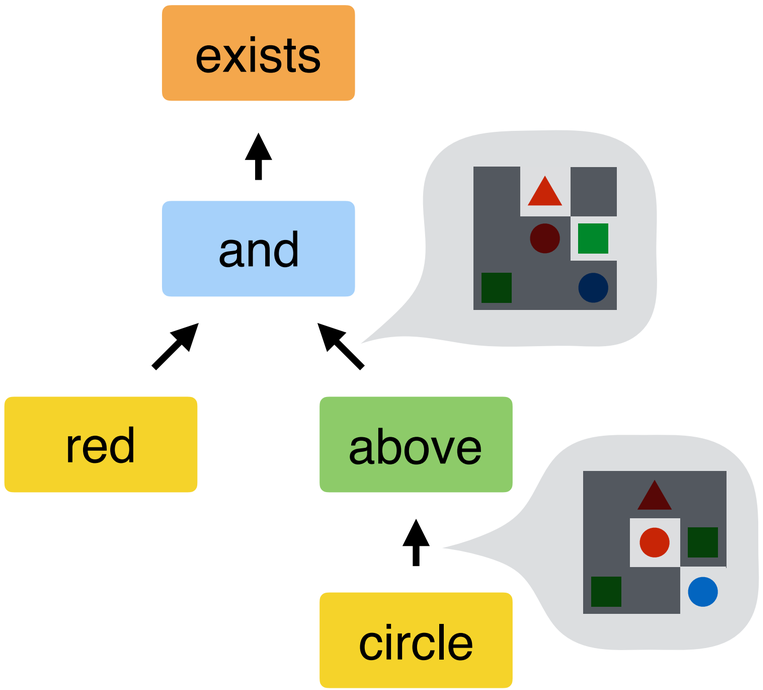


Is there a red shape above a circle?

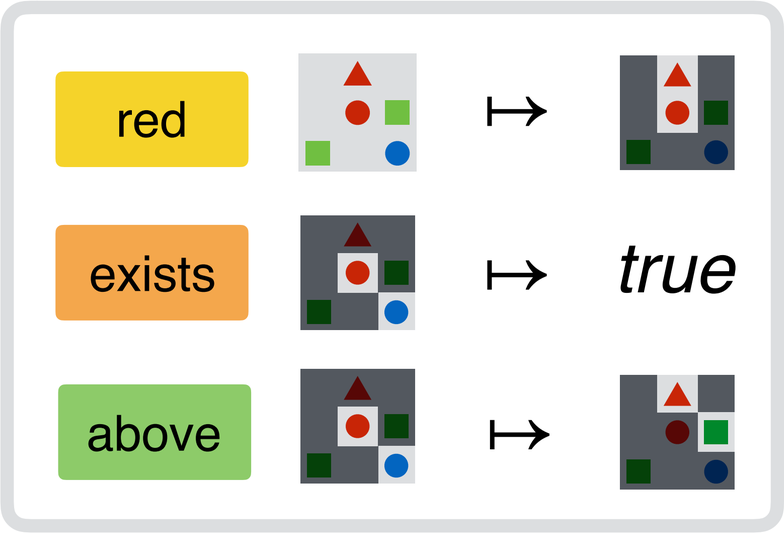


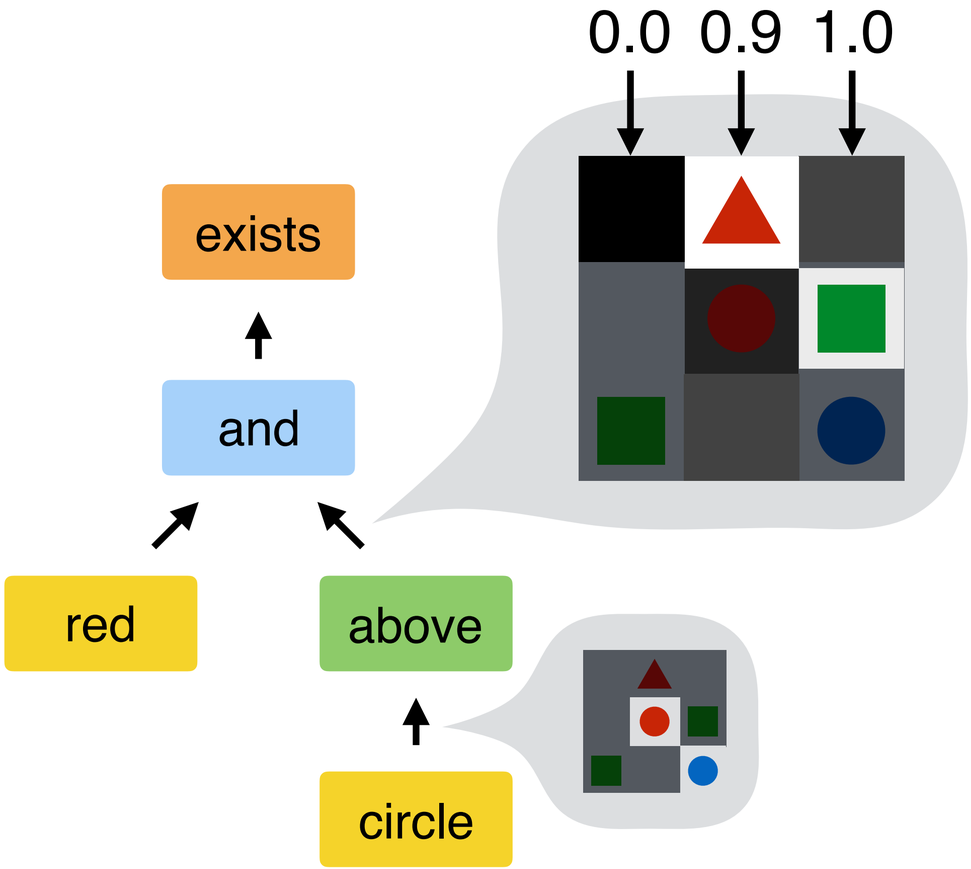
Computations are built from set functions



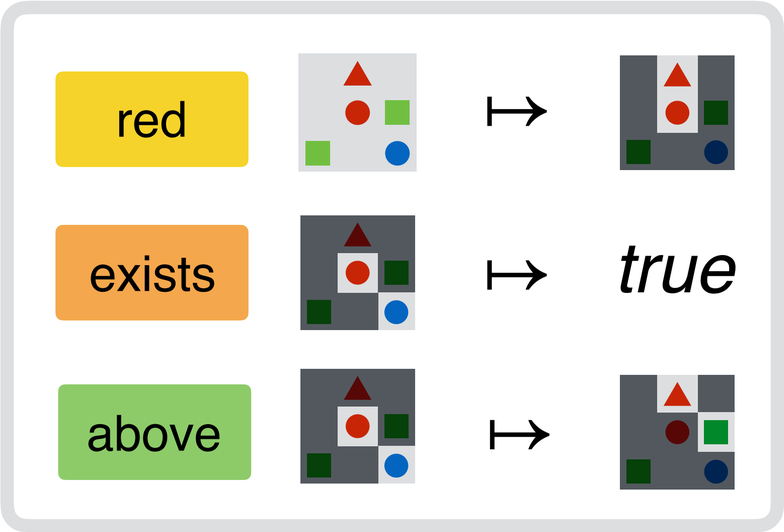


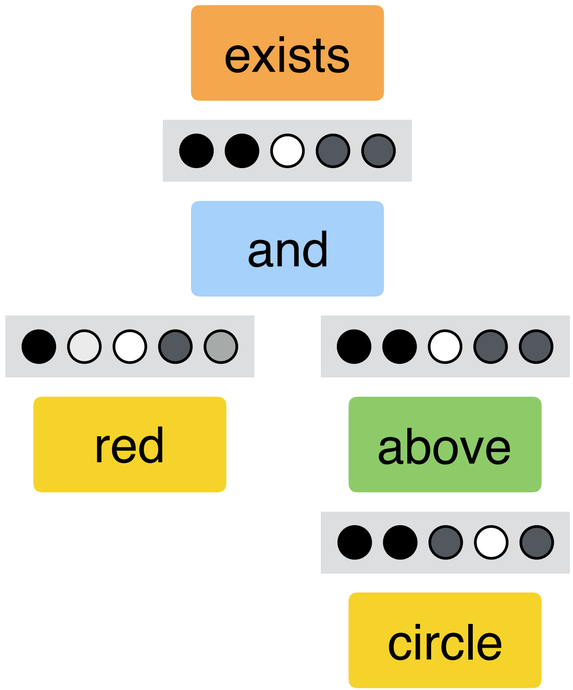
…or relaxed to vector functions



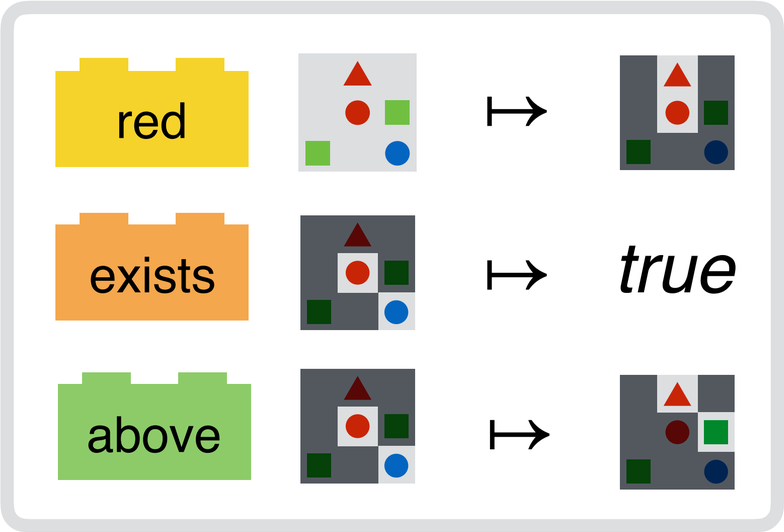


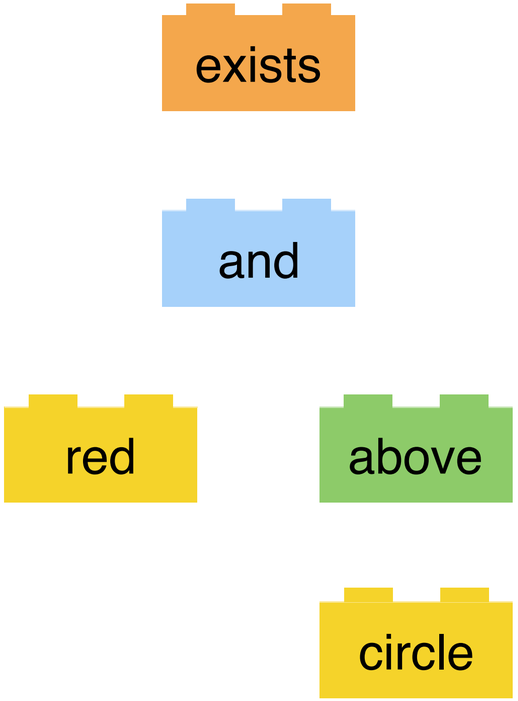
Composing vector functions



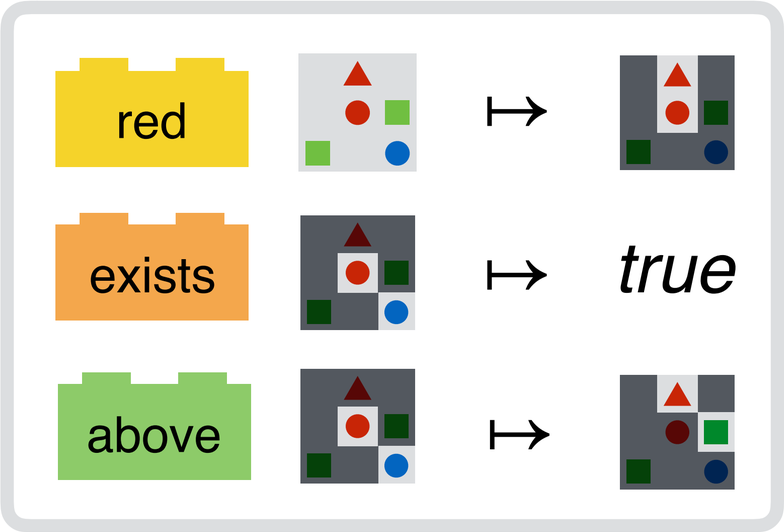


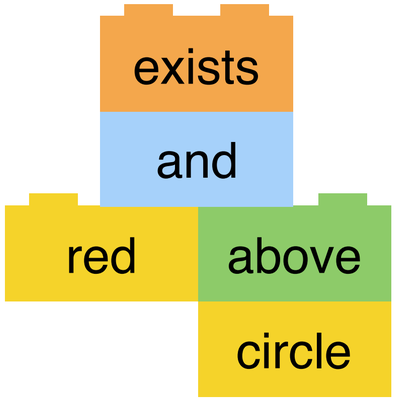
Composing vector functions



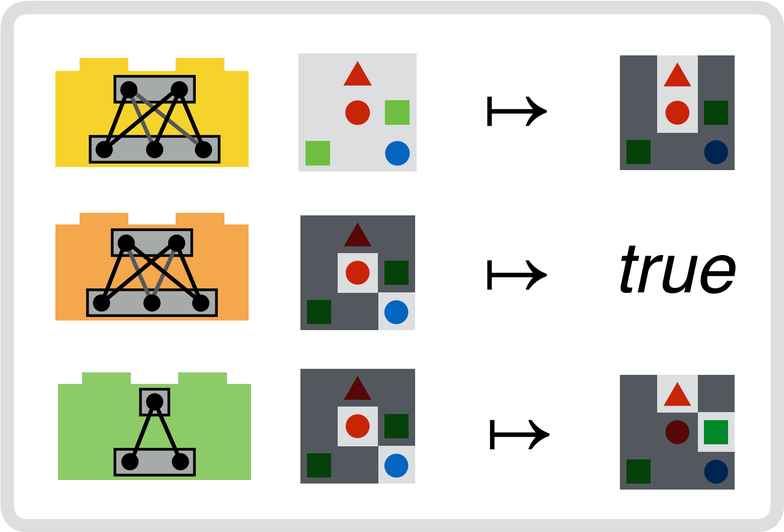


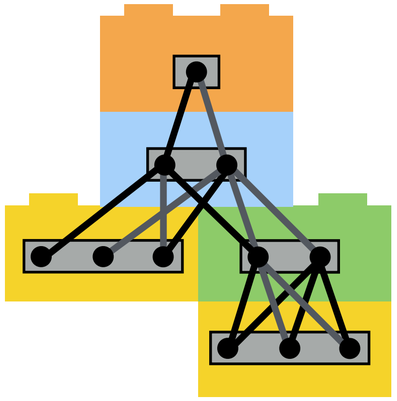
Composing vector functions



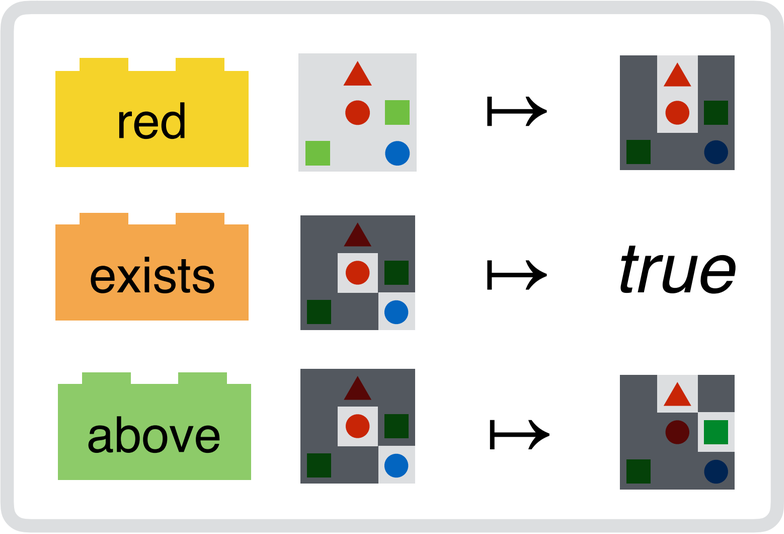


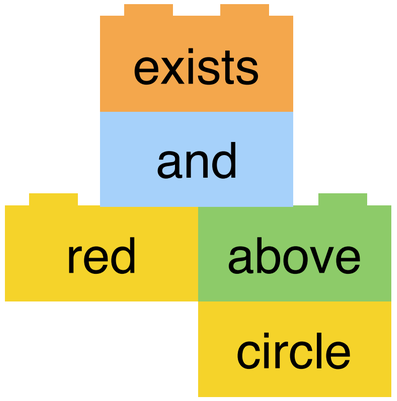
Compositions of vector functions are neural nets





