Lecture 20

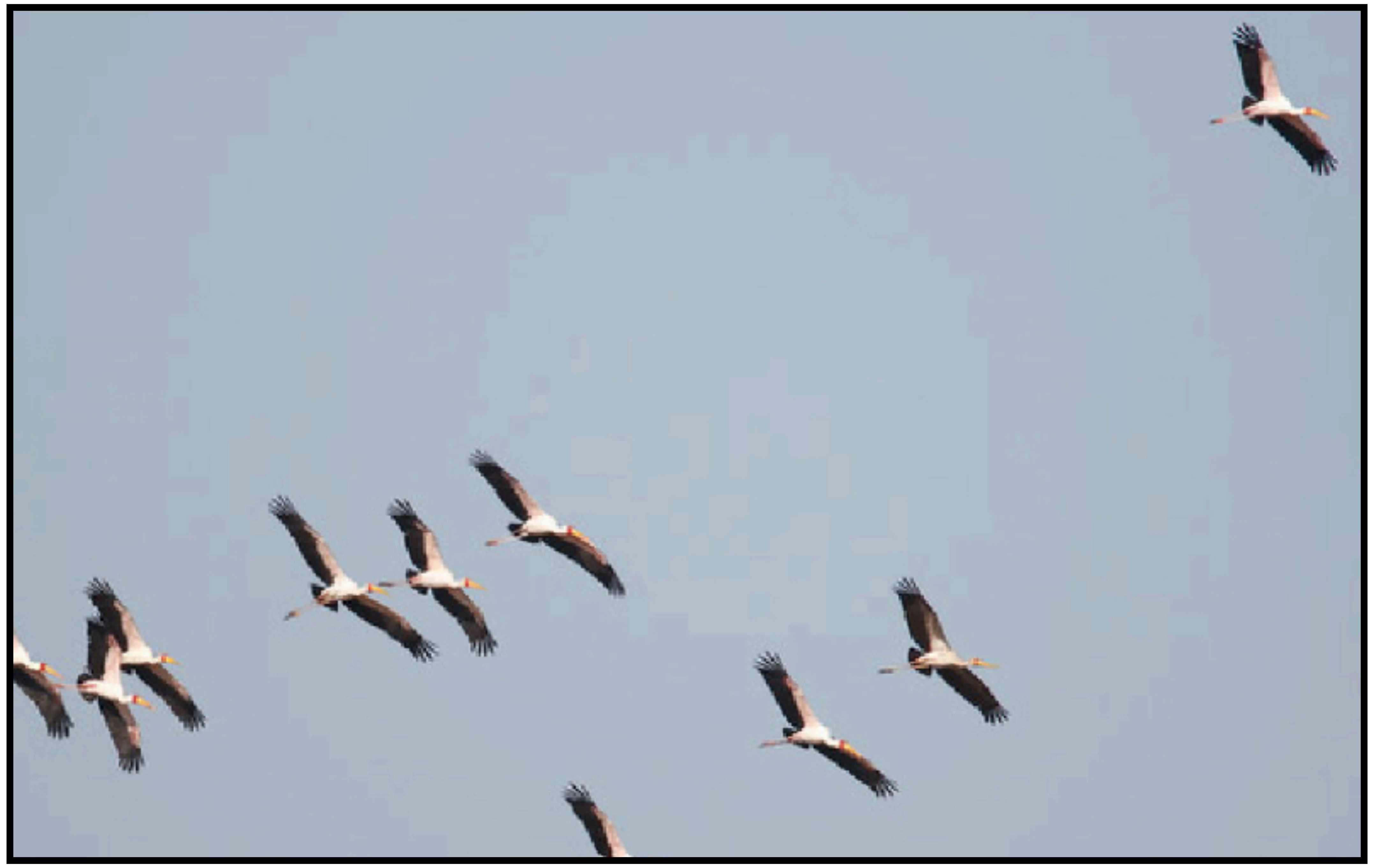
Vision and Language

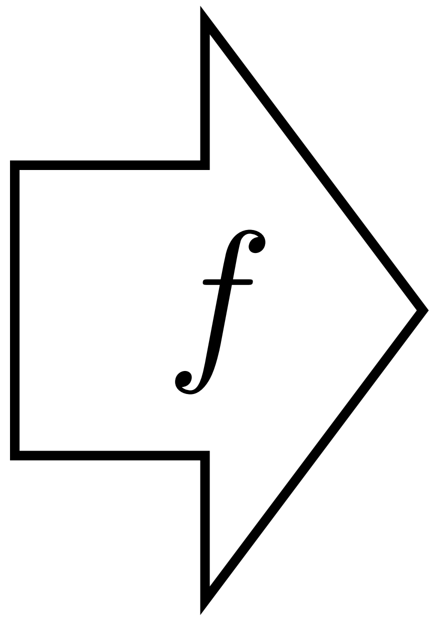
Bill Freeman, Antonio Torralba, Phillip Isola

20. Vision and Language

- Language as sequence modeling
- Image captioning
- Attention
- Visual Question Answering
- Neural module networks

Image captioning





“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)

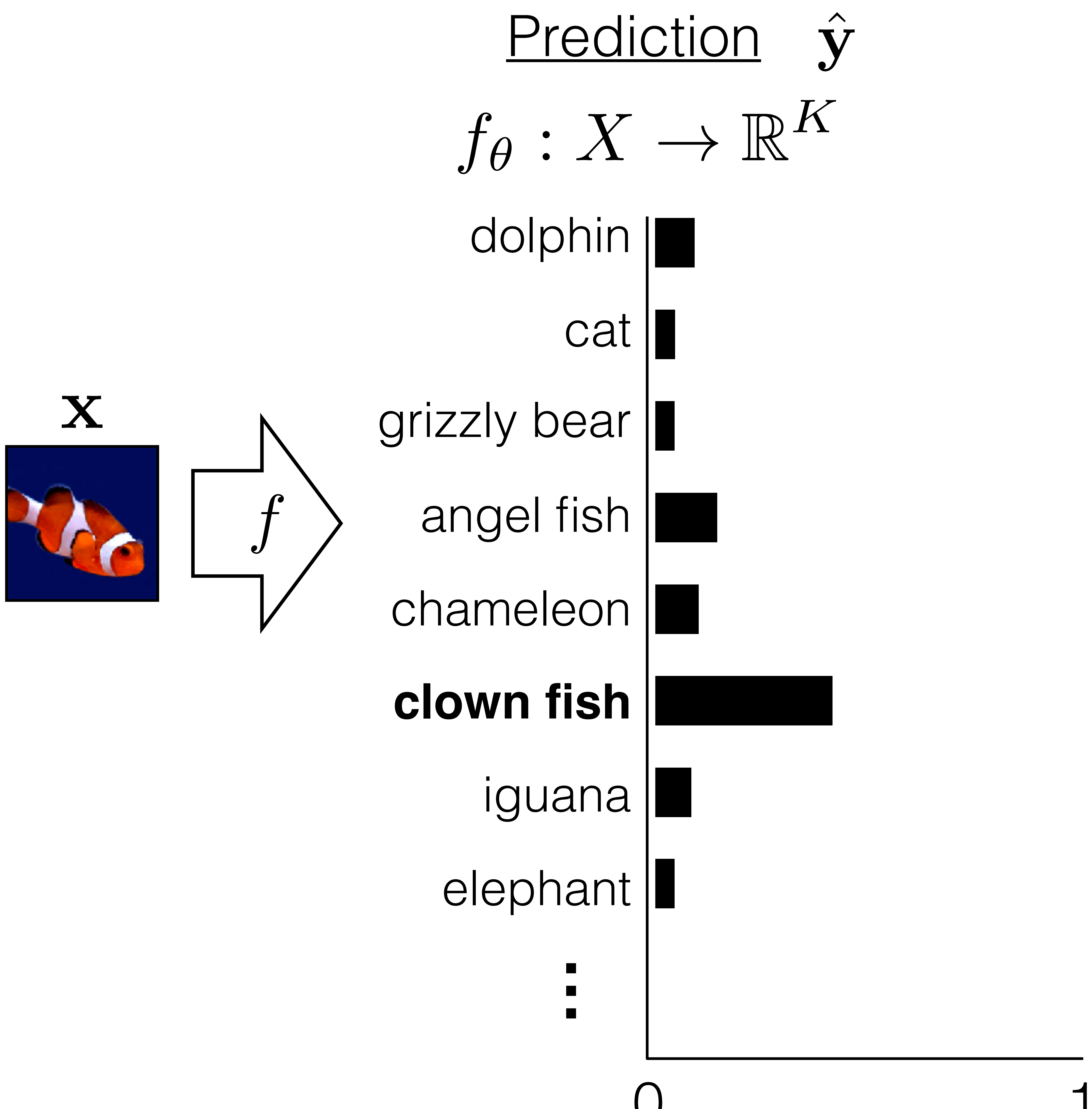
2. Transform your goal into a numerical measure (objective function)

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

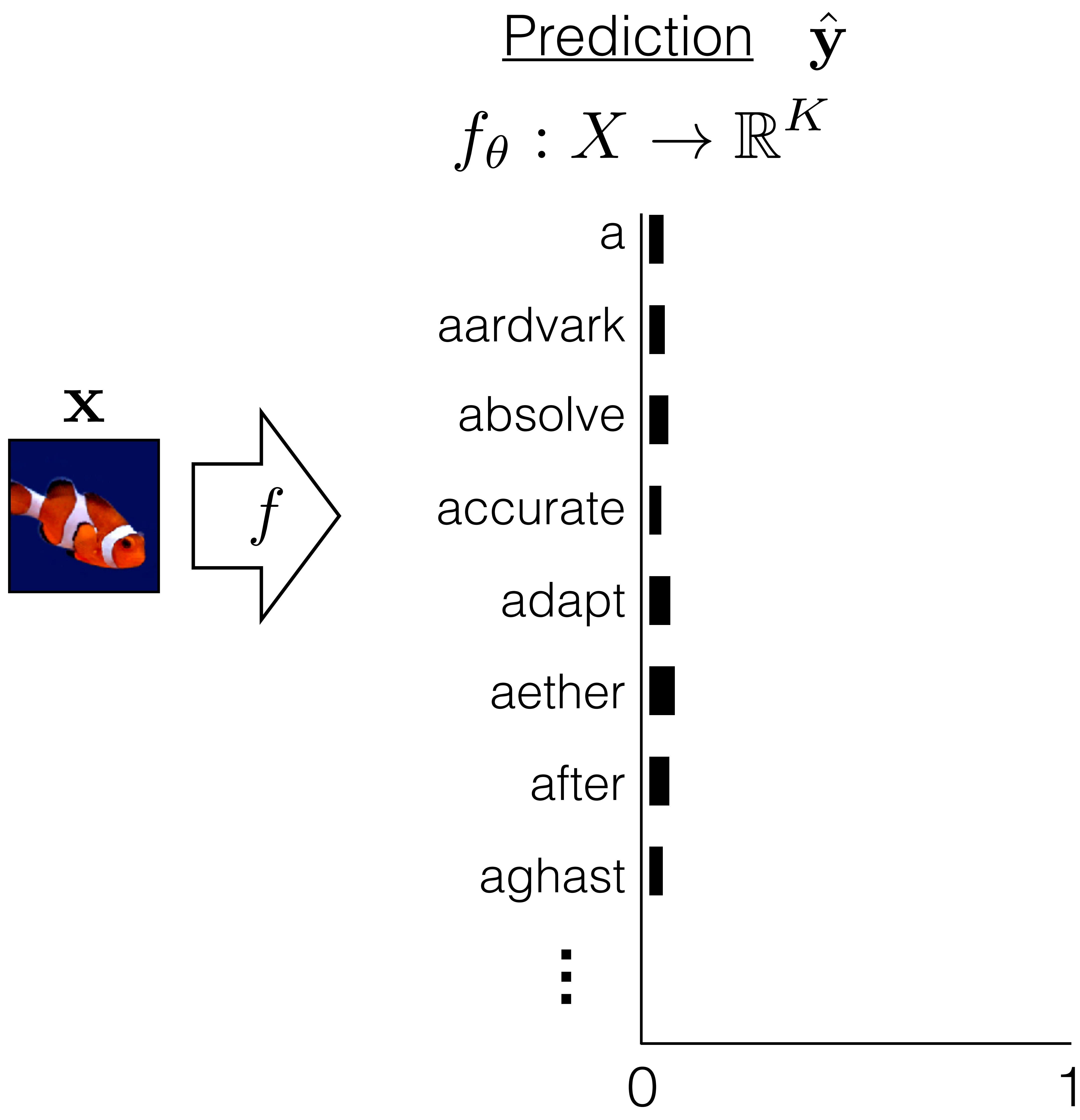
4. Use a generic optimizer (SGD) and an appropriate architecture (e.g., CNN or RNN) to solve the learning problem

How to represent words as numbers?

