Lecture 24: Language

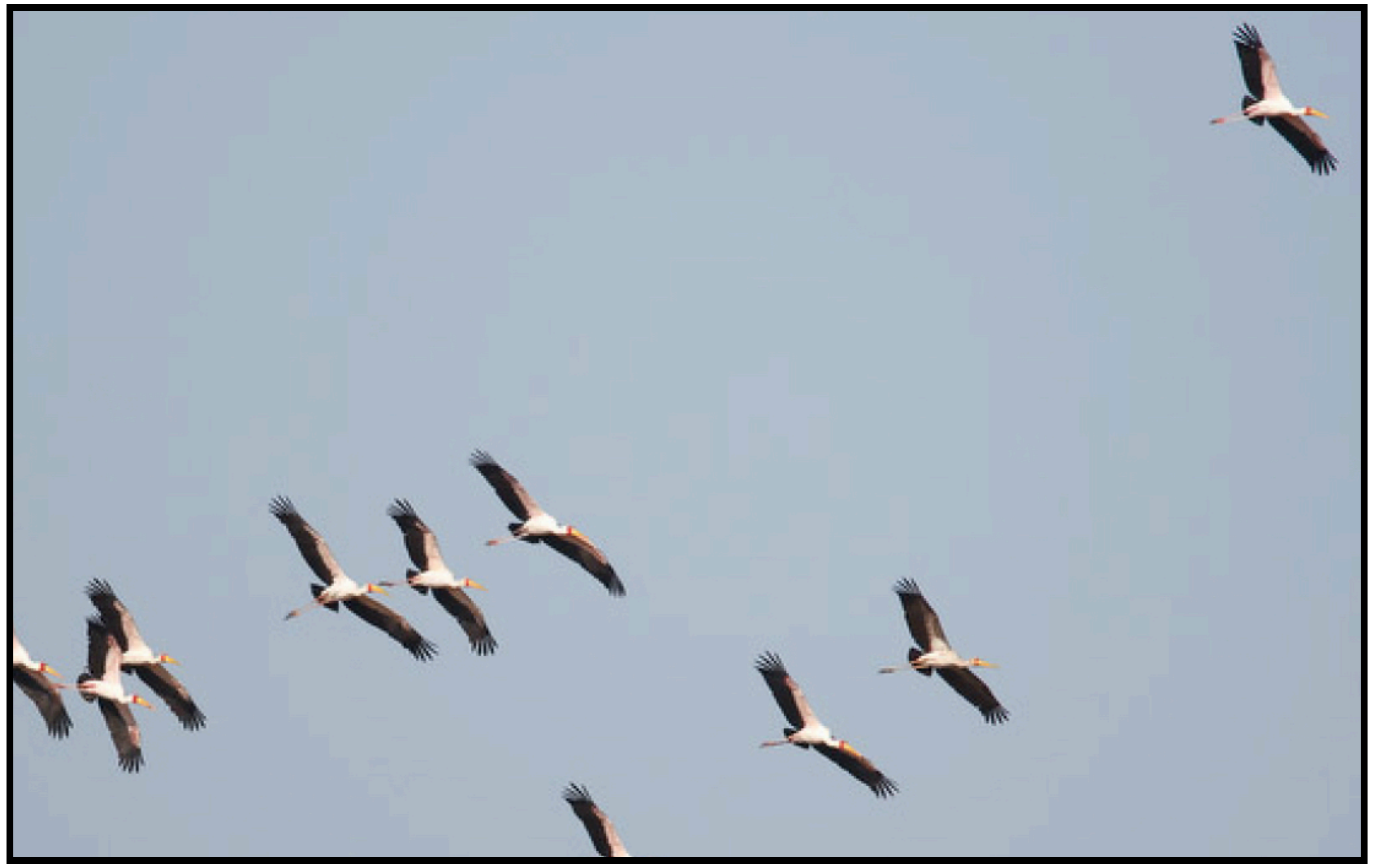
Announcements

- Project proposal comments out
- Chat with me at office hours if you have questions

Today

- Sequence modeling
- Image captioning
- Attention
- Visual Question Answering (VQA)
- Neural module networks

Image captioning



$\xrightarrow{f}$

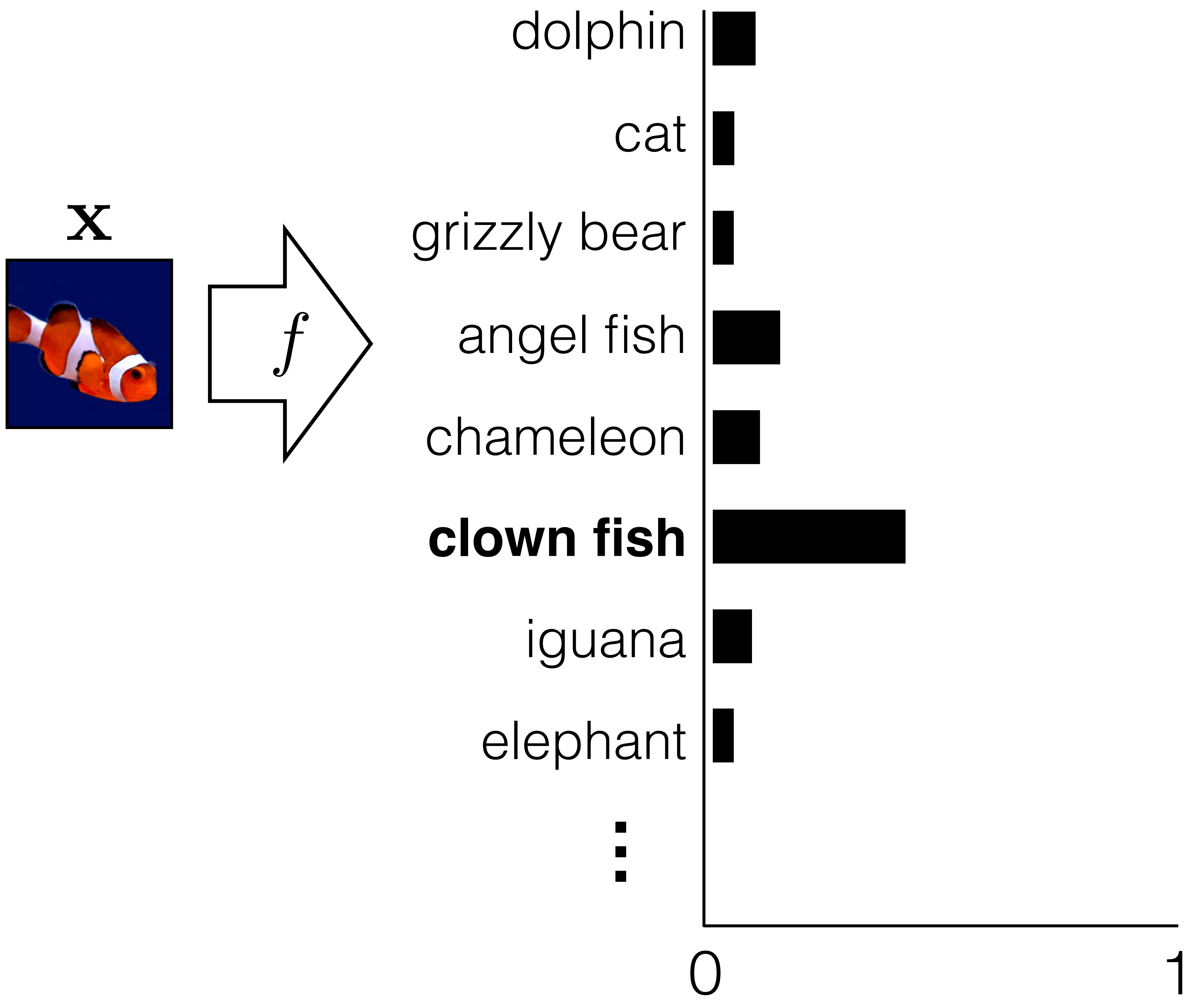
"A flock of birds against a gray sky"

Source: Isola, Torralba, Freeman

How to represent words as numbers?

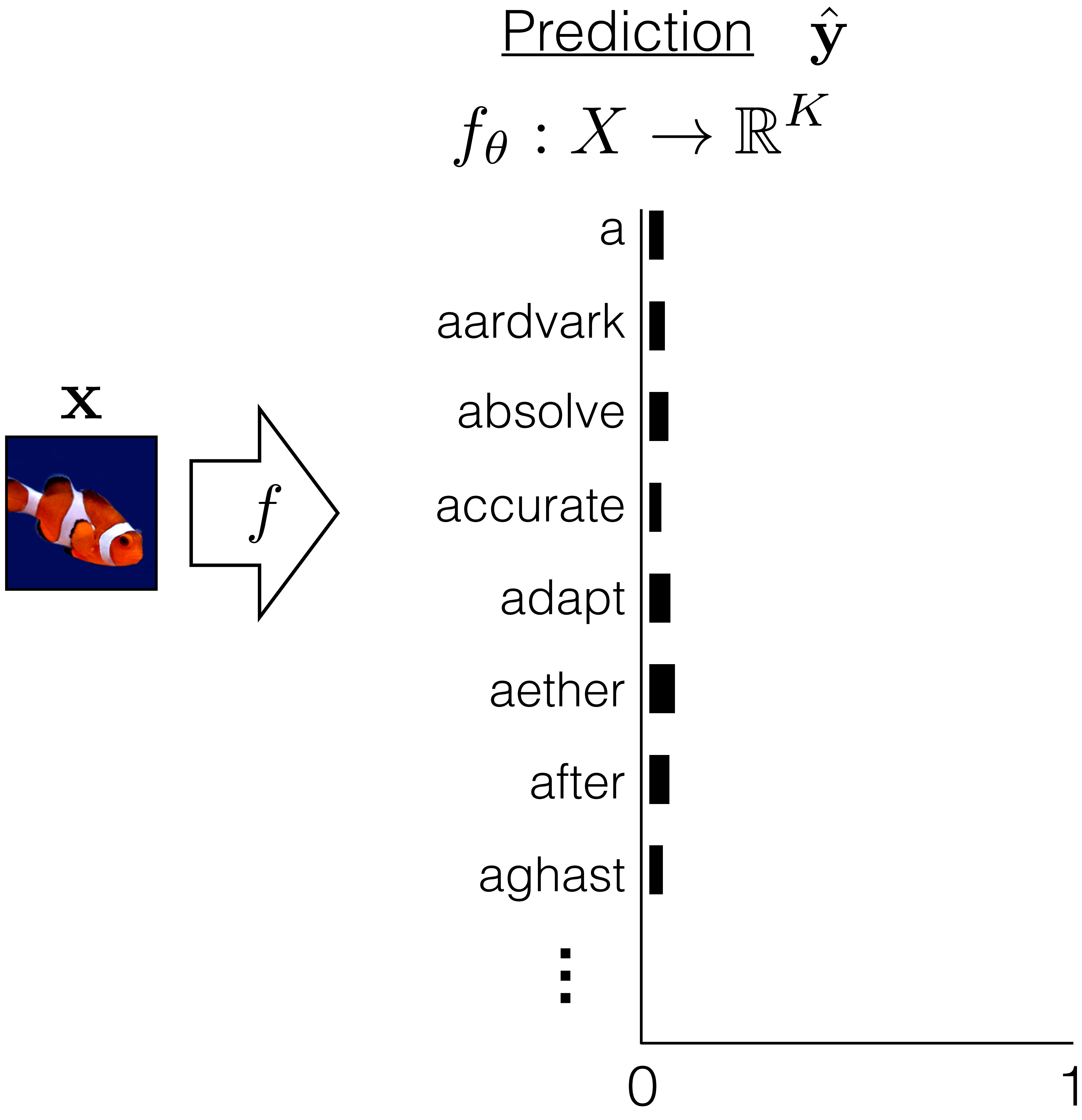
Prediction $\hat{\mathbf{y}}$

$$f_\theta : X \rightarrow \mathbb{R}^K$$



Source: Isola, Torralba, Freeman

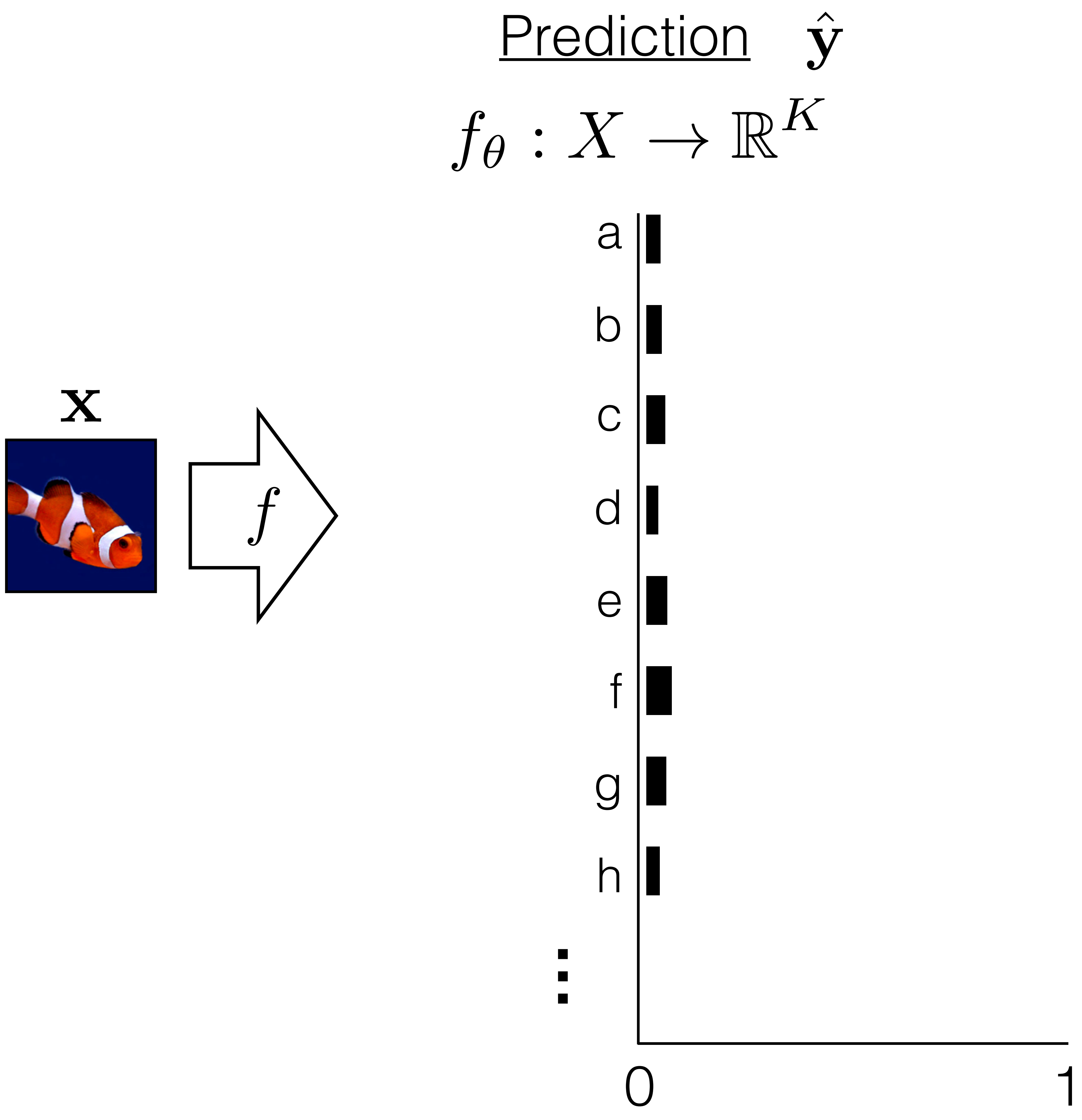
How to represent words as numbers?



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).

Source: Isola, Torralba, Freeman

How to represent words as numbers?



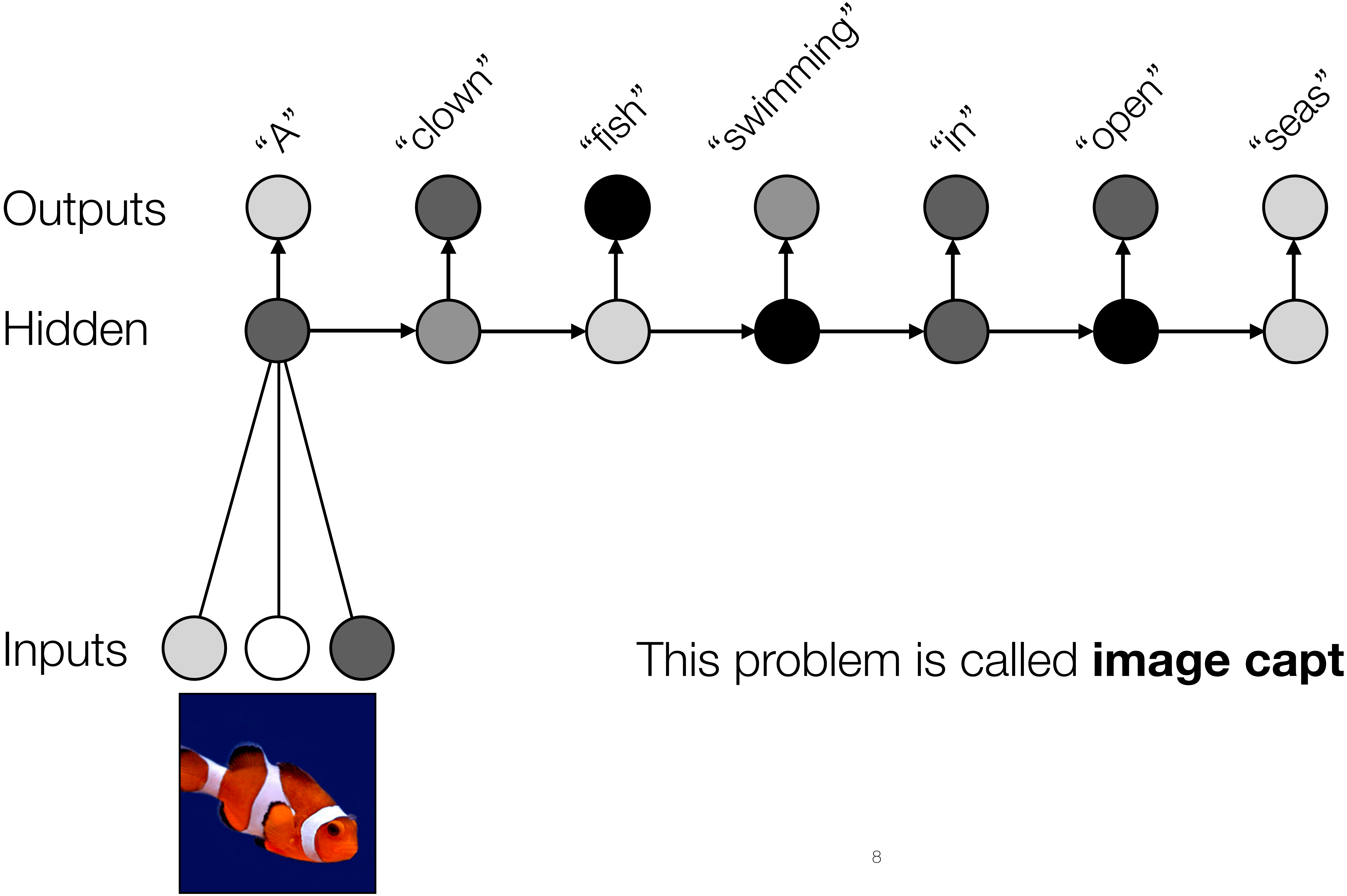
Prediction $\hat{\mathbf{y}}$

$f_\theta : X \rightarrow \mathbb{R}^K$

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

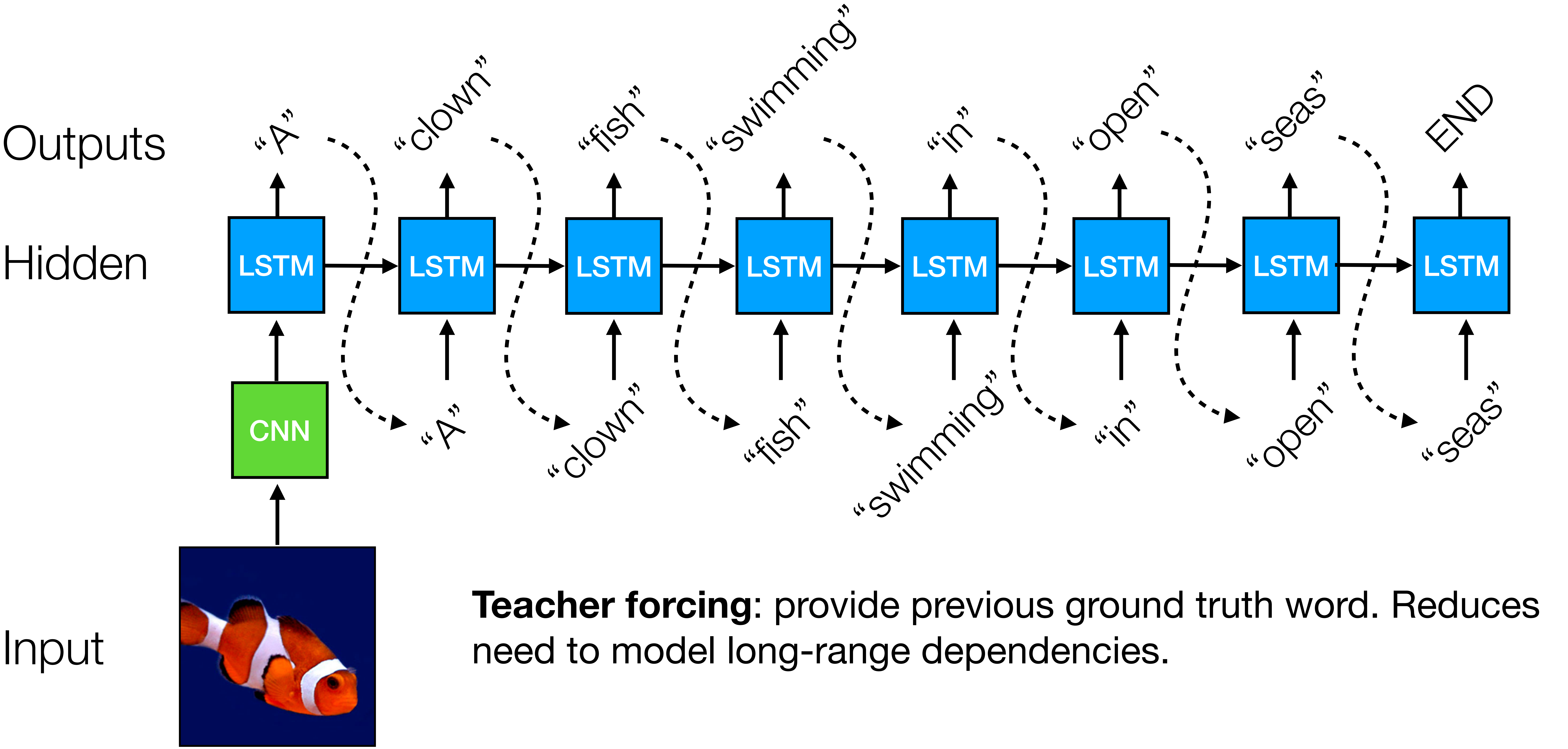
and represent words as a sequence of characters.

Source: Isola, Torralba, Freeman



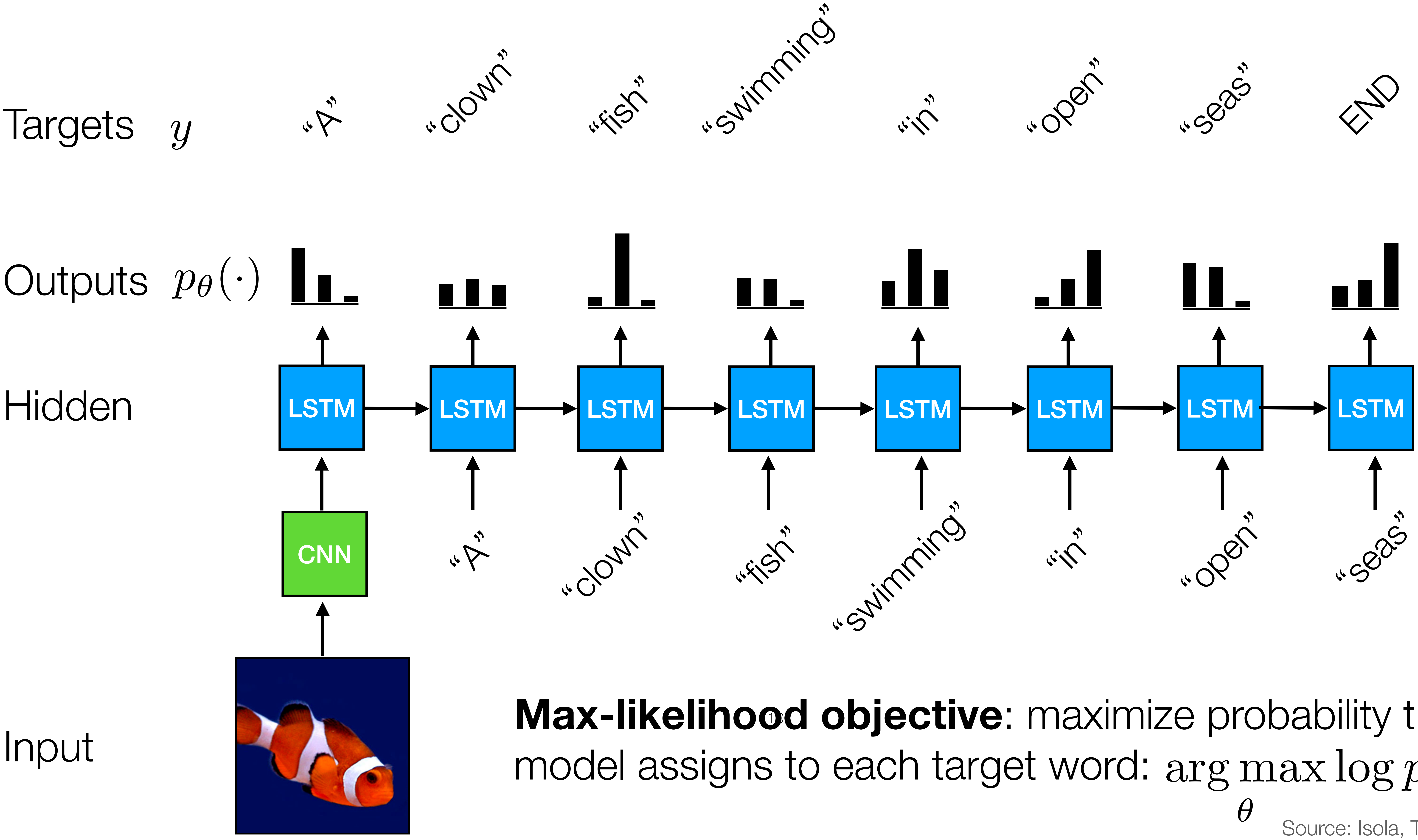
This problem is called **image captioning.**

Source: Isola, Torralba, Freeman



Teacher forcing: provide previous ground truth word. Reduces need to model long-range dependencies.