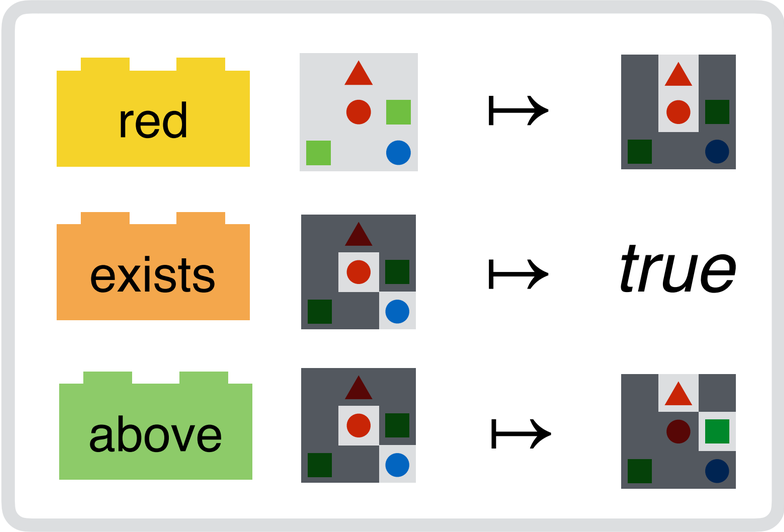
Compositions of vector functions are neural nets

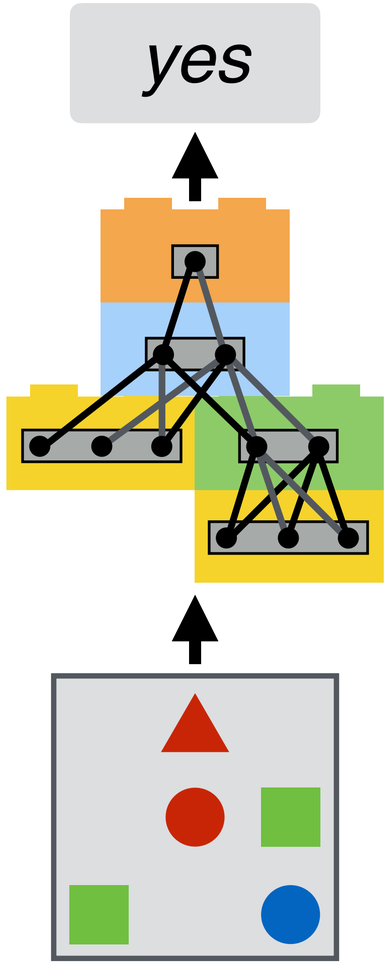




Outline

Is there a red shape above a circle?

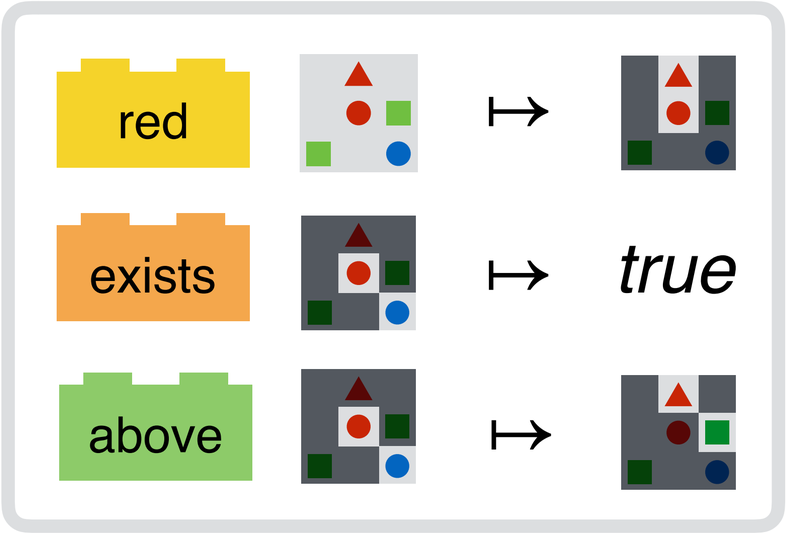




Outline

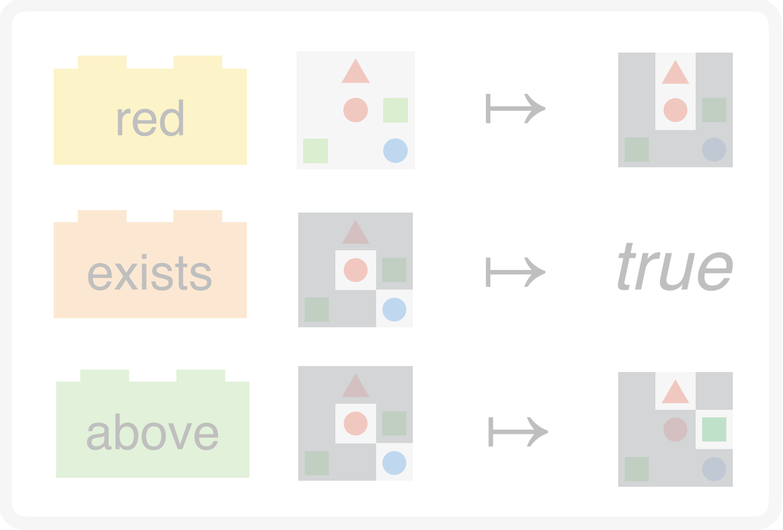
Is there a red shape above a circle?

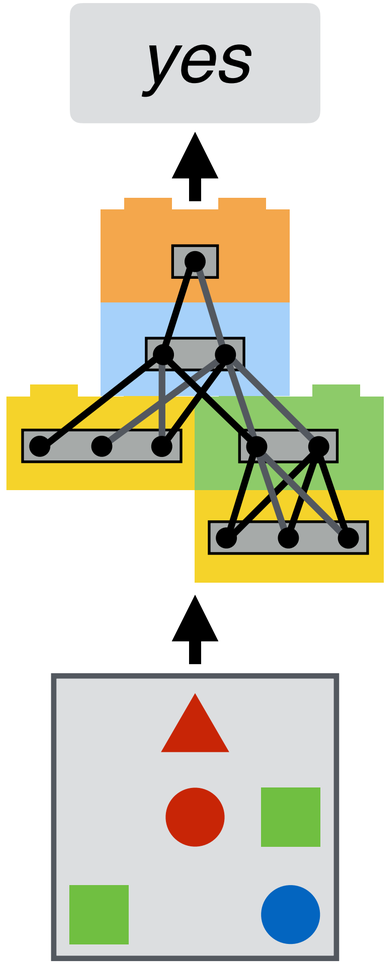
yes



Outline

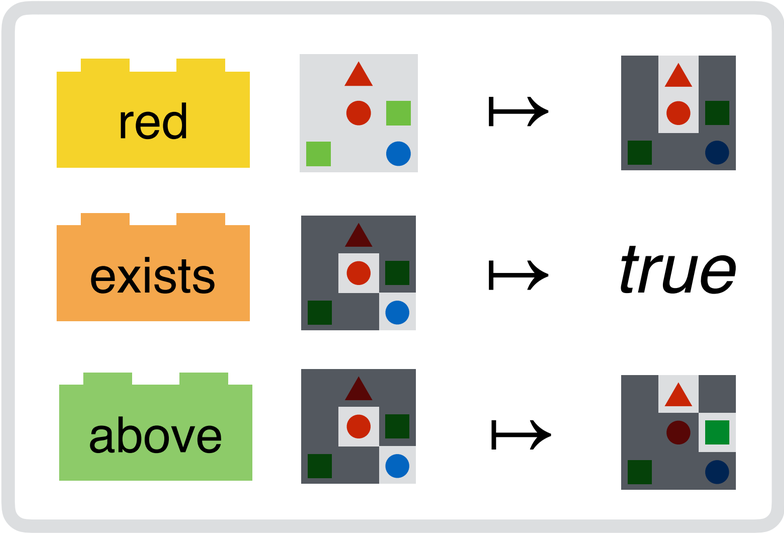
Is there a red shape above a circle?



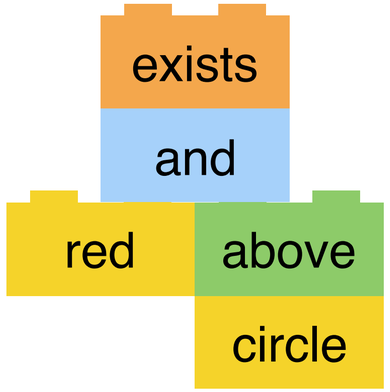


Outline

Is there a red shape above a circle?



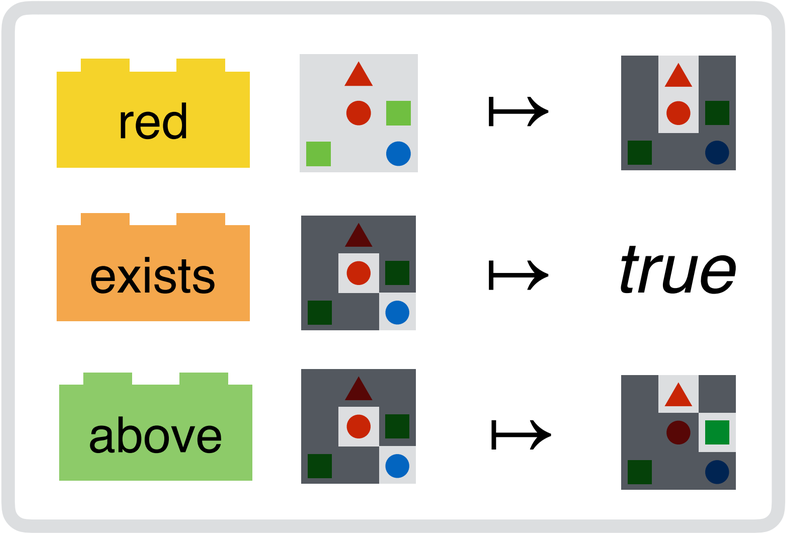
yes



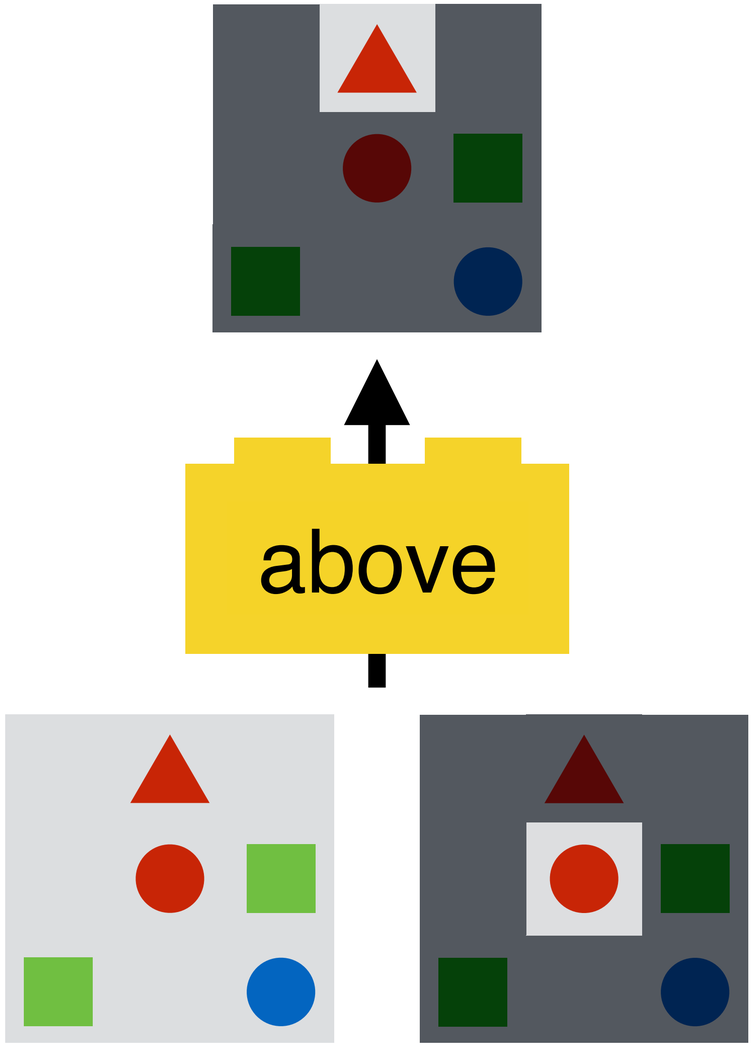
Outline

Is there a red shape above a circle?