How to represent words as numbers?



How to represent words as numbers?

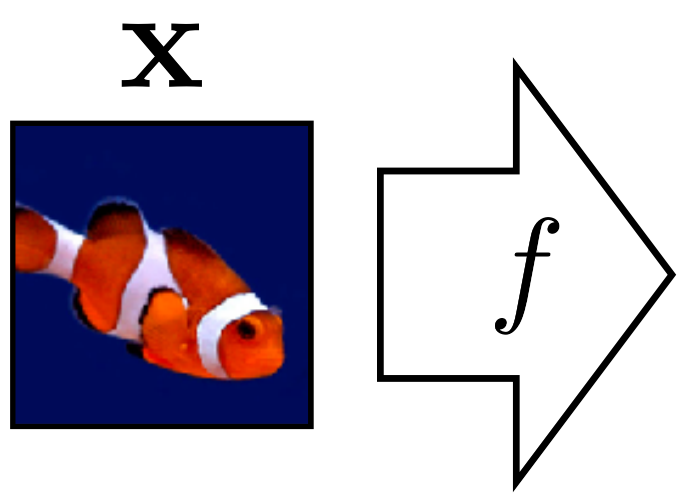


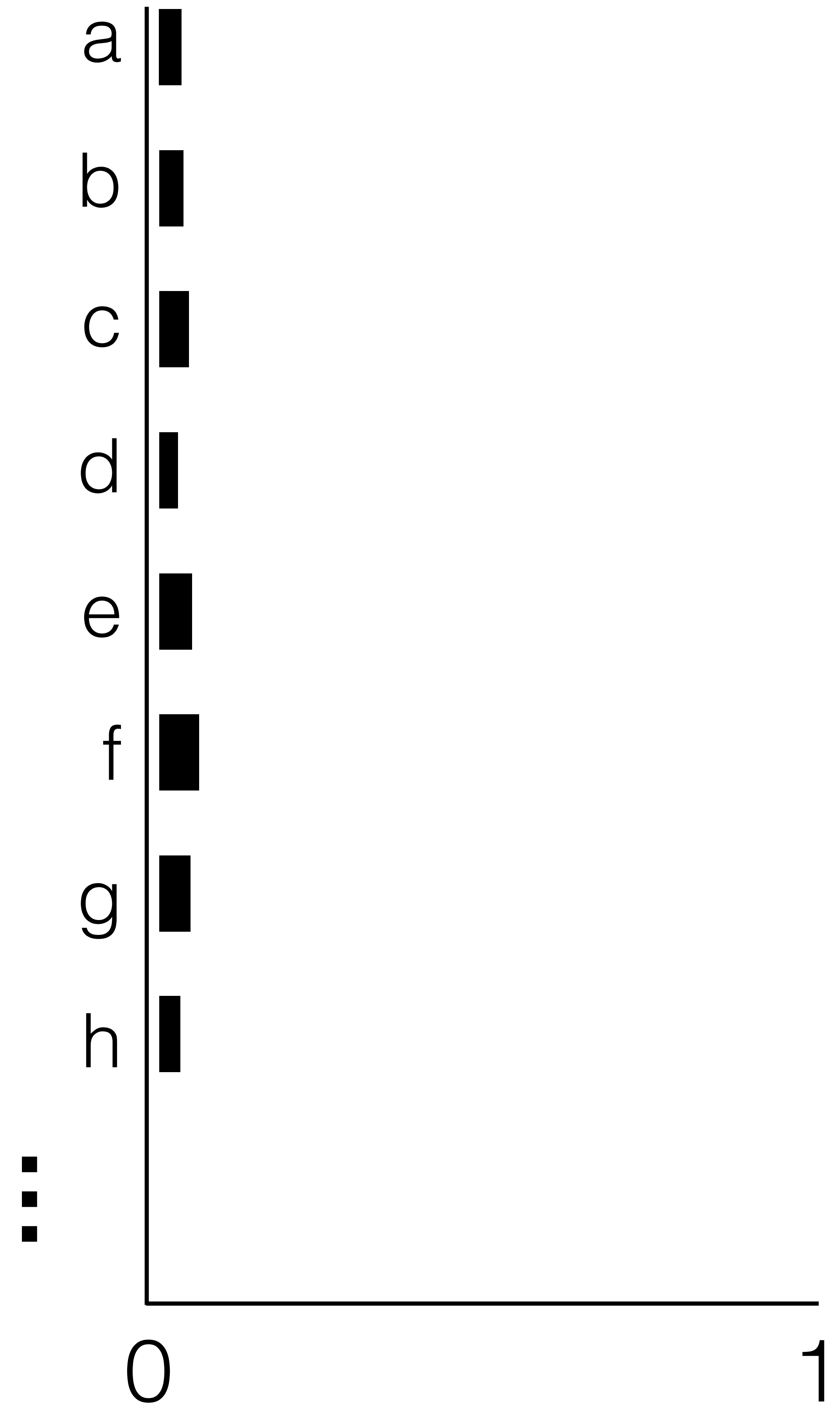
Rather than having just a handful of possible object classes, we can represent all words in a large vocabulary using a very large K (e.g., K=100,000).

How to represent words as numbers?

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

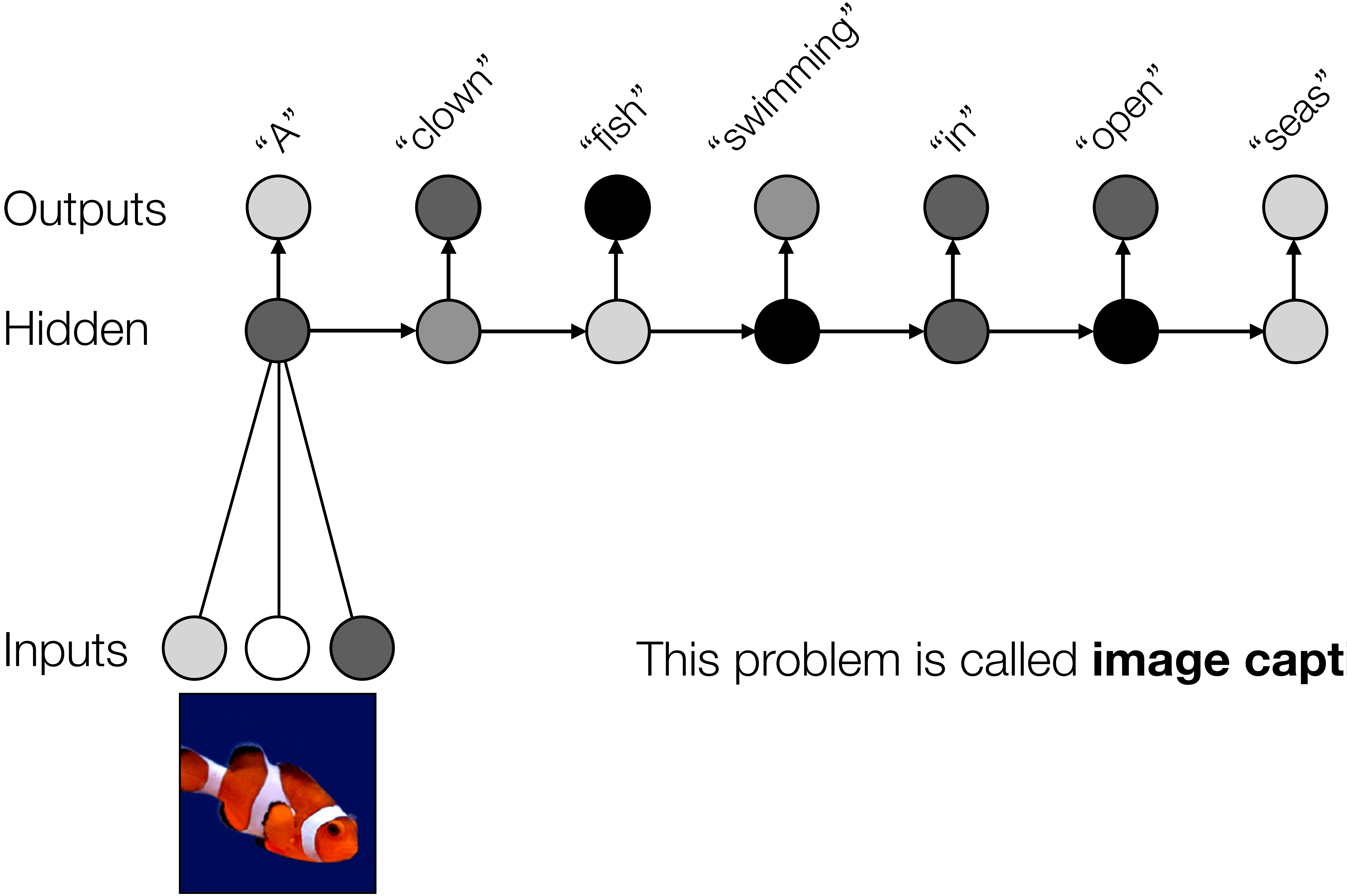
$$f_\theta : X \to \mathbb{R}^K$$



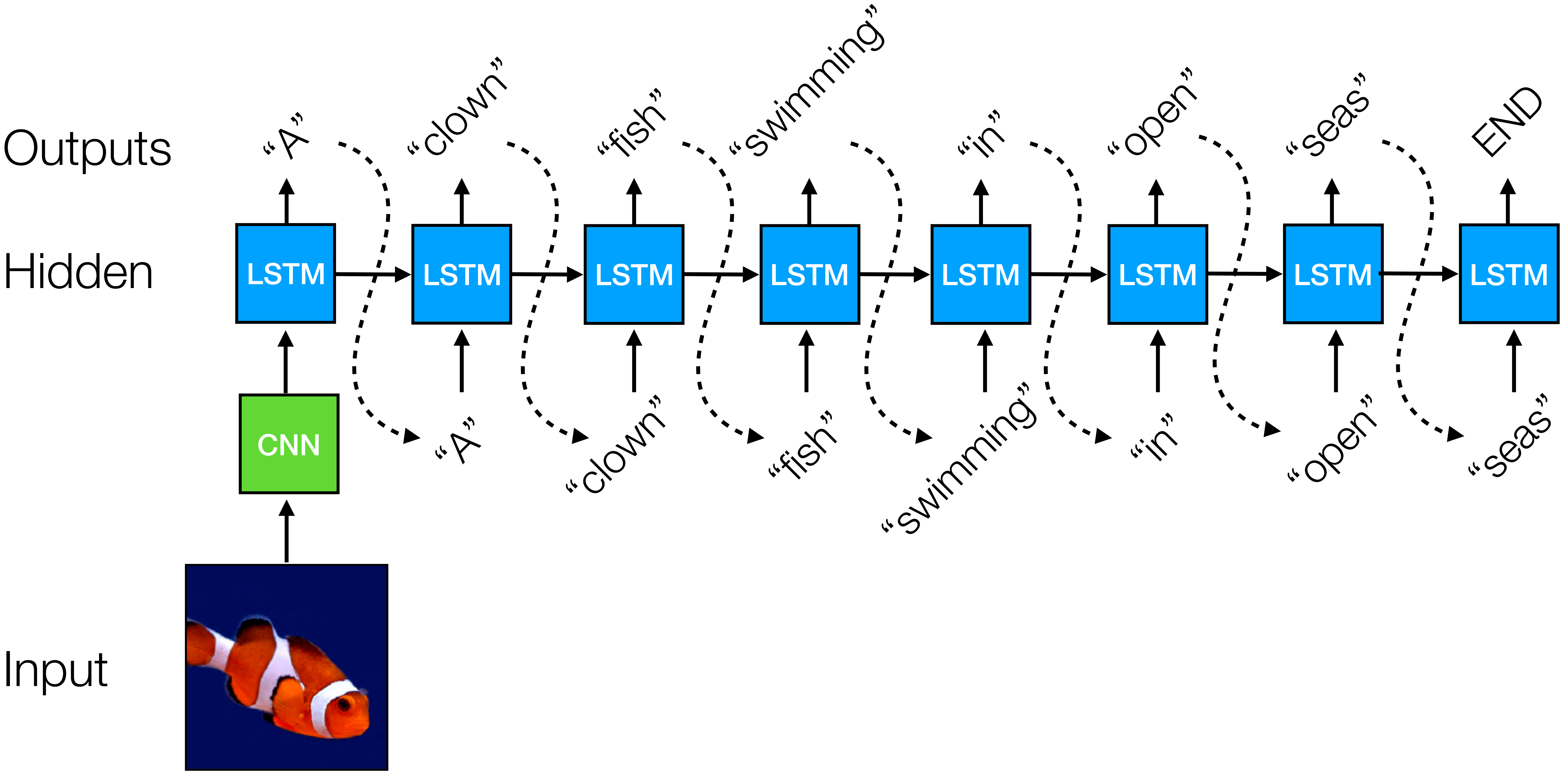


Or, represent each character as a class (e.g., K=26 for English letters),

and represent words as a sequence of characters.