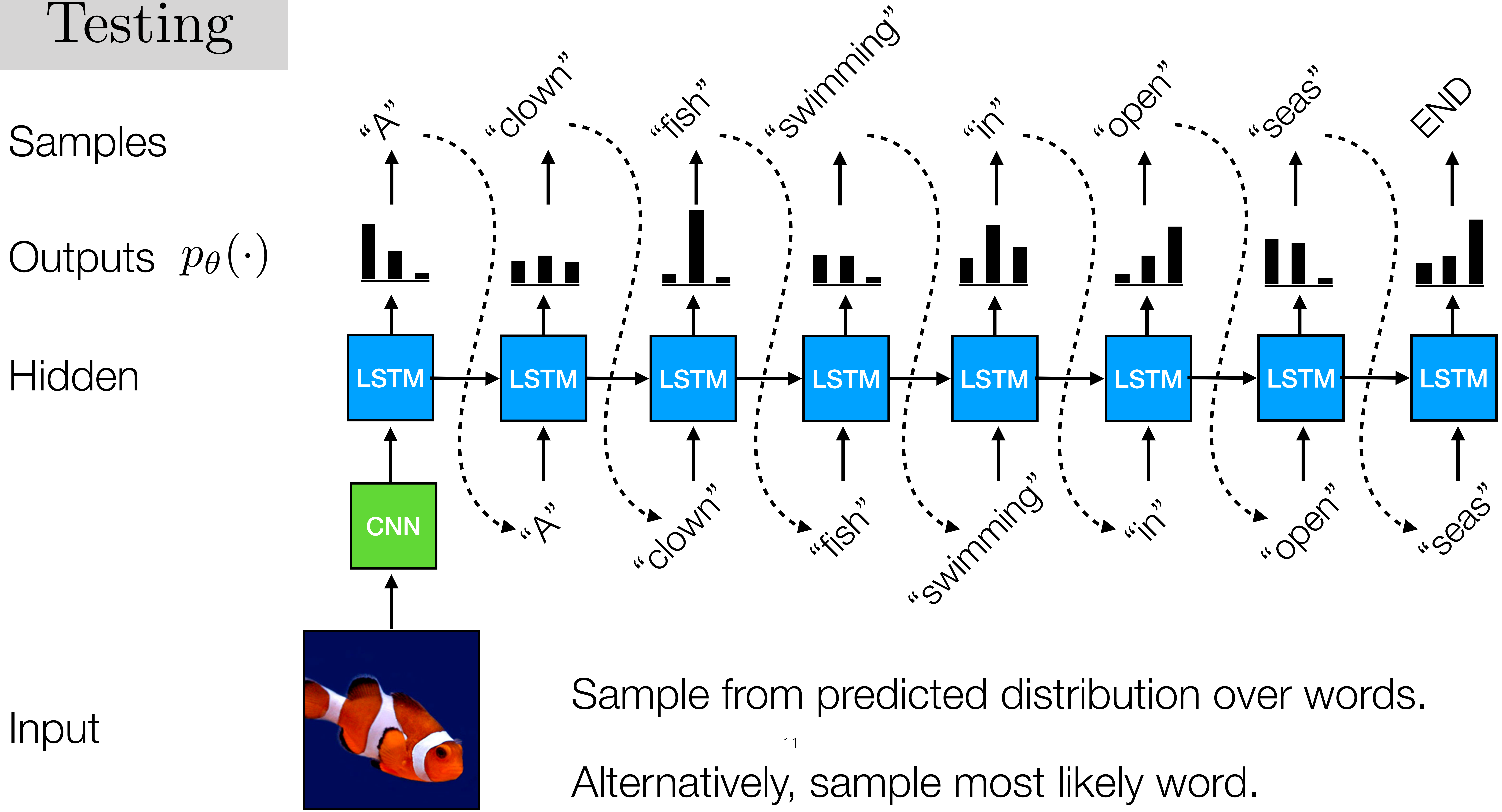
Source: Isola, Torralba, Freeman



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

Source: Isola, Torralba, Freeman

Testing

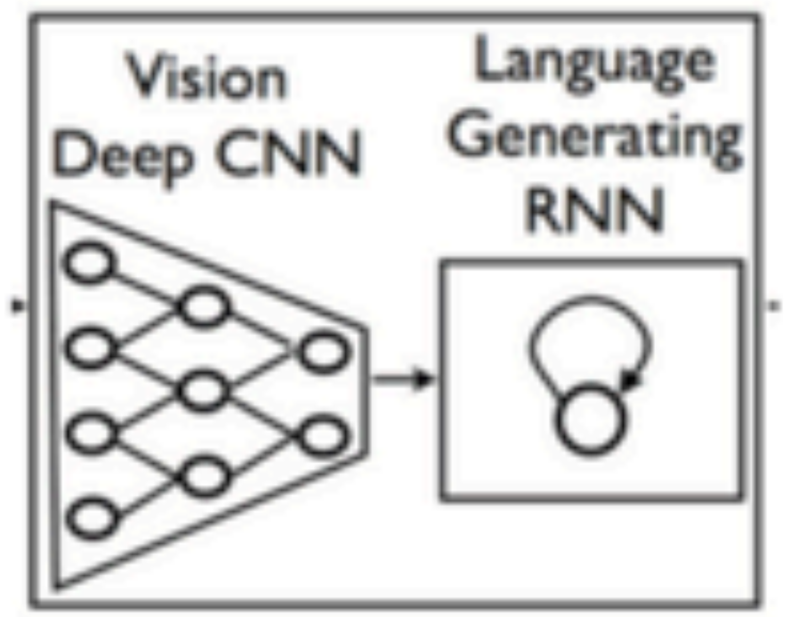


Sample from predicted distribution over words.

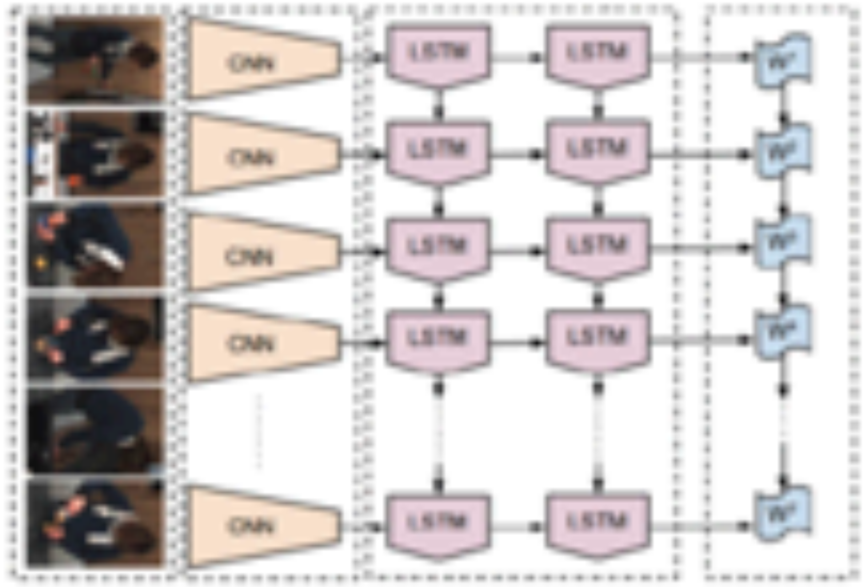
Alternatively, sample most likely word.

Source: Isola, Torralba, Freeman

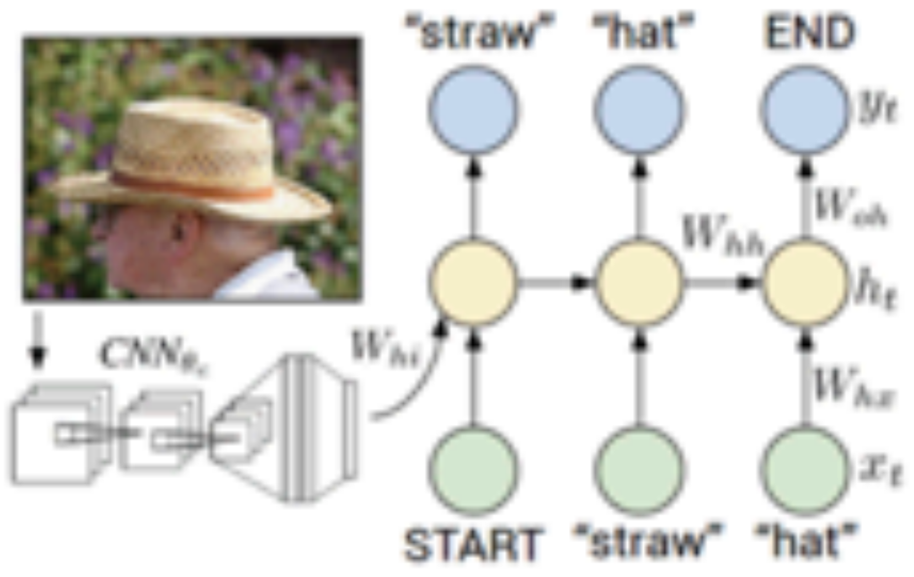
It was very popular a few years ago



Vinyals et al., 2015



Donahue et al., 2015



Karpathy and Fei-Fei, 2015

Hodosh et al., 2013

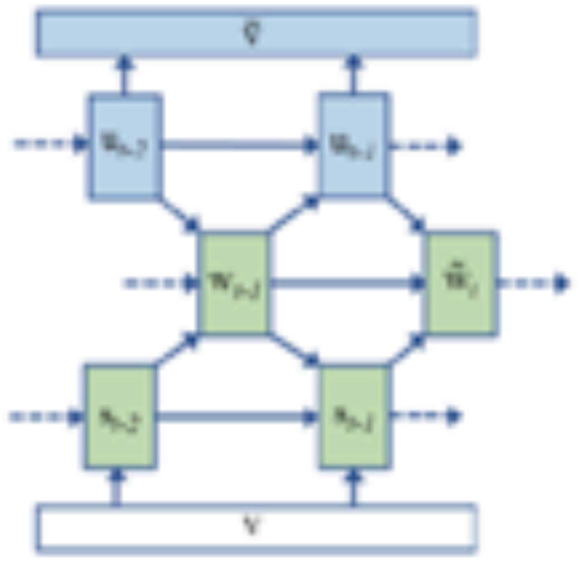
Fang et al., 2015

... and many more

Mao et al., 2015

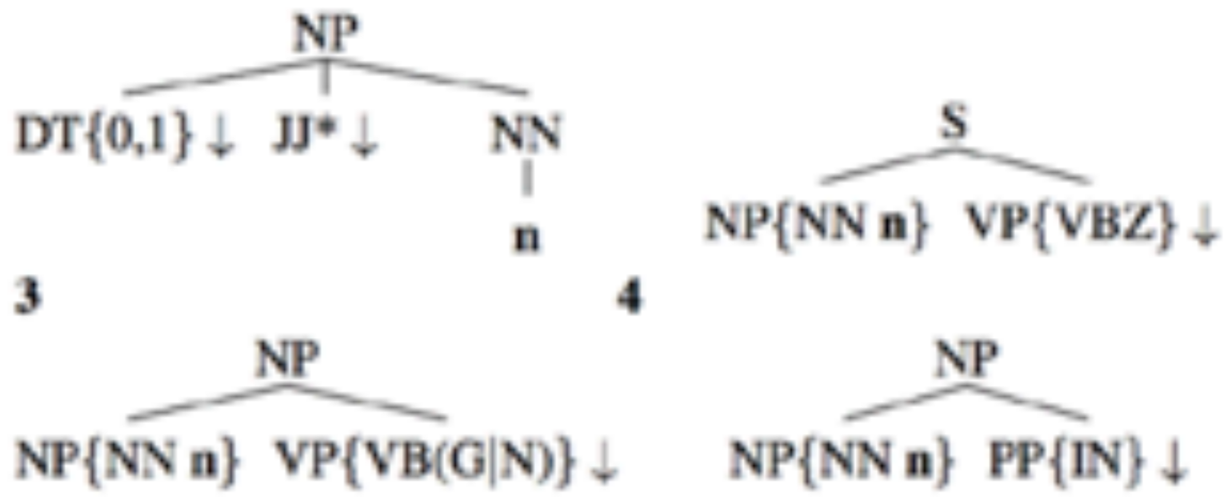
Ordonez et al., 2011

Kulkarni et al., 2011

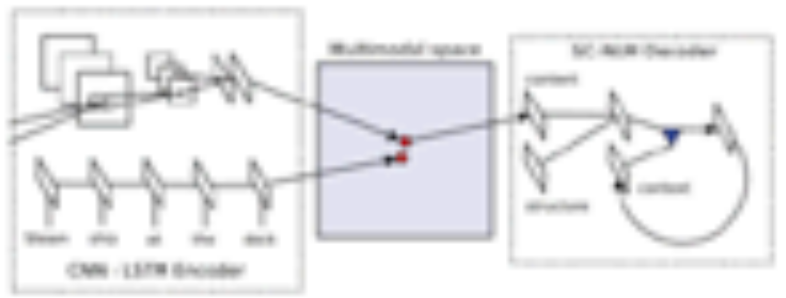


Chen and Zitnick, 2015

Farhadi et al., 2010



Mitchell et al., 2012

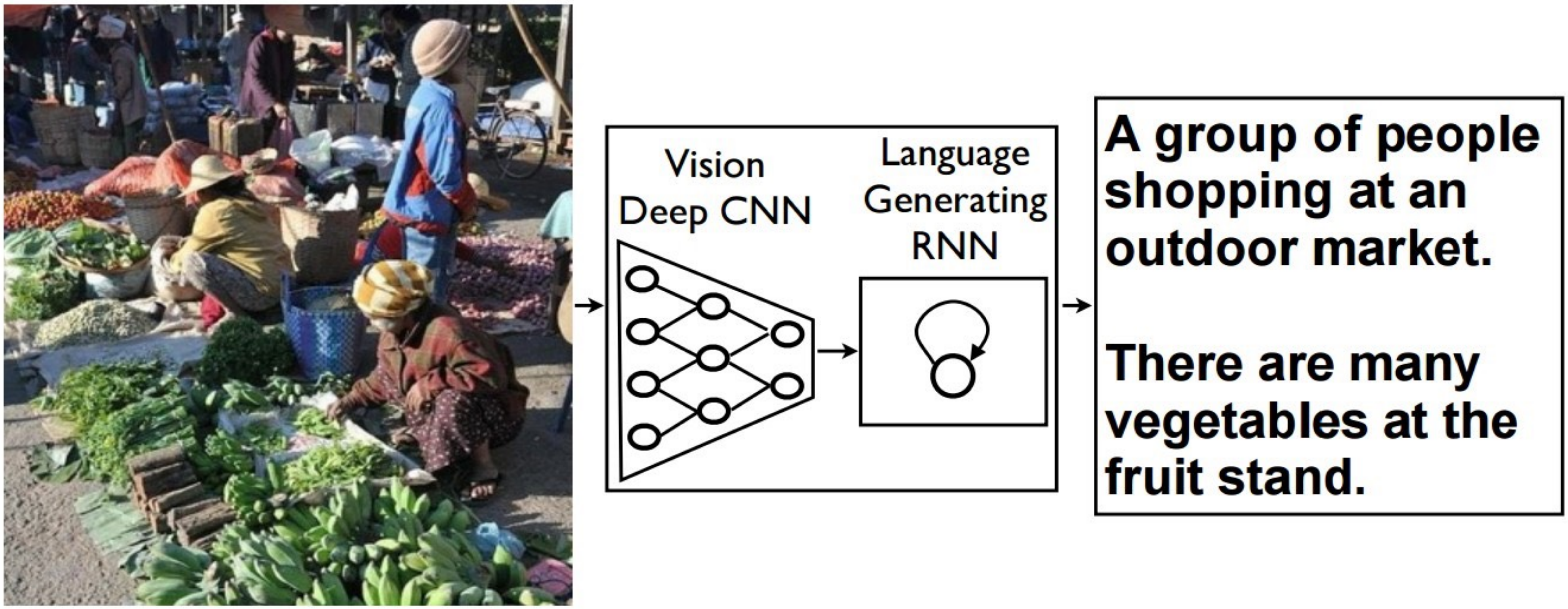


Kiros et al., 2015

Source: D. Parikh

Show and Tell: A Neural Image Caption Generator

[Vinyals et. al., CVPR 2015]



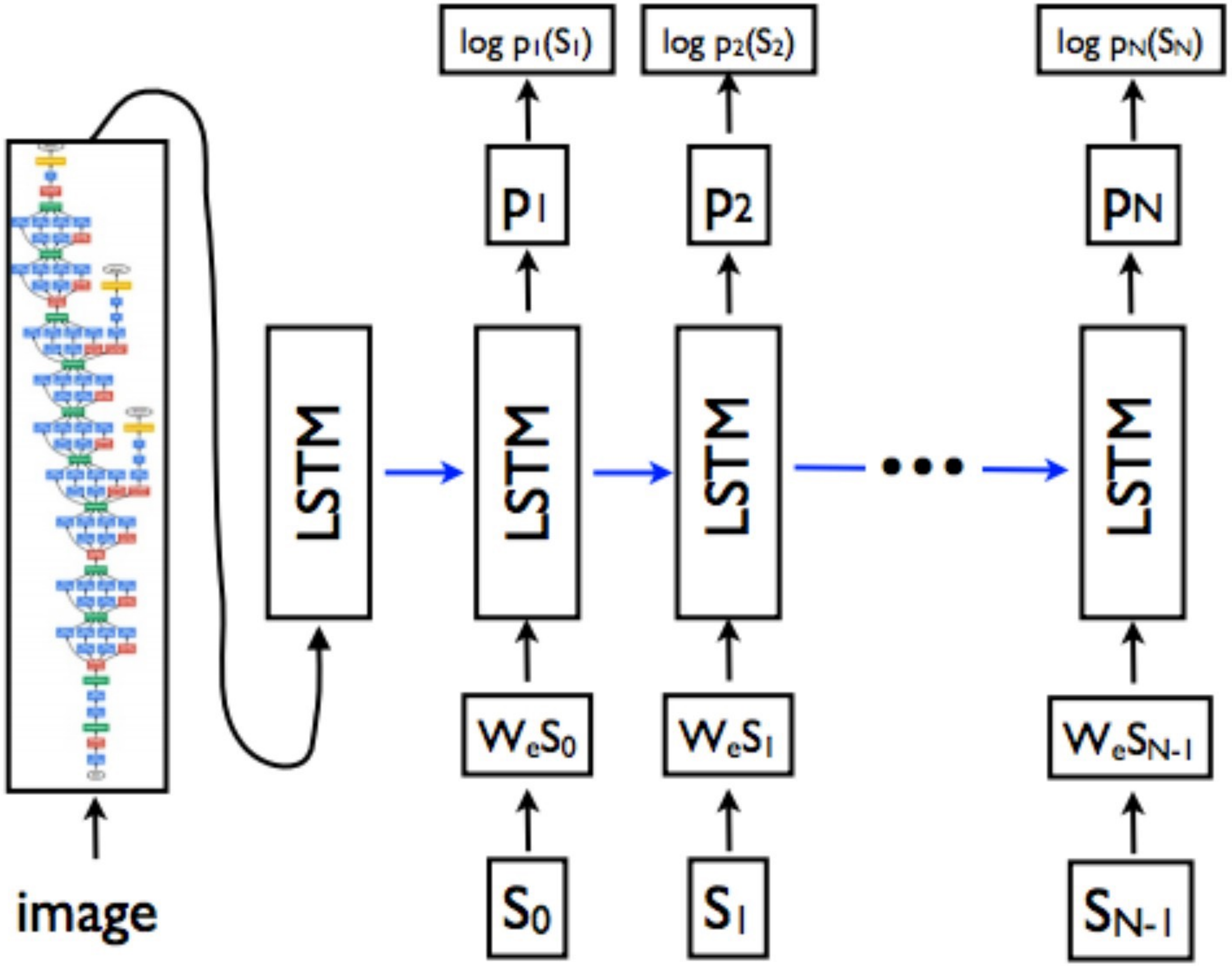
A group of people shopping at an outdoor market.

There are many vegetables at the fruit stand.

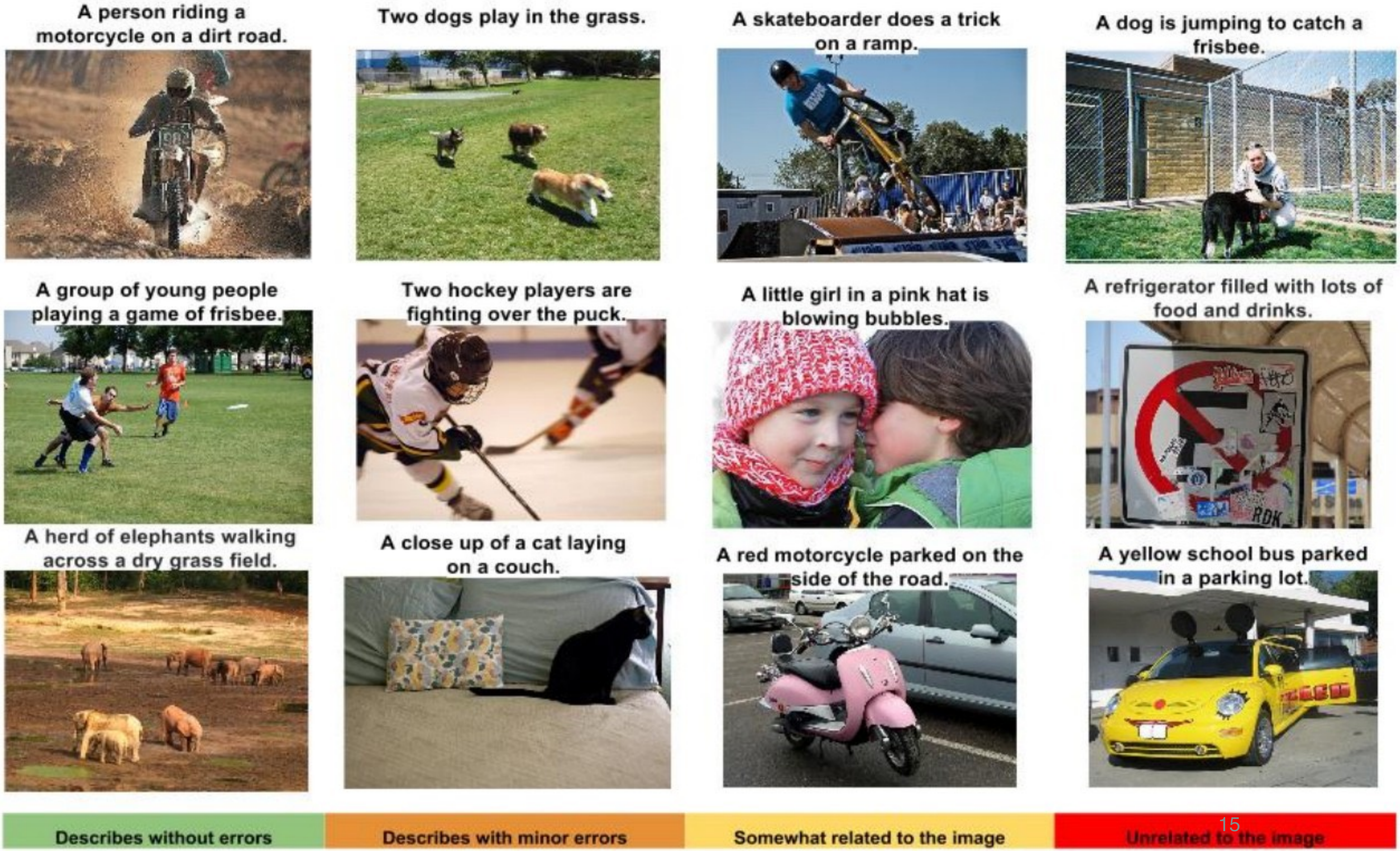
Source: Isola, Torralba, Freeman

Show and Tell: A Neural Image Caption Generator

[Vinyals et. al., CVPR 2015]



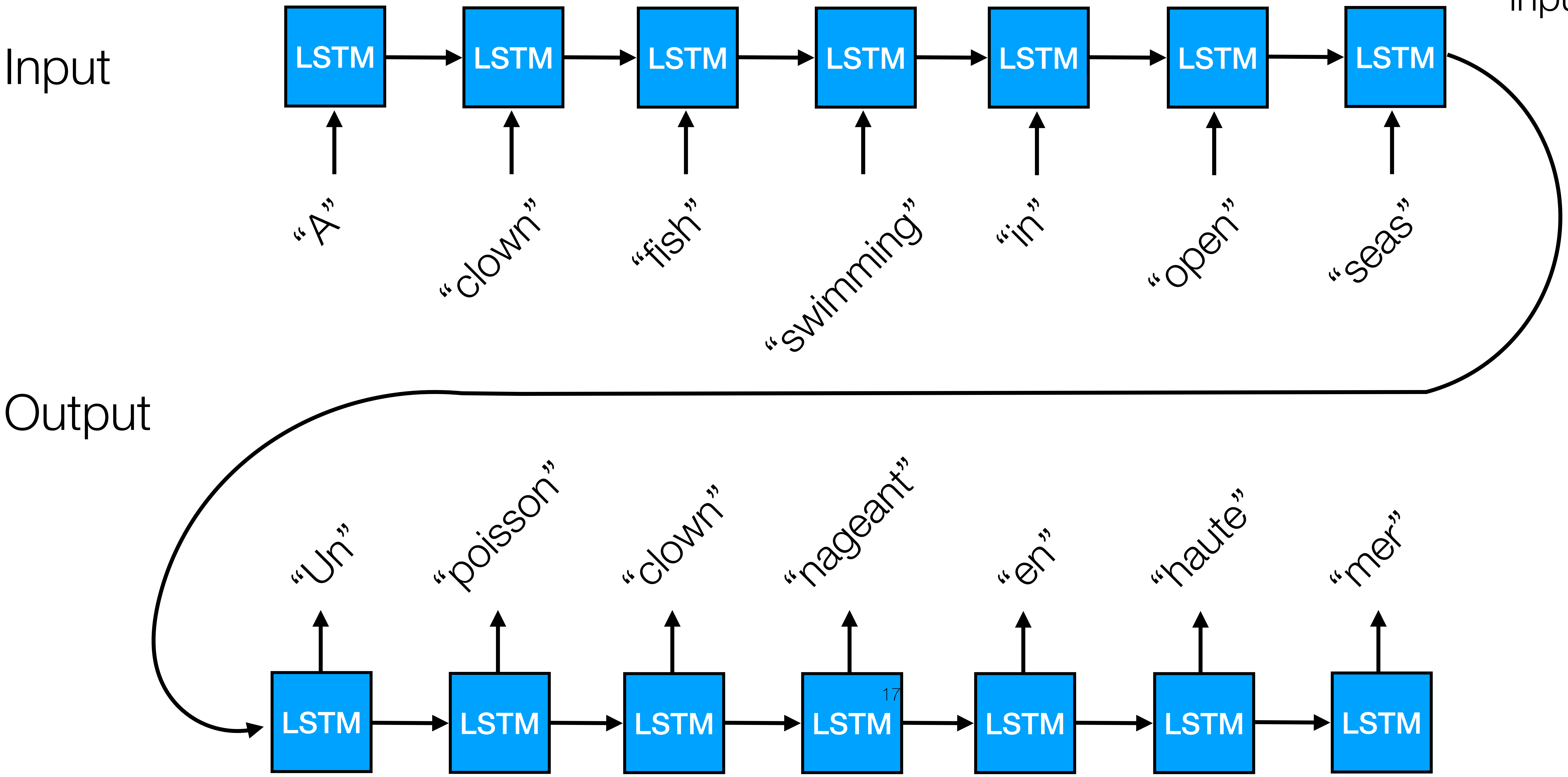
A person riding a motorcycle on a dirt road.