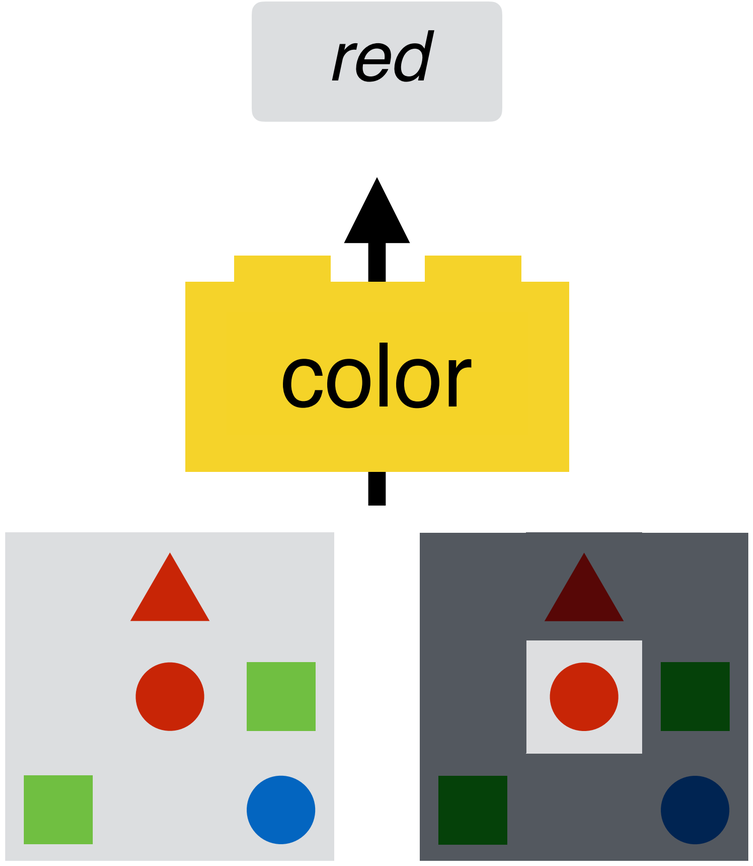
yes



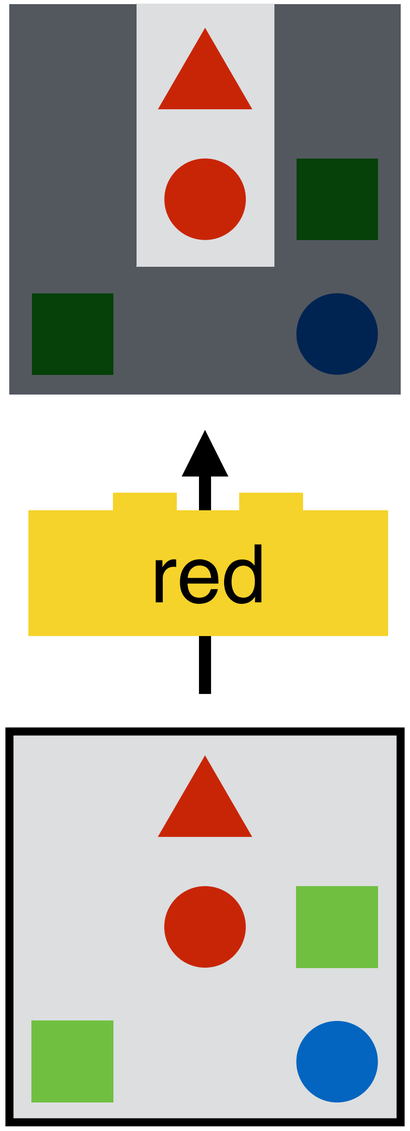
Anatomy of a module



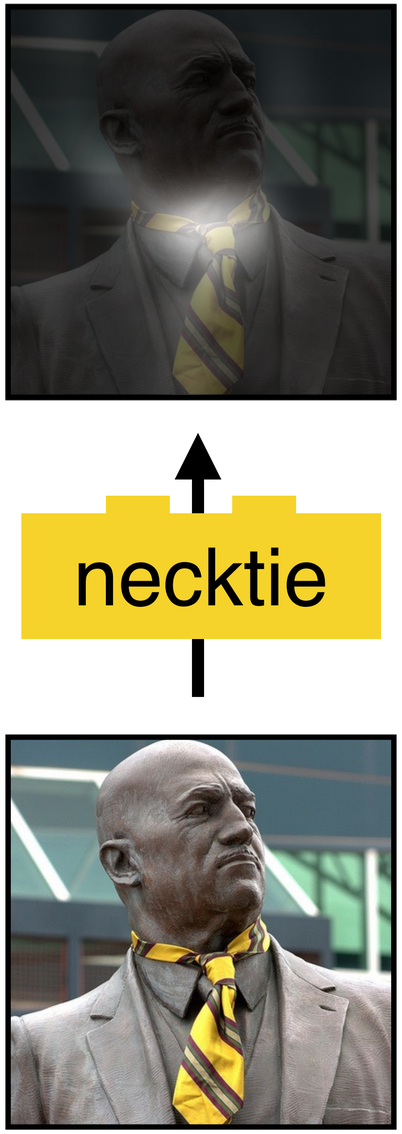
Anatomy of a module



The [find] module

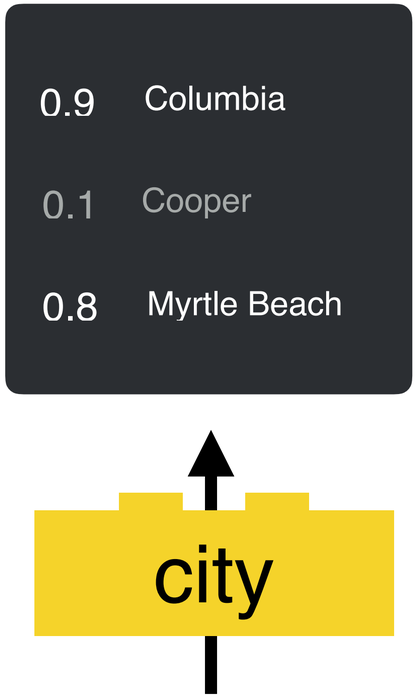


The [find] module



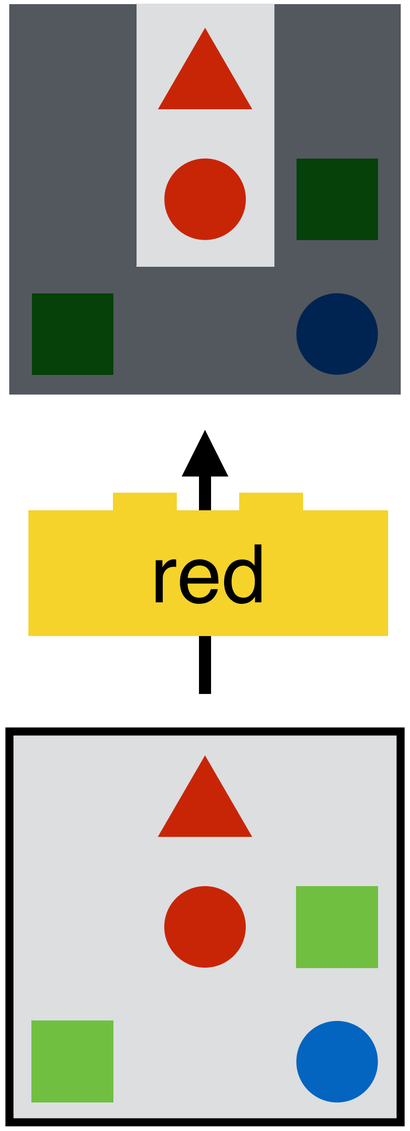
[Xu et al. 2015]