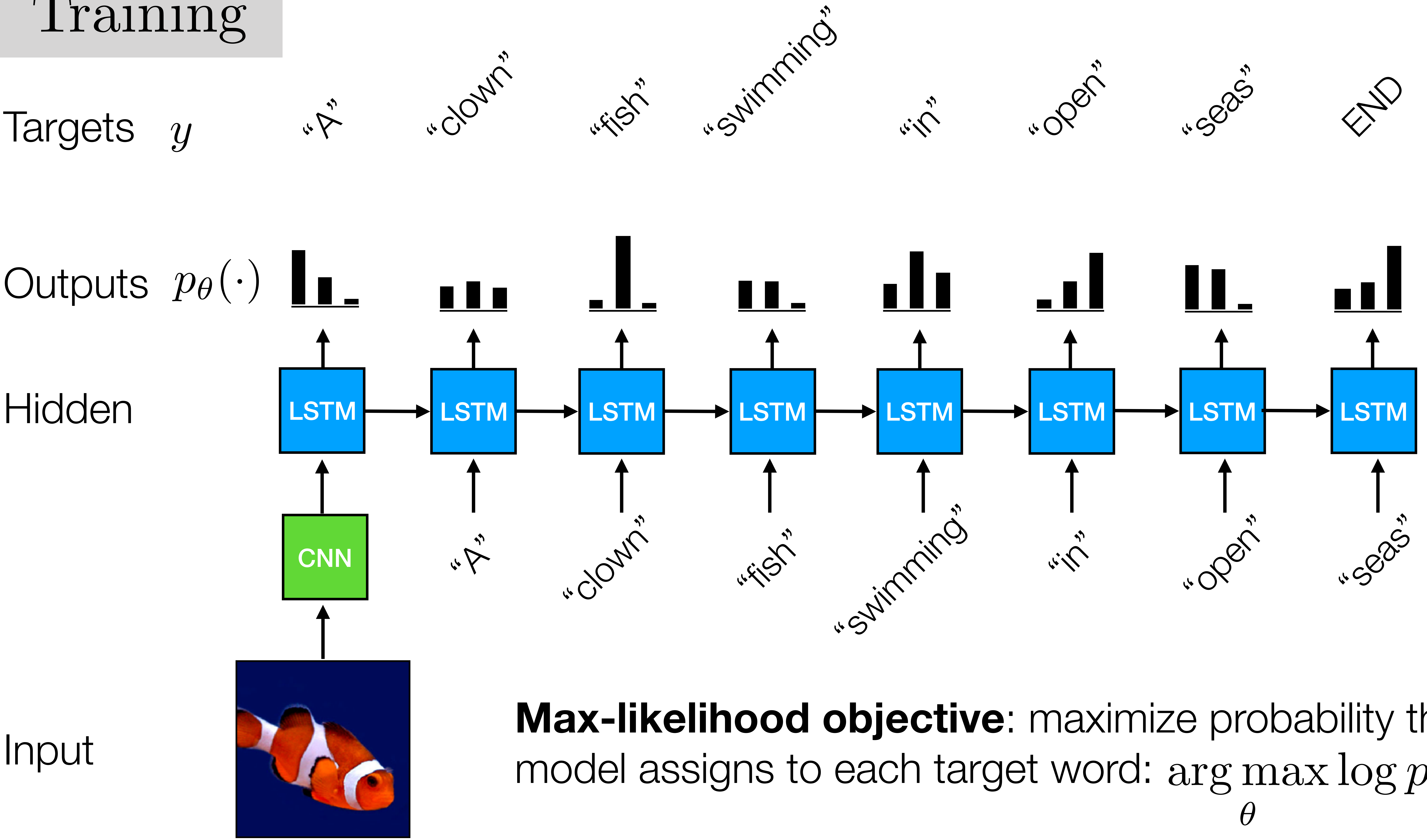


This problem is called **image captioning**



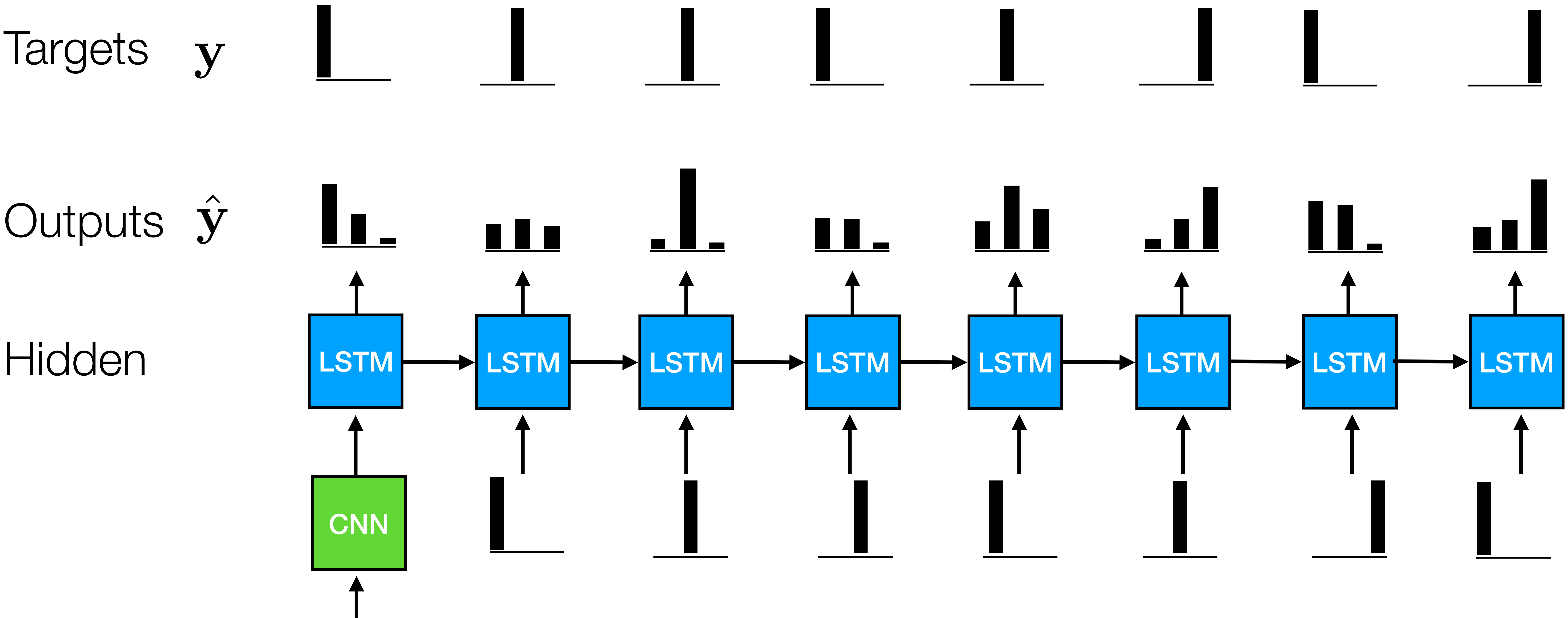
Outputs
Hidden
Input
"A"
"clown"
"fish"
"swimming"
"in"
"open"
"seas"
END
LSTM
LSTM
LSTM
LSTM
LSTM
LSTM
LSTM
LSTM
CNN
"A"
"clown"
"fish"
"swimming"
"in"
"open"
"seas"