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

Shortcomings of recurrent models

- The recurrent state needs to **remember** a lot
- Instead of remembering: look at the input data! This idea is often implemented using **attention**.
- Example: “sequence to sequence” language translation

Example: Seq2Seq Translation



Hidden state remembers whole input sentence!

See [Sutskever et al., “Sequence to sequence”, 2014]

Pooling

Outputs
“Un”
“poisson”
“clown”
“nageant”
“en”
“haute”
“mer”
…
LSTM
LSTM
LSTM
LSTM
LSTM
LSTM
LSTM
Σ
Hidden
Input
“A”
“clown”
“fish”
“swimming”
“in”
“open”
“seas”

Source: Isola, Torralba, Freeman

Attention

Outputs
"Un"
"poisson"
"clown"
"nageant"
"en"
"haute"
"mer"
...
LSTM
LSTM
LSTM
LSTM
LSTM
LSTM
LSTM
Σ
Hidden
Input
"A"
"clown"
"fish"
"swimming"
"in"
"open"
"seas"

Source: Isola, Torralba, Freeman

Attention

Outputs
“Un”
“poisson”
“clown”
“nageant”
“en”
“haute”
“mer”
. . .
LSTM
LSTM
LSTM
LSTM
LSTM
LSTM
LSTM
Σ
Hidden
Input
“A”
“clown”
“fish”
“swimming”
“in”
“open”
“seas”

Source: Isola, Torralba, Freeman

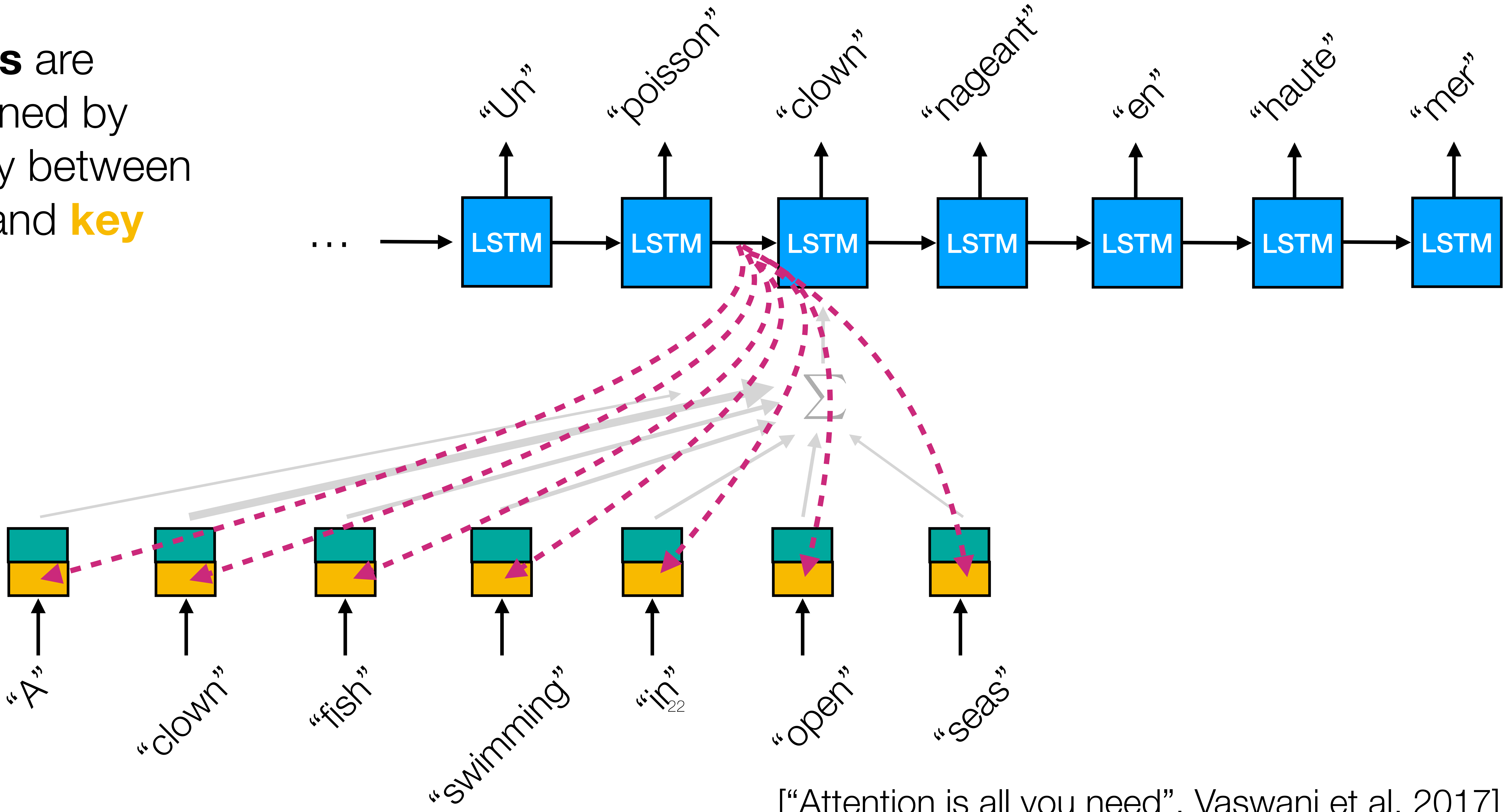
Attention

Outputs
“Un”
“poisson”
“clown”
“nageant”
“en”
“haute”
“mer”
. . .
LSTM
LSTM
LSTM
LSTM
LSTM
LSTM
LSTM
Σ
Hidden
Input
“A”
“clown”
“fish”
“swimming”
“in”
“open”
“seas”

Source: Isola, Torralba, Freeman

Attention

Weights are determined by similarity between **query** and **key**

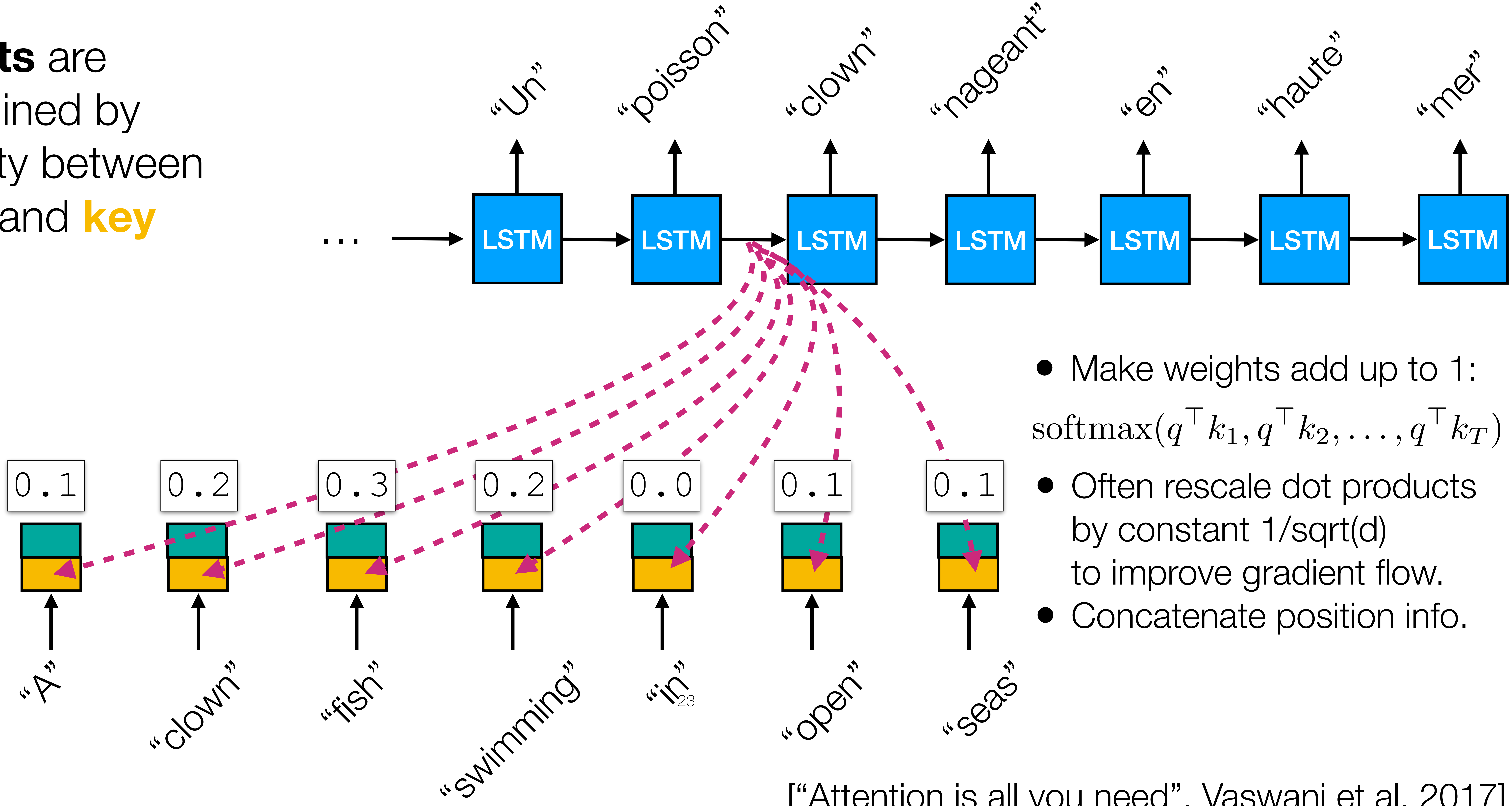


[“Attention is all you need”, Vaswani et al. 2017]

Source: Isola, Torralba, Freeman

Attention

Weights are determined by similarity between **query** and **key**



- Make weights add up to 1:

$$\mathrm{softmax}(q^\top k_1, q^\top k_2, \ldots, q^\top k_T)$$

- Often rescale dot products by constant 1/sqrt(d) to improve gradient flow.
- Concatenate position info.

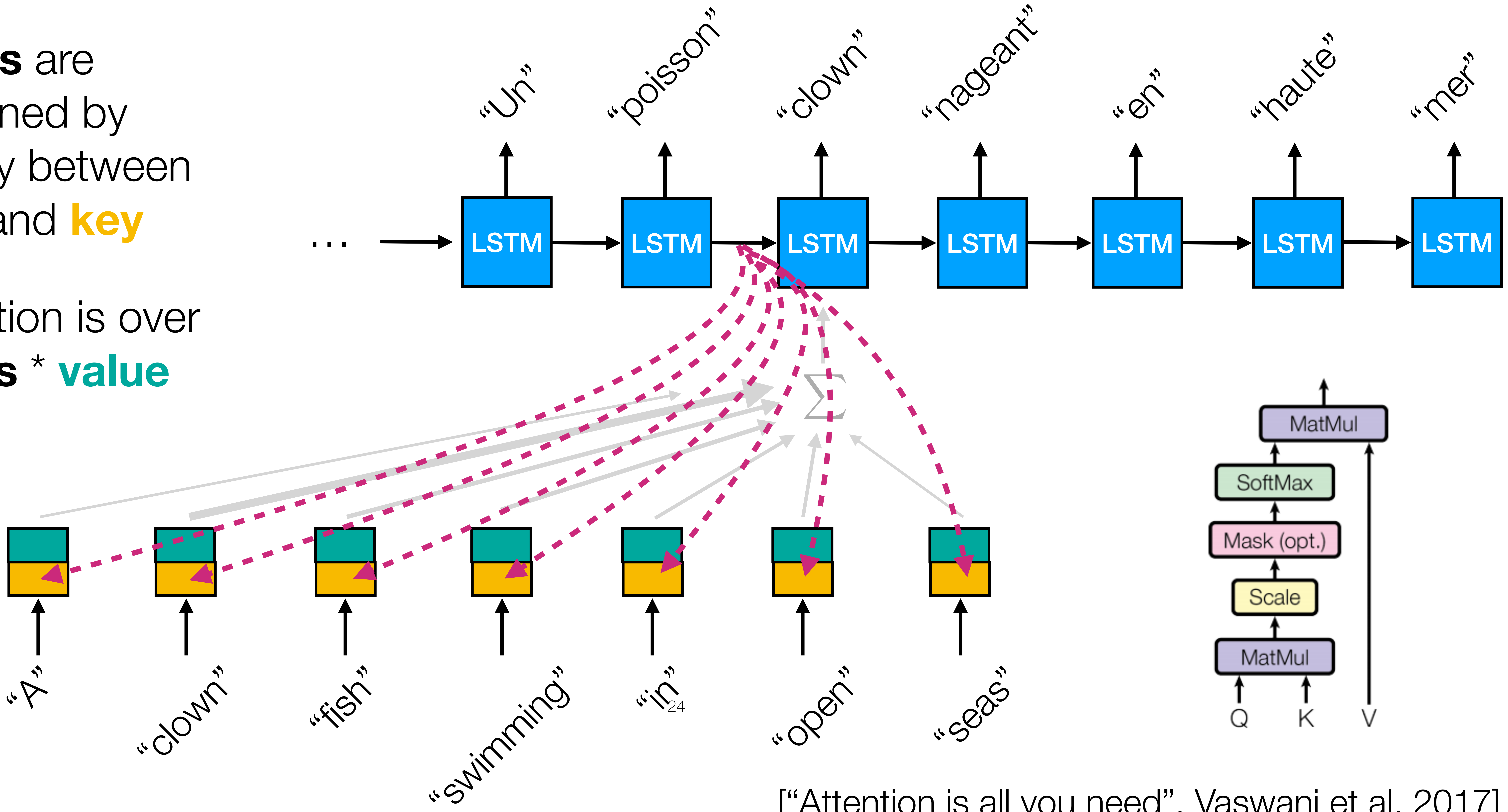
["Attention is all you need", Vaswani et al. 2017]

Source: Isola, Torralba, Freeman

Attention

Weights are determined by similarity between **query** and **key**

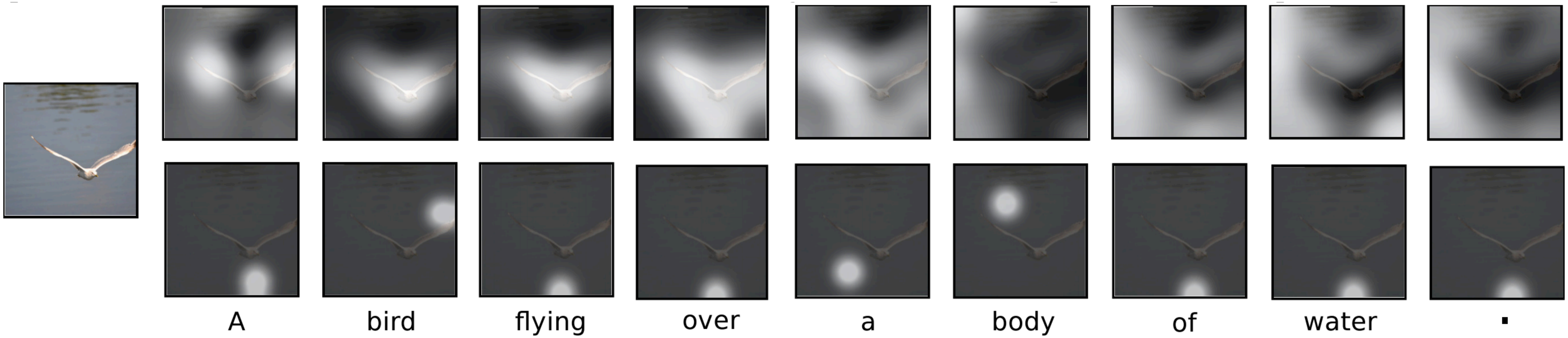
summation is over **weights** * **value**



[“Attention is all you need”, Vaswani et al. 2017]

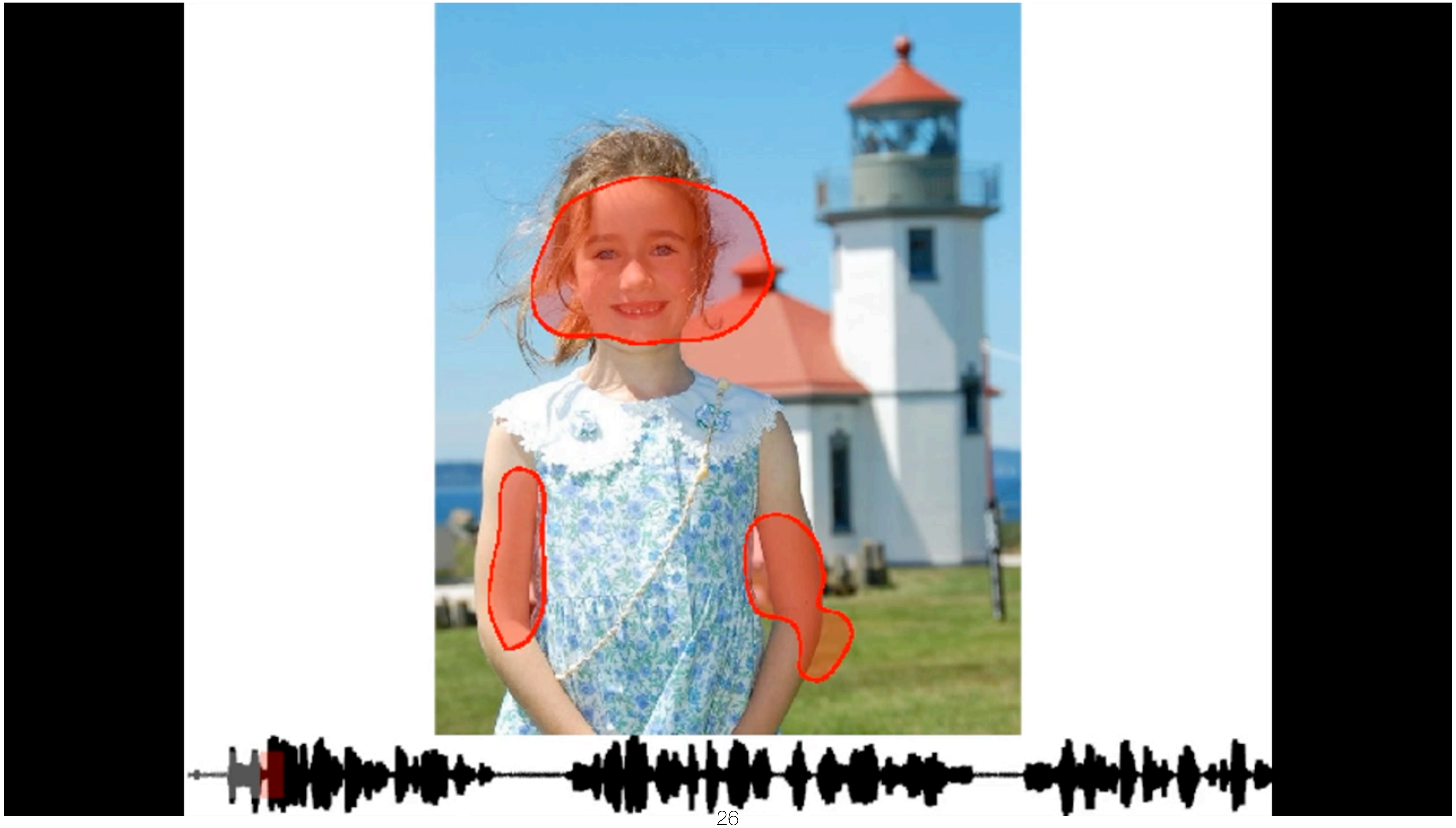
Source: Isola, Torralba, Freeman

Image captioning with attention



[Xu et al. “Show, Attend and Tell: Neural Image Caption Generation with Visual Attention”, 2016]

Discovering words in raw audio-visual data

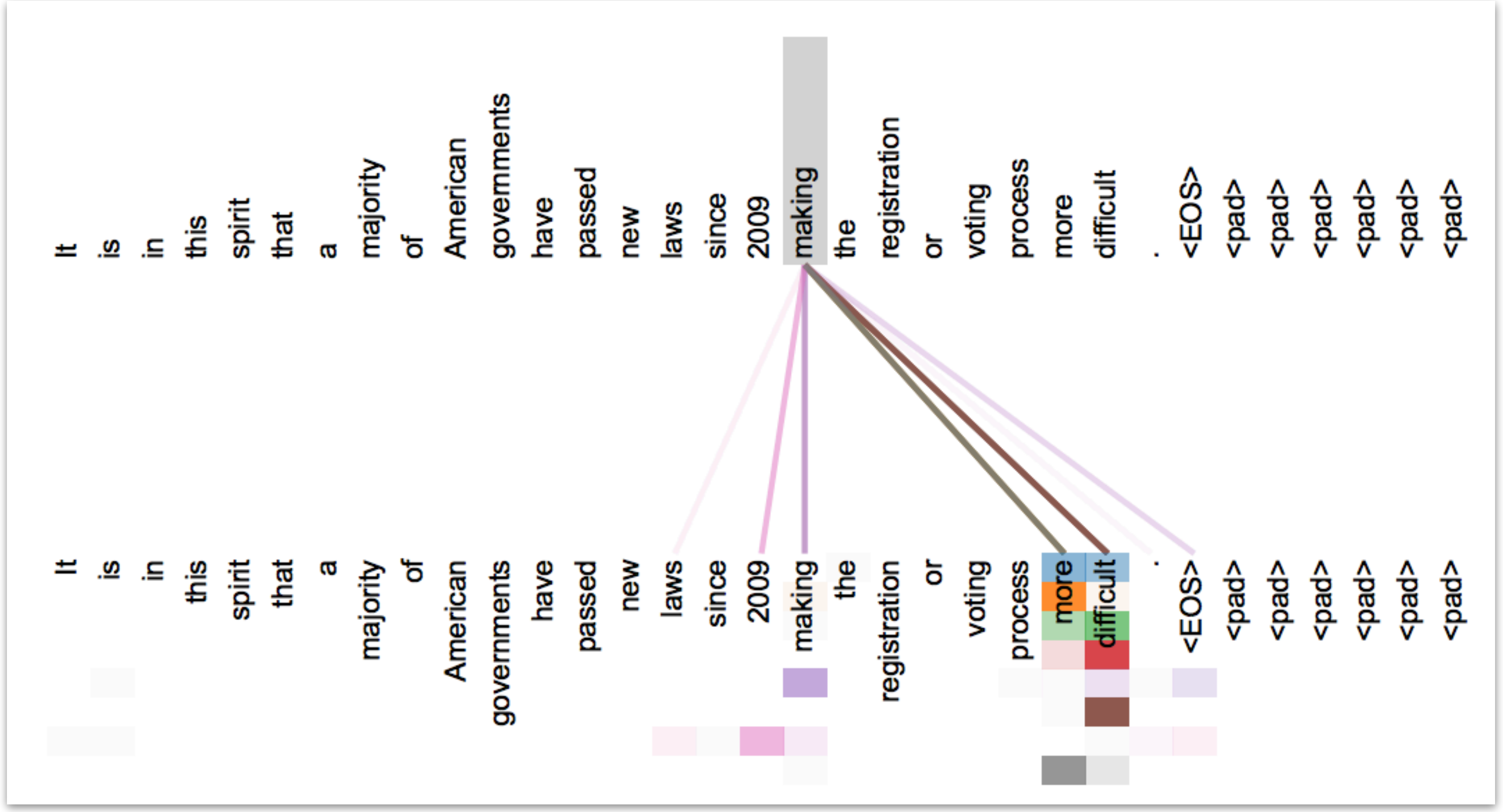


https://link.springer.com/chapter/10.1007%2F978-3-030-01231-1_40

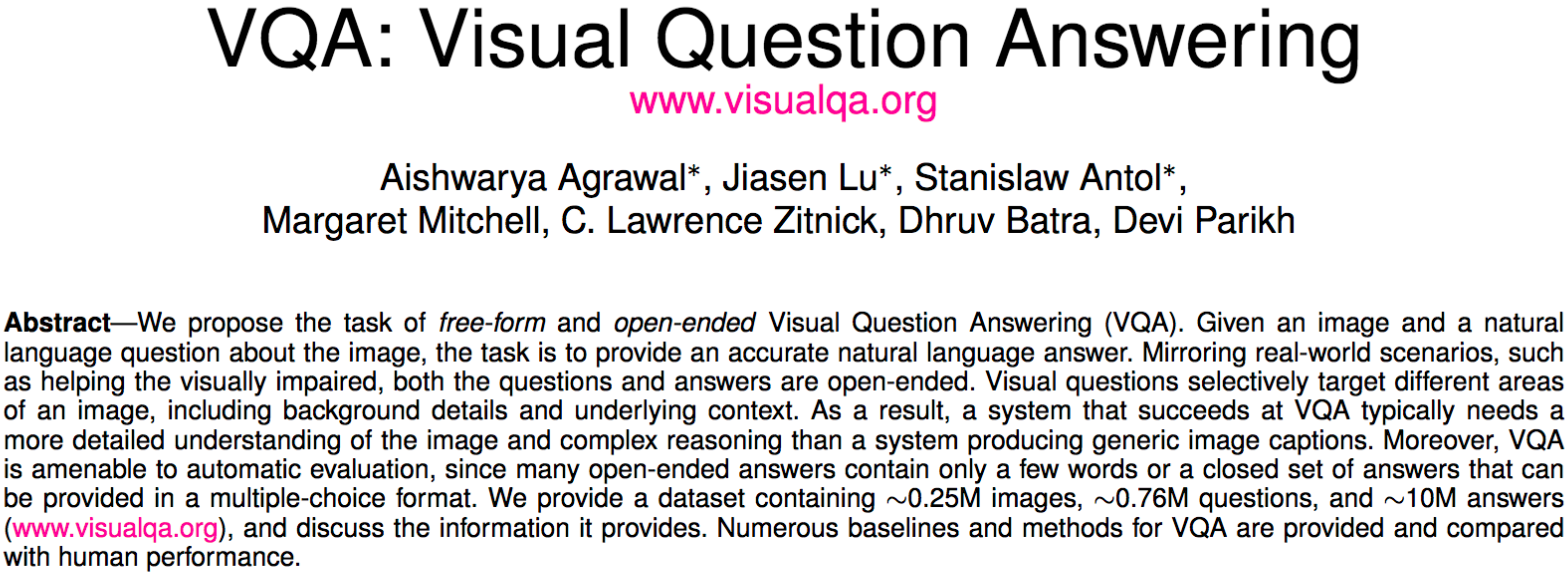
[Harwath et al. “Jointly Discovering Visual Objects and Spoken Words from Raw Sensory Input”, 2018]