The [find] module

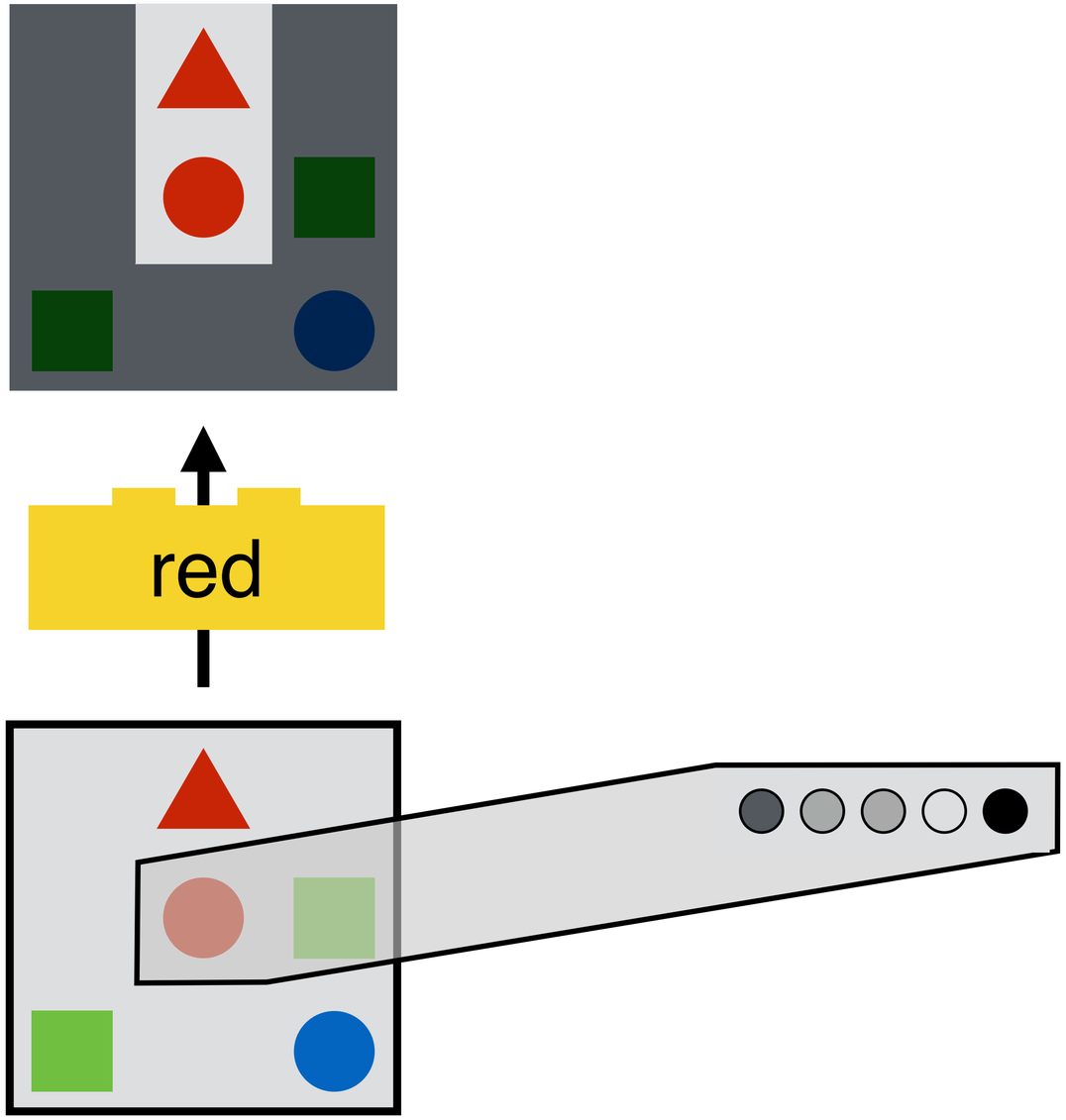


name	type	coastal
Columbia	city	no
Cooper	river	yes
Myrtle Beach	city	yes

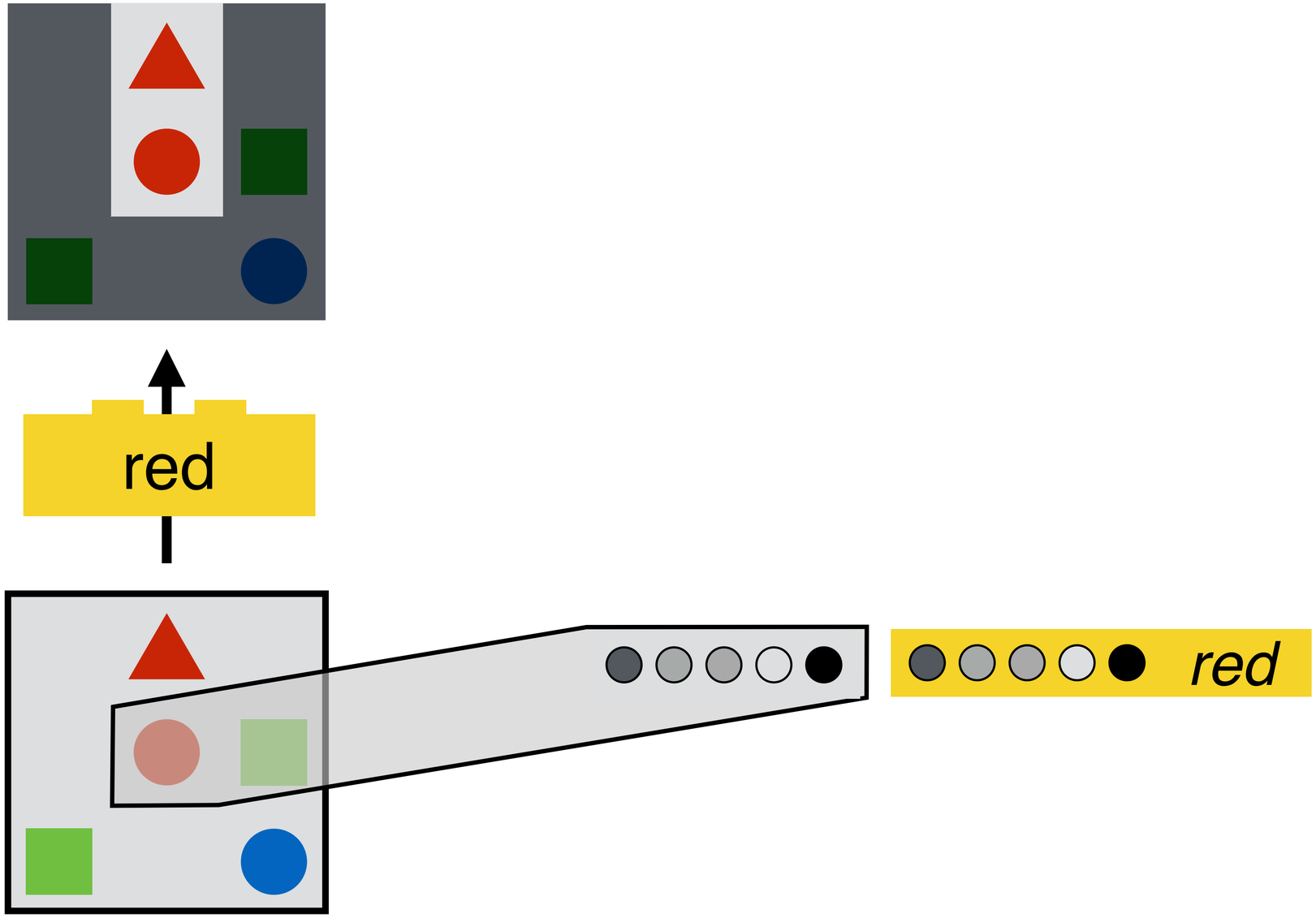
The [find] module



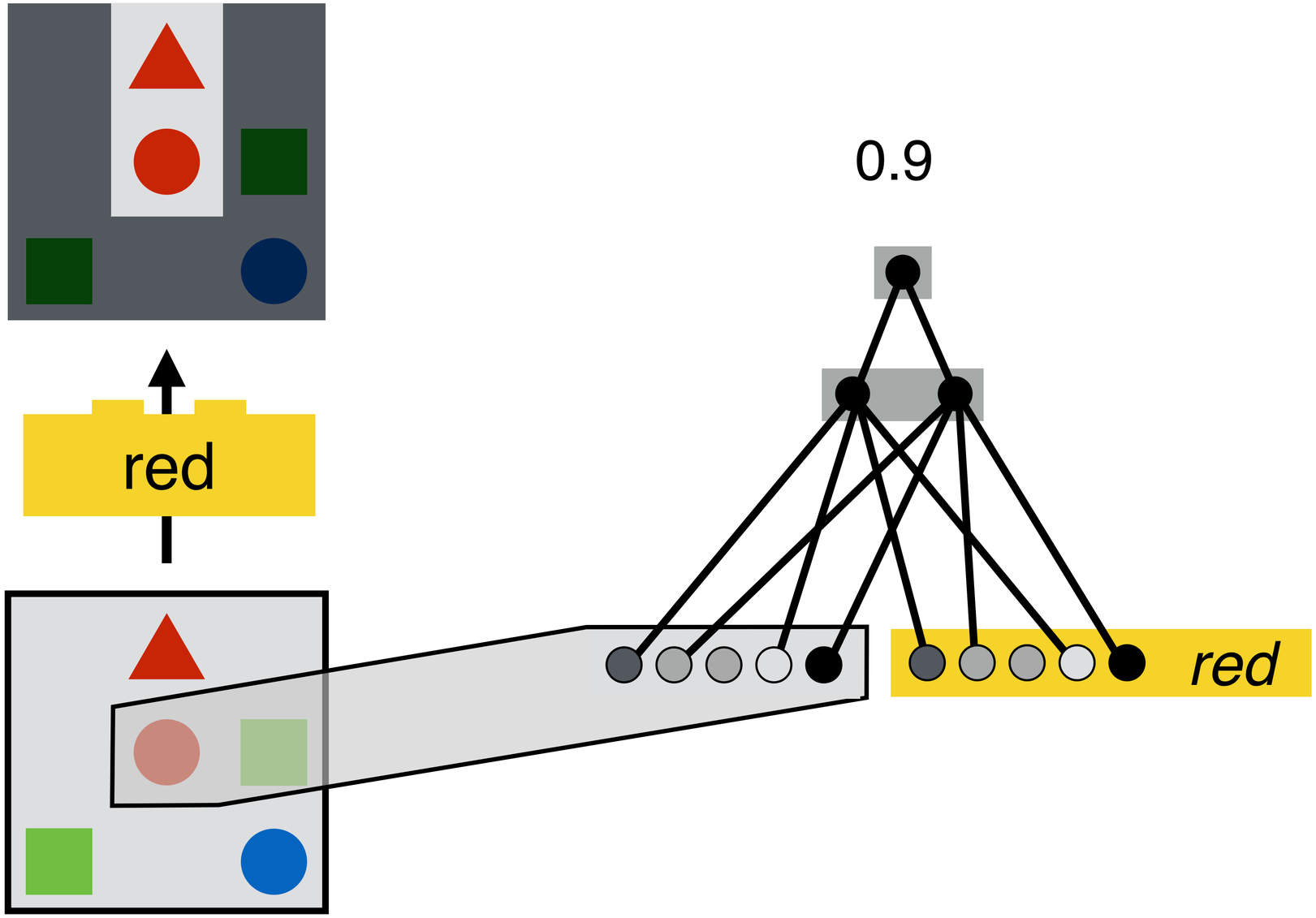
The [find] module



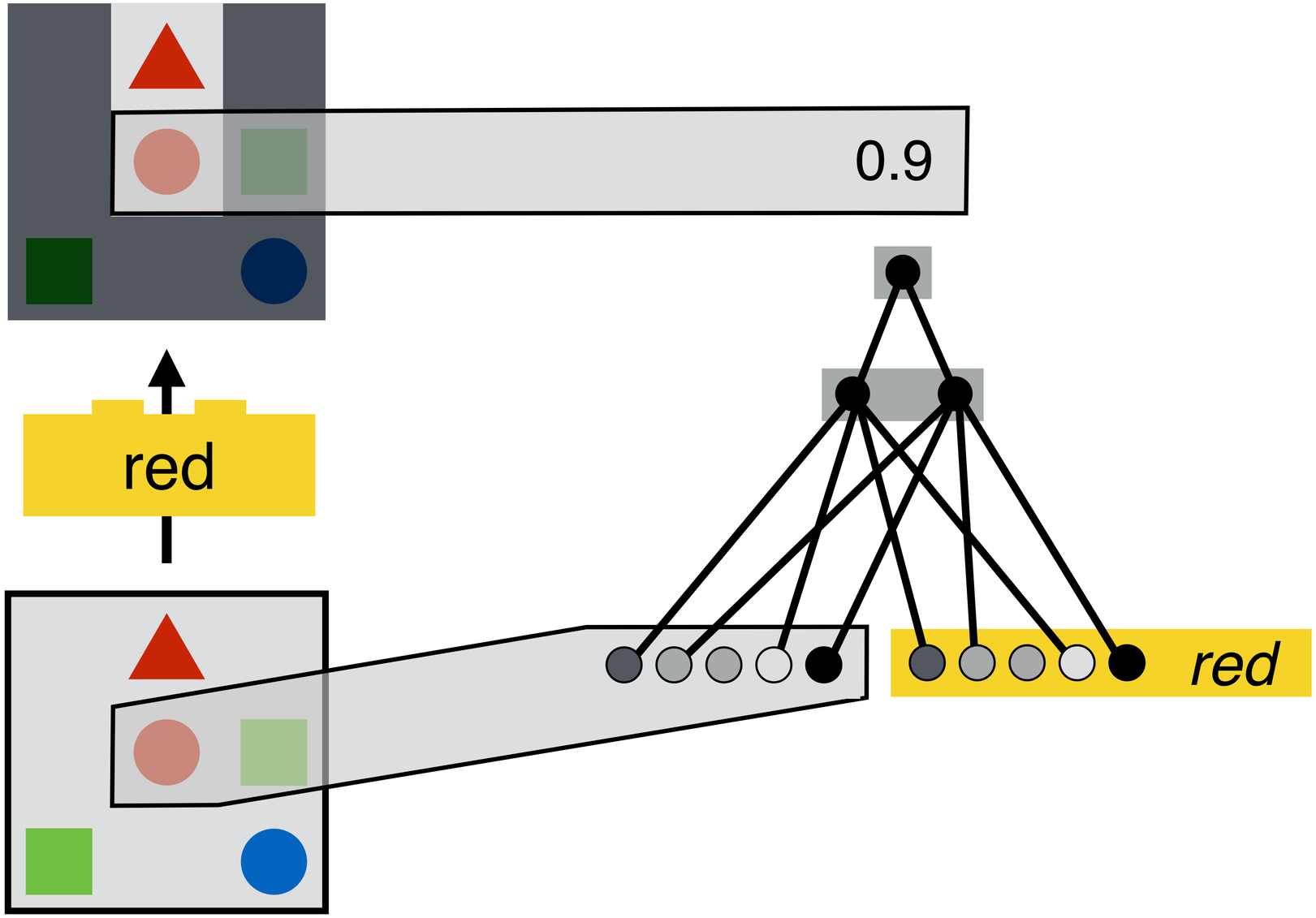
The [find] module



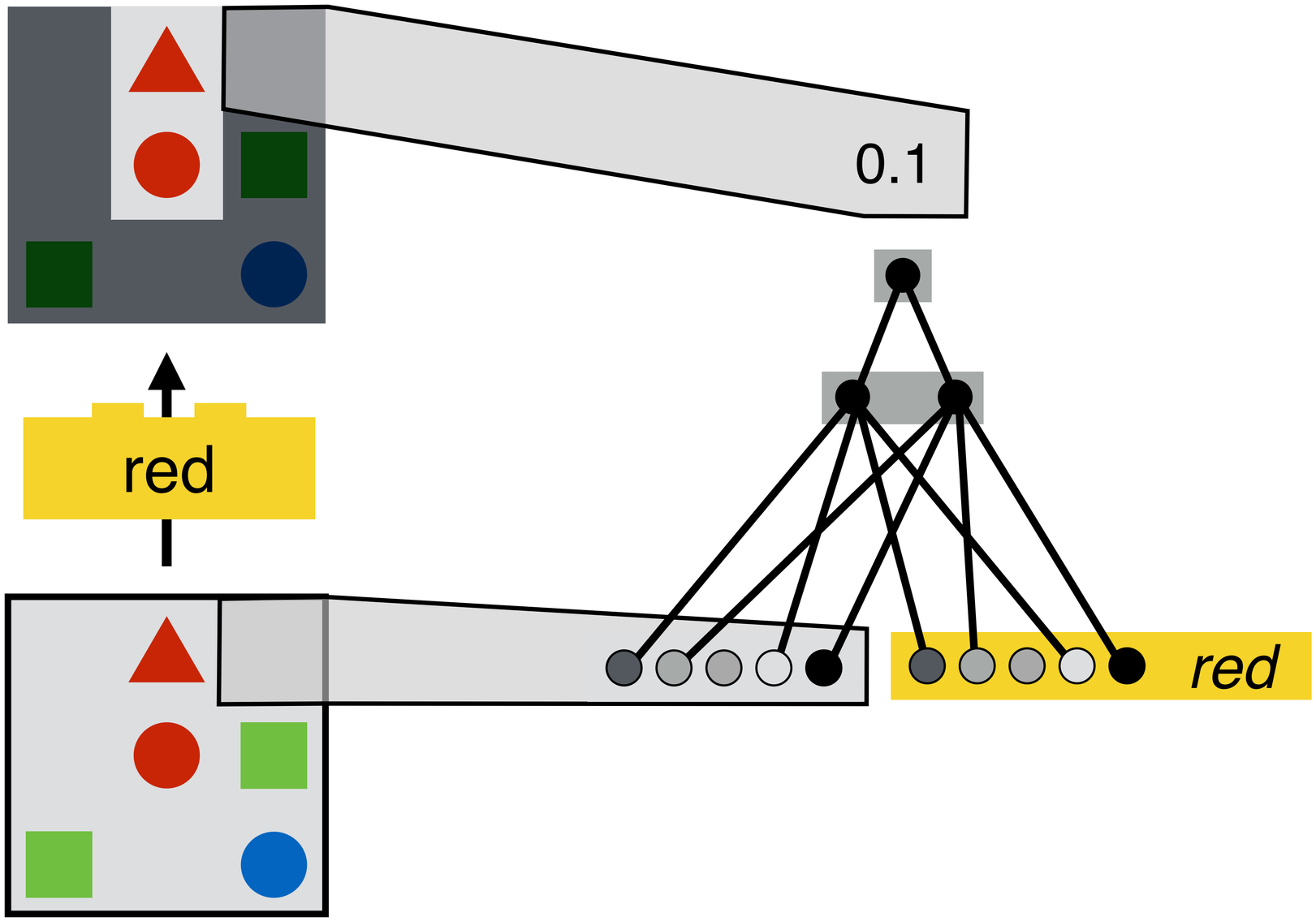
The [find] module



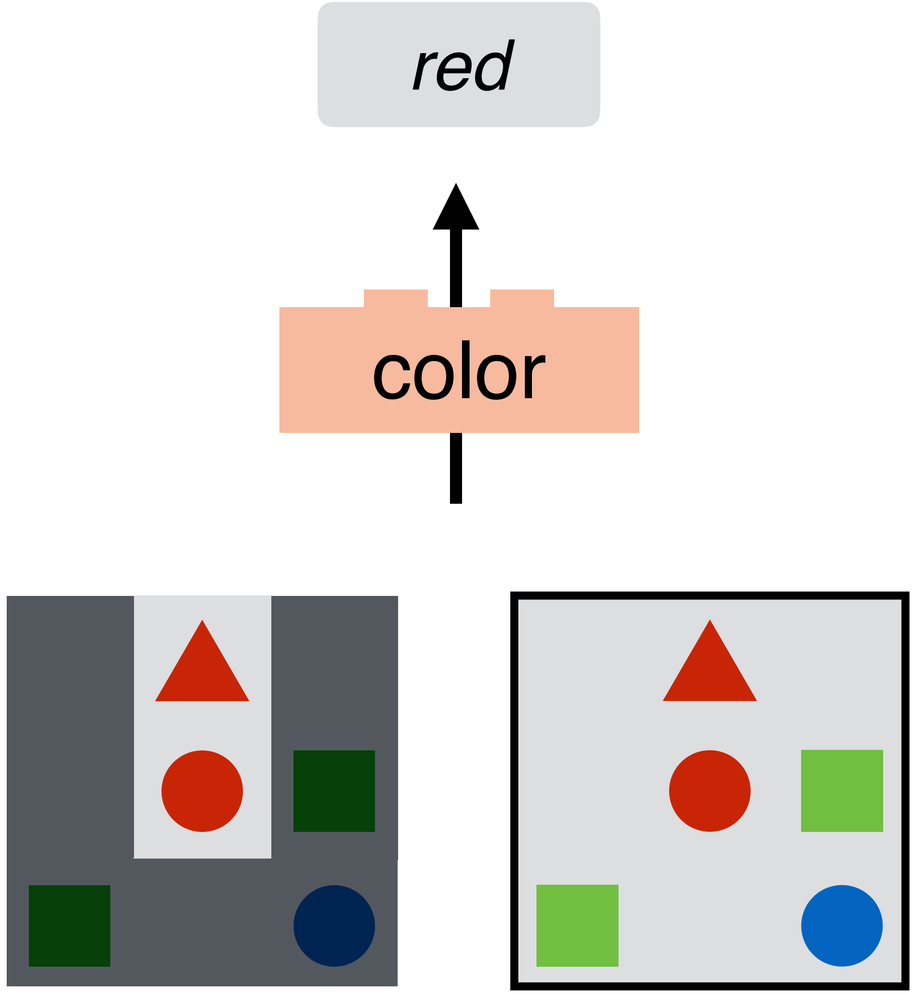
The [find] module



The [find] module

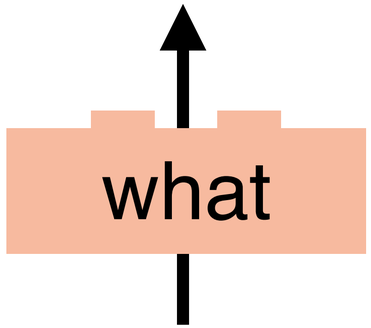


The [describe] module

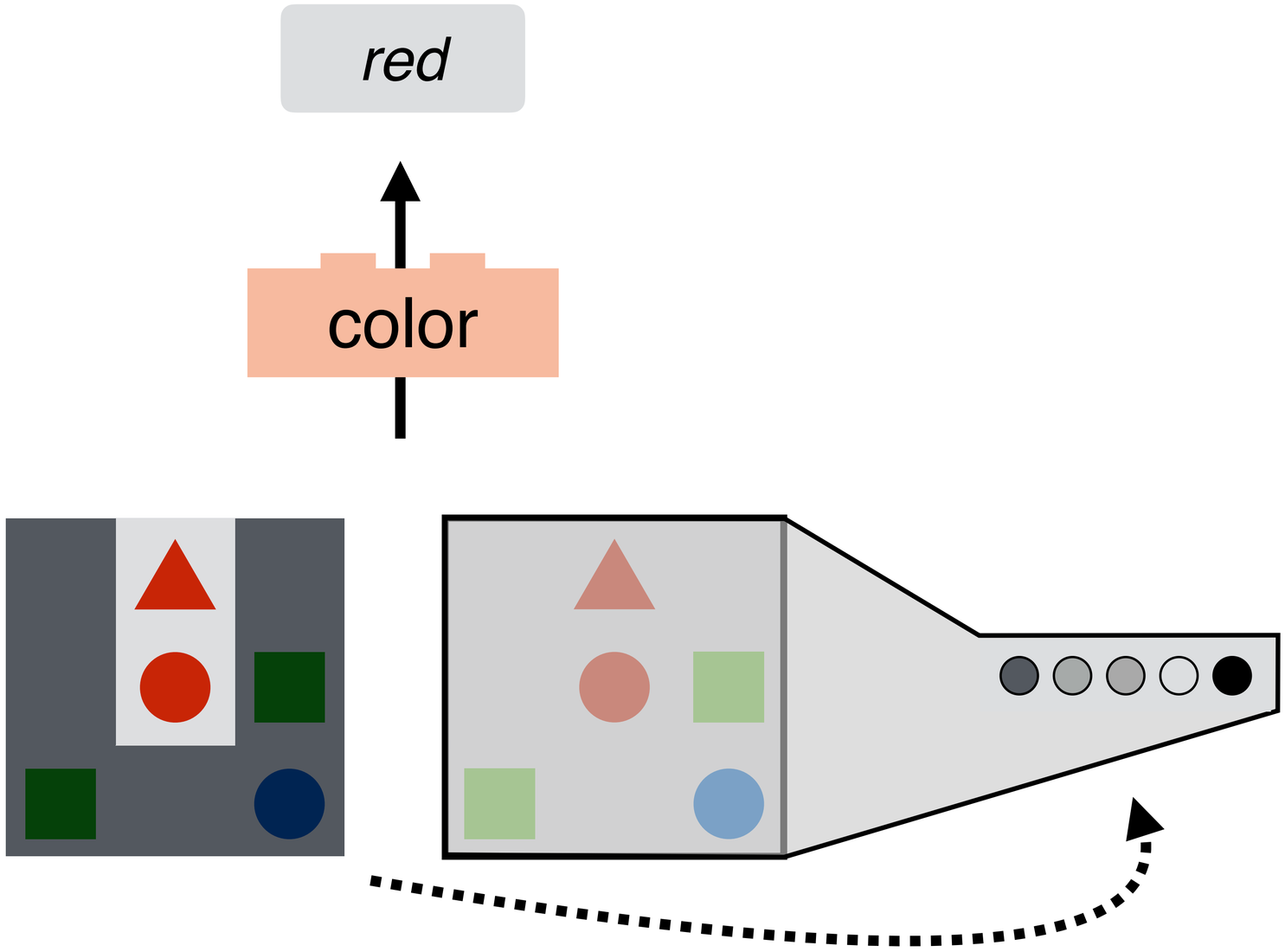


The [describe] module

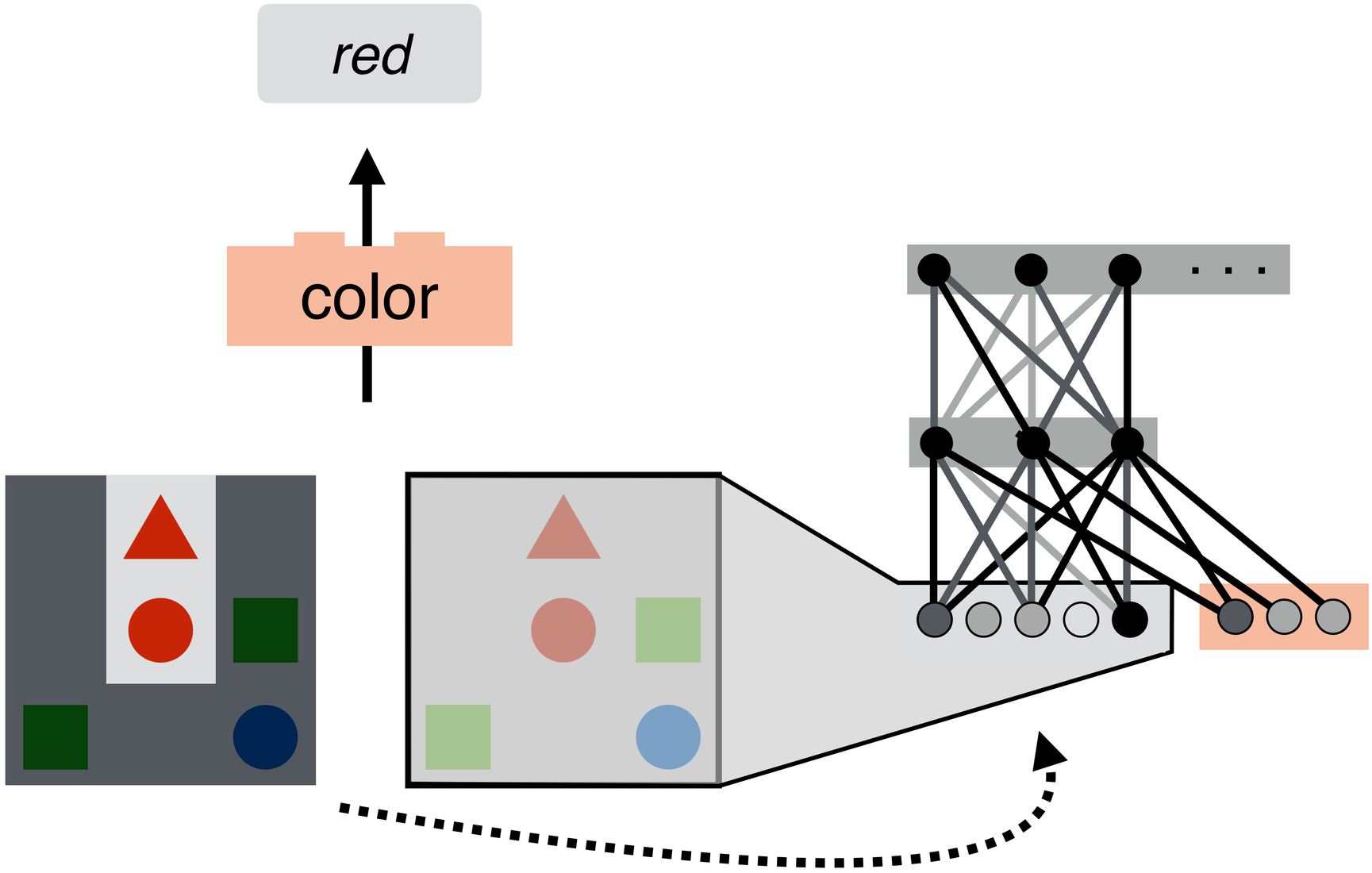
necktie



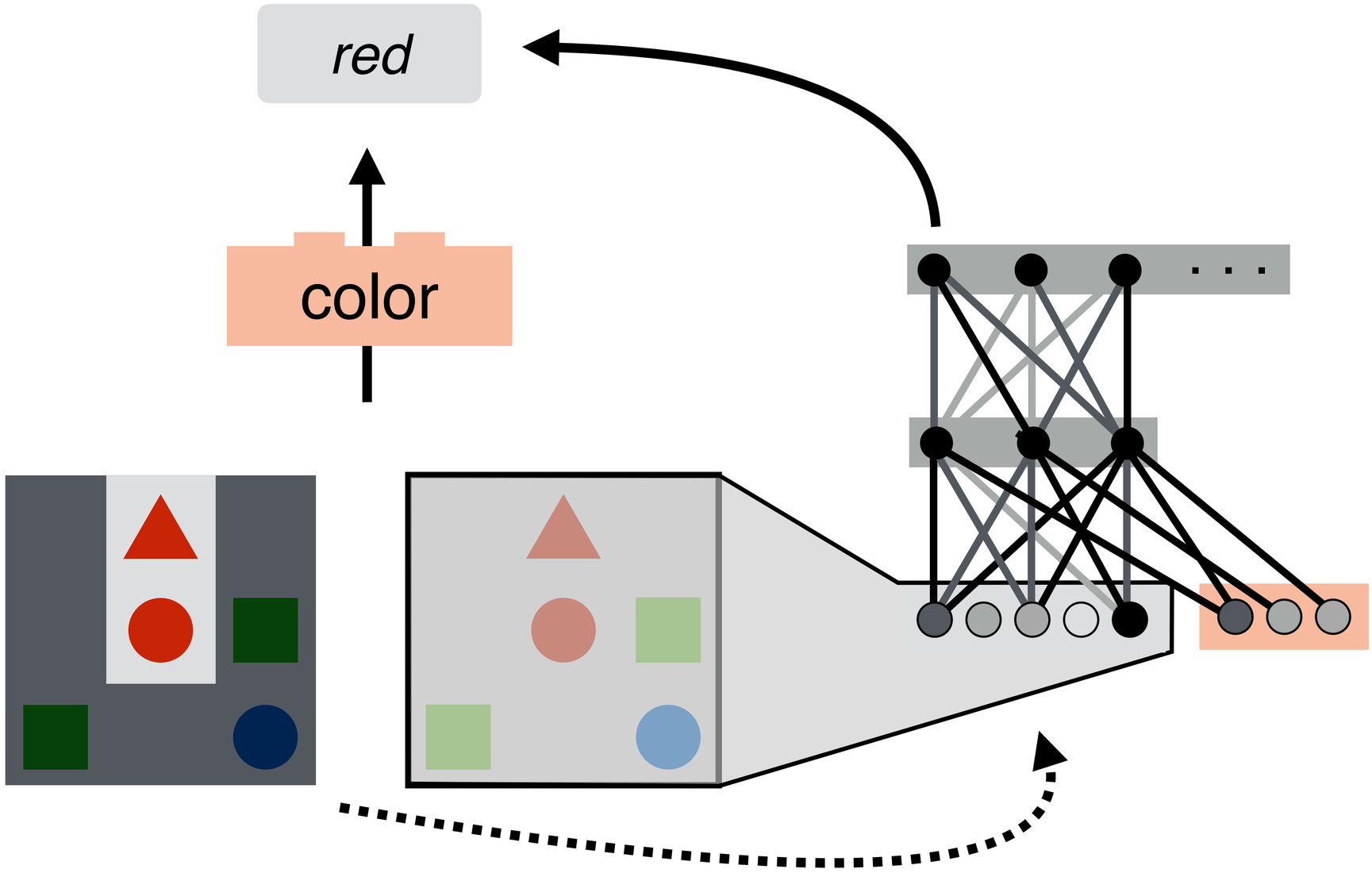