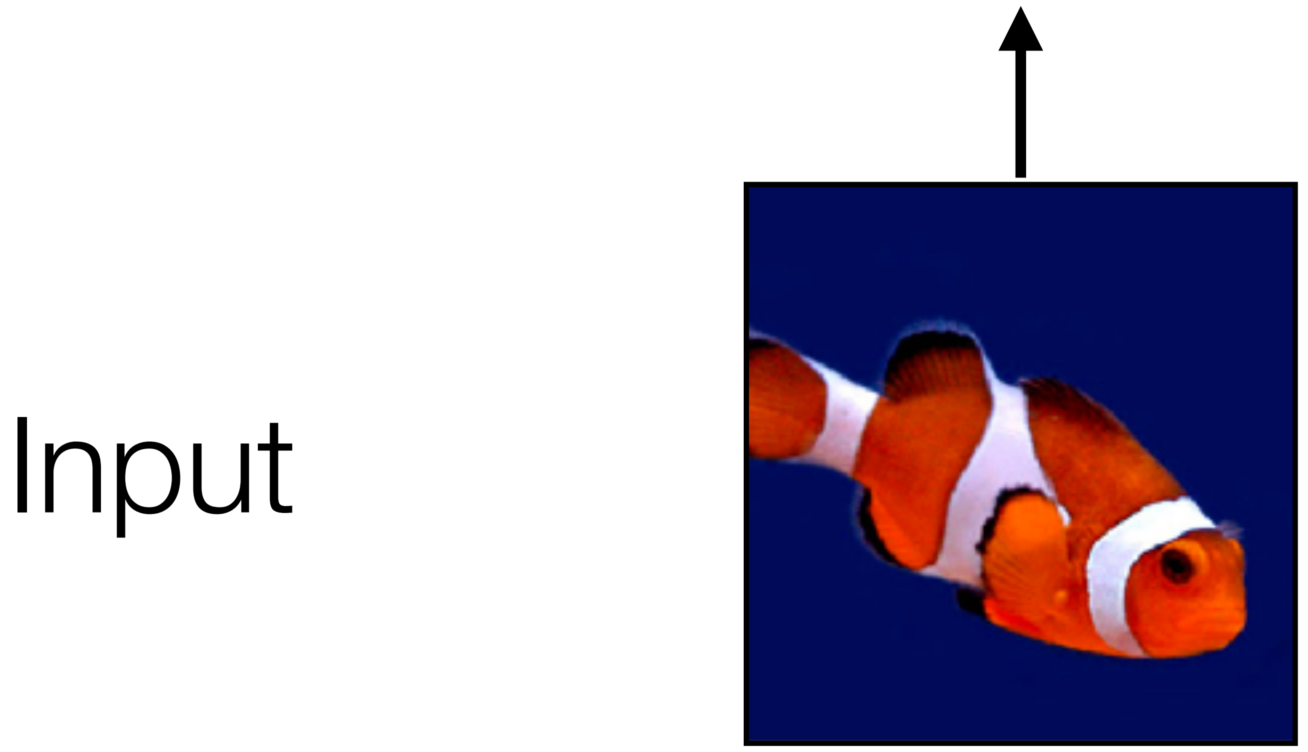
Training



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

Training

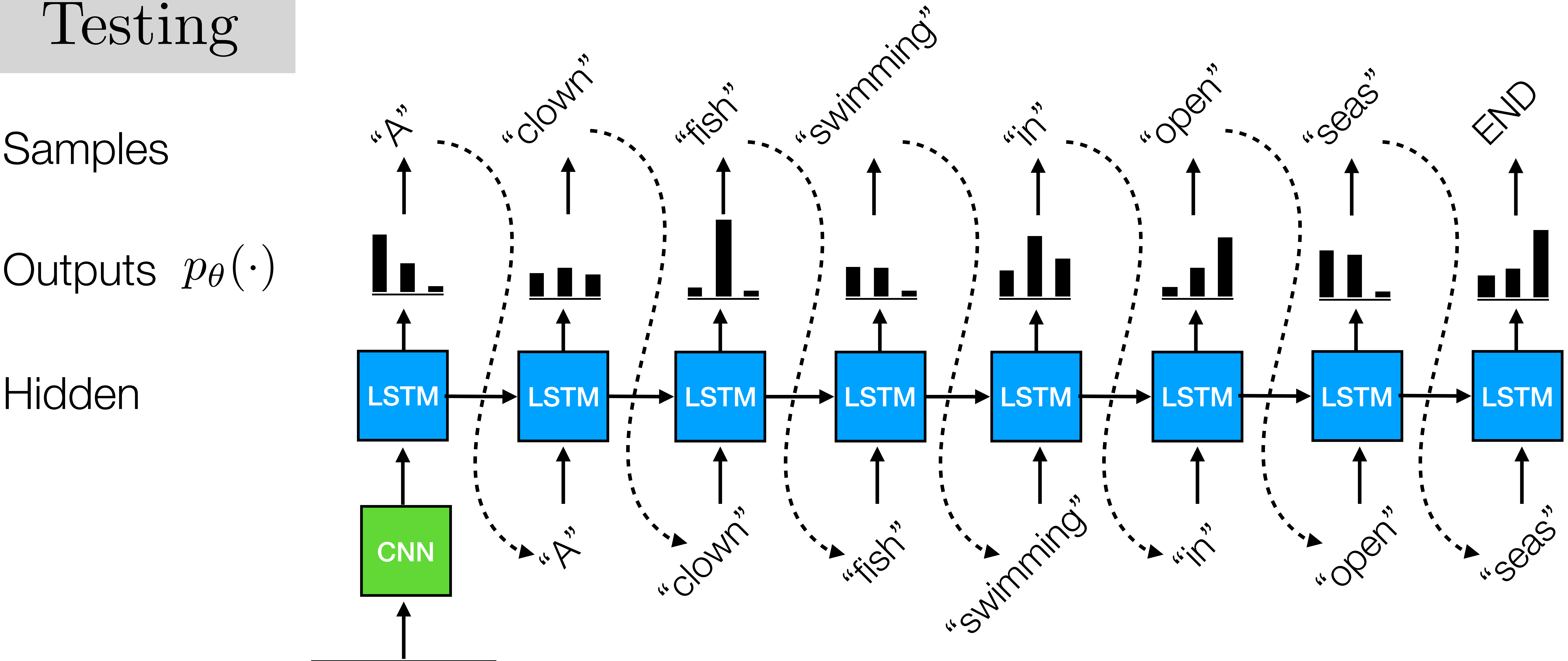




Max-likelihood objective:
minimize cross-entropy between model outputs and one-hot encoded targets.

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

Testing

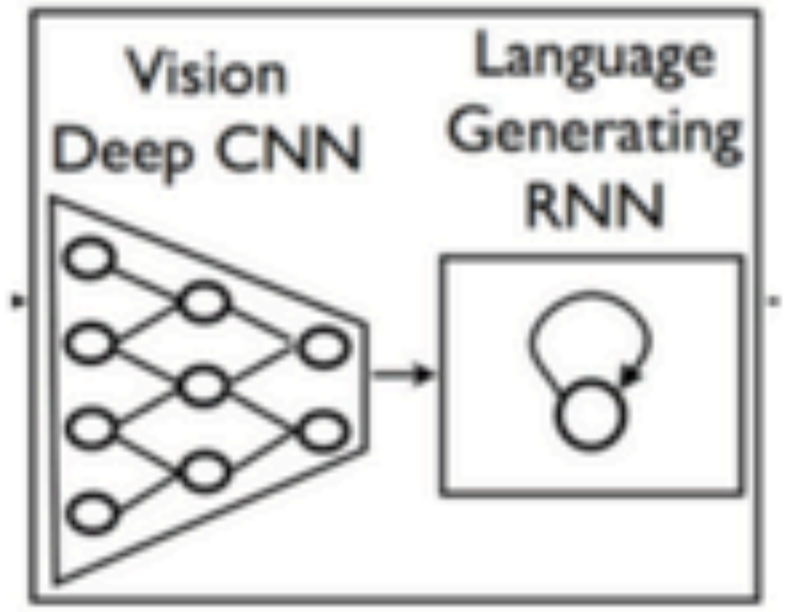


Input

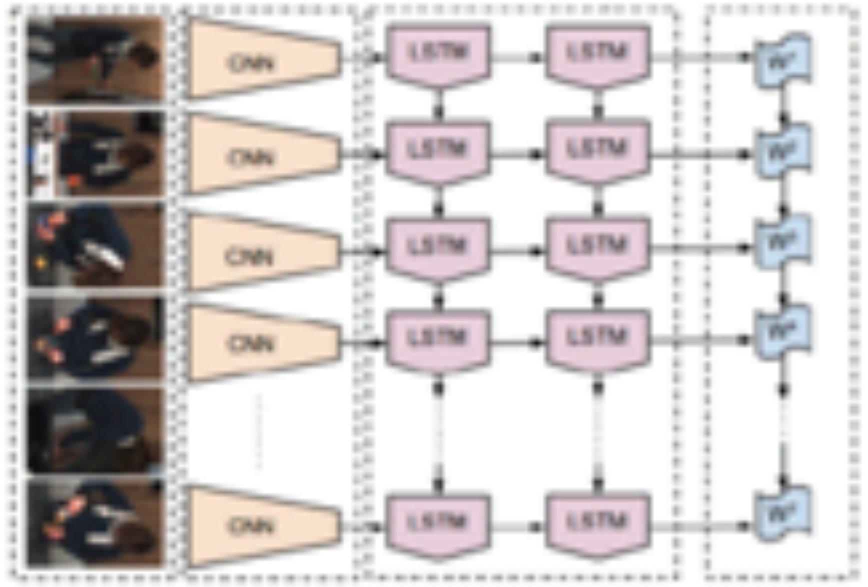
Sample from predicted distribution over words.

Alternatively, sample most likely word.