Transformers

- Get rid of the recurrent net!
- Just stack many layers of attention.
- Use multiple keys/values per layer.
- Powerful model for natural language processing. Used pretty much everywhere now…


It is in this spirit that a majority of American governments have passed new laws since 2009 making the registration or voting process more difficult . <EOS> <pad> <pad> <pad> <pad> <pad> <pad>
It is in this spirit that a majority of American governments have passed new laws since 2009 making the registration or voting process more difficult . <EOS> <pad> <pad> <pad> <pad> <pad> <pad>

[“Attention is all you need”, Vaswani et al. 2017]



VQA: Visual Question Answering

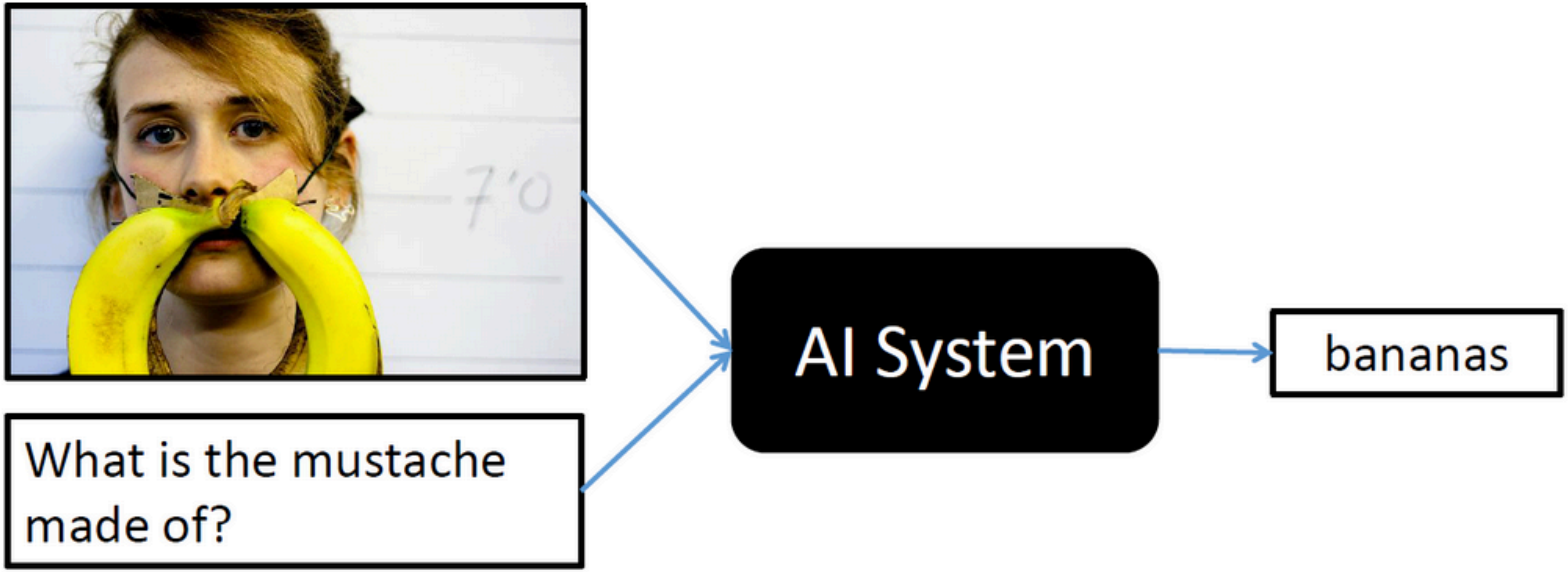
www.visualqa.org

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]

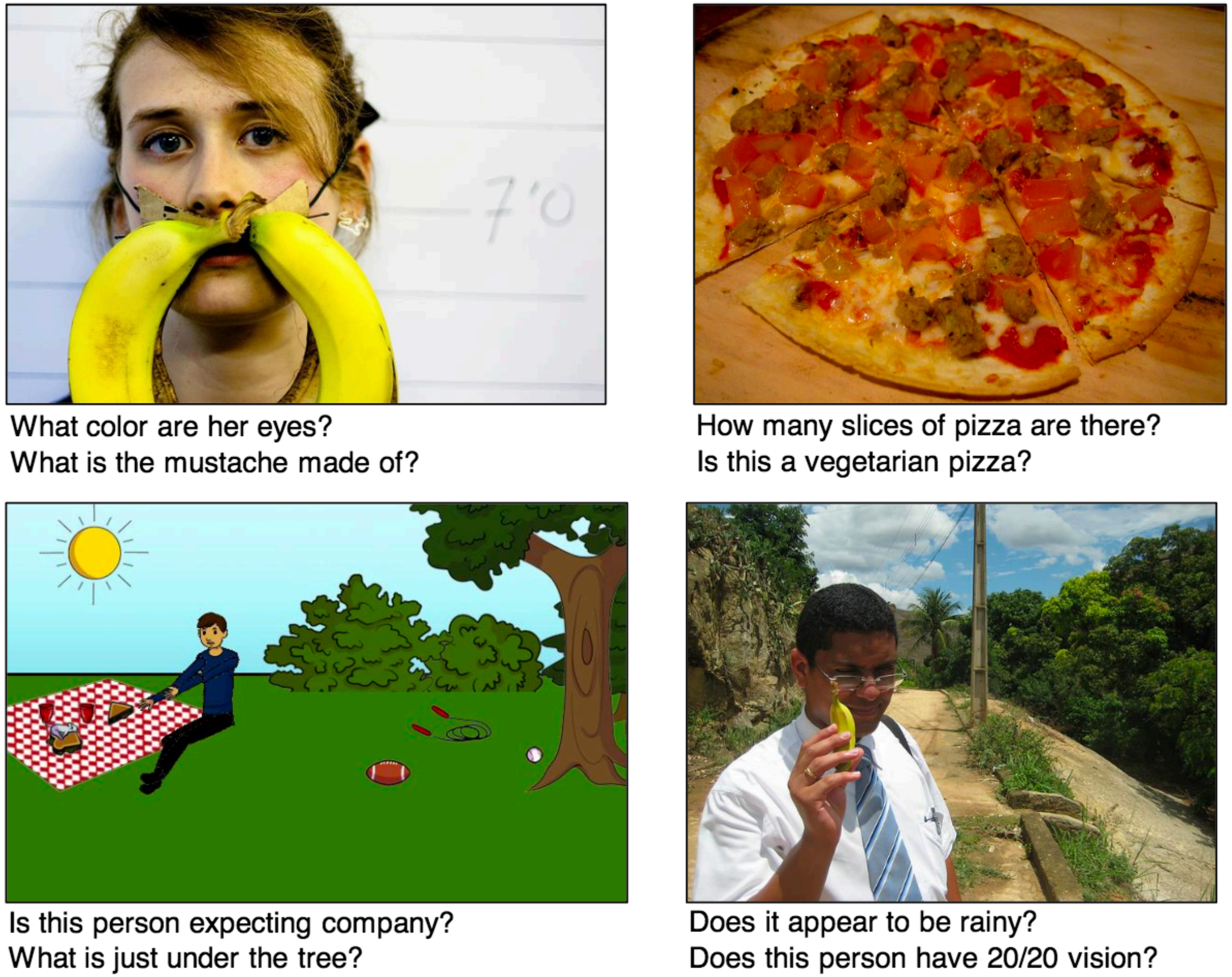


What is the mustache made of?

AI System

bananas

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



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

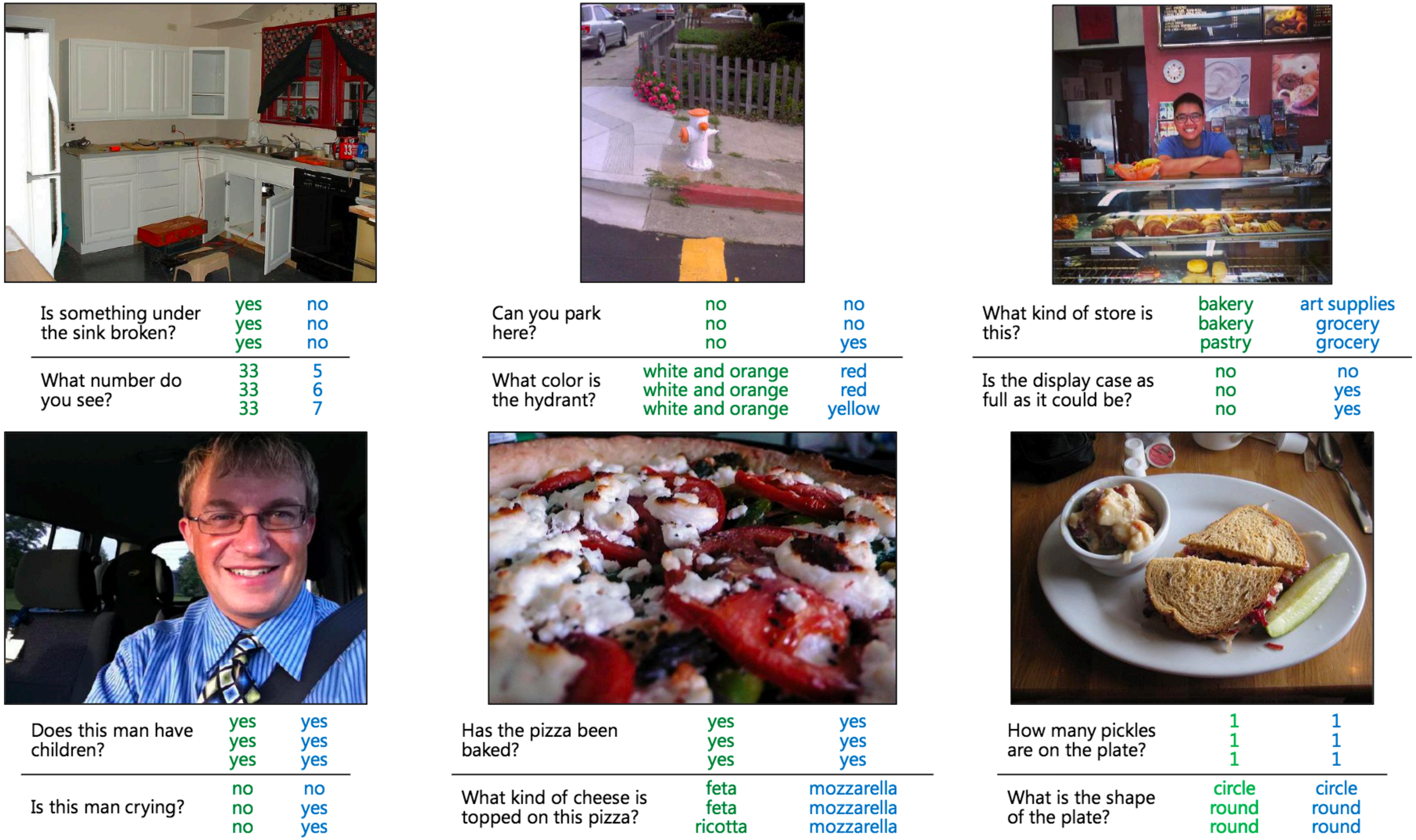
How many slices of pizza are there?
Is this a vegetarian pizza?

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.

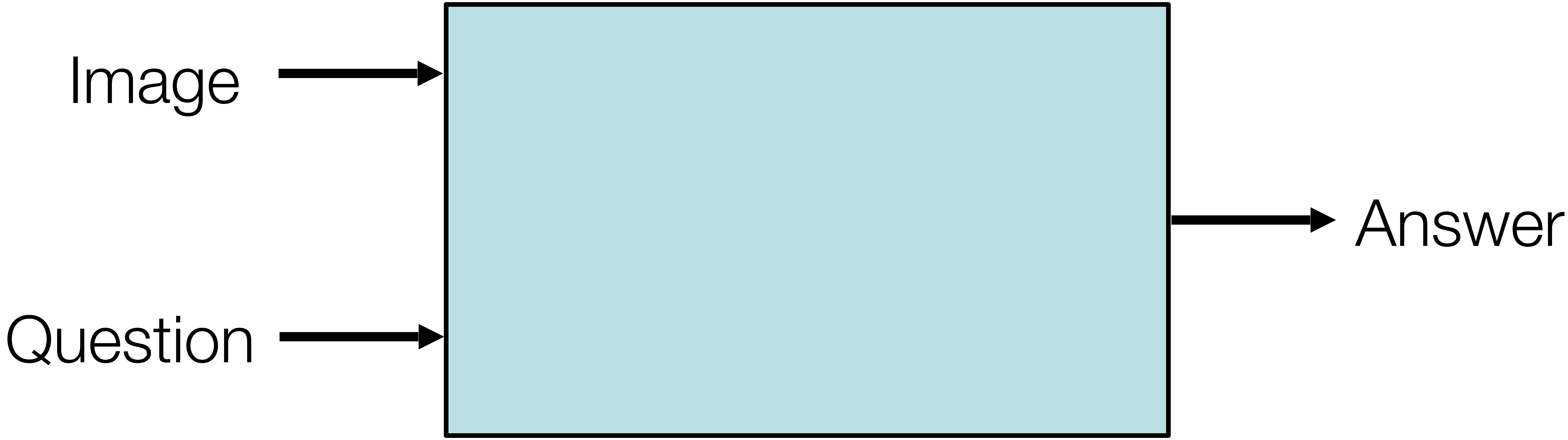
Questions and answers collected with Amazon Mechanical Turk



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

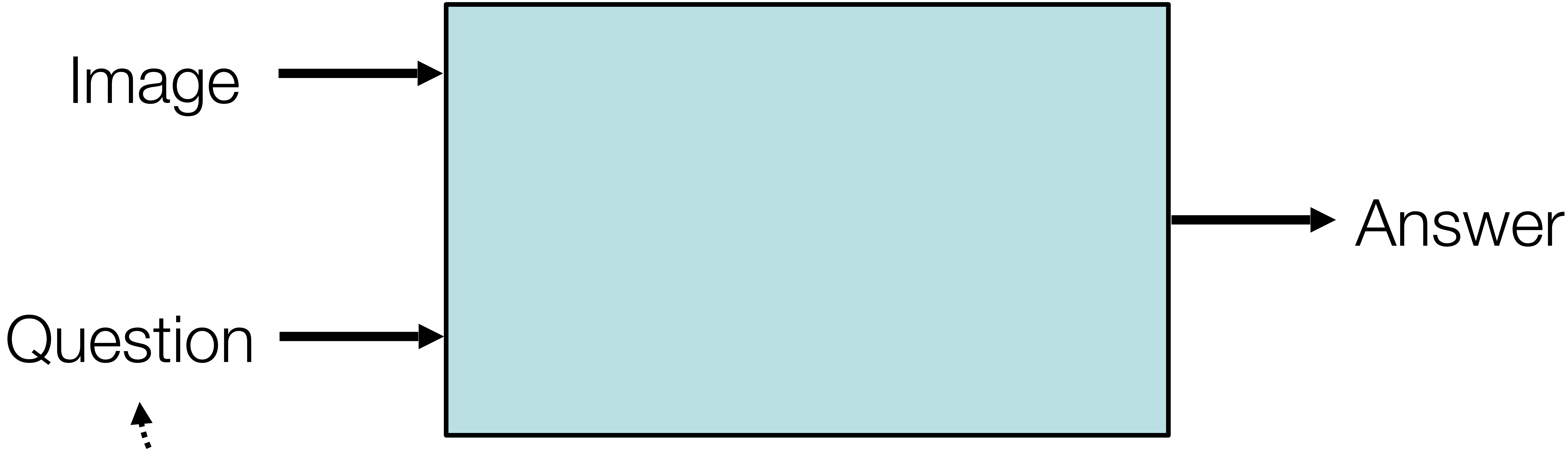
Fig. 2: Examples of questions (black), (a subset of the) answers given when looking at the image (green), and answers given when not looking at the image (blue) for numerous representative examples of the dataset. See the appendix for more examples.

Architecture



Source: Isola, Torralba, Freeman

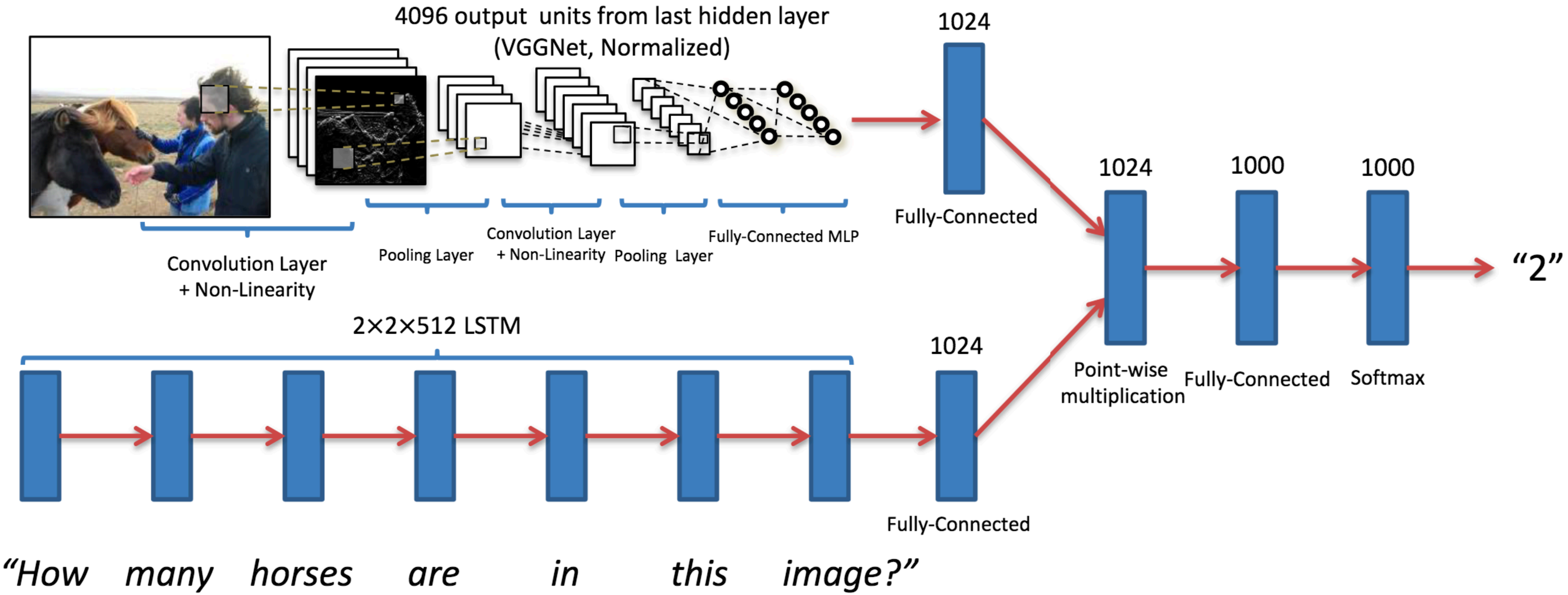
Architecture



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

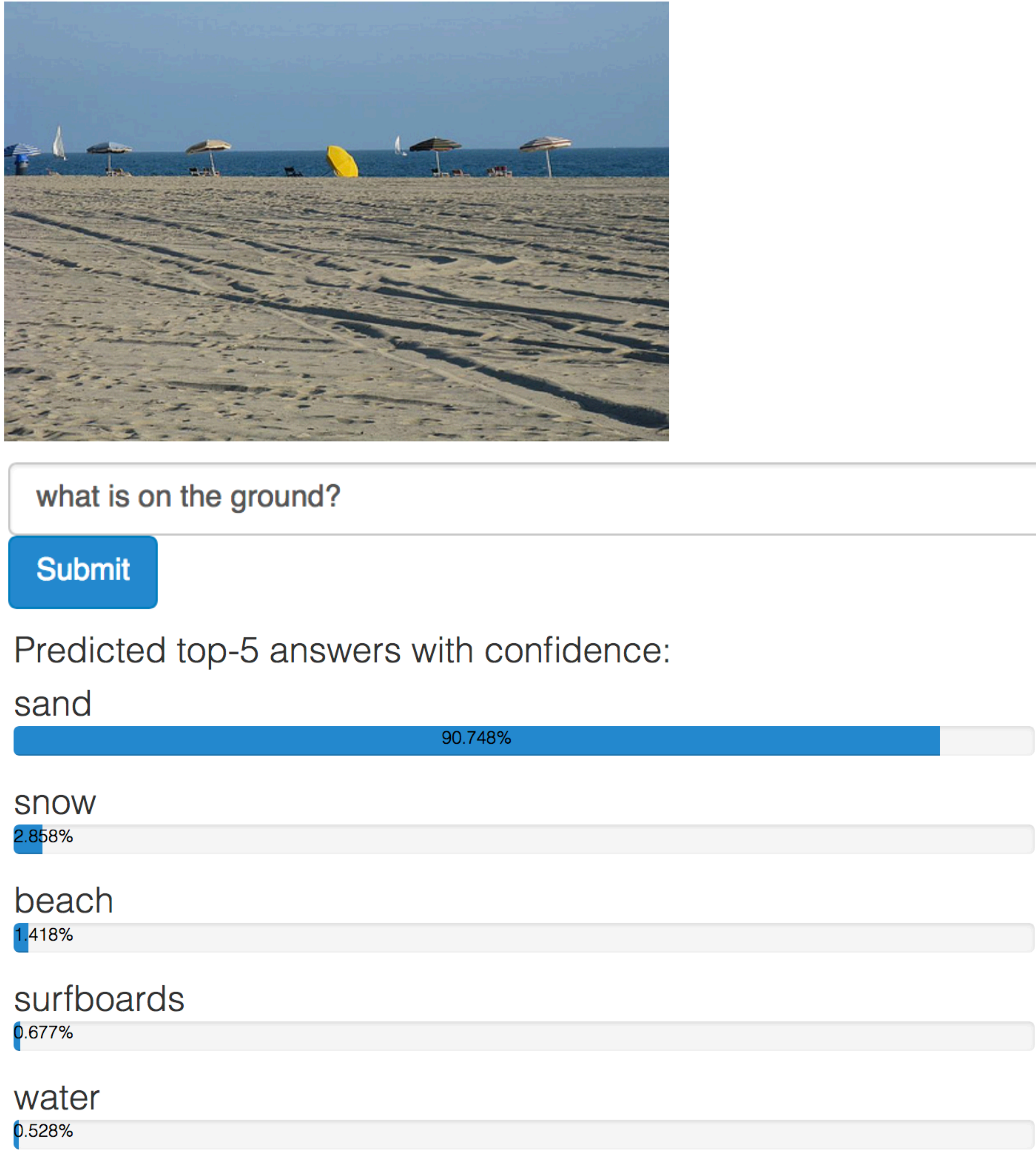
Source: Isola, Torralba, Freeman

Architecture



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

[Agrawal et al., "VQA: Visual Question Answering" 2016]



what is on the ground?

Predicted top-5 answers with confidence:

sand

90.748%

snow

2.858%

beach

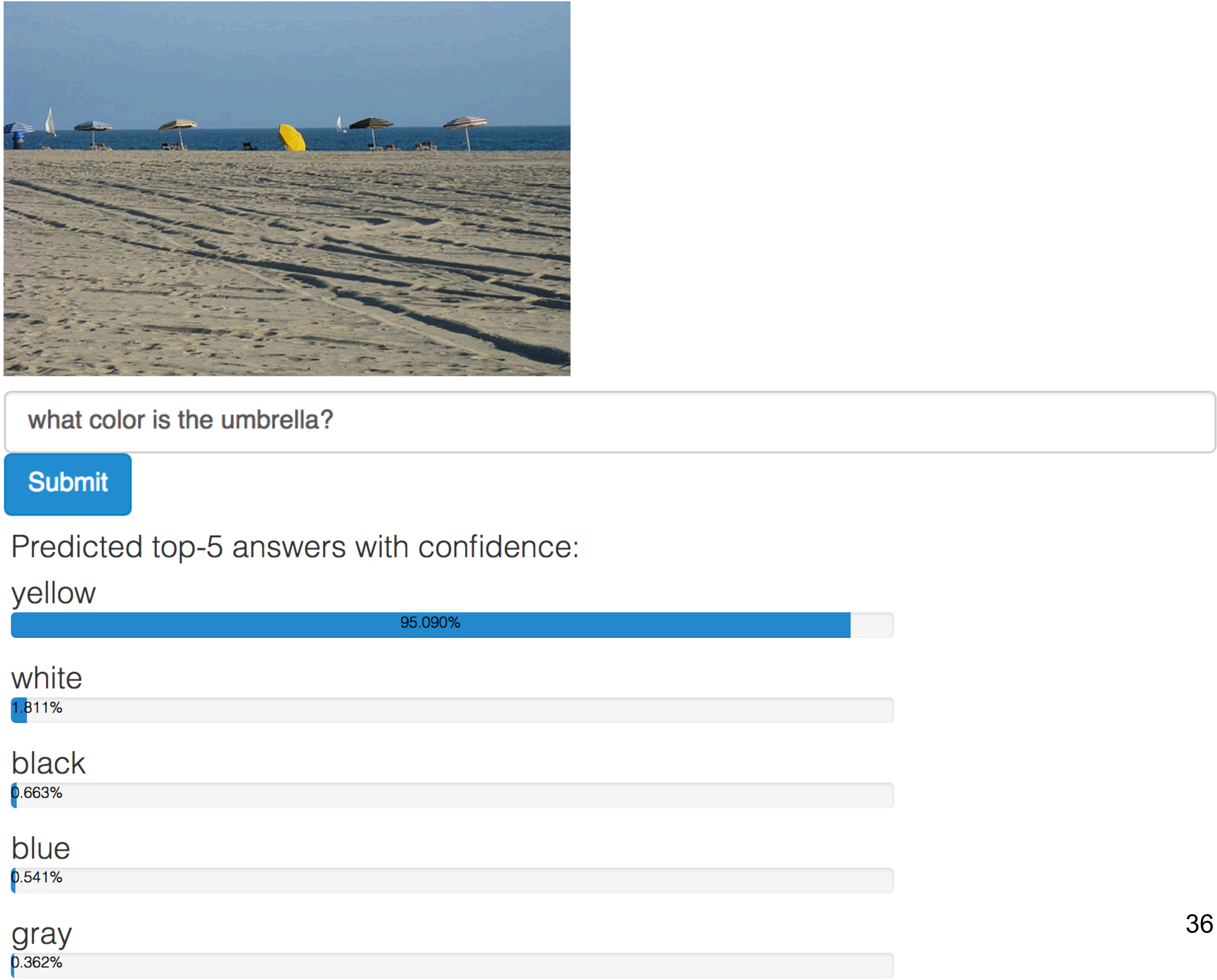
1.418%

surfboards

0.677%

water

0.528%



what color is the umbrella?