The [describe] module



The [describe] module



The [describe] module



What modules do we need?

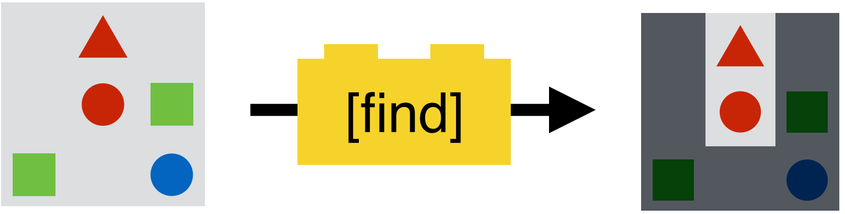
Is there a red shape above a circle?

What color is the triangle?

Who is running in the grass?

What cities are south of San Diego?

A module for predicates



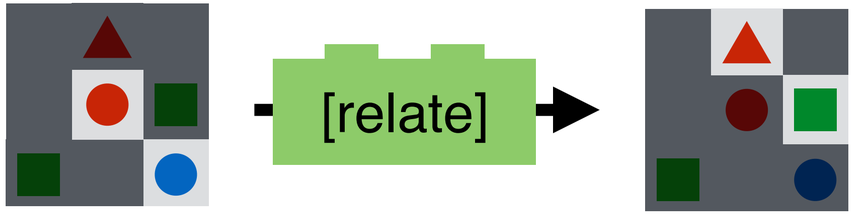
Is there a red shape above a

What color is the triangle?

Who is running in the grass?

What cities are south of San

A module for relations



Is there a red shape above a circle?

What color is the triangle?

Who is running in the grass?

[find]

What cities are south of San Diego?

Module inventory

Is there a red shape above a circle?

What color is the triangle?