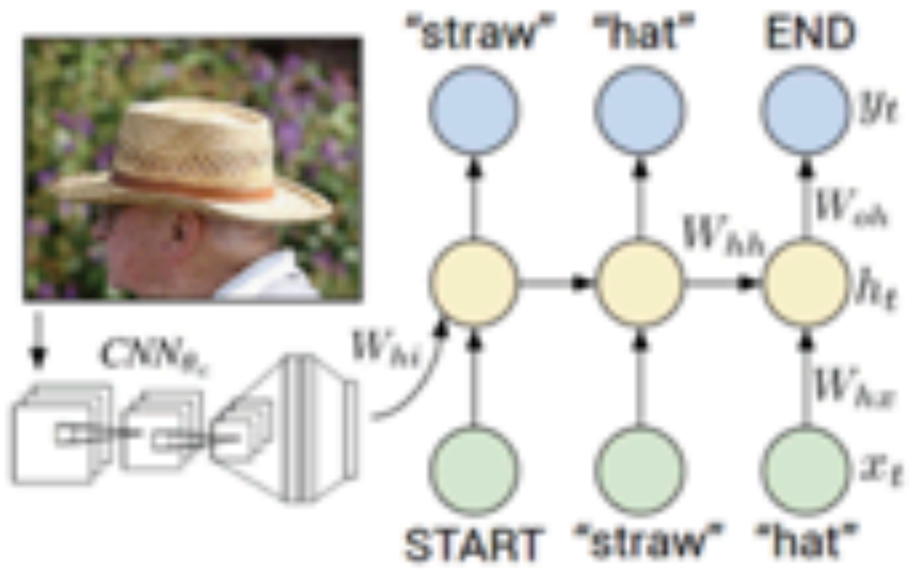
It was very popular a few years ago



Vinyals et al., 2015



Donahue et al., 2015



Karpathy and Fei-Fei, 2015

Hodosh et al., 2013

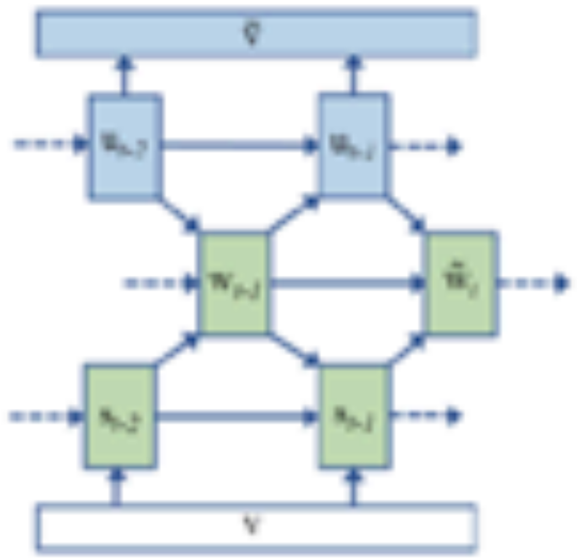
Fang et al., 2015

... and many more

Mao et al., 2015

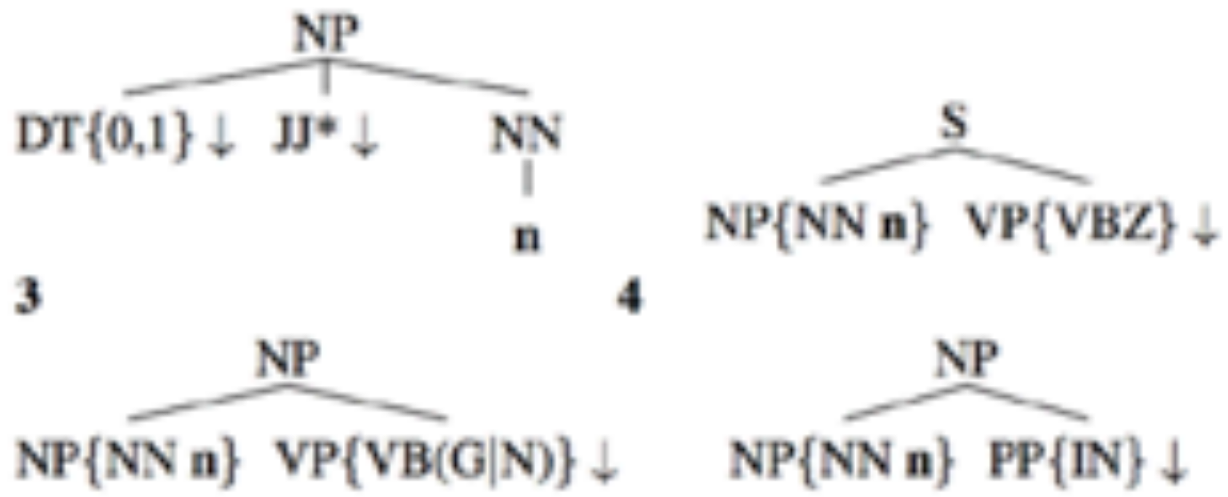
Ordonez et al., 2011

Kulkarni et al., 2011

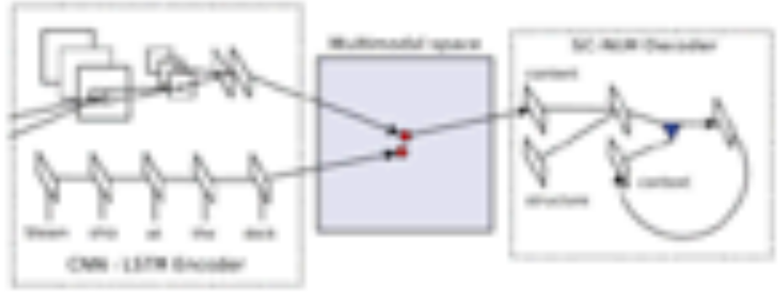


Chen and Zitnick, 2015

Farhadi et al., 2010



Mitchell et al., 2012

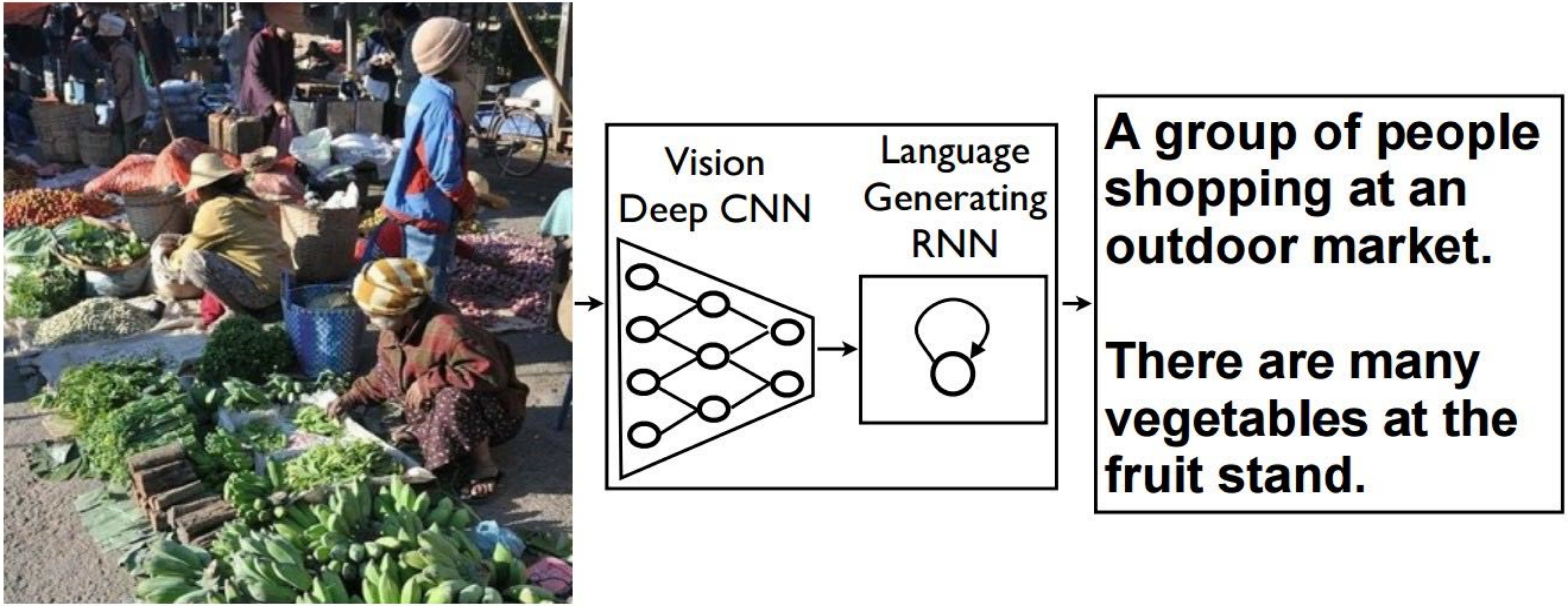


Kiros et al., 2015

Slide credit: Devi Parikh

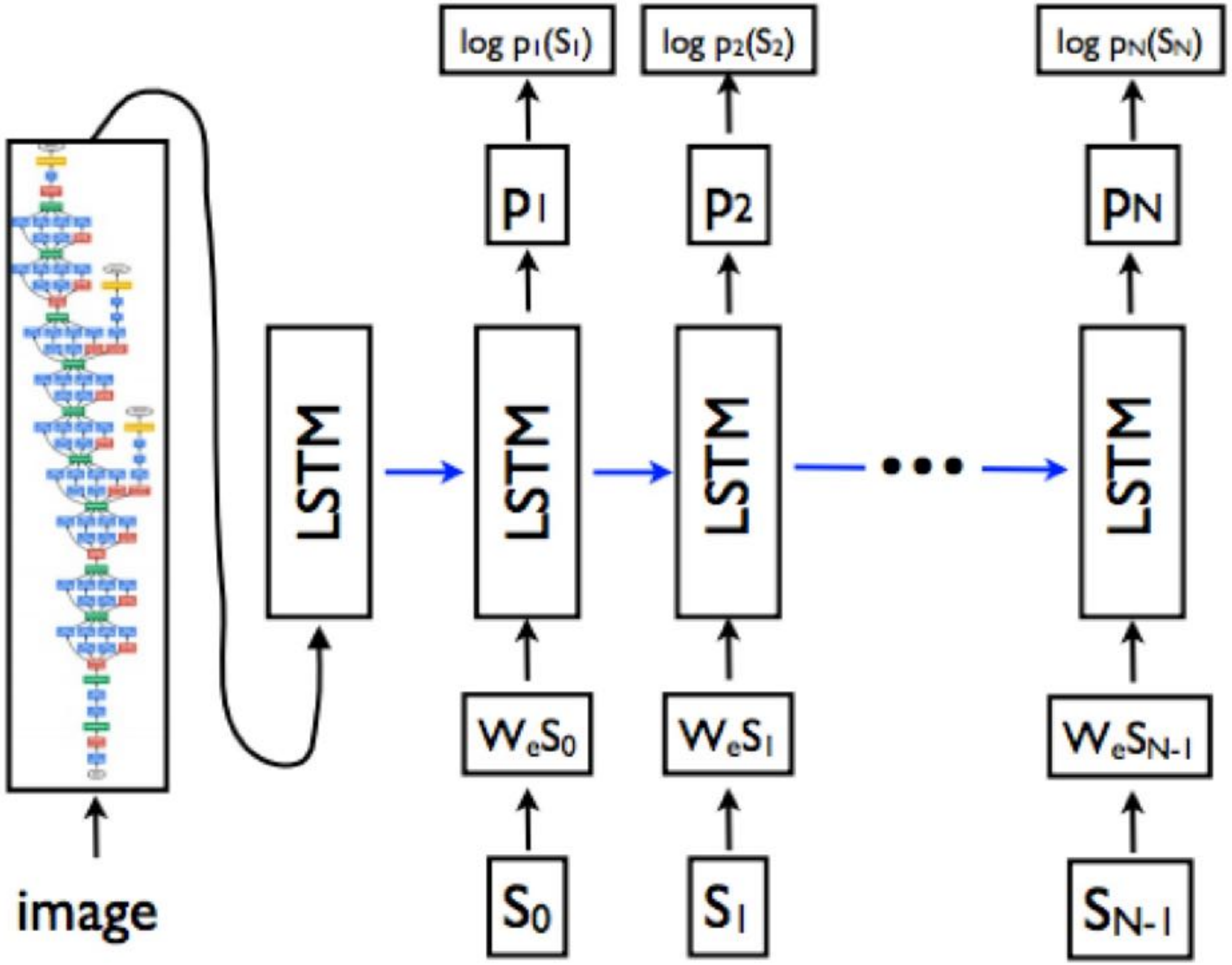
Show and Tell: A Neural Image Caption Generator

[Vinyals et. al., CVPR 2015]



Show and Tell: A Neural Image Caption Generator

[Vinyals et. al., CVPR 2015]

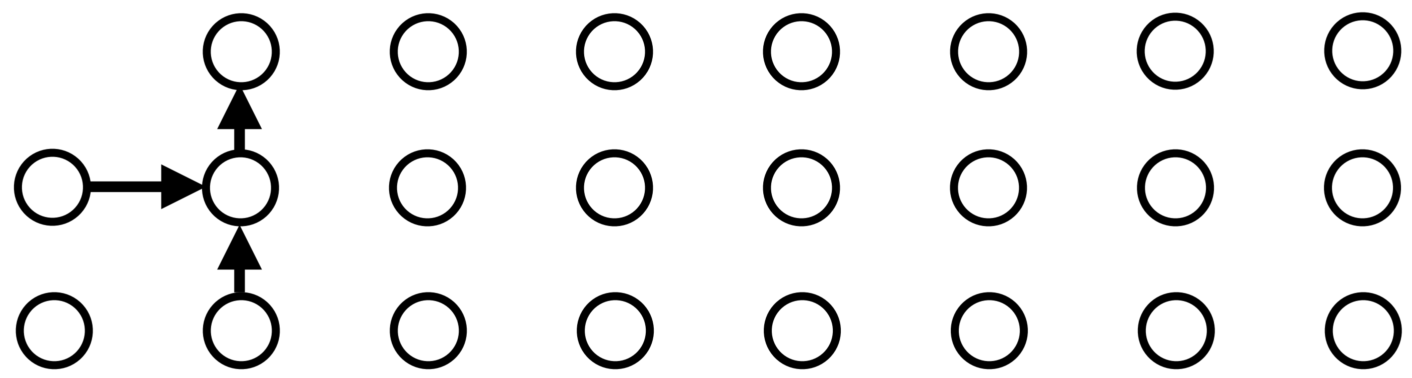


A person riding a motorcycle on a dirt road.

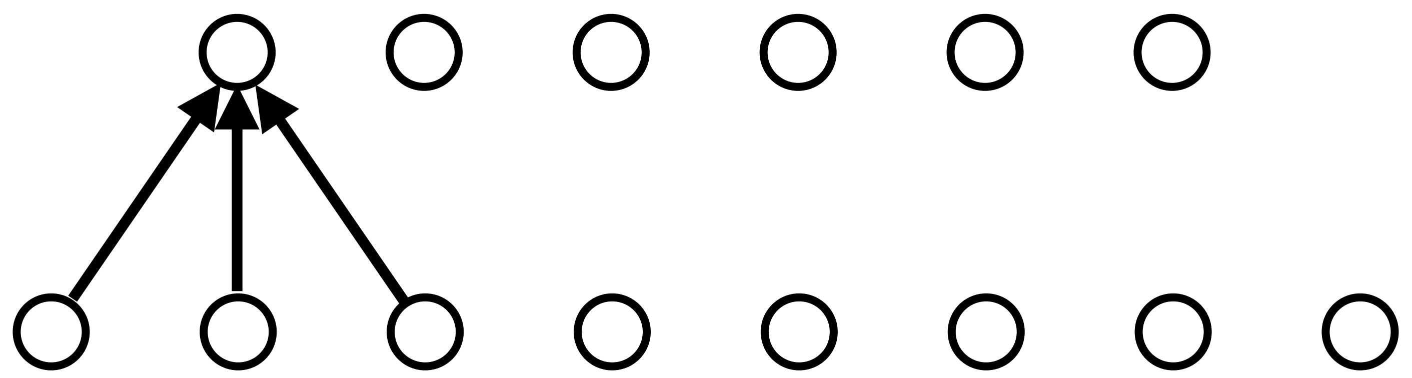

A person riding a motorcycle on a dirt road.
Two dogs play in the grass.
A skateboarder does a trick on a ramp.
A dog is jumping to catch a frisbee.
A group of young people playing a game of frisbee.
Two hockey players are fighting over the puck.
A little girl in a pink hat is blowing bubbles.
A refrigerator 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