Predicted top-5 answers with confidence:

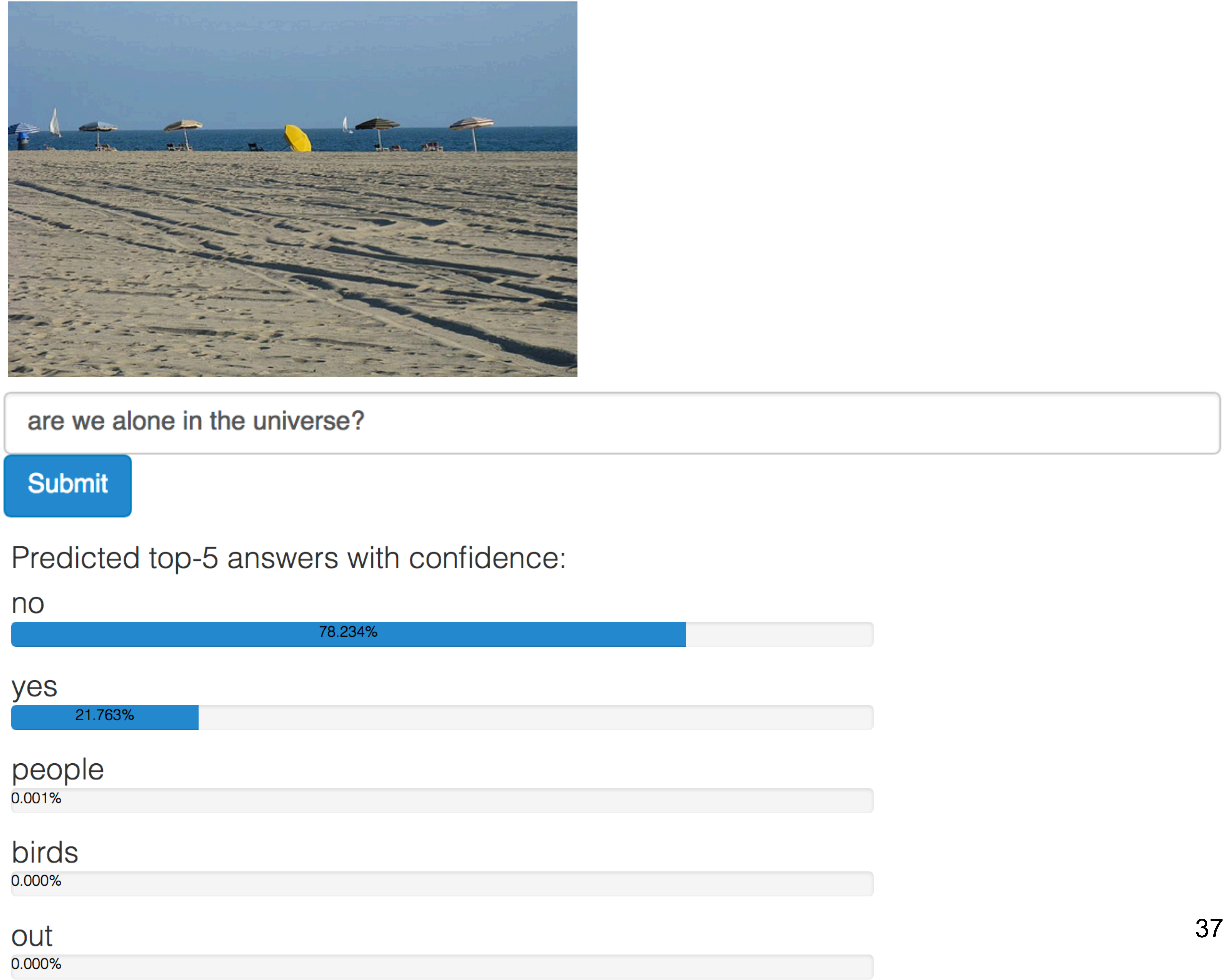
yellow
95.090%

white
1.811%

black
0.663%

blue
0.541%

gray
0.362%



are we alone in the universe?

Predicted top-5 answers with confidence:

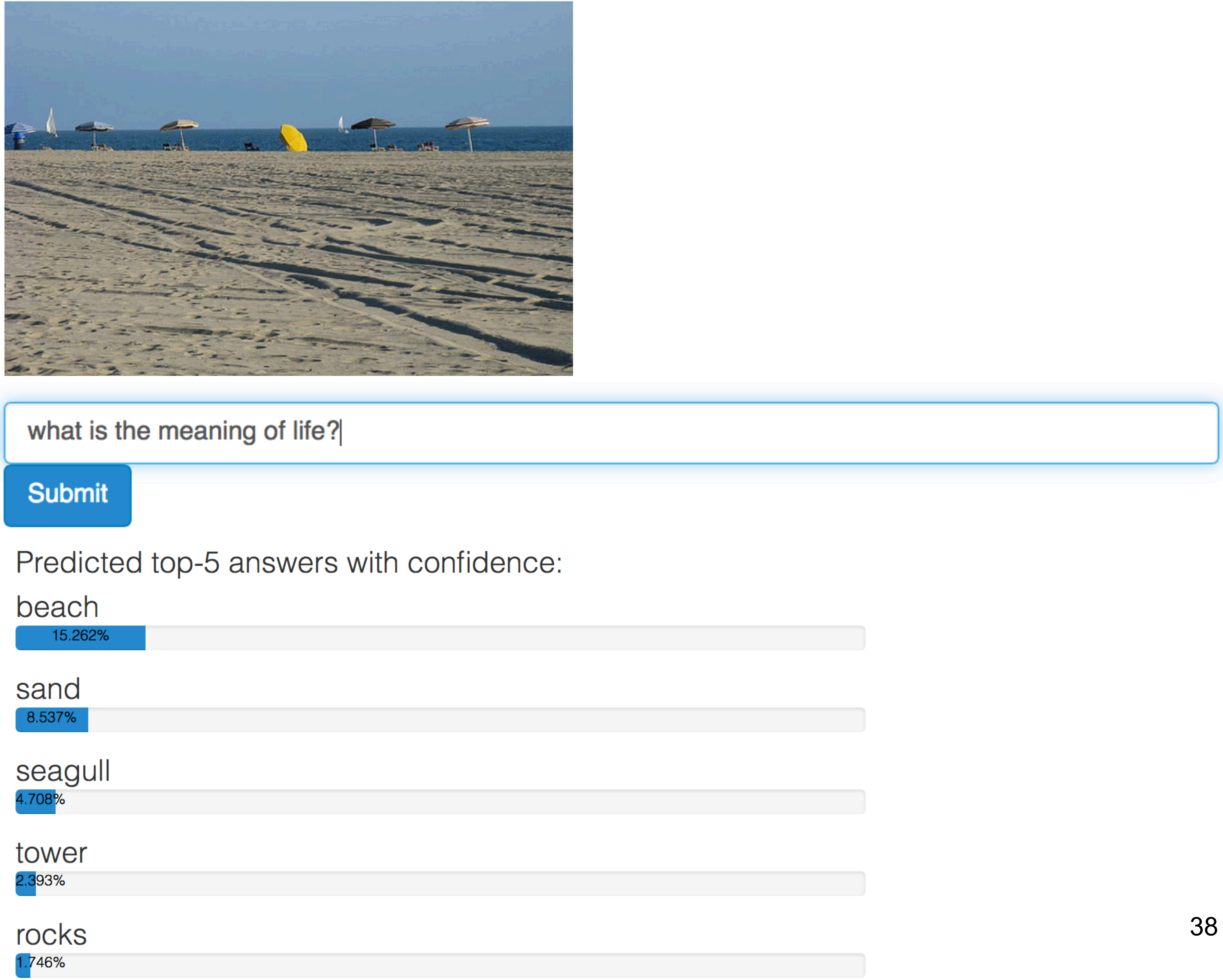
no
78.234%

yes
21.763%

people
0.001%

birds
0.000%

out
0.000%



what is the meaning of life?

Predicted top-5 answers with confidence:

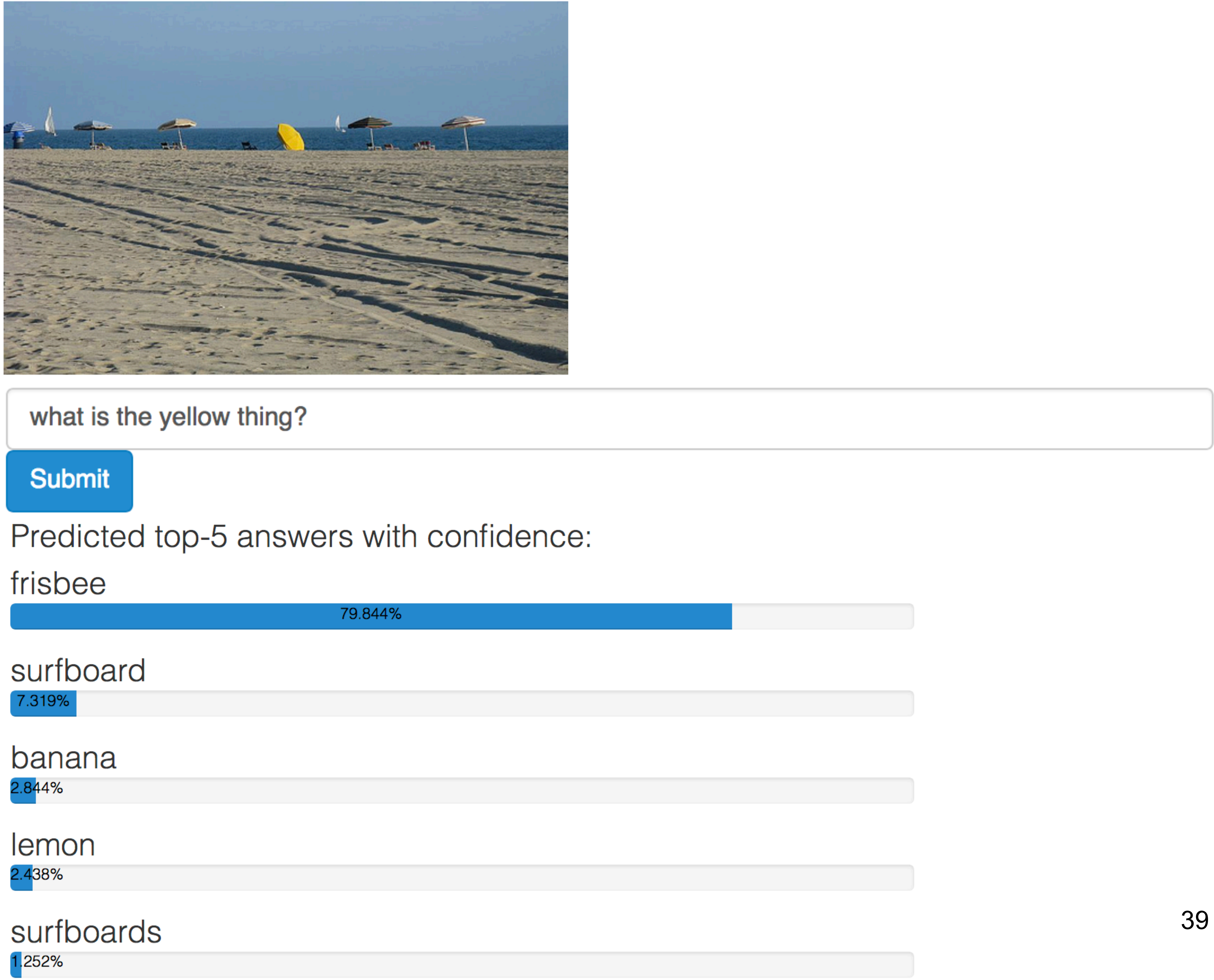
beach
15.262%

sand
8.537%

seagull
4.708%

tower
2.393%

rocks
1.746%



what is the yellow thing?

Predicted top-5 answers with confidence:

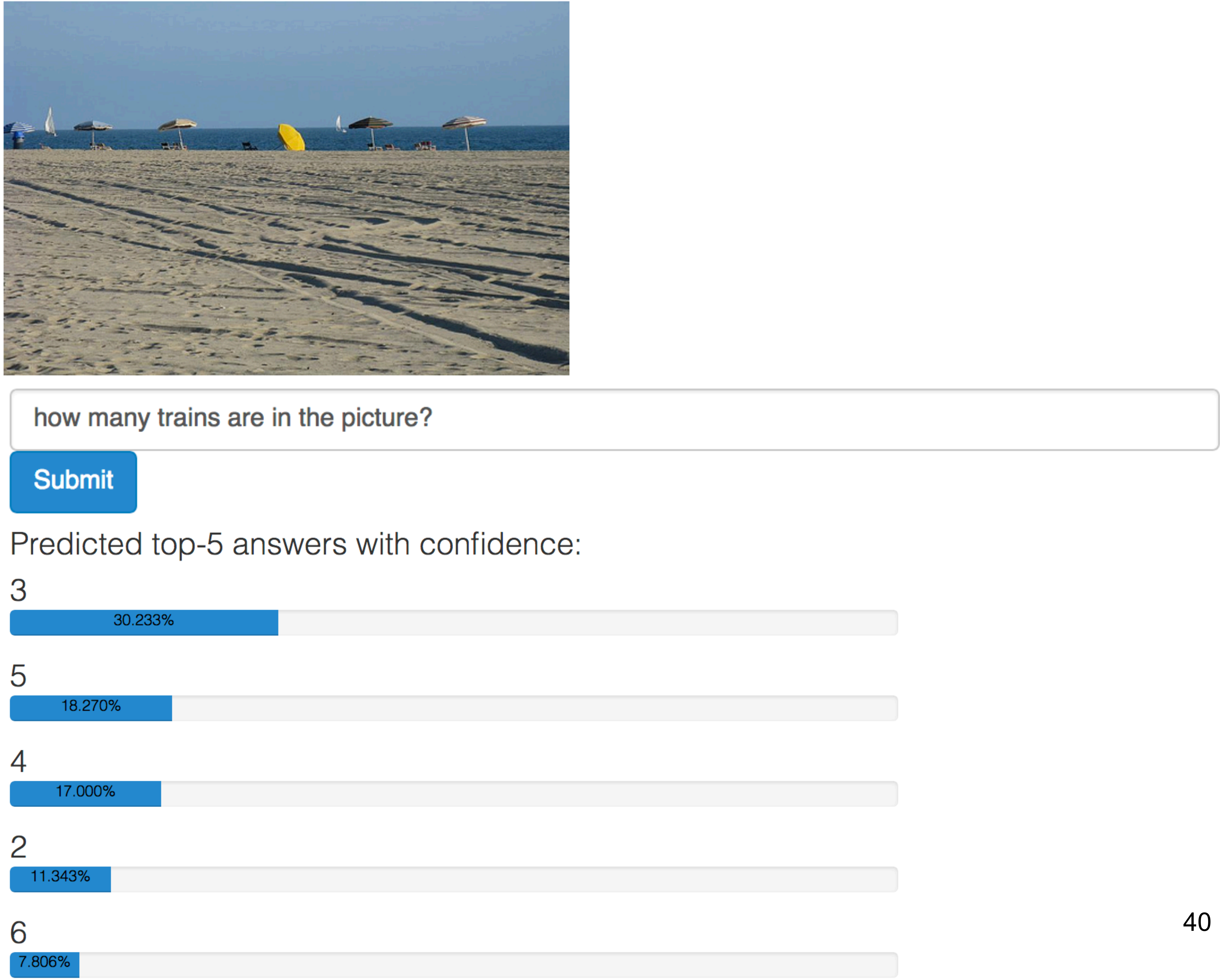
frisbee
79.844%

surfboard
7.319%

banana
2.844%

lemon
2.438%

surfboards
1.252%



how many trains are in the picture?

Predicted top-5 answers with confidence:

3

30.233%

5

18.270%

4

17.000%

2

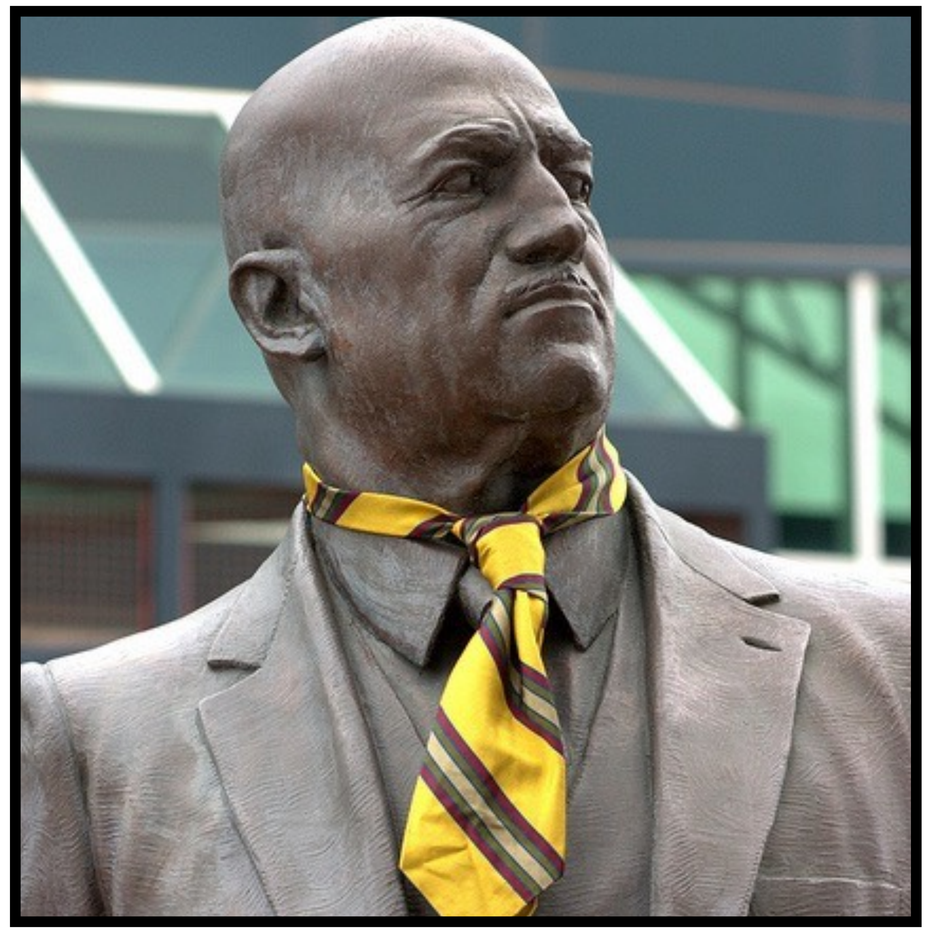
11.343%

6

7.806%

Grounded question answering

What color is the necktie?

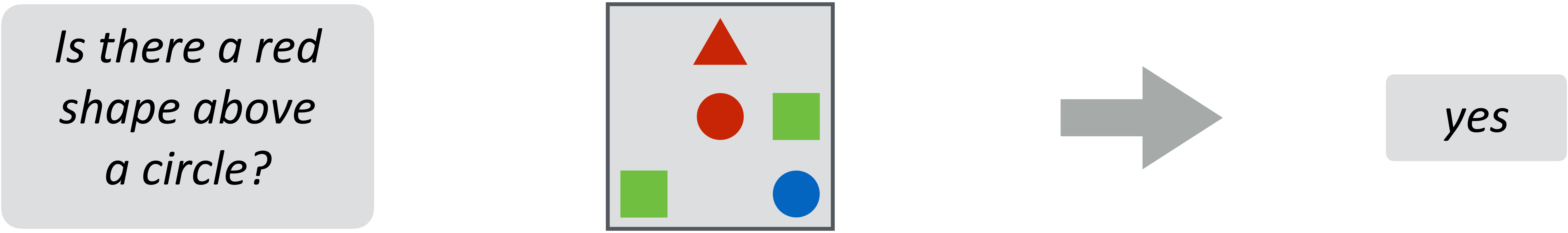


yellow

Neural module networks: a compositional language-understanding model

[Slides credit: Jacob Andreas]

Grounded question answering

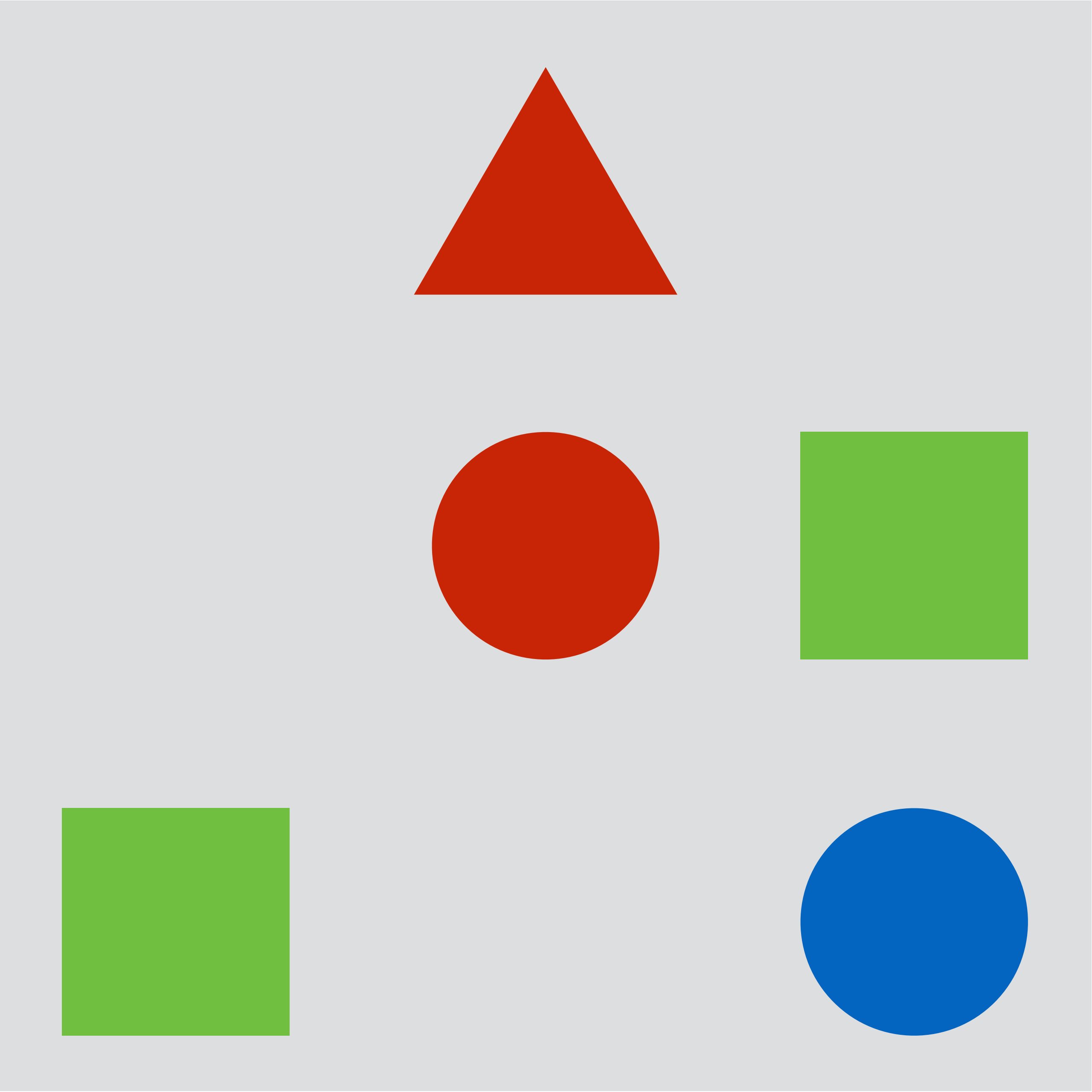


Is there a red shape above a circle?

yes

[Slides credit: Jacob Andreas]

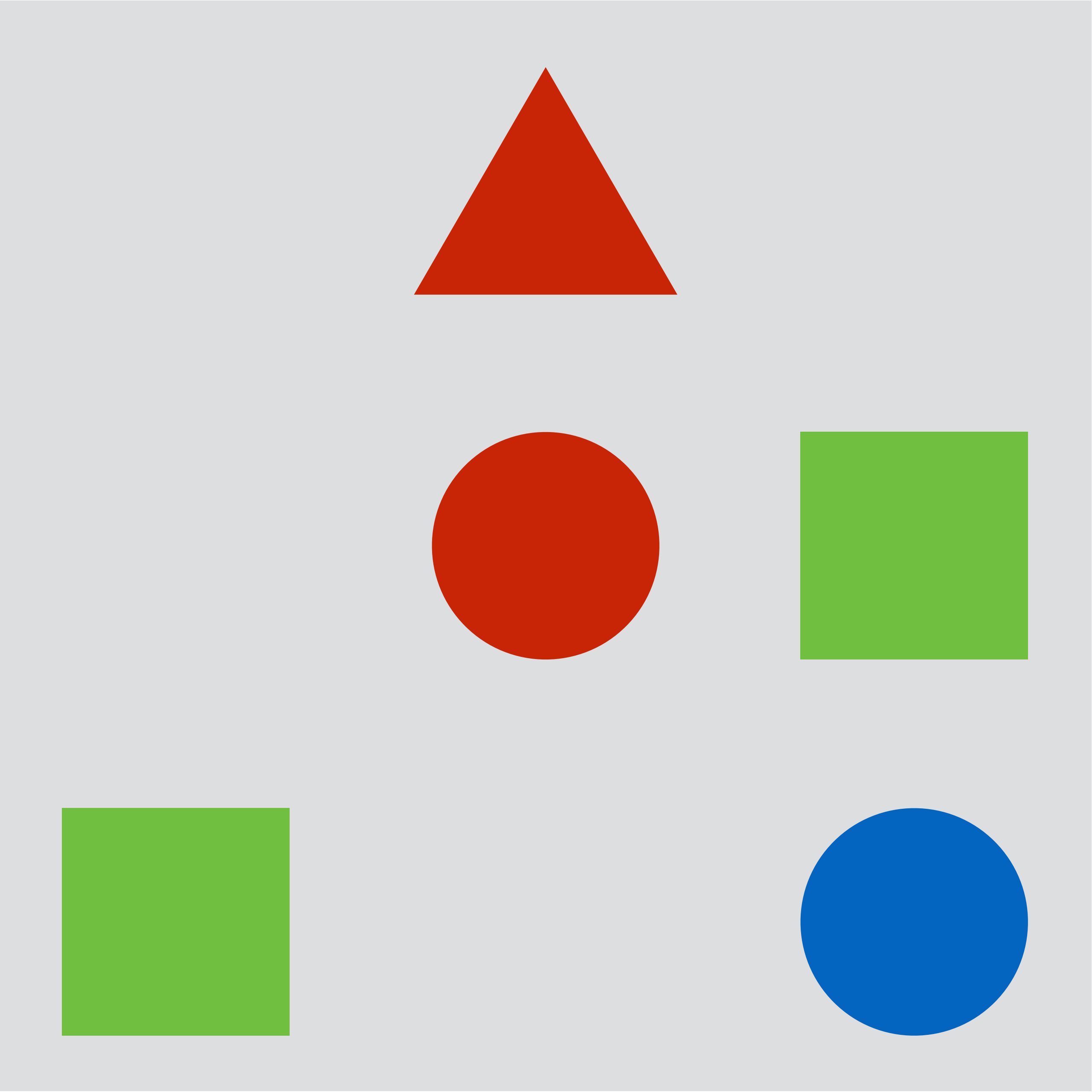
Representing meaning



Is there a red shape above a circle?