[describe] [and]

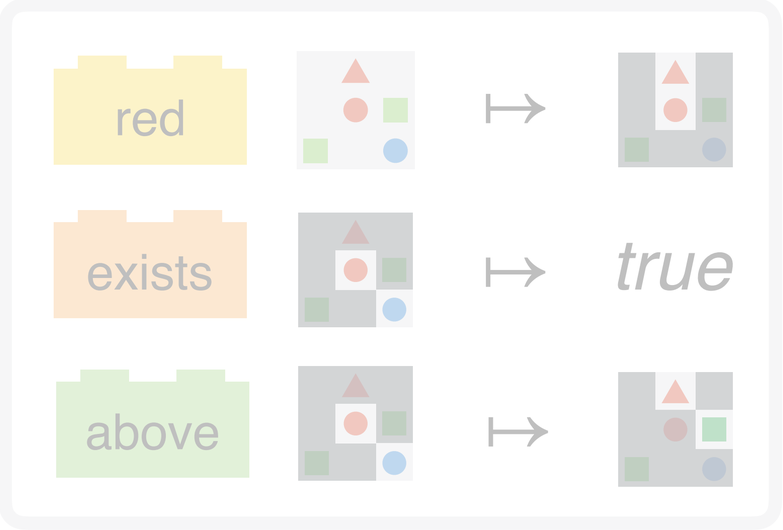
Who is running in the grass?

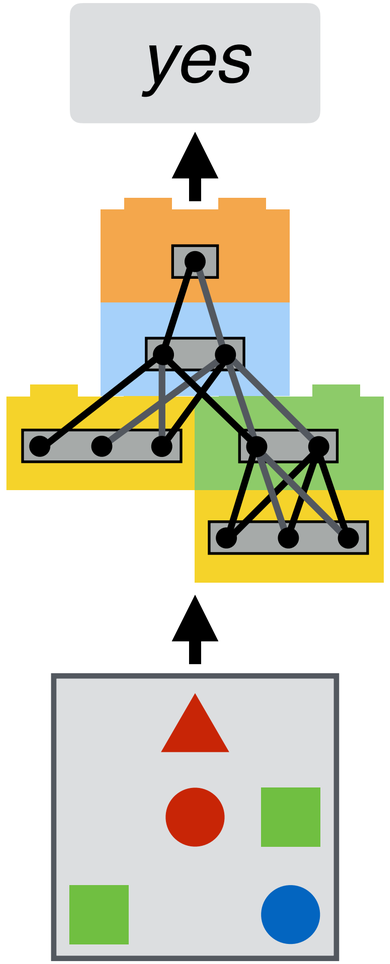
[find] [relate] [exists]

What cities are south of San Diego?

Outline

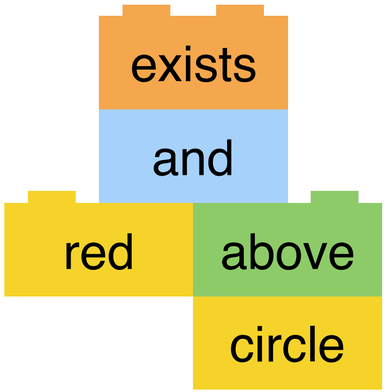
Is there a red shape above a circle?

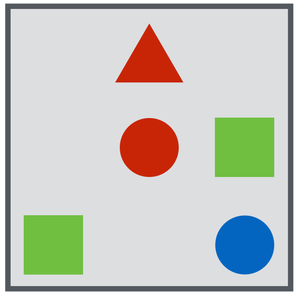




Learning

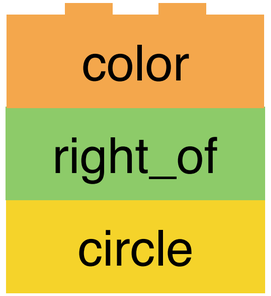
yes

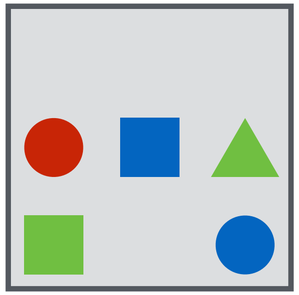




Is there a red shape above a circle?

blue

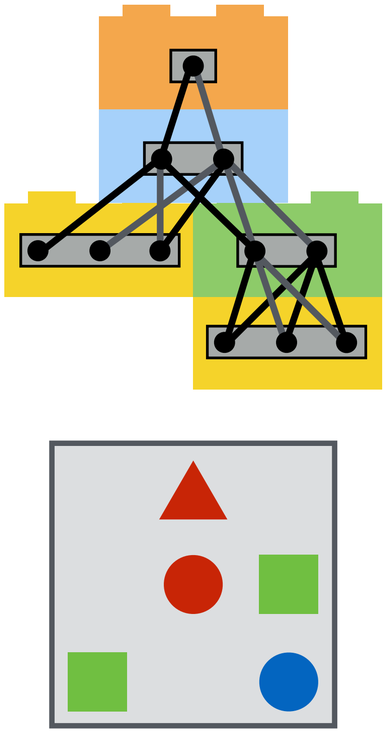




What color is the shape right of a circle?

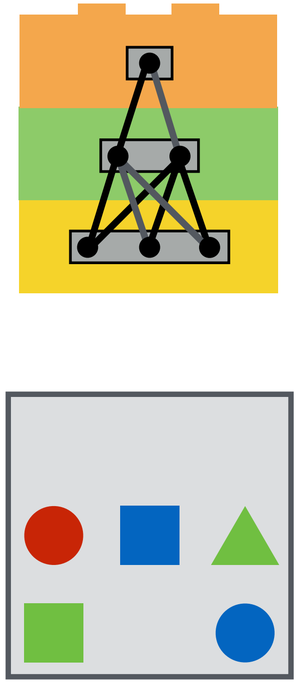
Learning

yes



Is there a red shape above a circle?

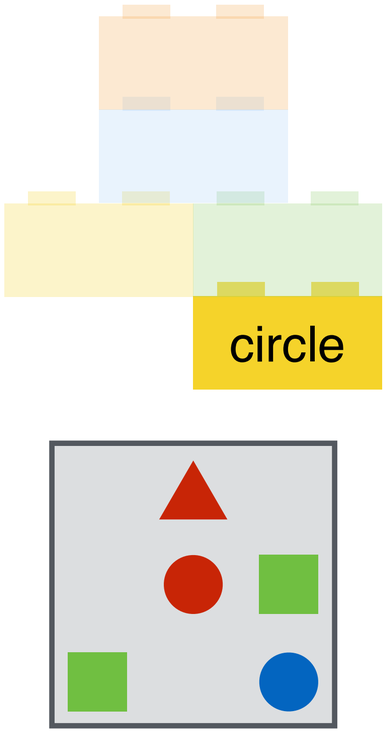
blue



What color is the shape right of a circle?

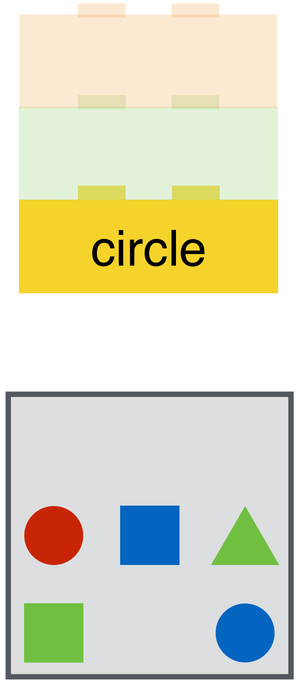
Parameter tying

yes



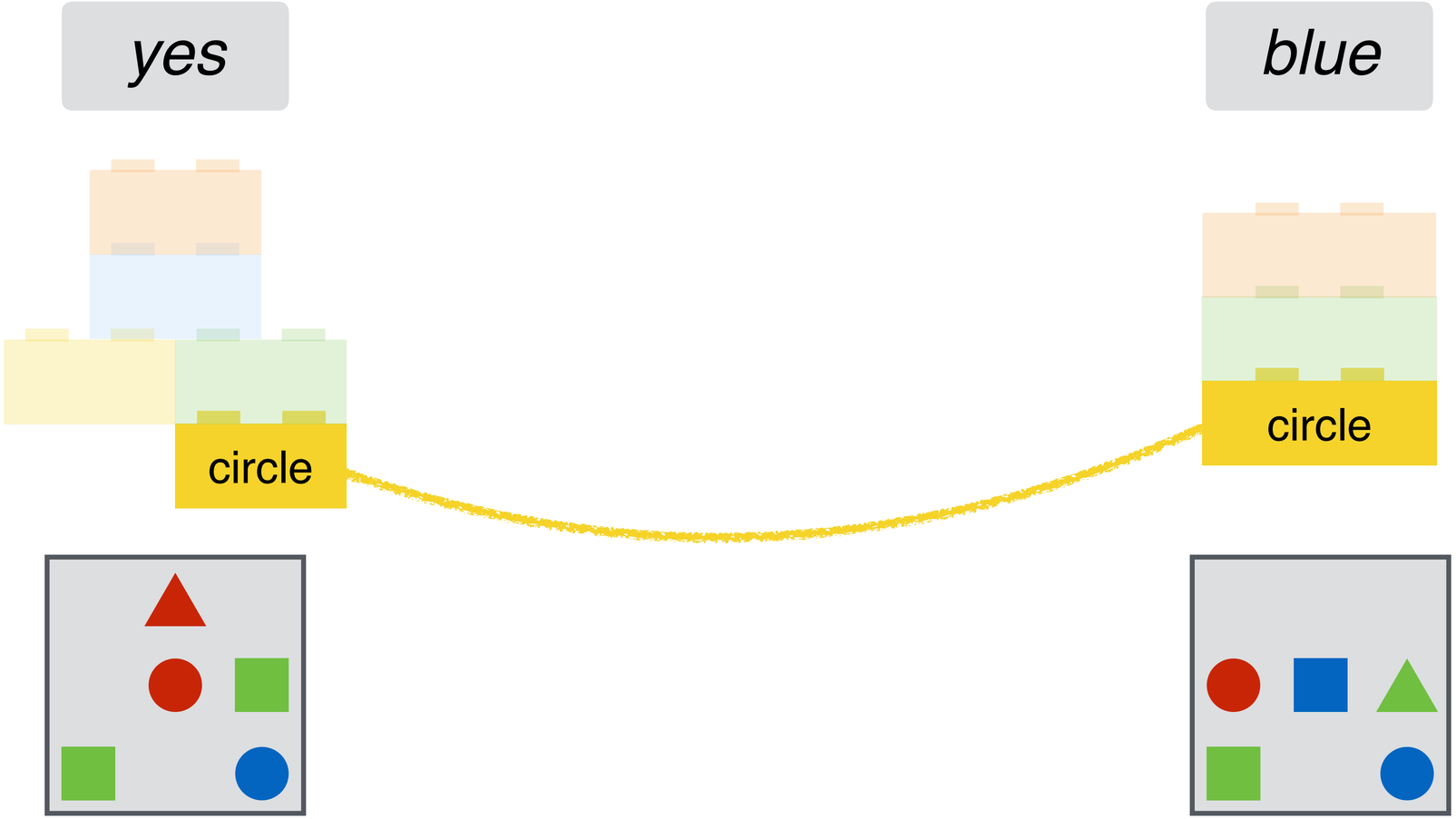
Is there a red shape above a circle?

blue



What color is the shape right of a circle?

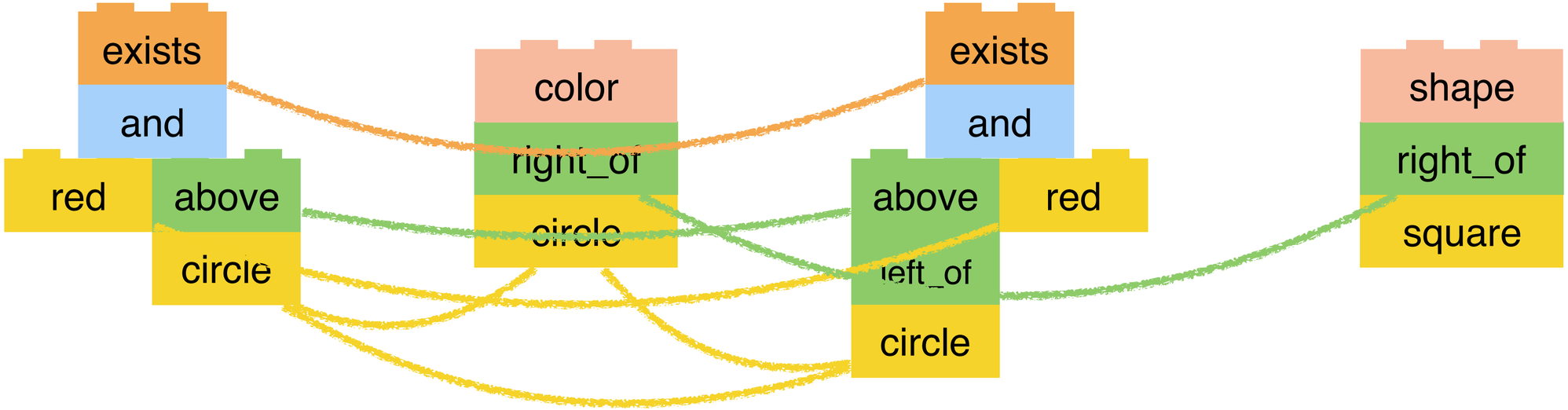
Parameter tying



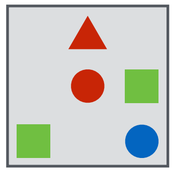
Is there a red shape above a circle?

What color is the shape right of a circle?

Extreme parameter tying



Learning with fixed layouts is easy!

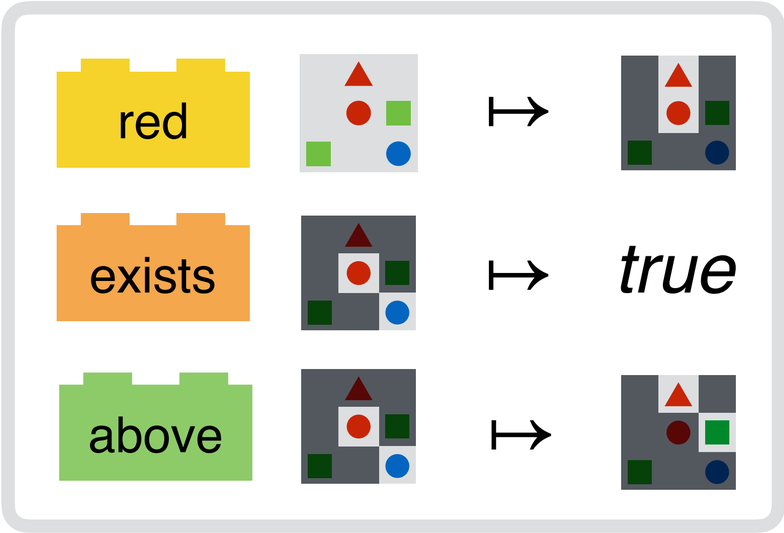
$$\arg\max_{W} \sum p(\text{yes} \mid \text{  }, \text{  }; W)$$

(where every root module outputs a distribution over answers and W is the set of all module parameters)

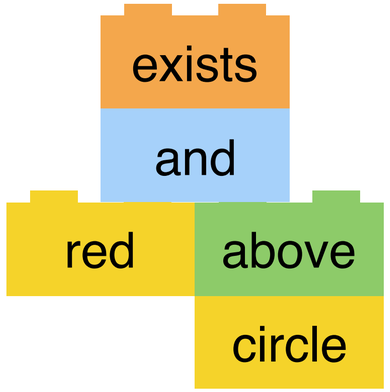
Maximum likelihood estimation

Outline

Is there a red shape above a circle?