Methods for handling unknown length sequences

- Parameter sharing
  - RNNs — recurrent weights are shared across time

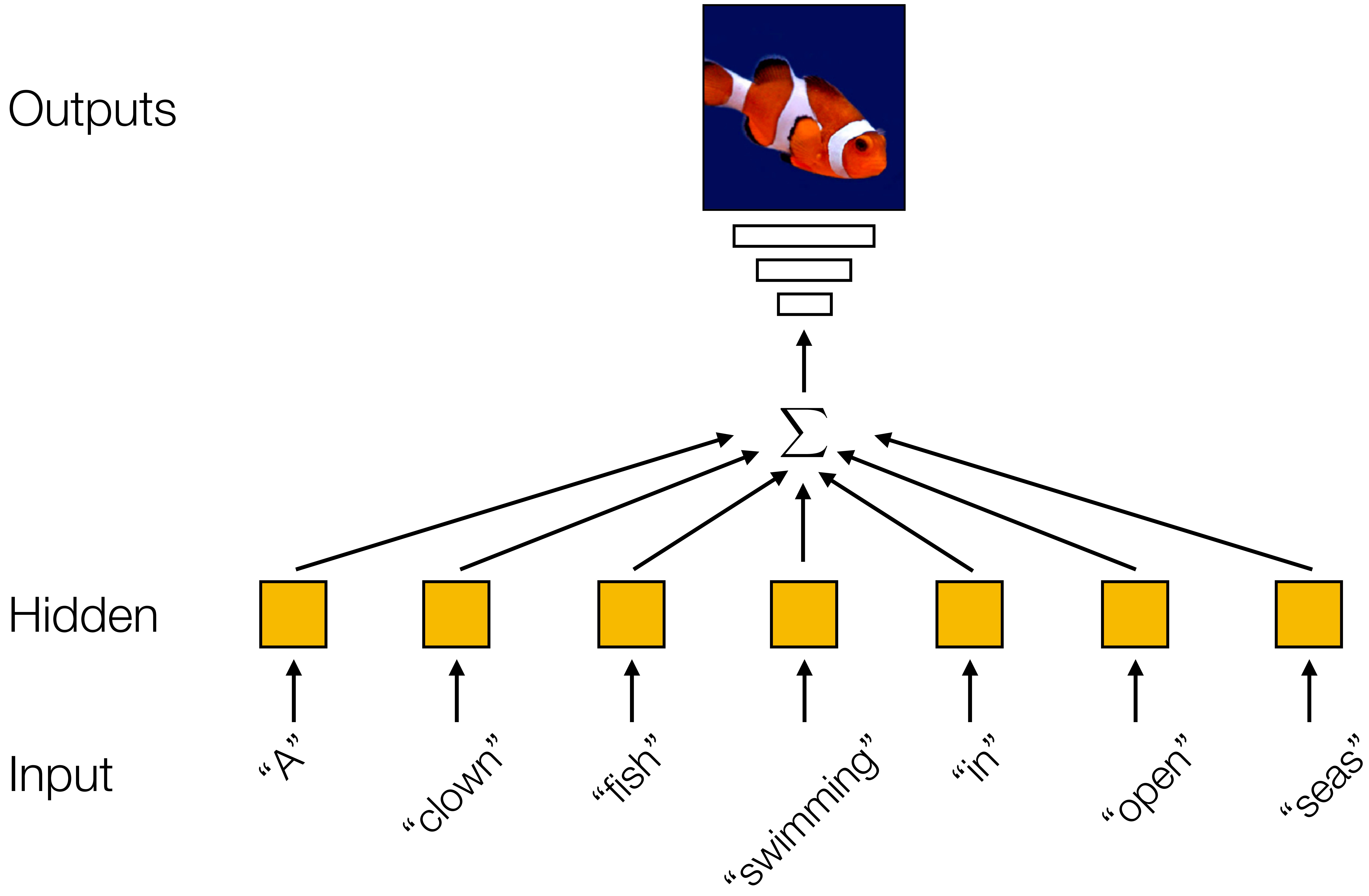


  - Convolution — conv weights are shared across time



- Pooling (e.g., attention)

Pooling



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”

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”

Attention


Outputs
"Un"
"poisson"
"clown"
"nageant"
"en"
"haute"
"mer"
LSTM
LSTM
LSTM
LSTM
LSTM
LSTM
LSTM
$\sum$
Hidden
Input
"A"
"clown"
"fish"
"swimming"
"in"
"open"
"seas"

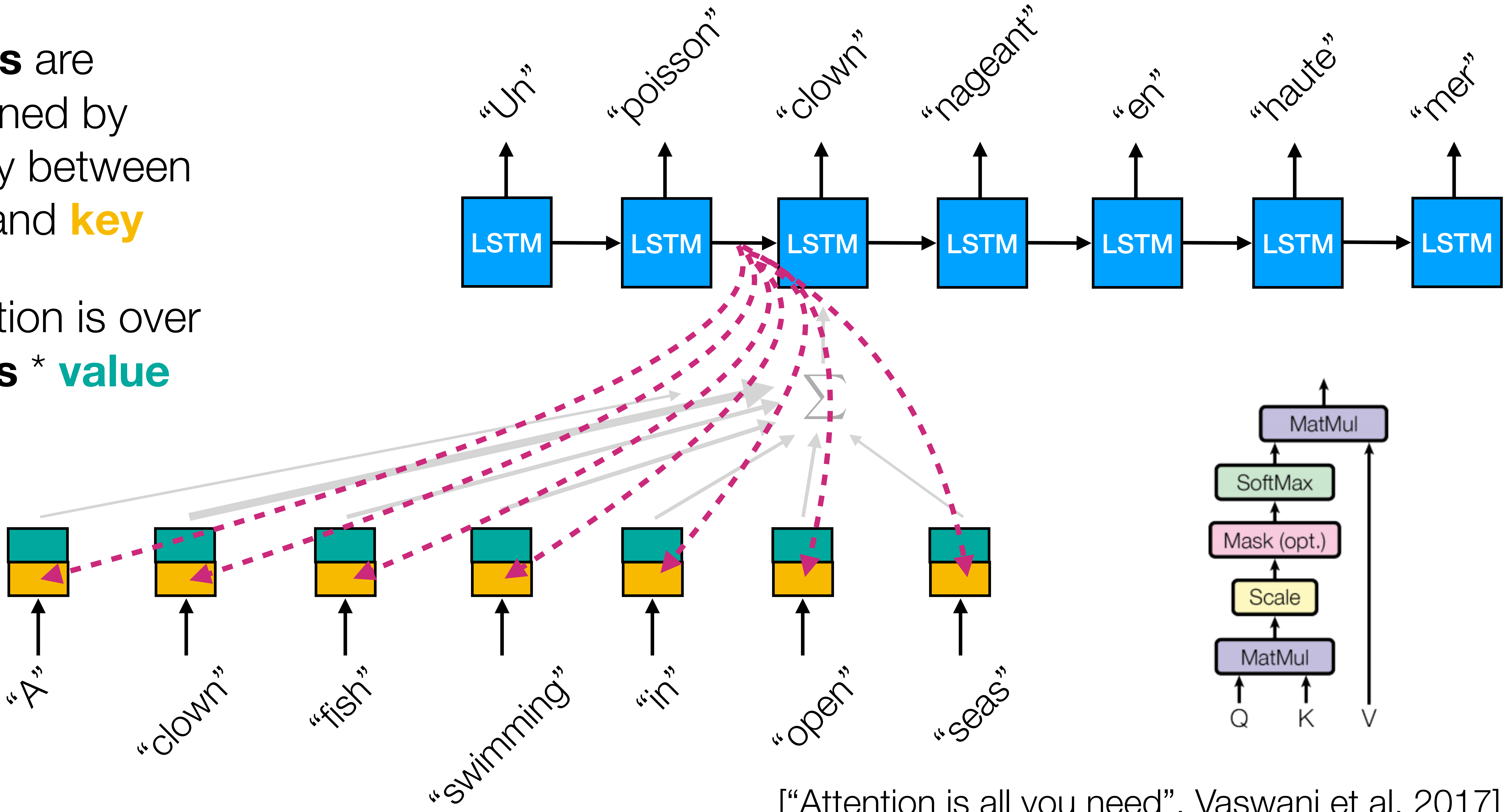
Attention

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”

Attention

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

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



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

arXiv:1706.03762v5 [cs.CL] 6 Dec 2017

Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com
Noam Shazeer* Google Brain noam@google.com
Niki Parmar* Google Research nikip@google.com
Jakob Uszkoreit* Google Research usz@google.com
Llion Jones* Google Research llion@google.com
Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu
Łukasz Kaiser* Google Brain lukaszkaiser@google.com
Illia Polosukhin* ‡ illia.polosukhin@gmail.com

Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.8 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature. We show that the Transformer generalizes well to other tasks by applying it successfully to English constituency parsing both with large and limited training data.

1 Introduction

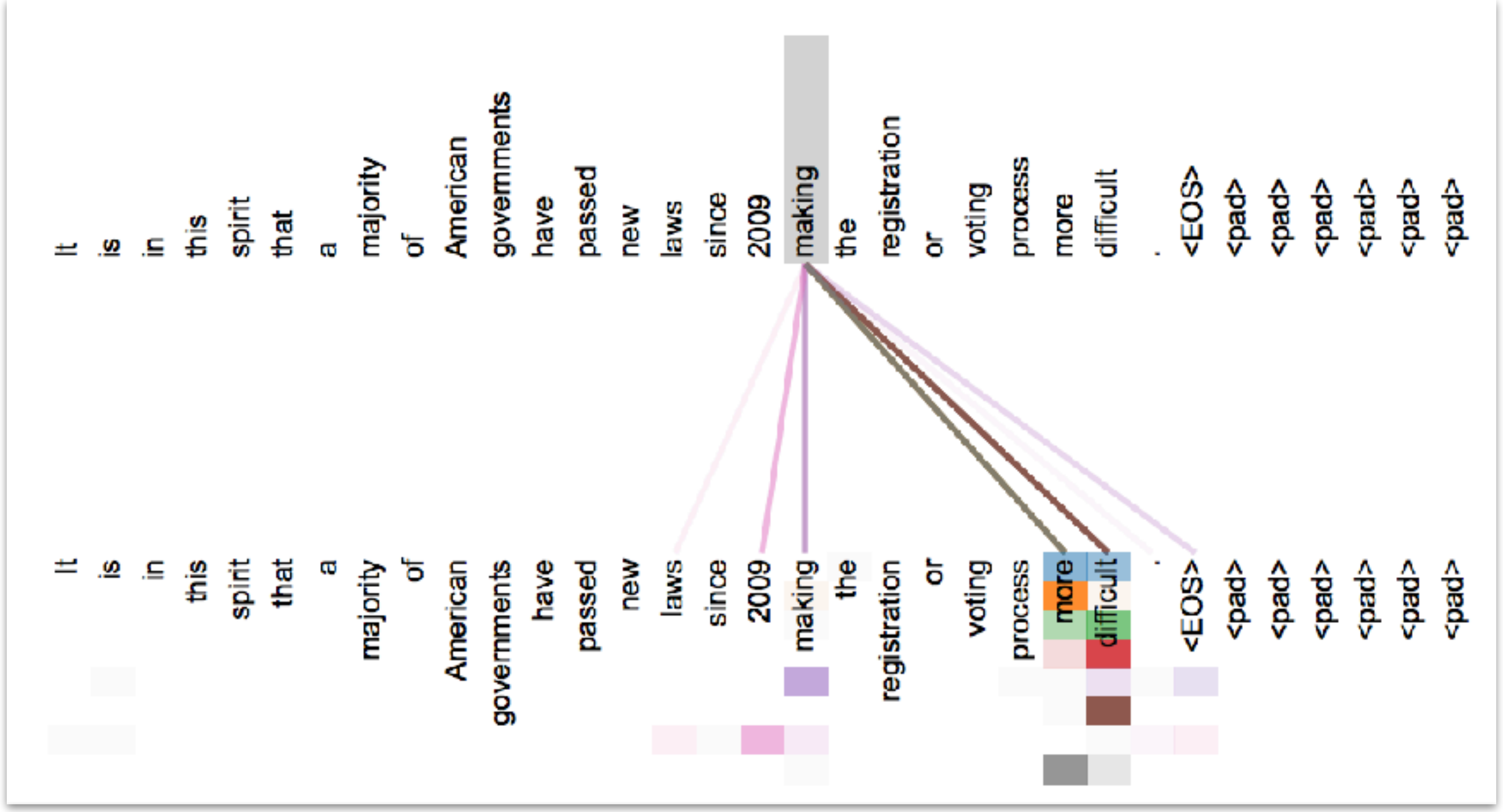
Recurrent neural networks, long short-term memory [13] and gated recurrent [7] neural networks in particular, have been firmly established as state of the art approaches in sequence modeling and

*Equal contribution. Listing order is random. Jakob proposed replacing RNNs with self-attention and started the effort to evaluate this idea. Ashish, with Illia, designed and implemented the first Transformer models and has been crucially involved in every aspect of this work. Noam proposed scaled dot-product attention, multi-head attention and the parameter-free position representation and became the other person involved in nearly every detail. Niki designed, implemented, tuned and evaluated countless model variants in our original codebase and tensor2tensor. Llion also experimented with novel model variants, was responsible for our initial codebase, and efficient inference and visualizations. Lukasz and Aidan spent countless long days designing various parts of and implementing tensor2tensor, replacing our earlier codebase, greatly improving results and massively accelerating our research.