[Slides credit: Jacob Andreas]

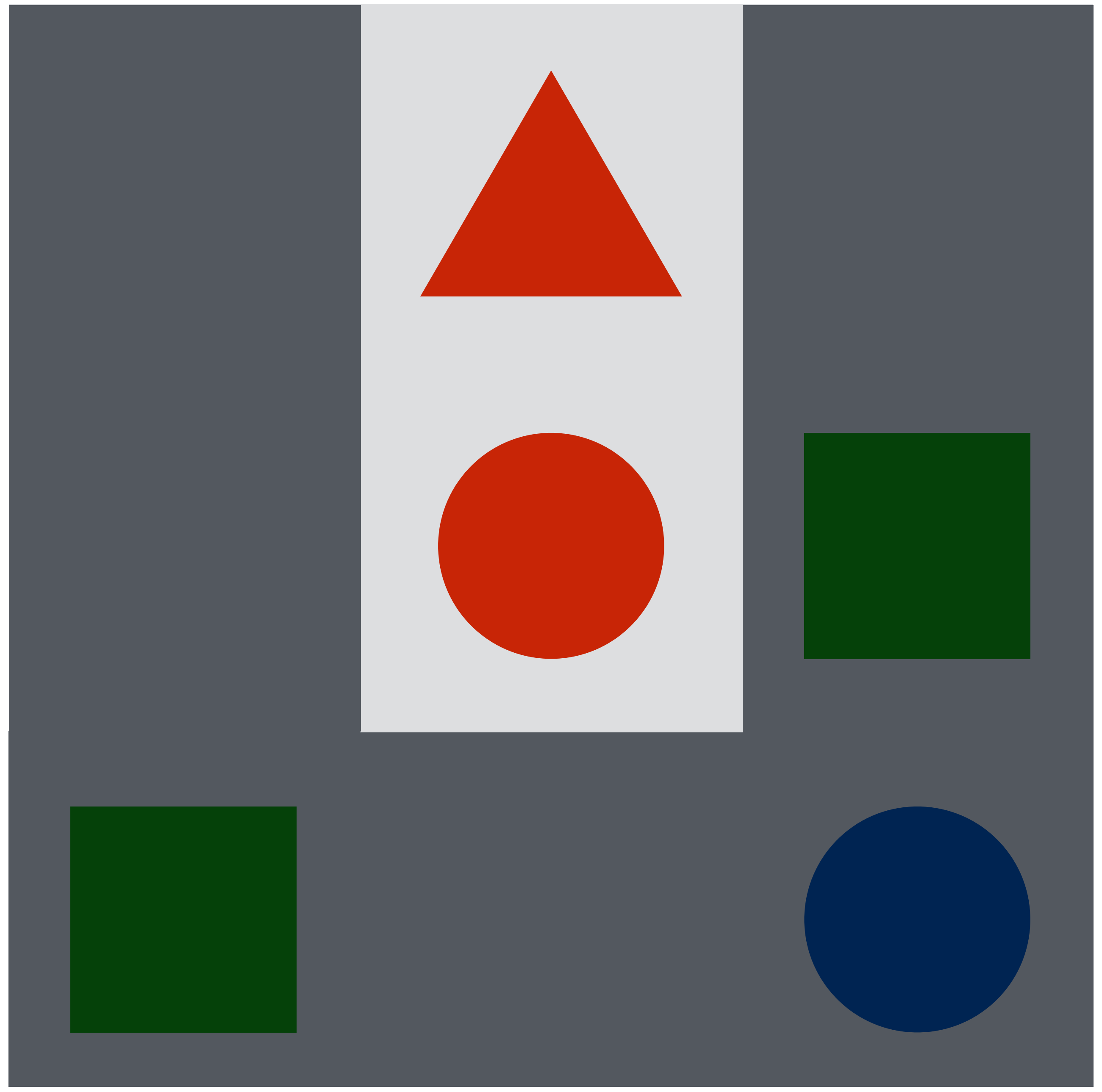
Representing meaning



Is there a red shape above a circle?

[Slides credit: Jacob Andreas]

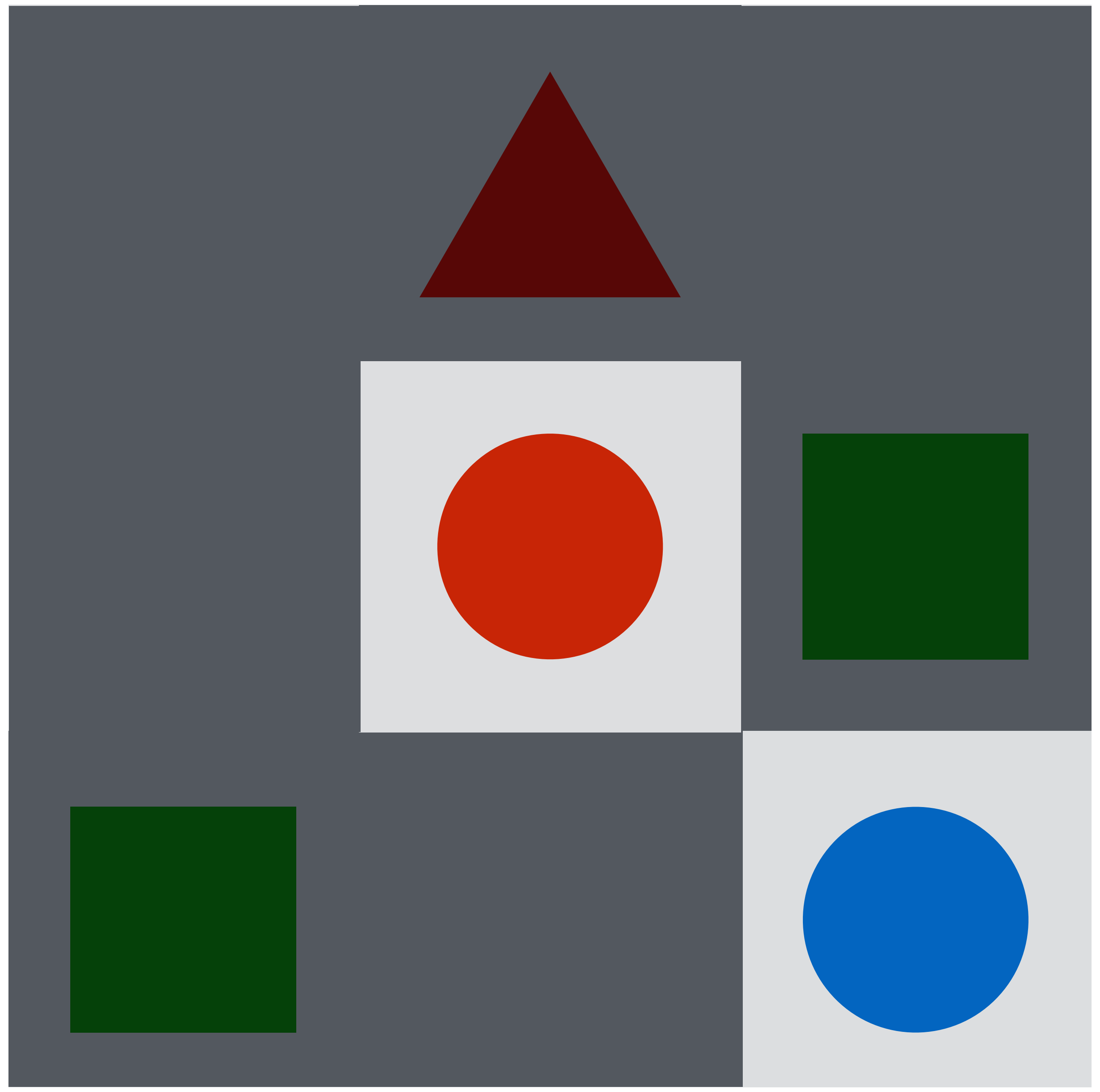
Sets encode meaning



Is there a red shape above a circle?

[Slides credit: Jacob Andreas]

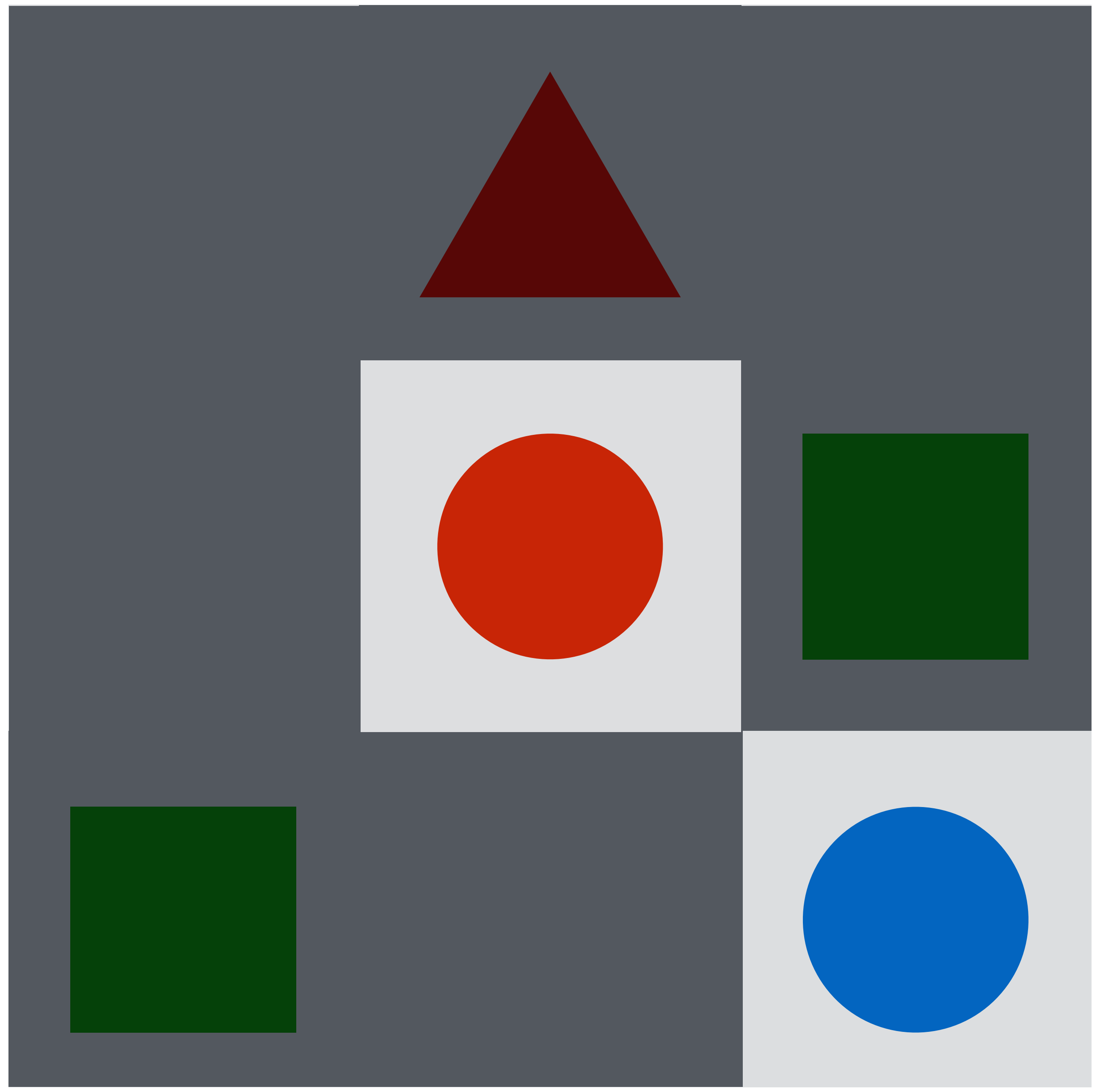
Sets encode meaning



Is there a red shape above a circle?

[Slides credit: Jacob Andreas]

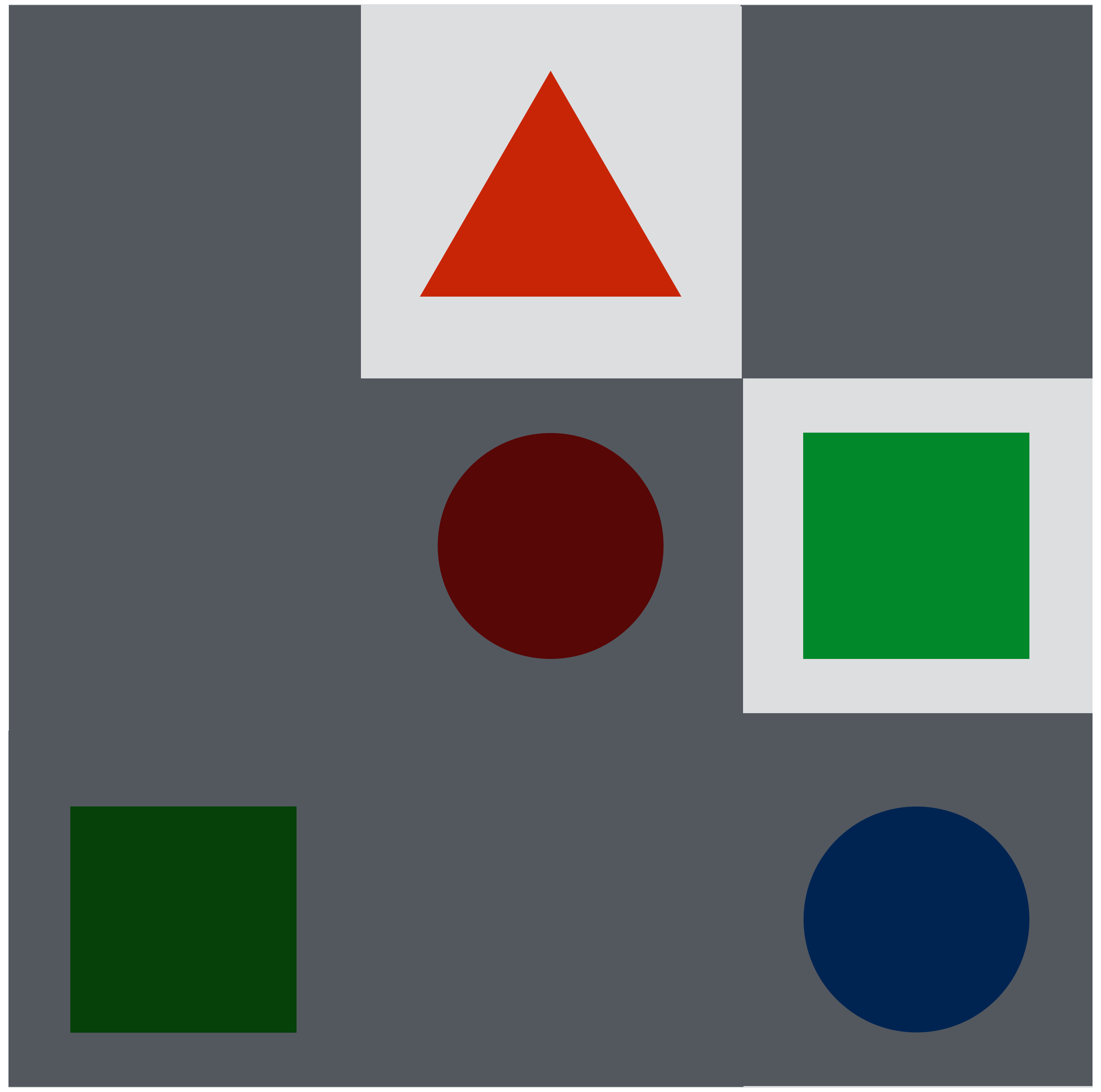
Set transformations encode meaning



Is there a red shape above a circle?

[Slides credit: Jacob Andreas]

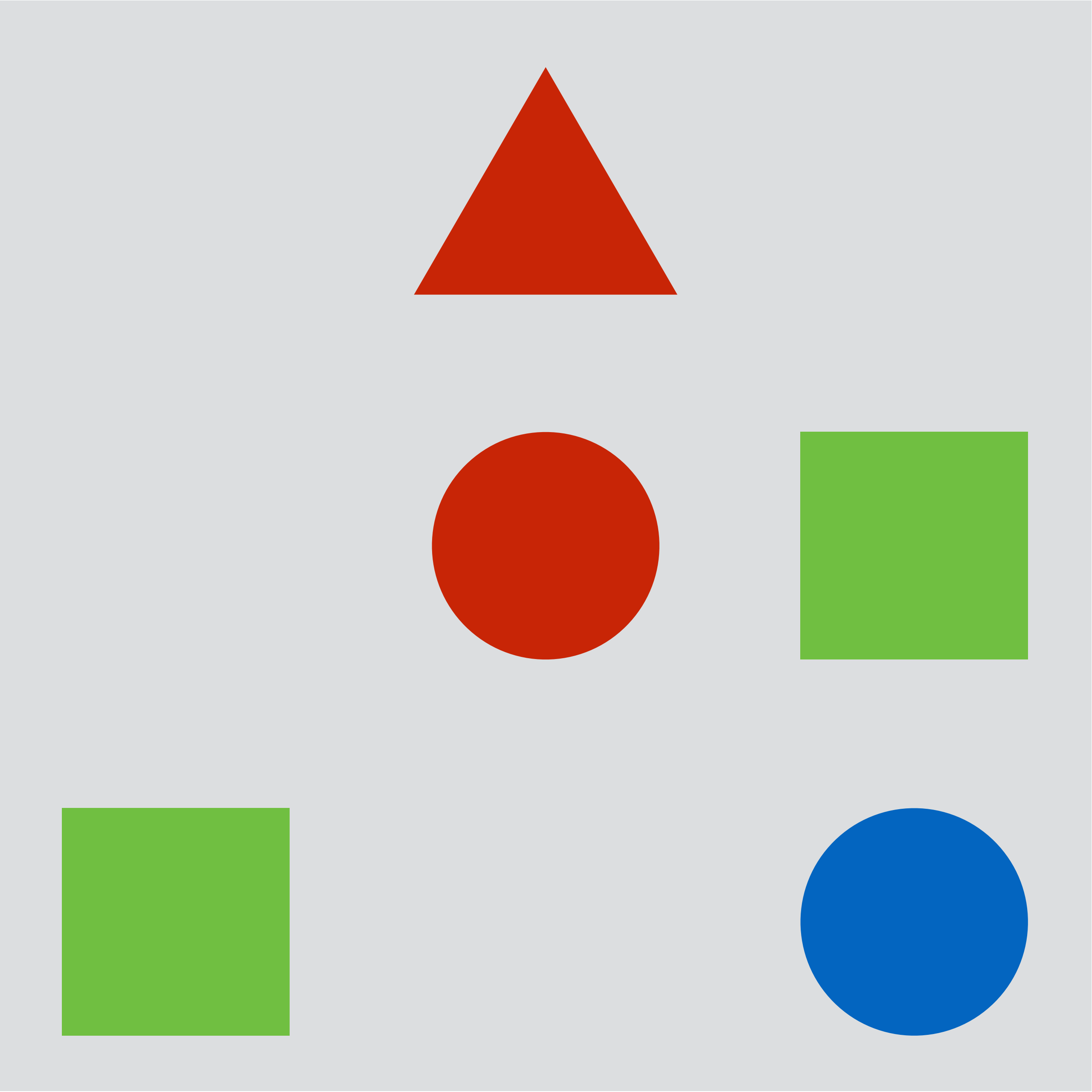
Set transformations encode meaning



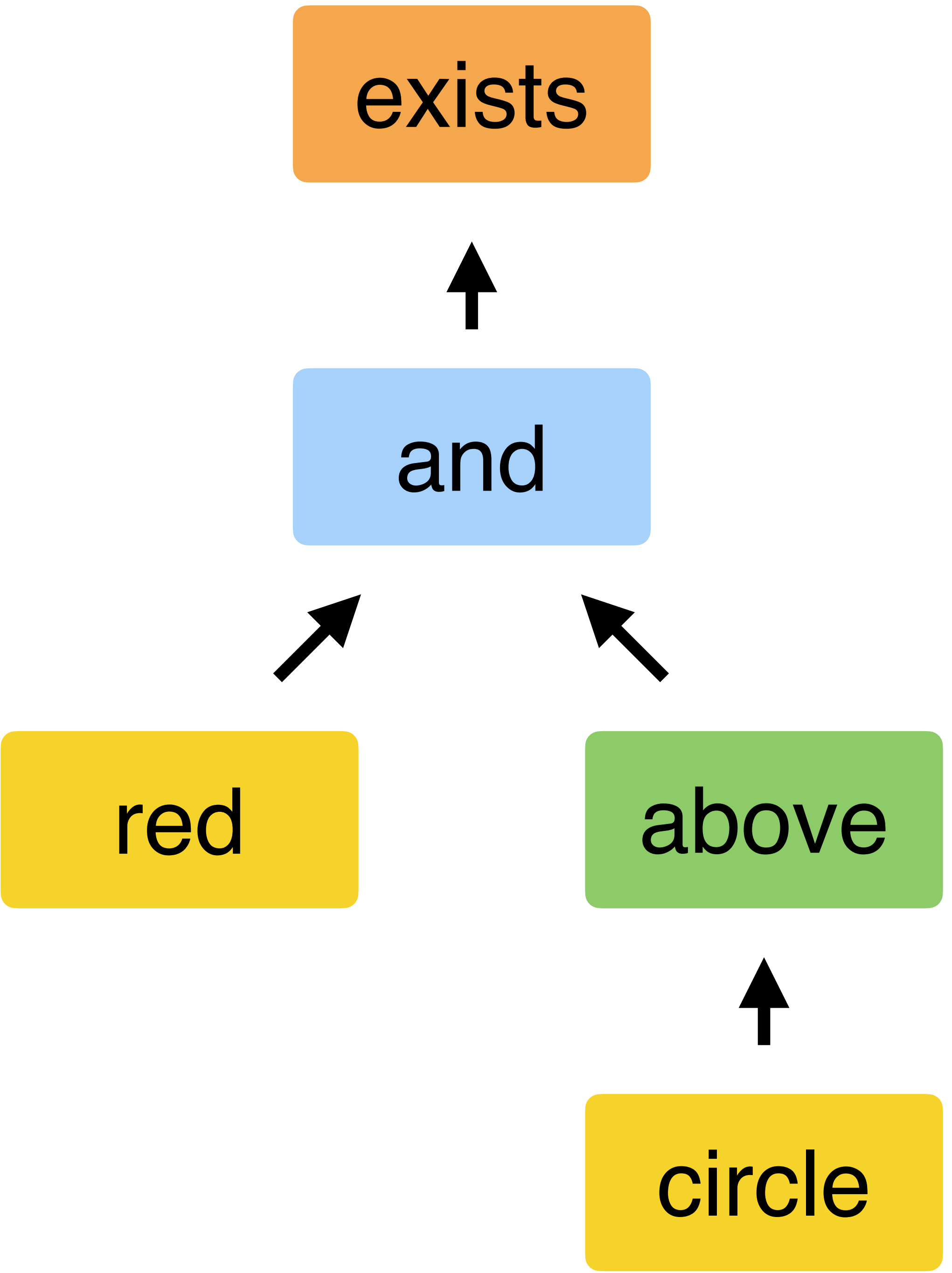
Is there a red shape above a circle?