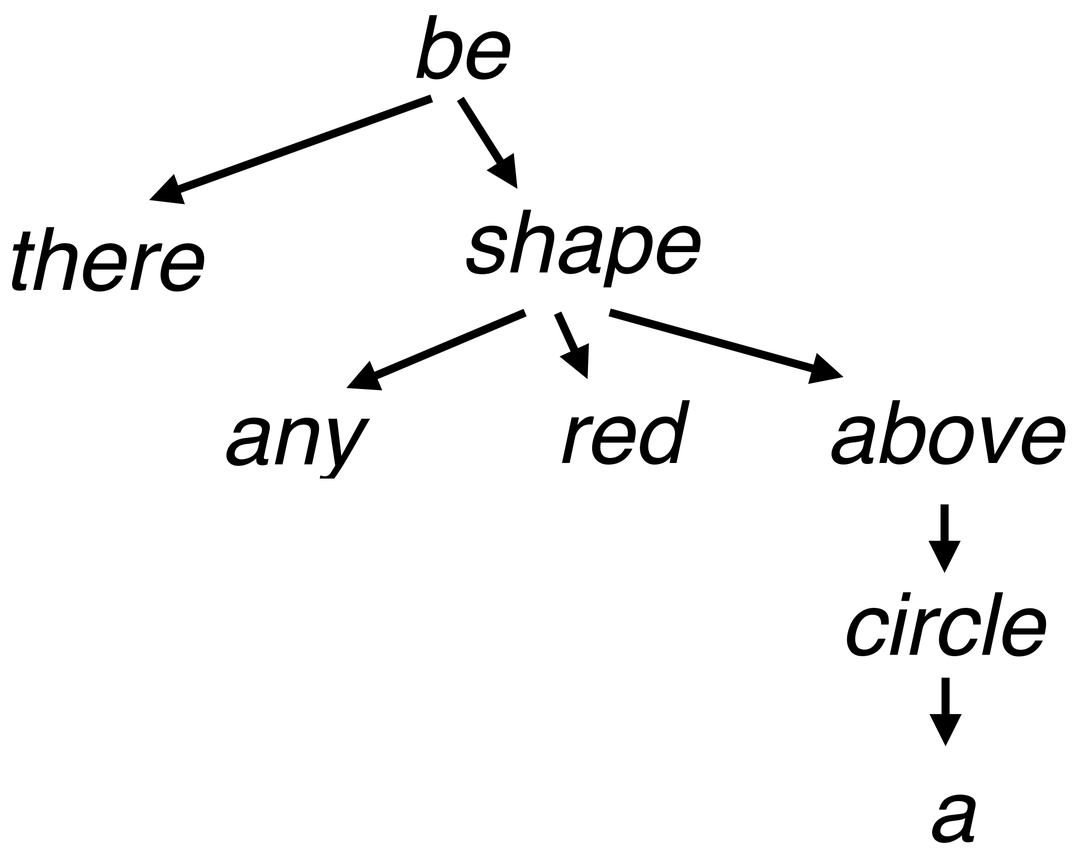


yes



Where do layouts come from?

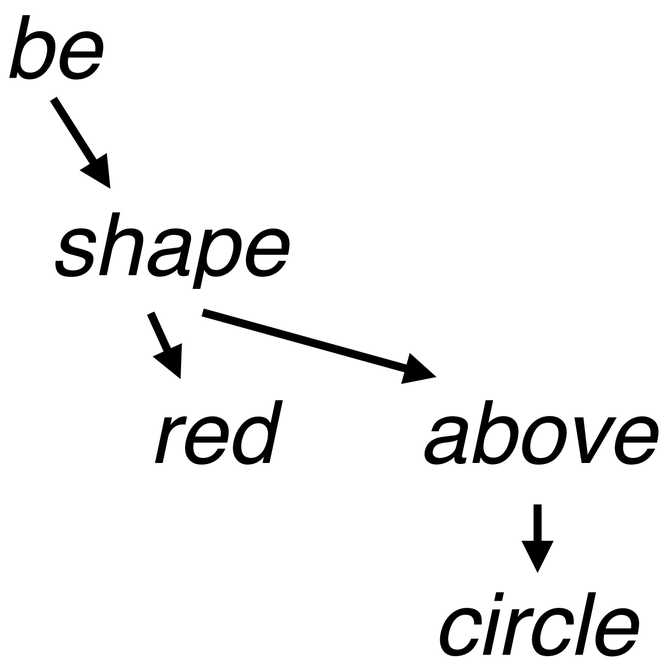
Is there a red shape above a circle?



[Reddy et al. 2016]

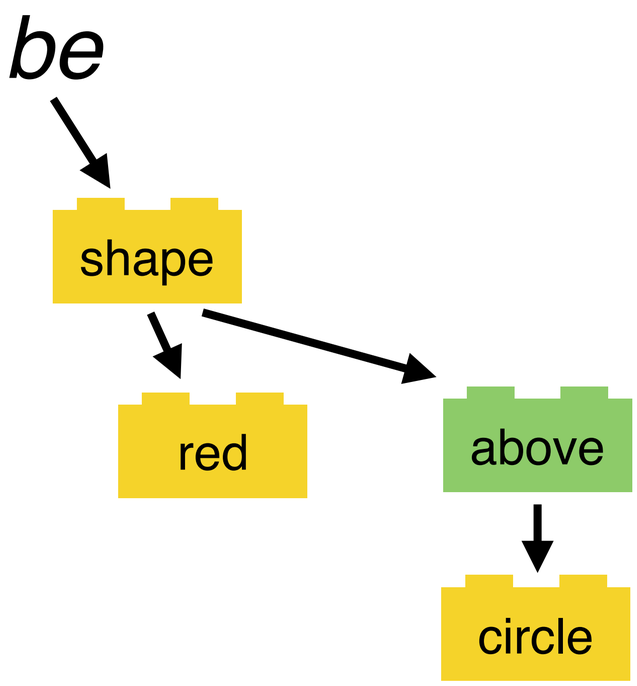
Where do layouts come from?

Is there a red shape above a circle?



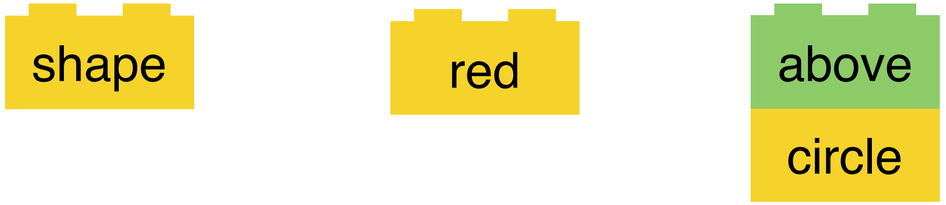
Where do layouts come from?

Is there a red shape above a circle?



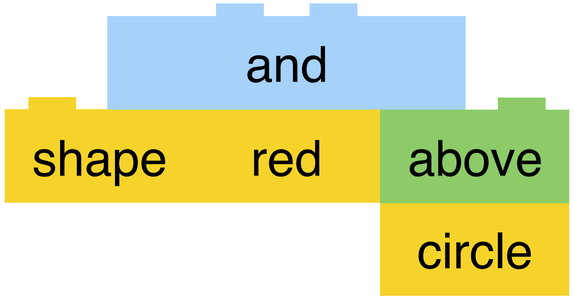
Where do layouts come from?

Is there a red shape above a circle?

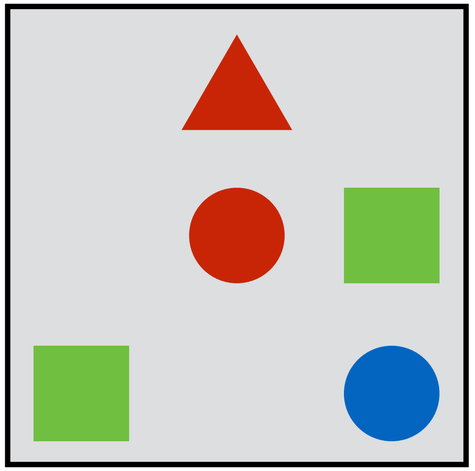


Where do layouts come from?

Is there a red shape above a circle?



Experiments



name	type	coastal
Columbia	city	no
Cooper	river	yes
Charleston	city	yes

Experiments: VQA dataset

What color is the necktie?

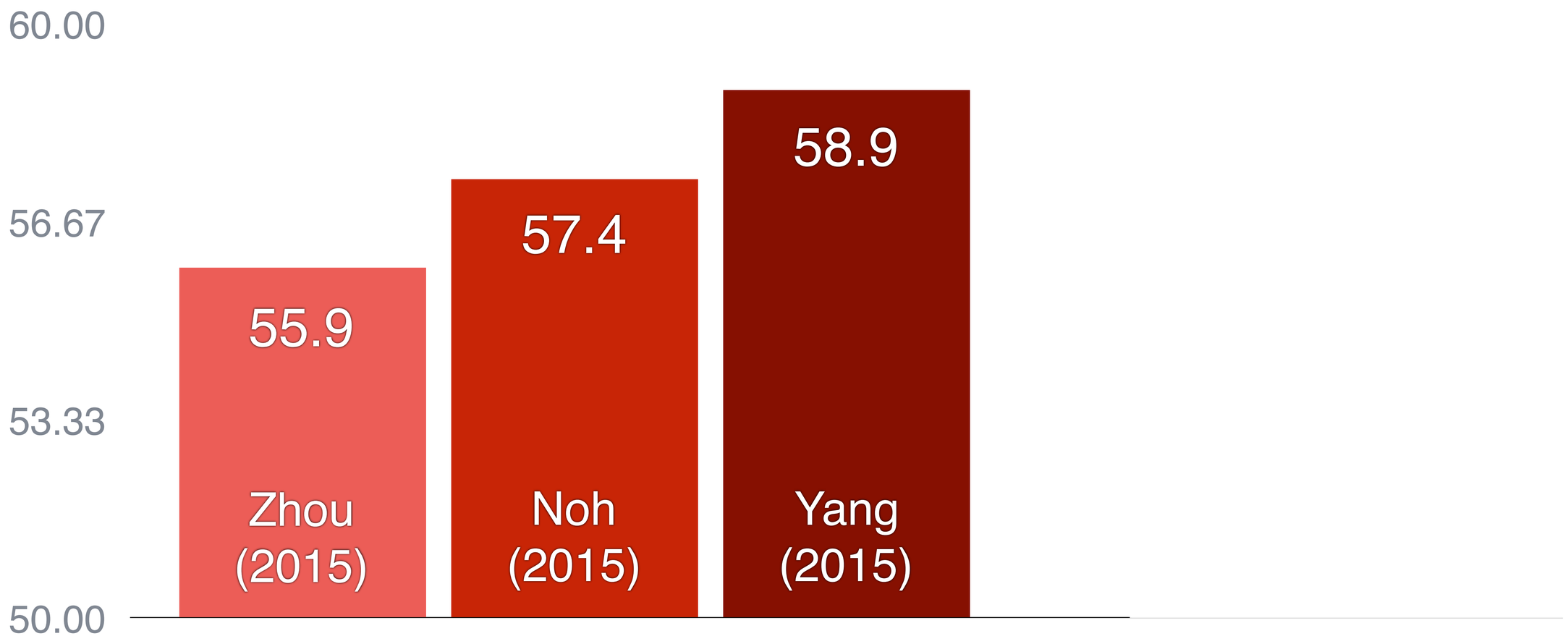
yellow

What is in the sheep's ear?

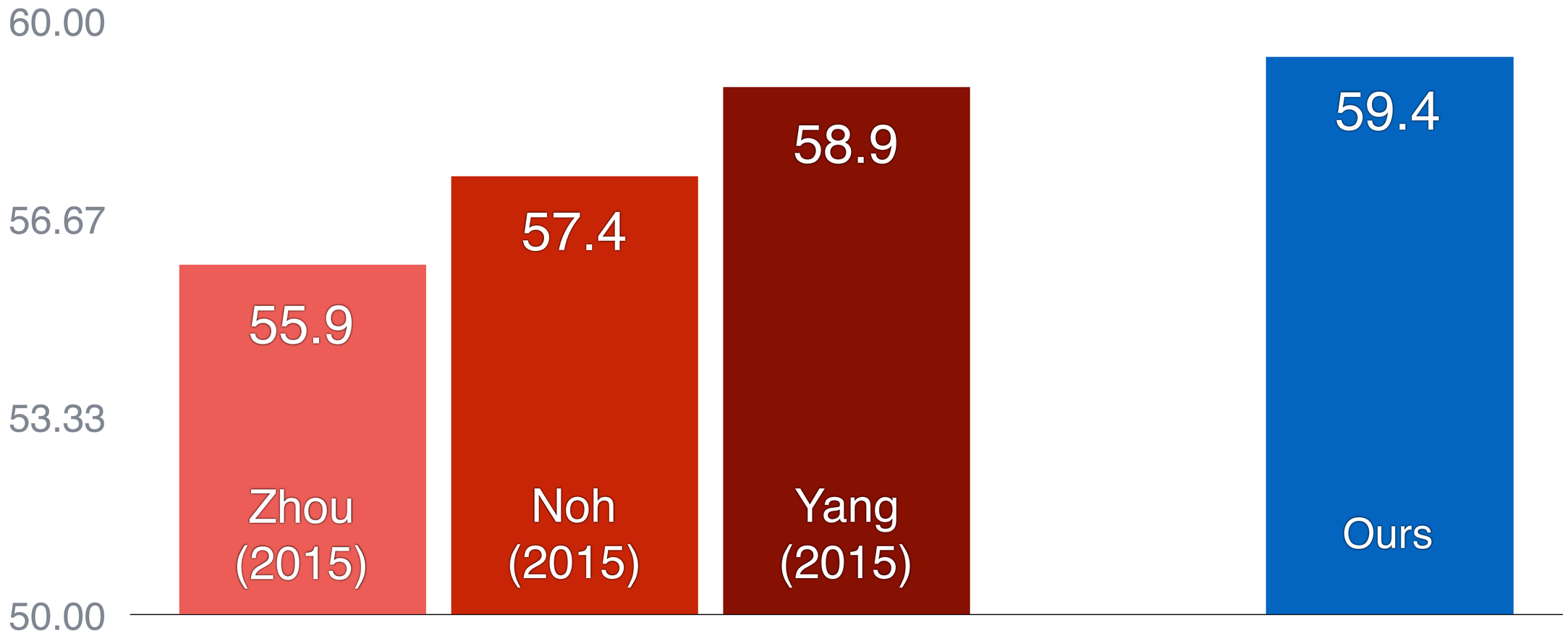
tag

[Antol et al. 2015]