†Work performed while at Google Brain.

‡Work performed while at Google Research.

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.



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

VQA: Visual Question Answering

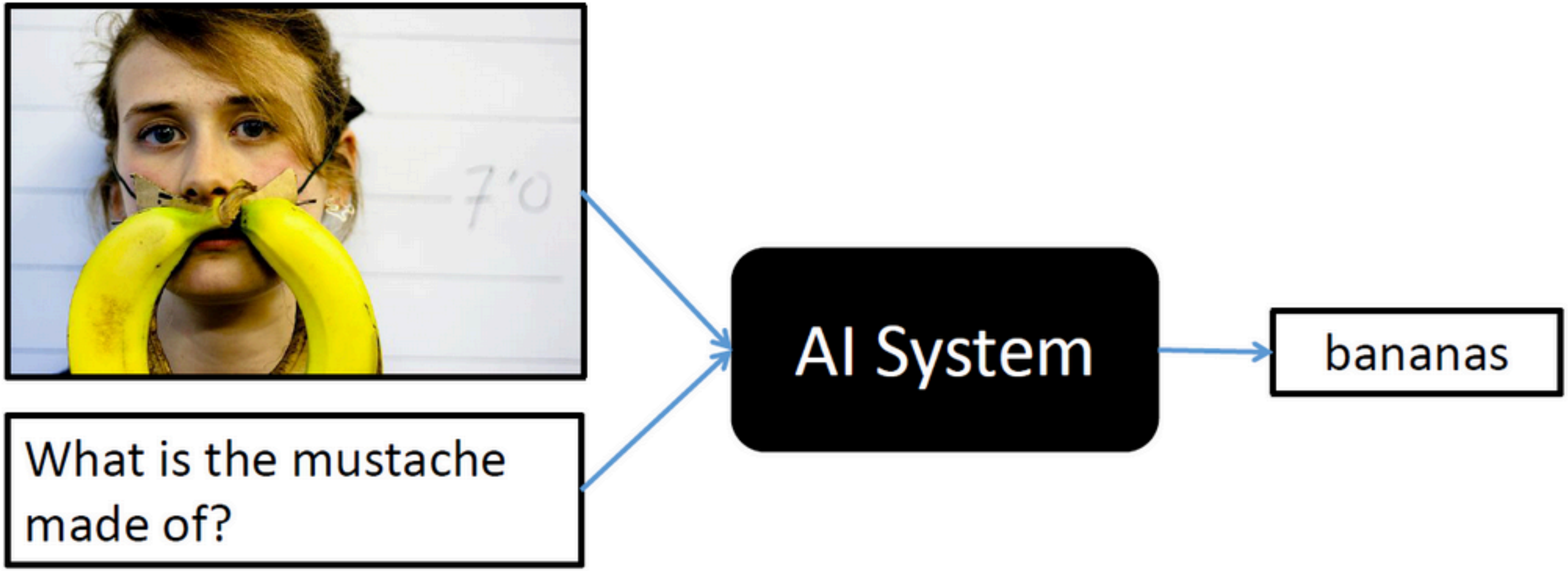
www.visualqa.org

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

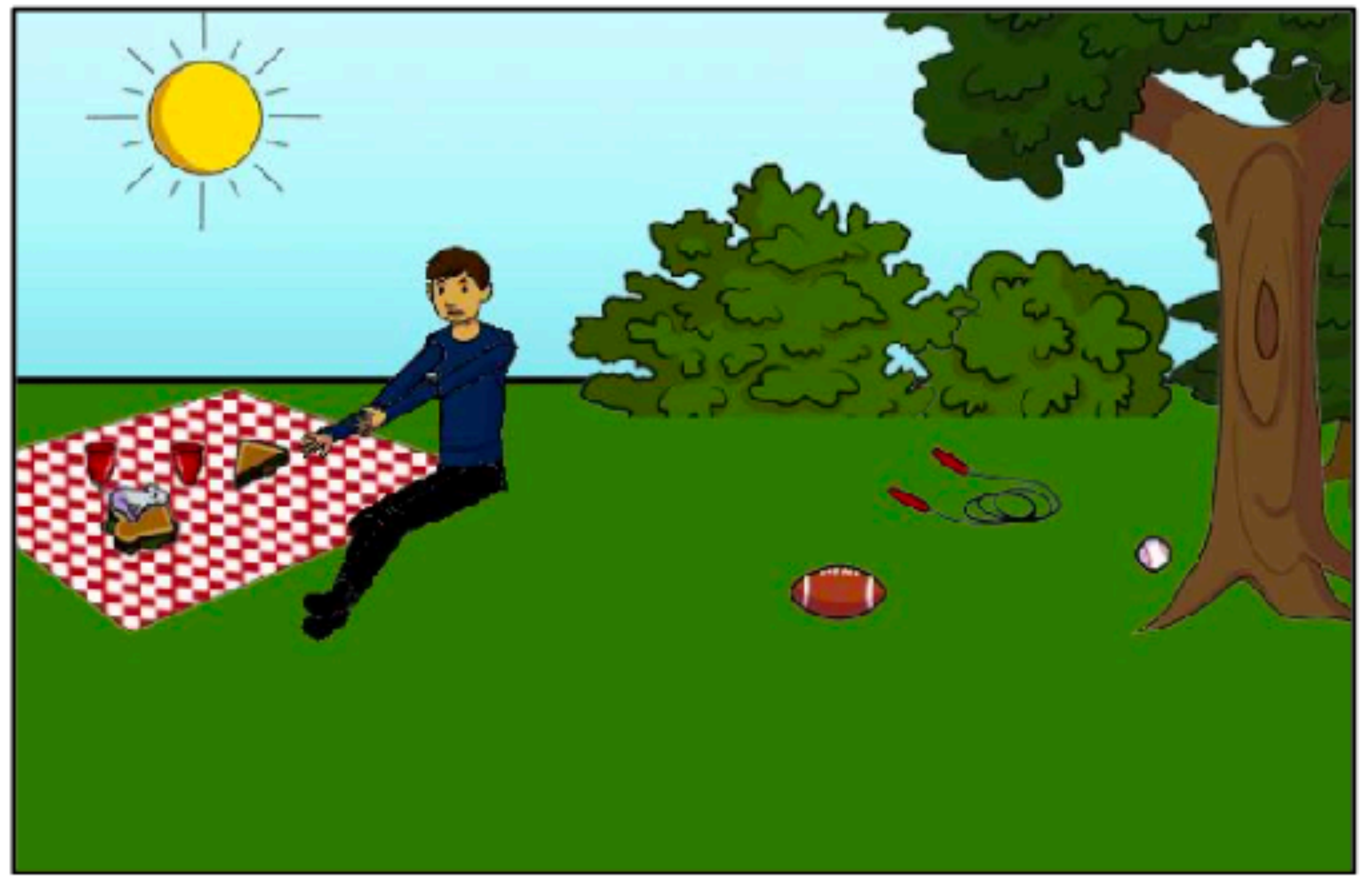
[https://arxiv.org/pdf/1505.00468v6.pdf]



[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?



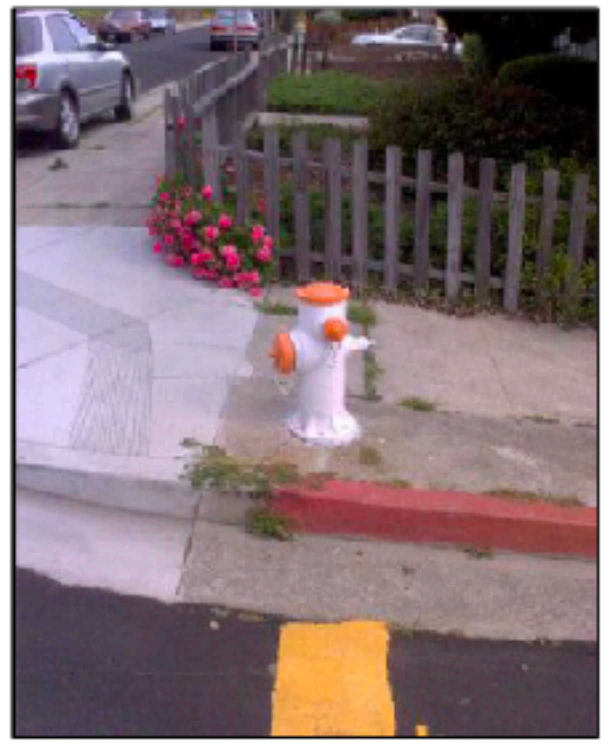
Is this person expecting company?
What is just under the tree?

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

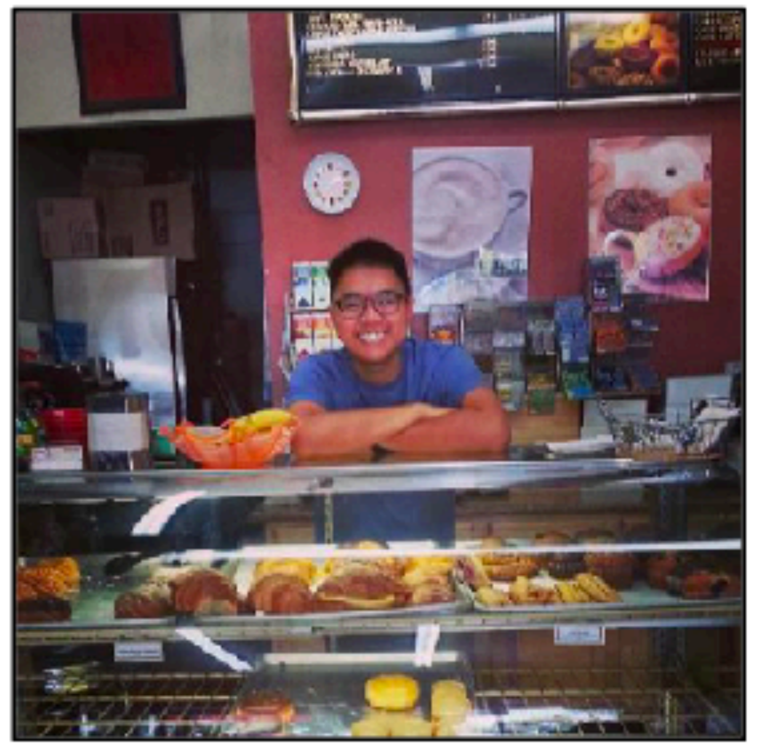
Fig. 1: Examples of free-form, open-ended questions collected for images via Amazon Mechanical Turk. Note that commonsense knowledge is needed along with a visual understanding of the scene to answer many questions.