[Slides credit: Jacob Andreas]

Sentence meanings are computations

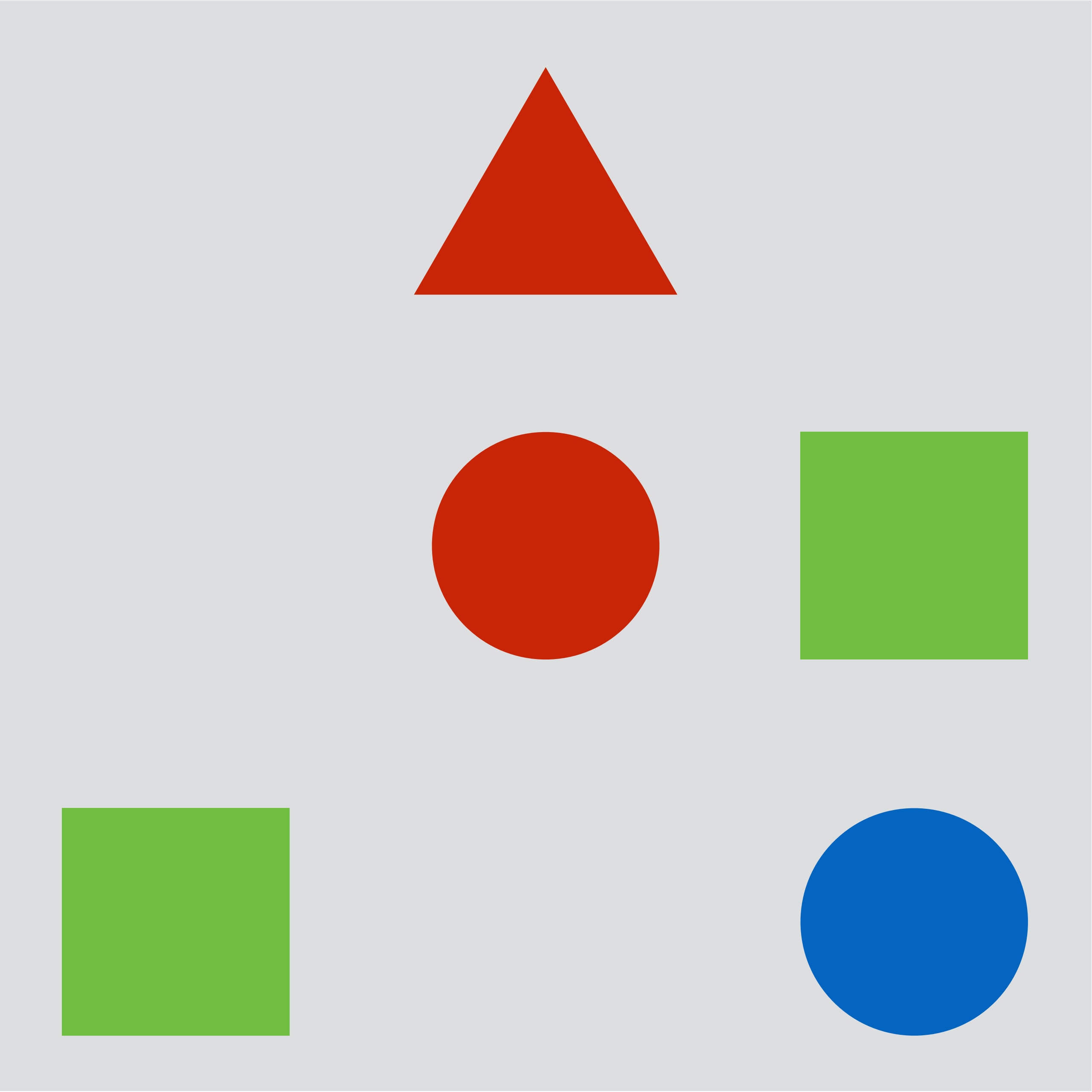


Is there a red shape above a circle?

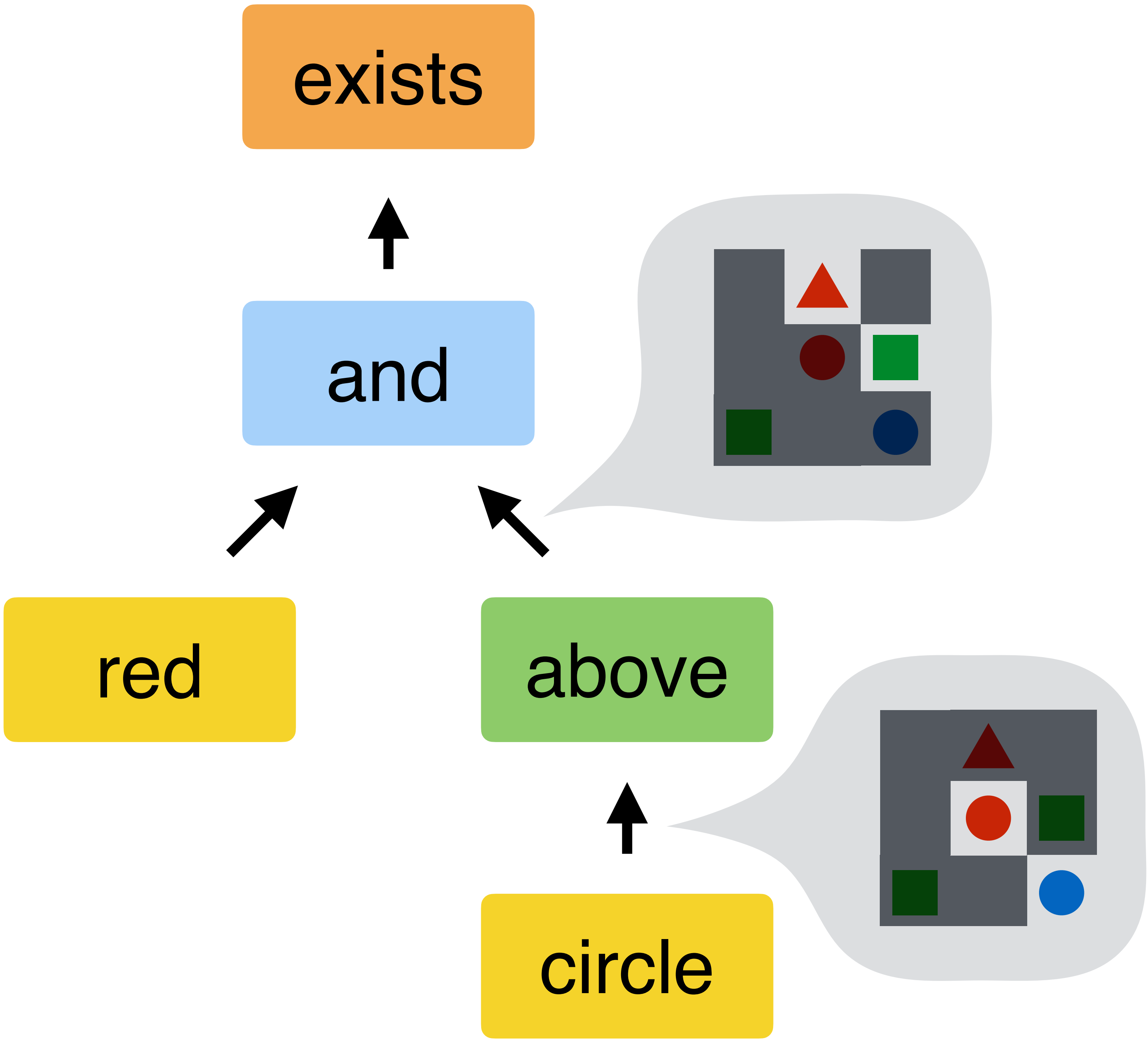


[Slides credit: Jacob Andreas]

Sentence meanings are computations



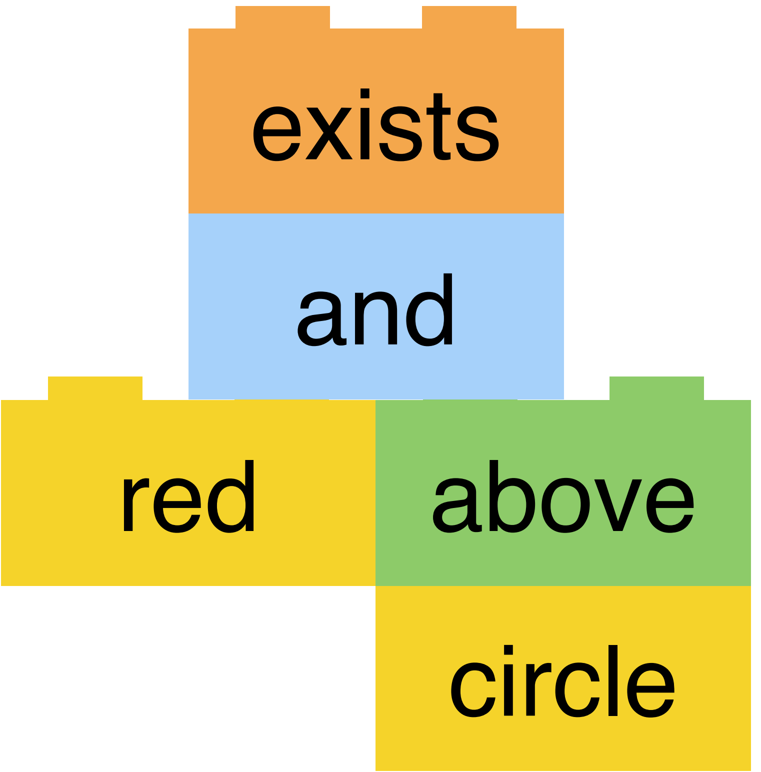
Is there a red shape above a circle?

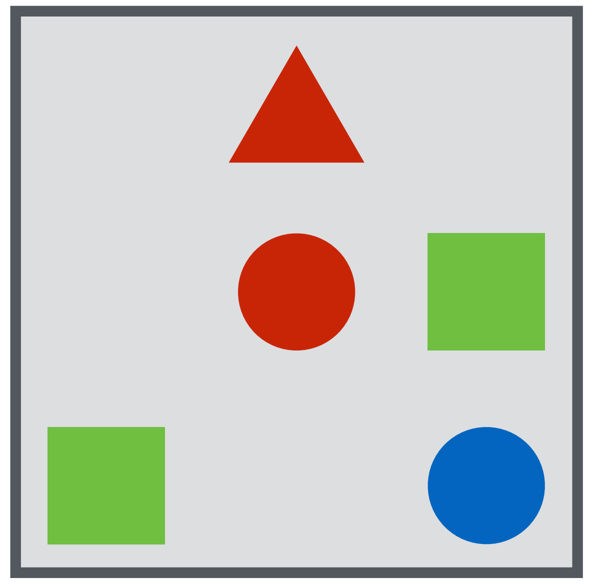


[Slides credit: Jacob Andreas]

Learning

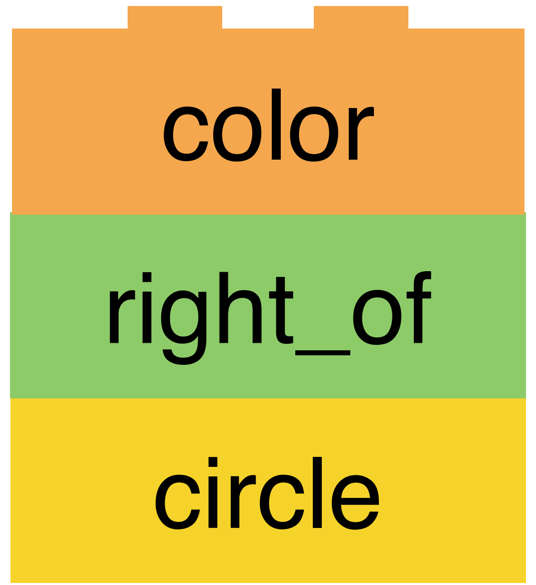
yes

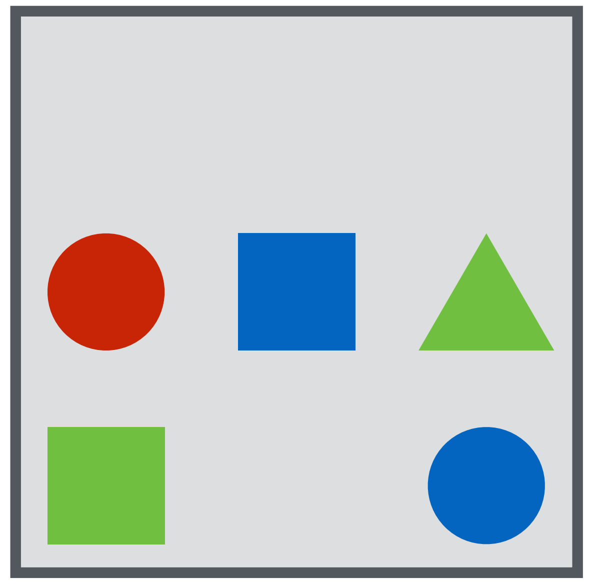




Is there a red shape above a circle?

blue

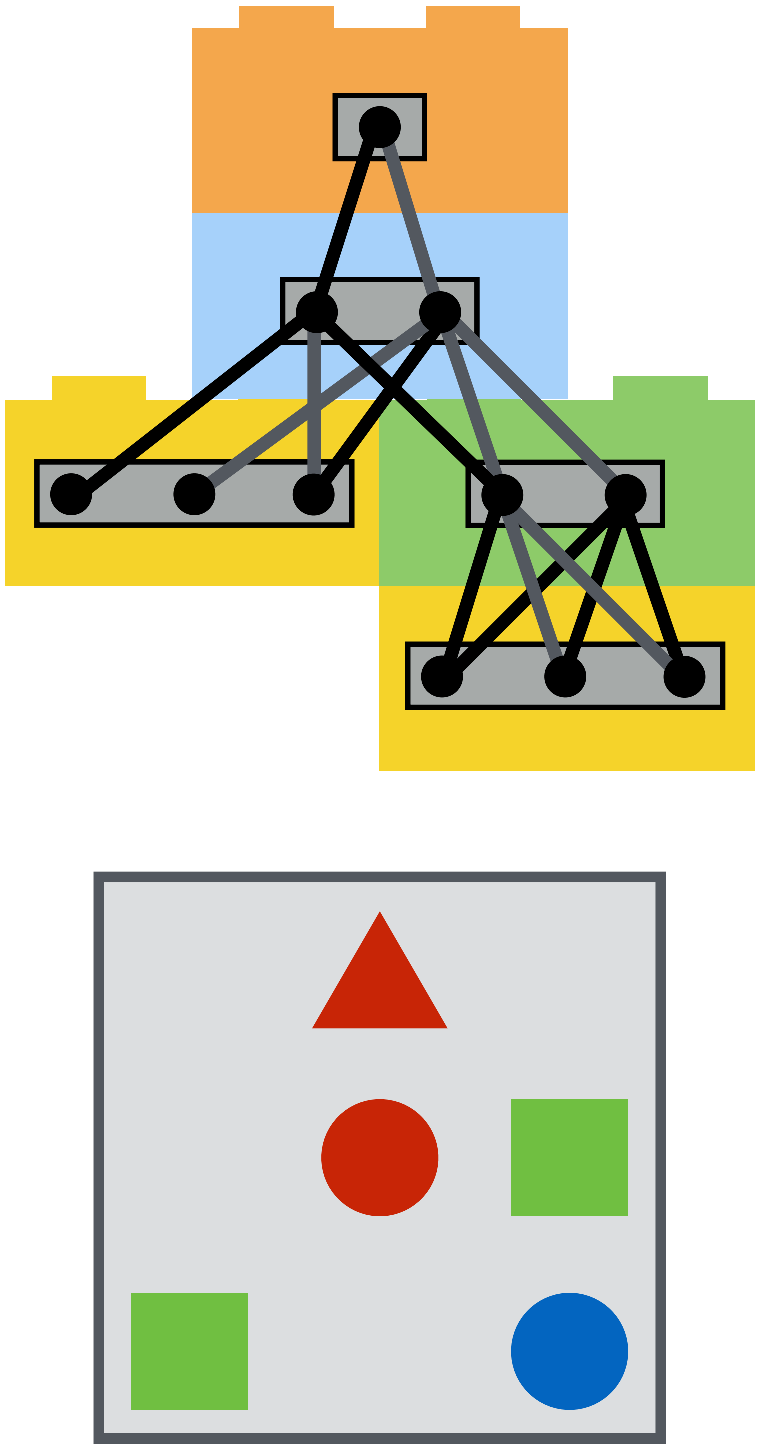




What color is the shape right of a circle?

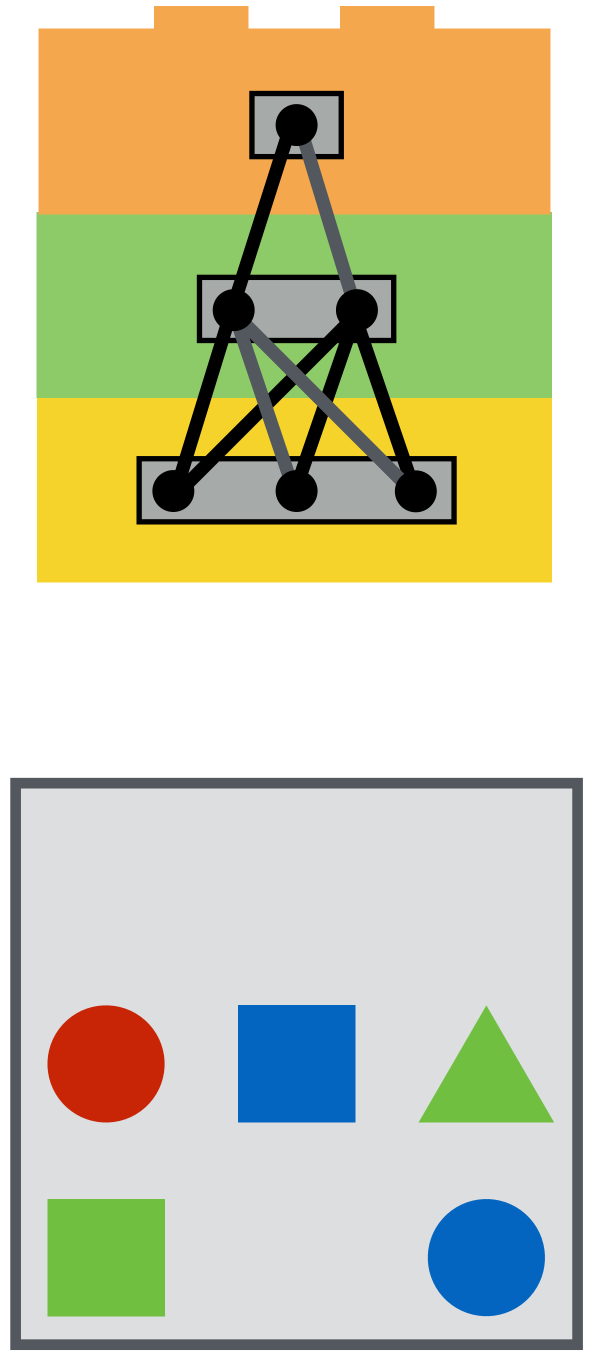
Learning

yes



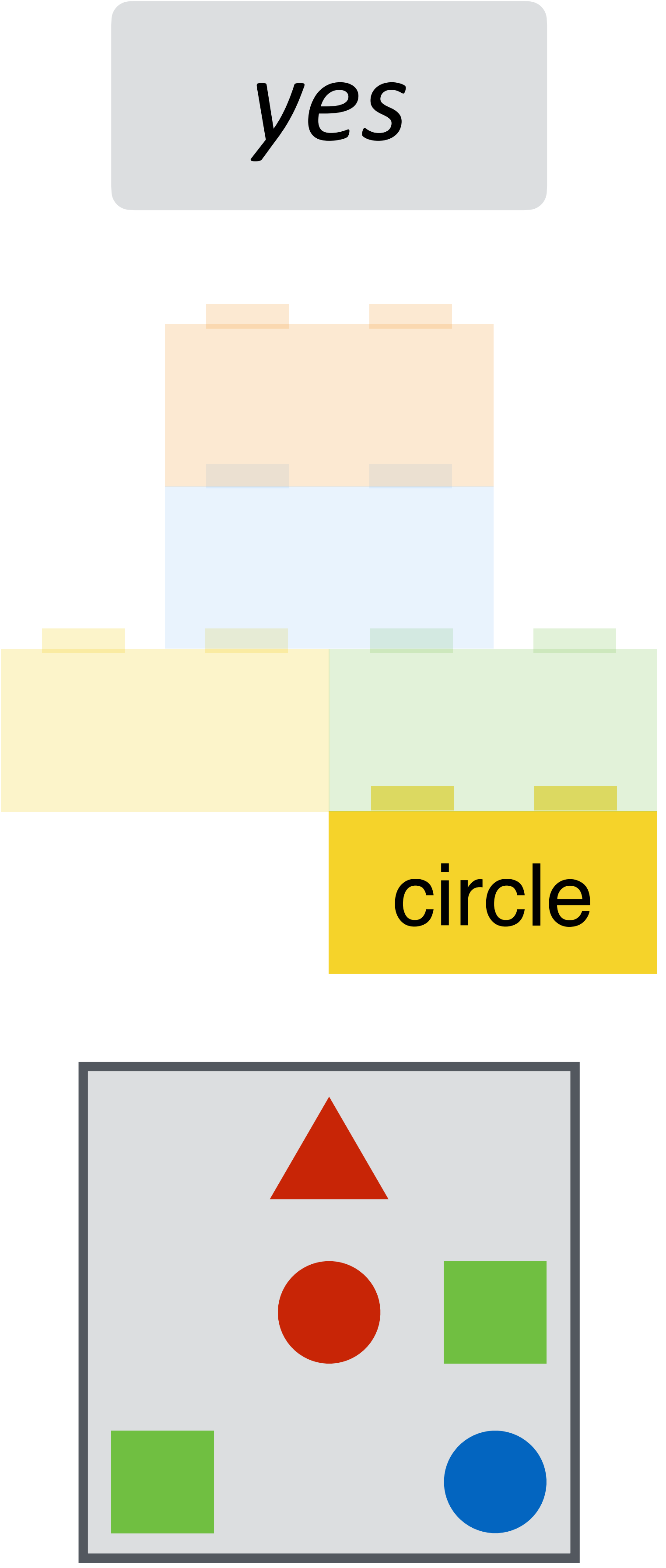
Is there a red shape above a circle?

blue

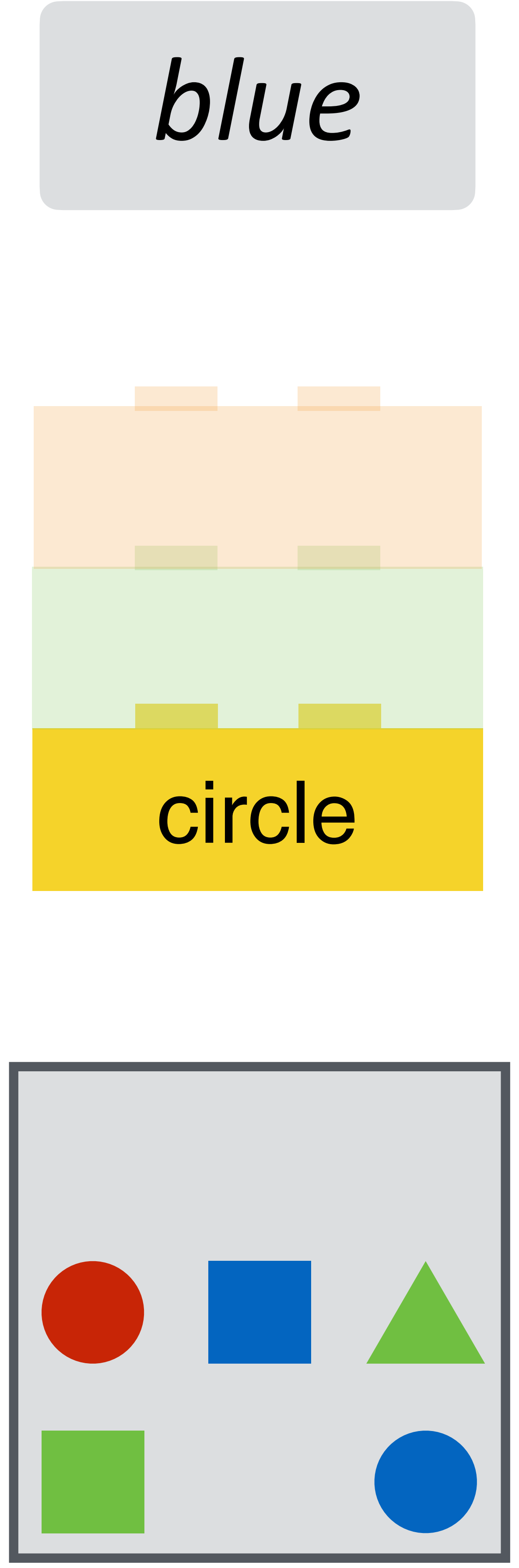


What color is the shape right of a circle?

Parameter tying

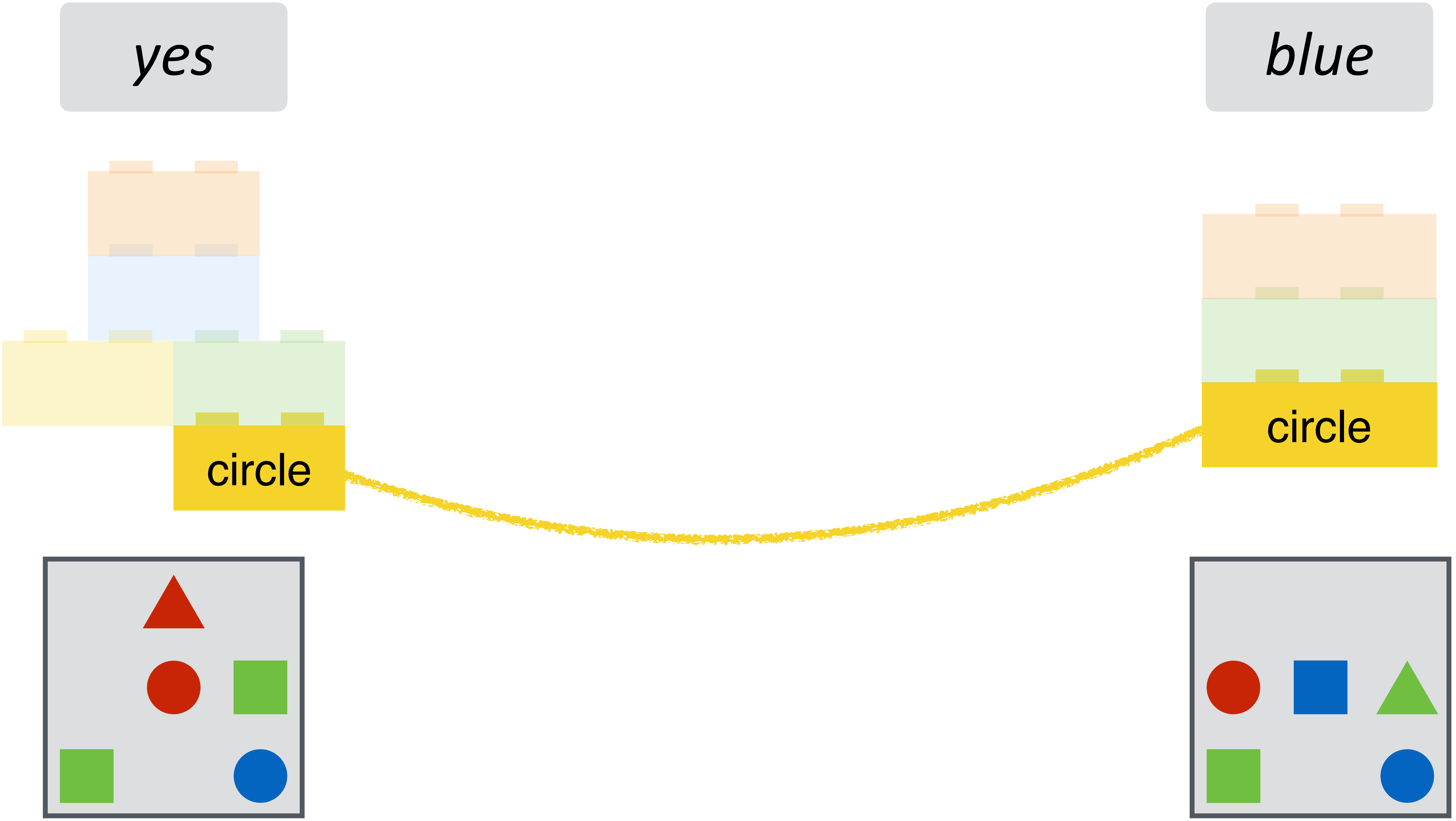


Is there a red shape above a circle?