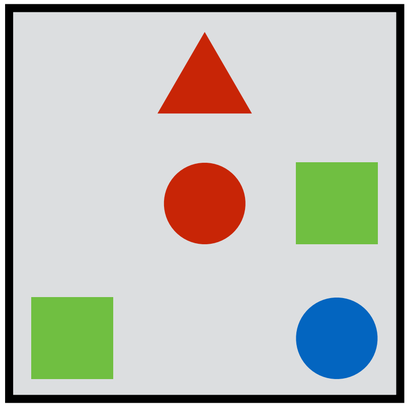
Experiments: VQA dataset

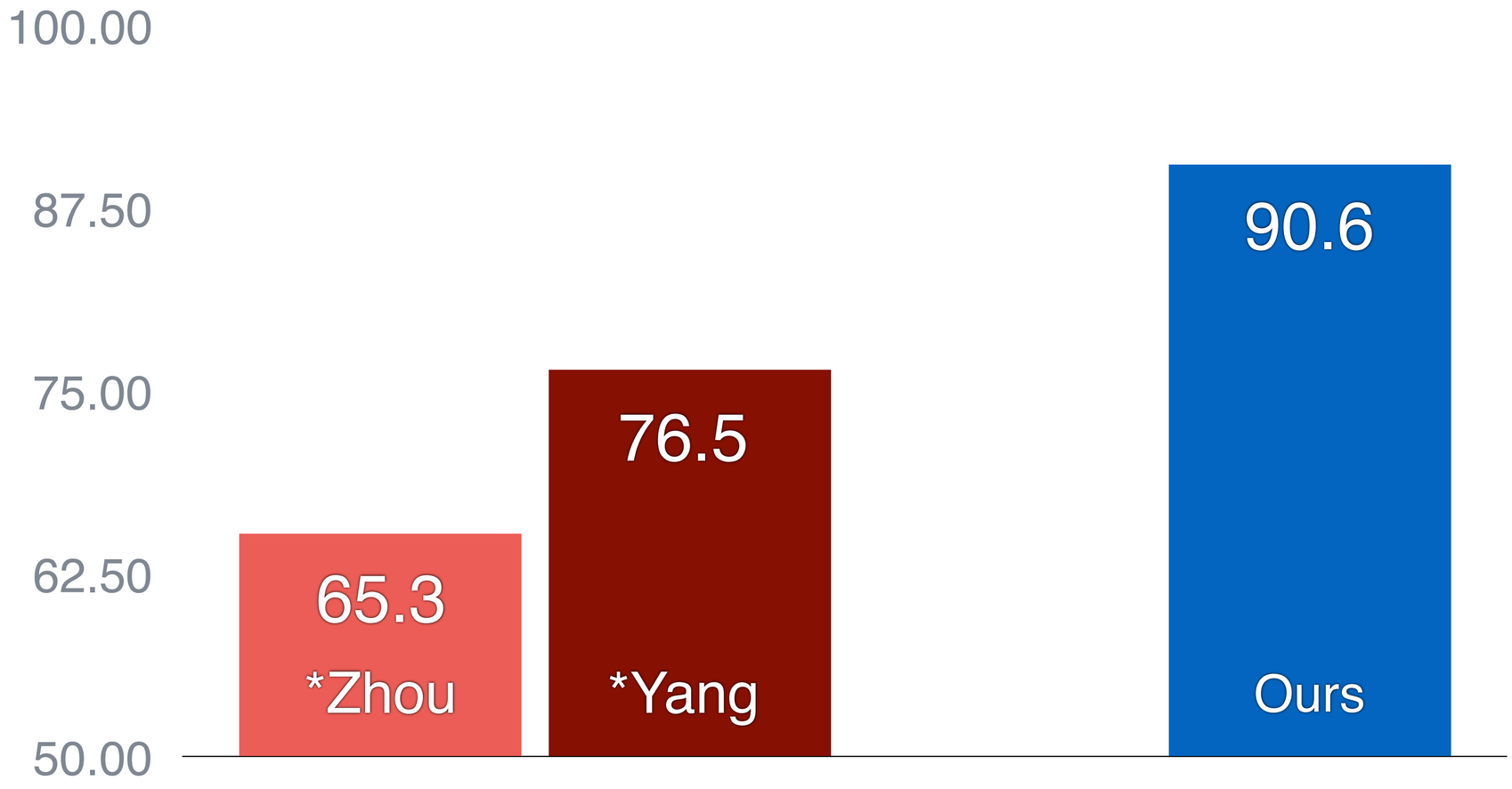


Experiments: VQA dataset



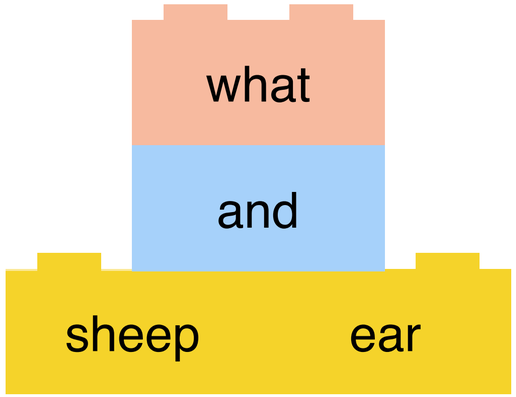
Experiments: SHAPES dataset





Experiments: VQA Dataset

What is in the sheep's ear?



tag

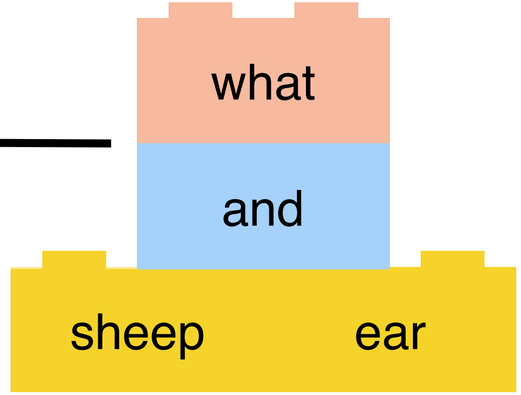
Experiments: VQA Dataset

What i
sheep

tag

Experiments: VQA Dataset

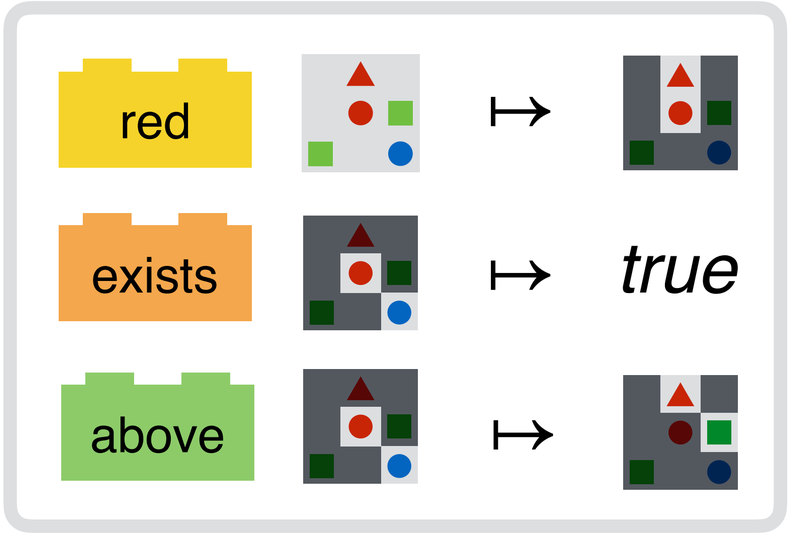
What is in the sheep's ear?

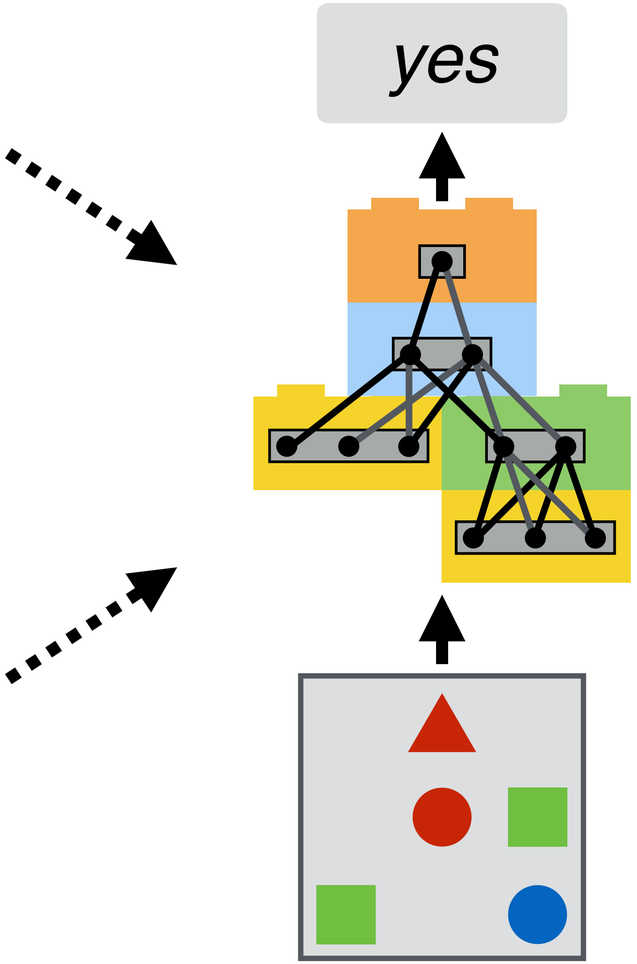


tag

Neural module networks

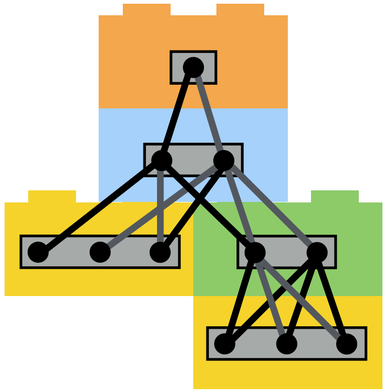
Is there a red shape above a circle?





Neural module networks

Linguistic structure dynamically generates model structure



Combines advantages of:

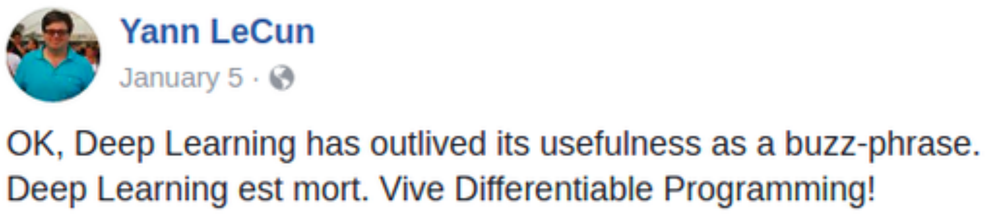
- Representation learning (like a neural net)
- Compositionality (like a semantic parser)

Differentiable programming

Deep nets are popular for a few reasons:

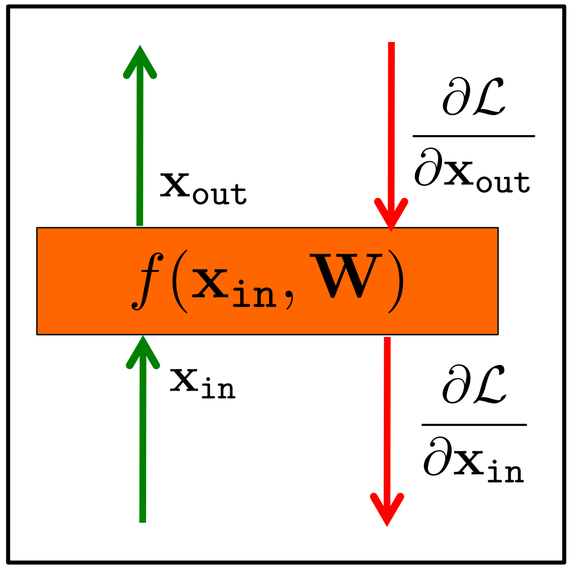
1. High capacity
2. Easy to optimize (differentiable)
3. Compositional “block based programming”

An emerging term for general models with these properties is **differentiable programming**.

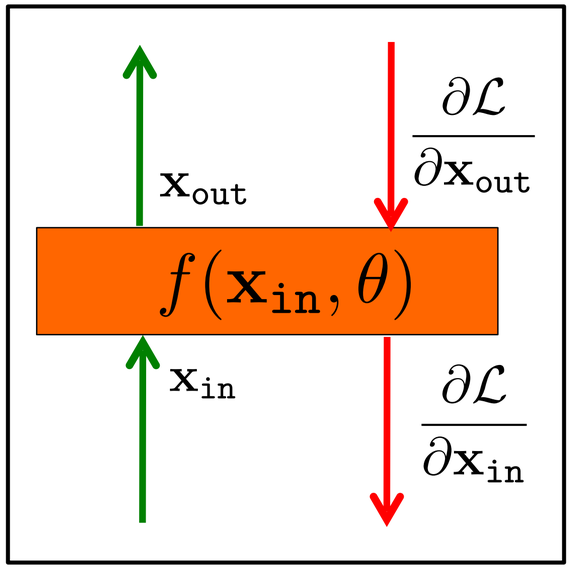


Differentiable programming

Deep learning

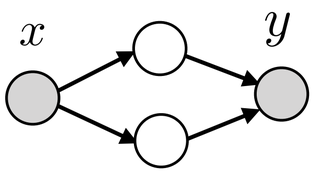


Differentiable programming



PyTorch

TensorFlow ™




```
for i, data in enumerate(dataset):
  iter_start_time = time.time()
  if total_steps % opt.print_freq == 0:
      t_data = iter_start_time - iter_data_time
  visualizer.reset()
  total_steps += opt.batch_size
  epoch_iter += opt.batch_size
  model.set_input(data)
  model.optimize_parameters()
```

Differentiable programming