What color is the shape right of a circle?

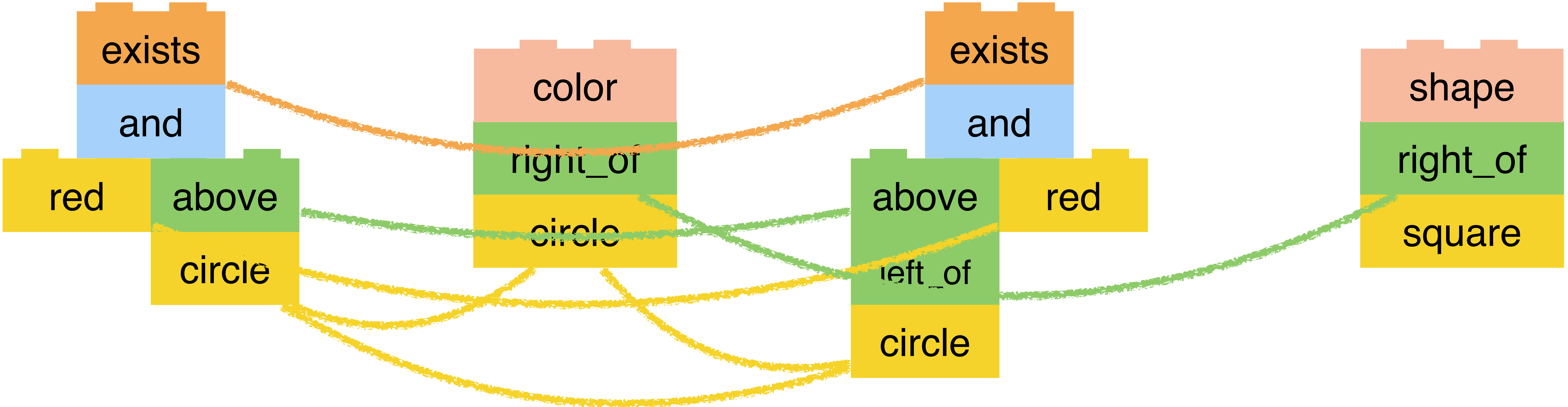
Parameter tying



Is there a red shape above a circle?

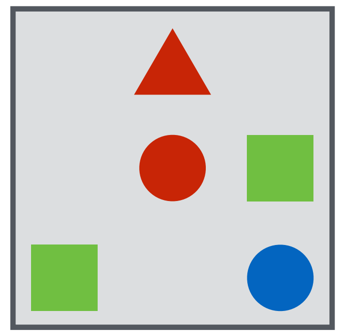
What color is the shape right of a circle?

Extreme parameter tying



[Slides credit: Jacob Andreas]

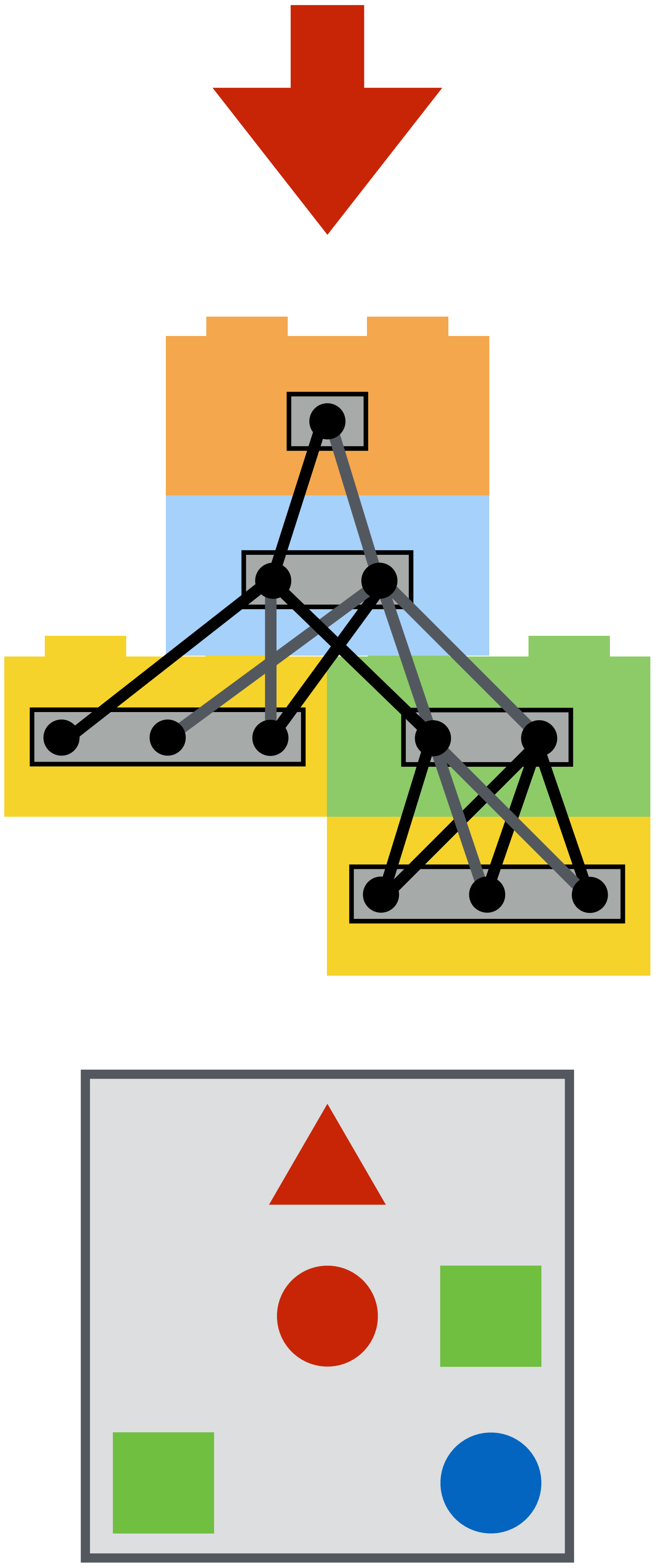
Learning with fixed layouts is easy!

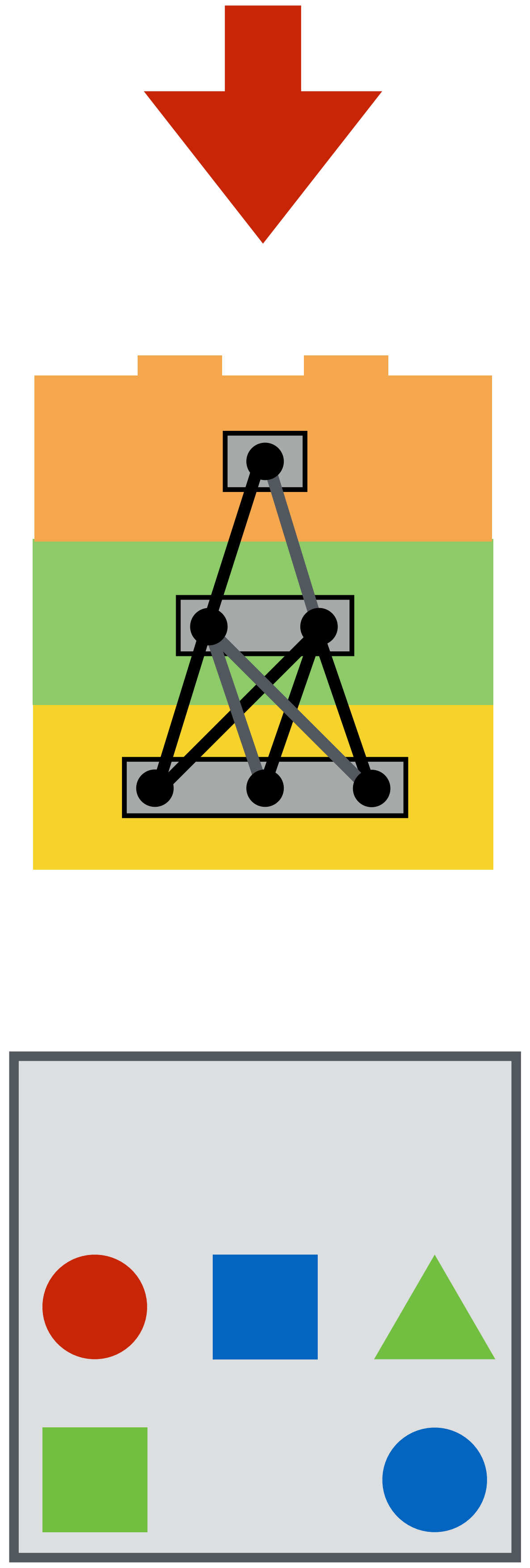
$\underset{W}{\arg\max} \sum p($ *yes* $|$  , $; W)$

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

[Slides credit: Jacob Andreas]

Maximum likelihood estimation

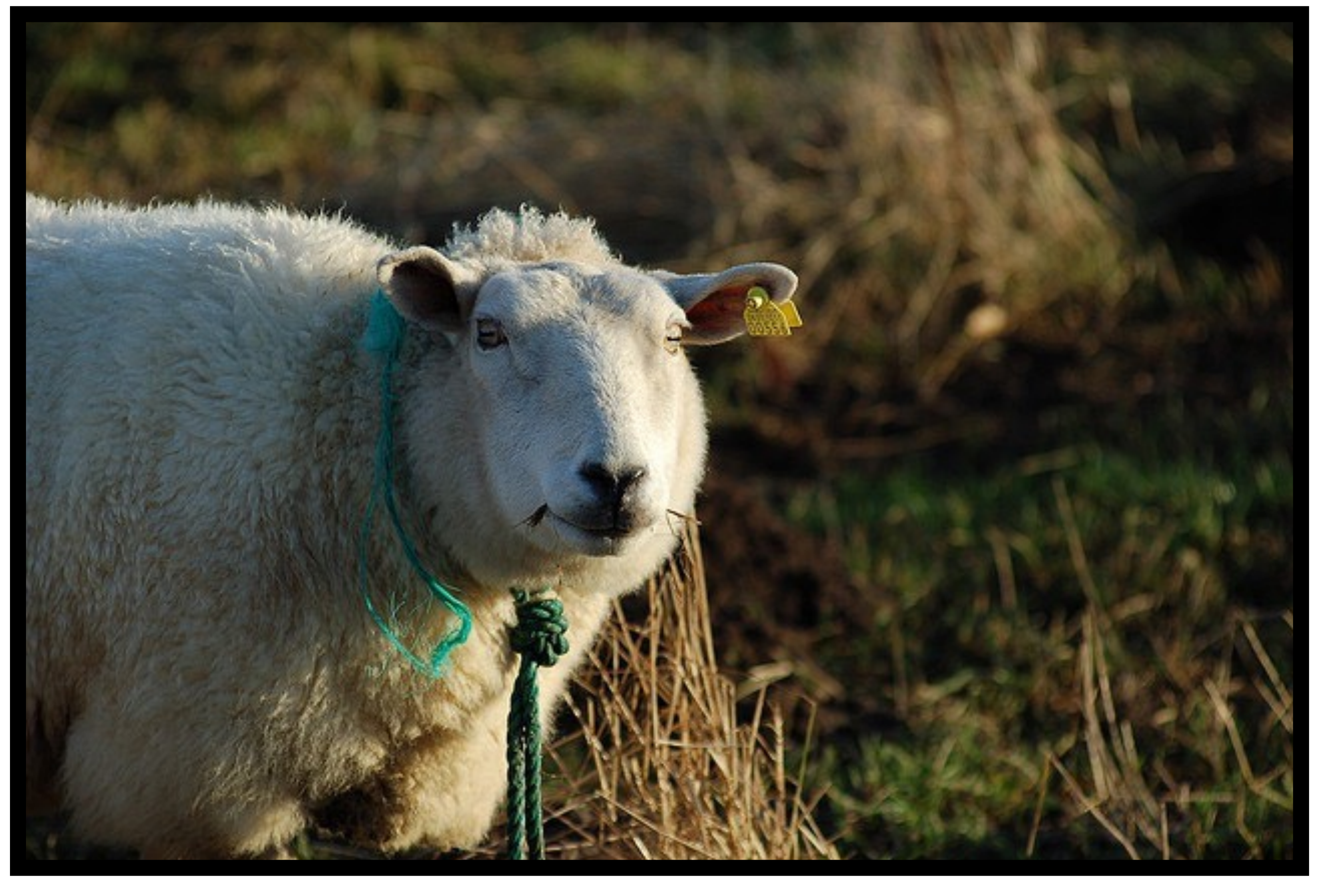


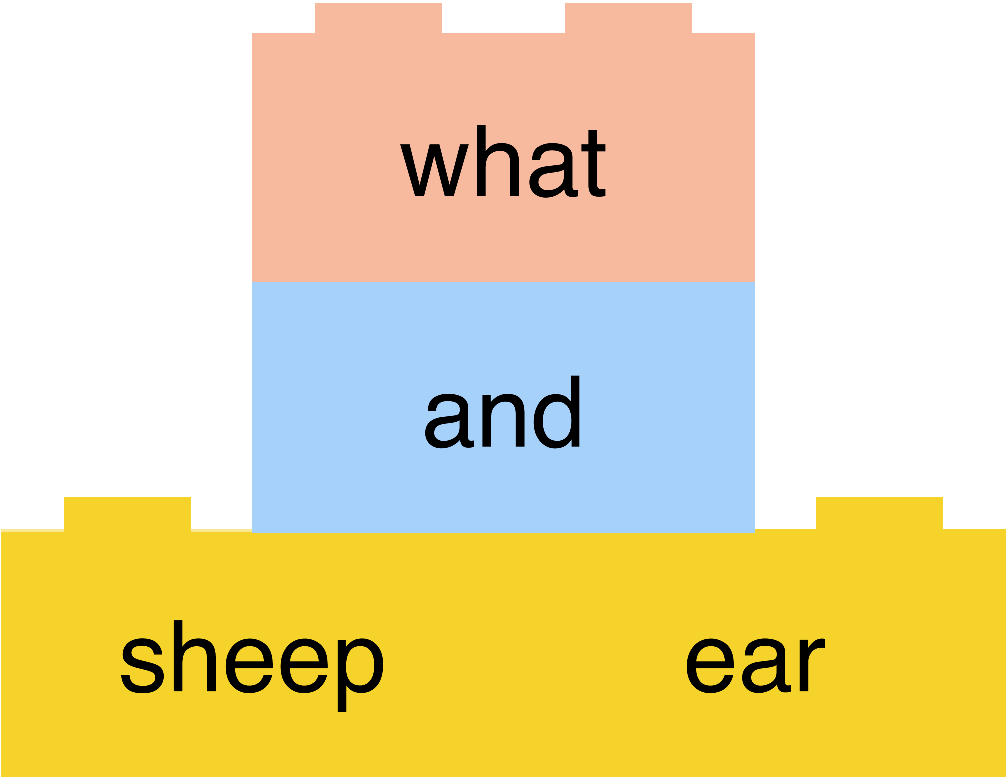


[Slides credit: Jacob Andreas]

Experiments: VQA Dataset

What is in the sheep’s ear?



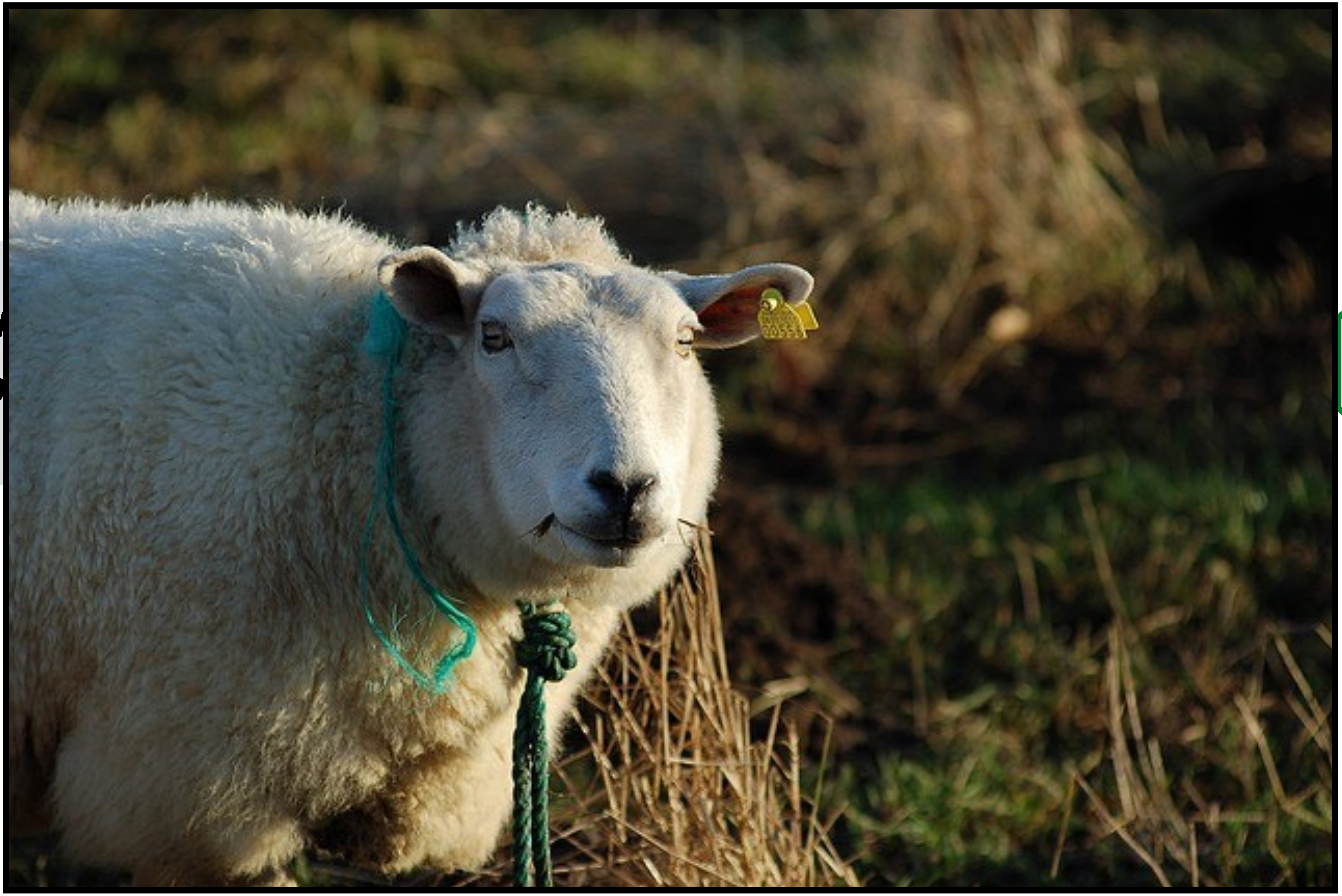


tag

[Slides credit: Jacob Andreas]

Experiments: VQA Dataset

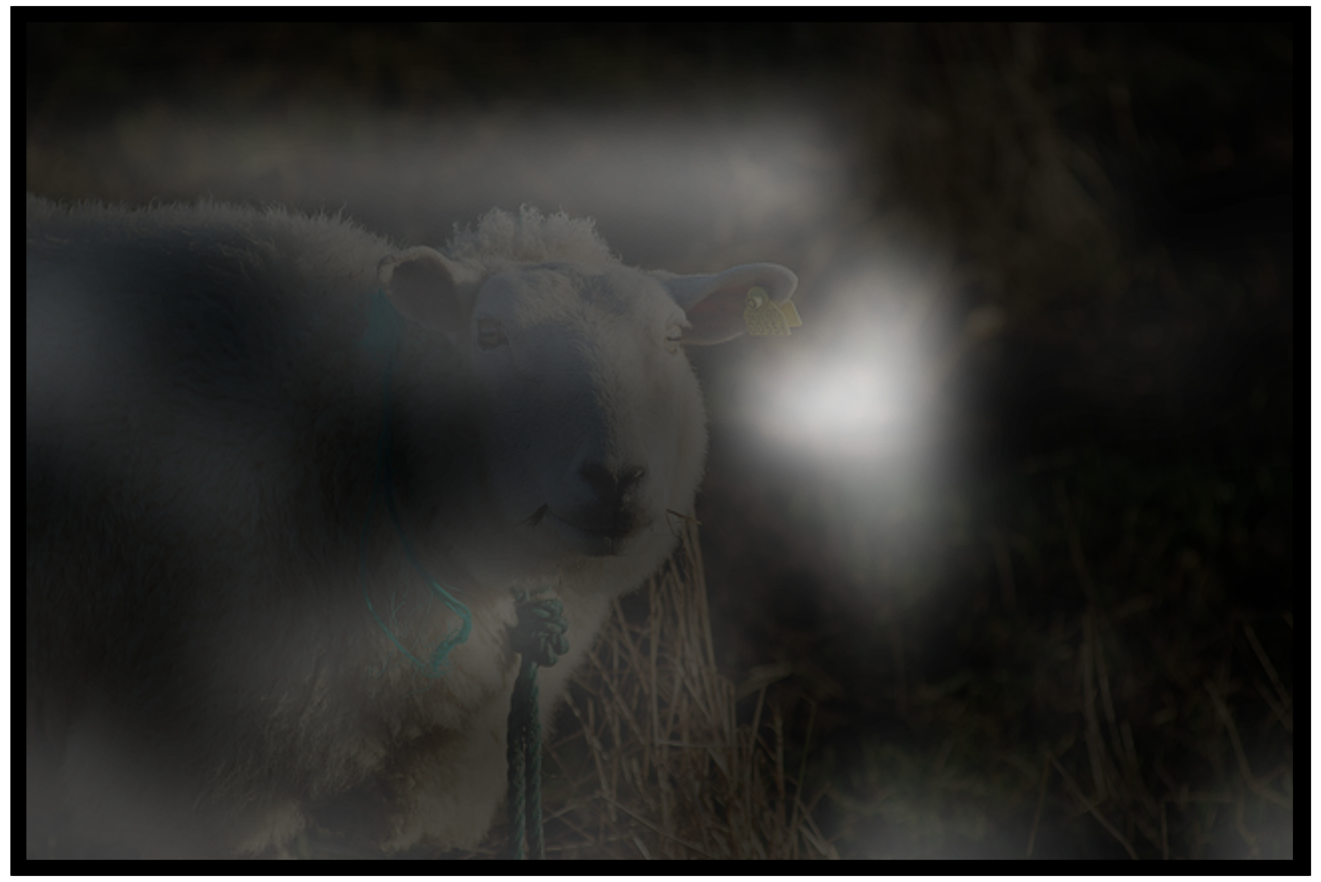
What i
sheep

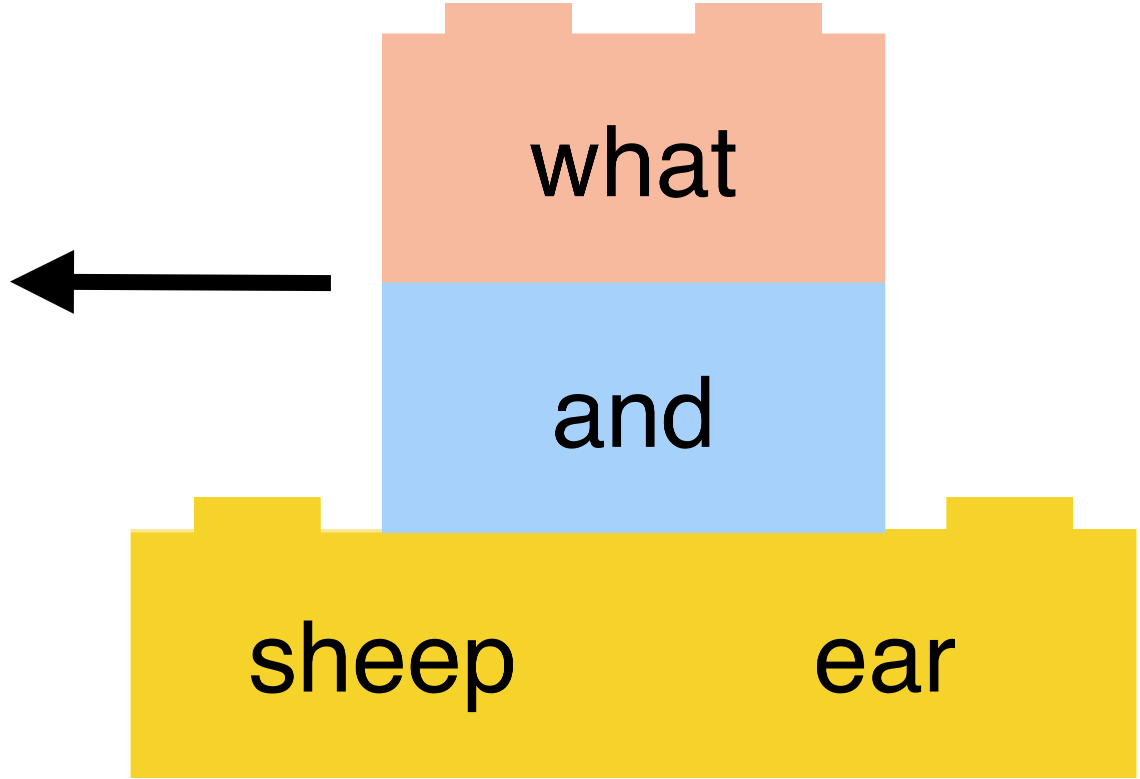


tag

Experiments: VQA Dataset

What is in the sheep's ear?





tag

[Slides credit: Jacob Andreas]