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

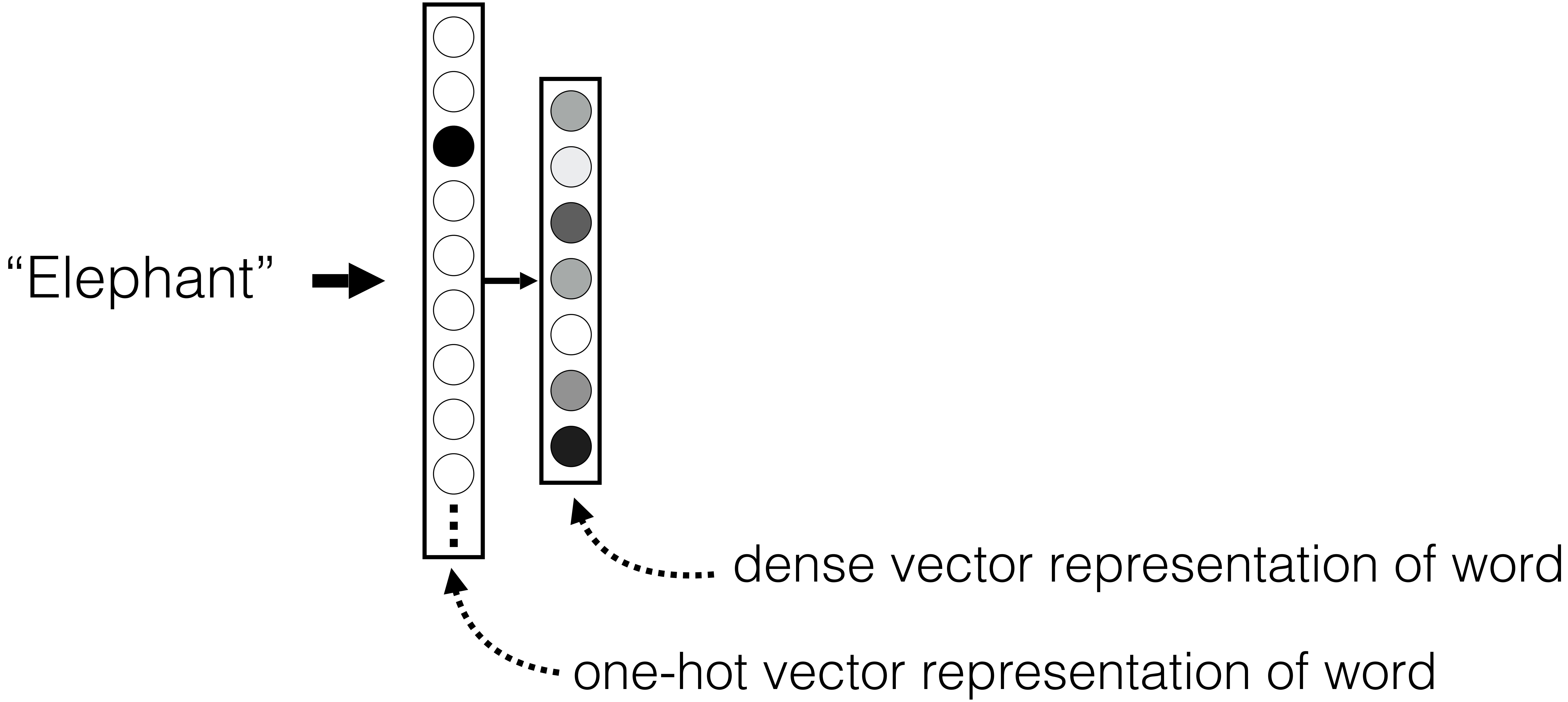
Has the pizza been baked?	yes yes yes	yes yes yes
What kind of cheese is topped on this pizza?	feta feta ricotta	mozzarella mozzarella mozzarella

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

word2vec

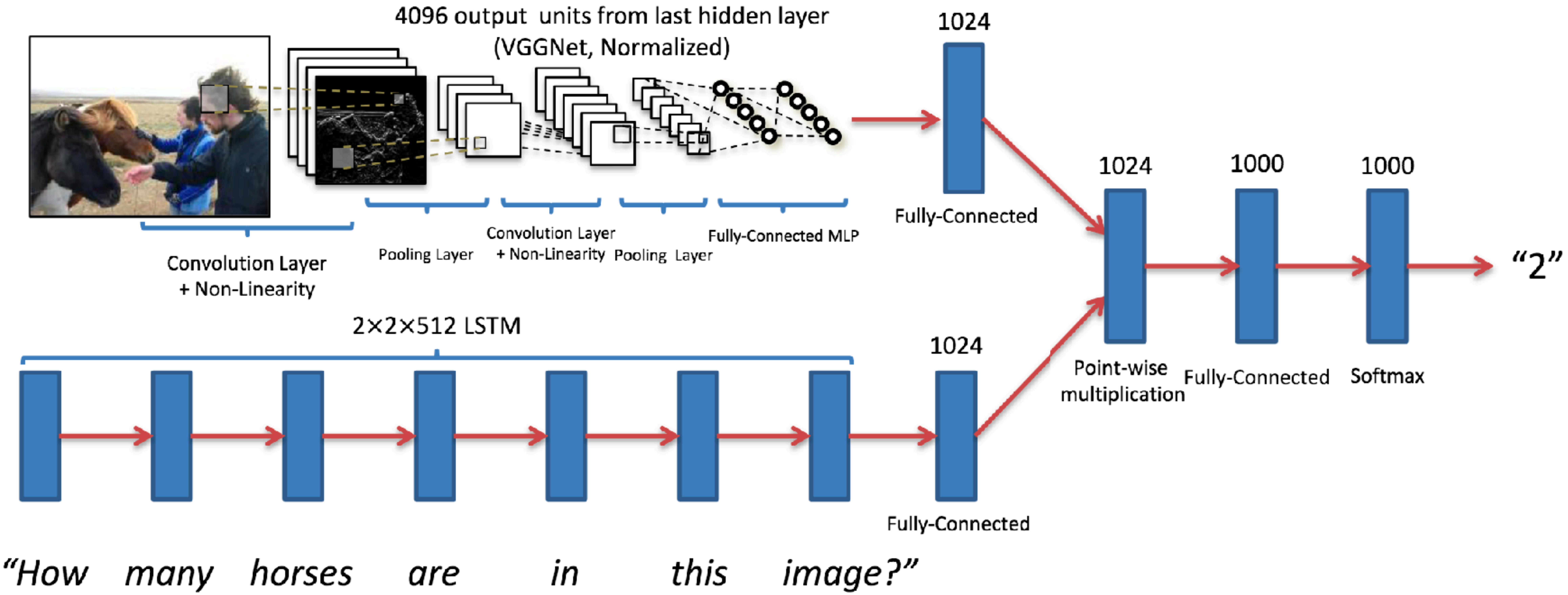


X2vec methods are also called embeddings of X, e.g., a **word embedding**

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

Berkeley N L P

Jacob Andreas

(with Dan Klein, Marcus Rohrbach and Trevor Darrell)

[Slides credit: Jacob Andreas]

Grounded question answering

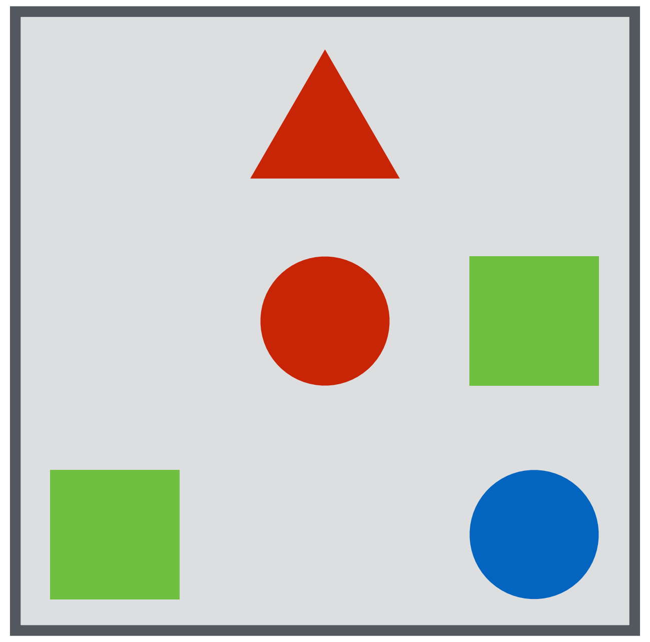
What color is the necktie?

yellow

[Slides credit: Jacob Andreas]

Grounded question answering

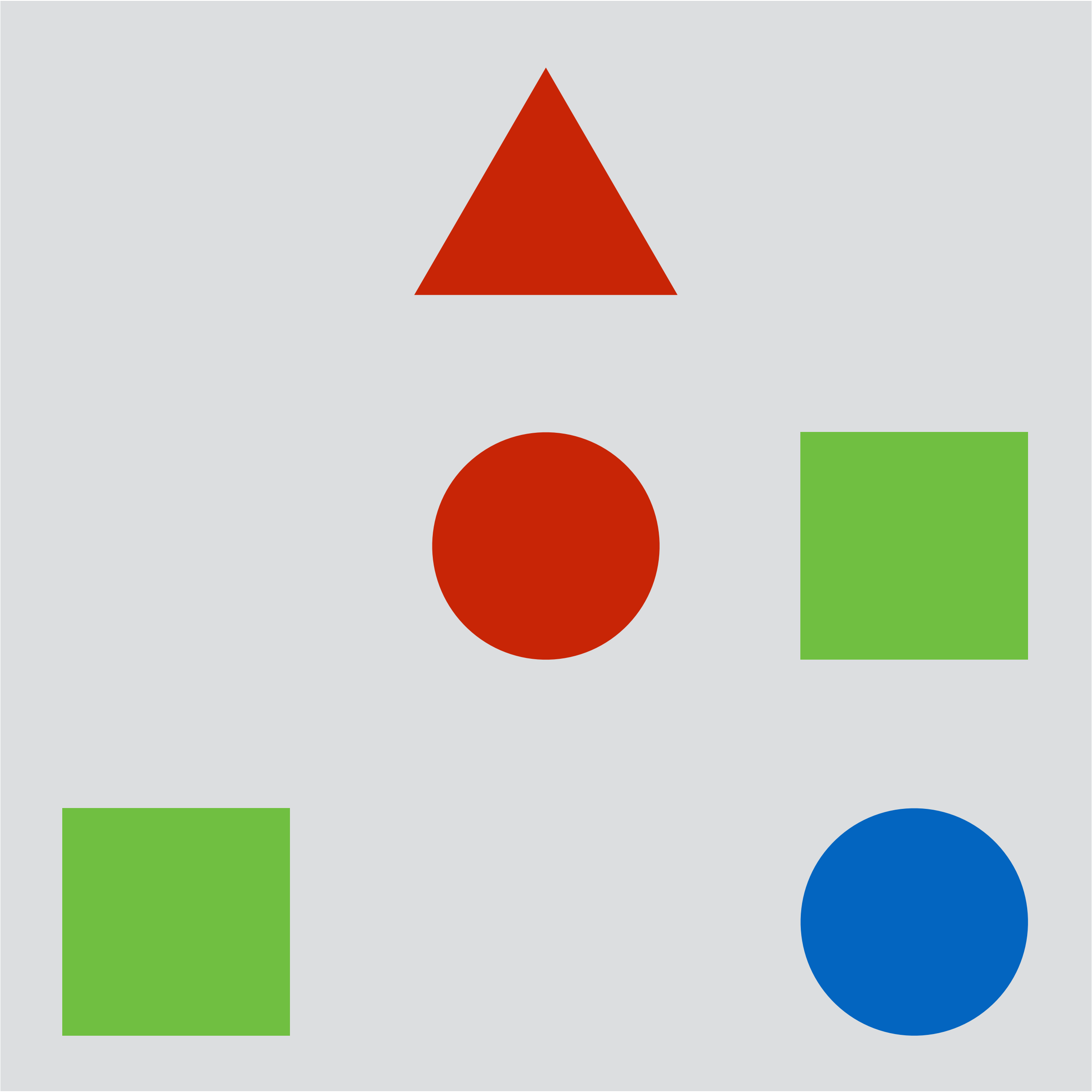
Is there a red shape above a circle?



yes

[Slides credit: Jacob Andreas]

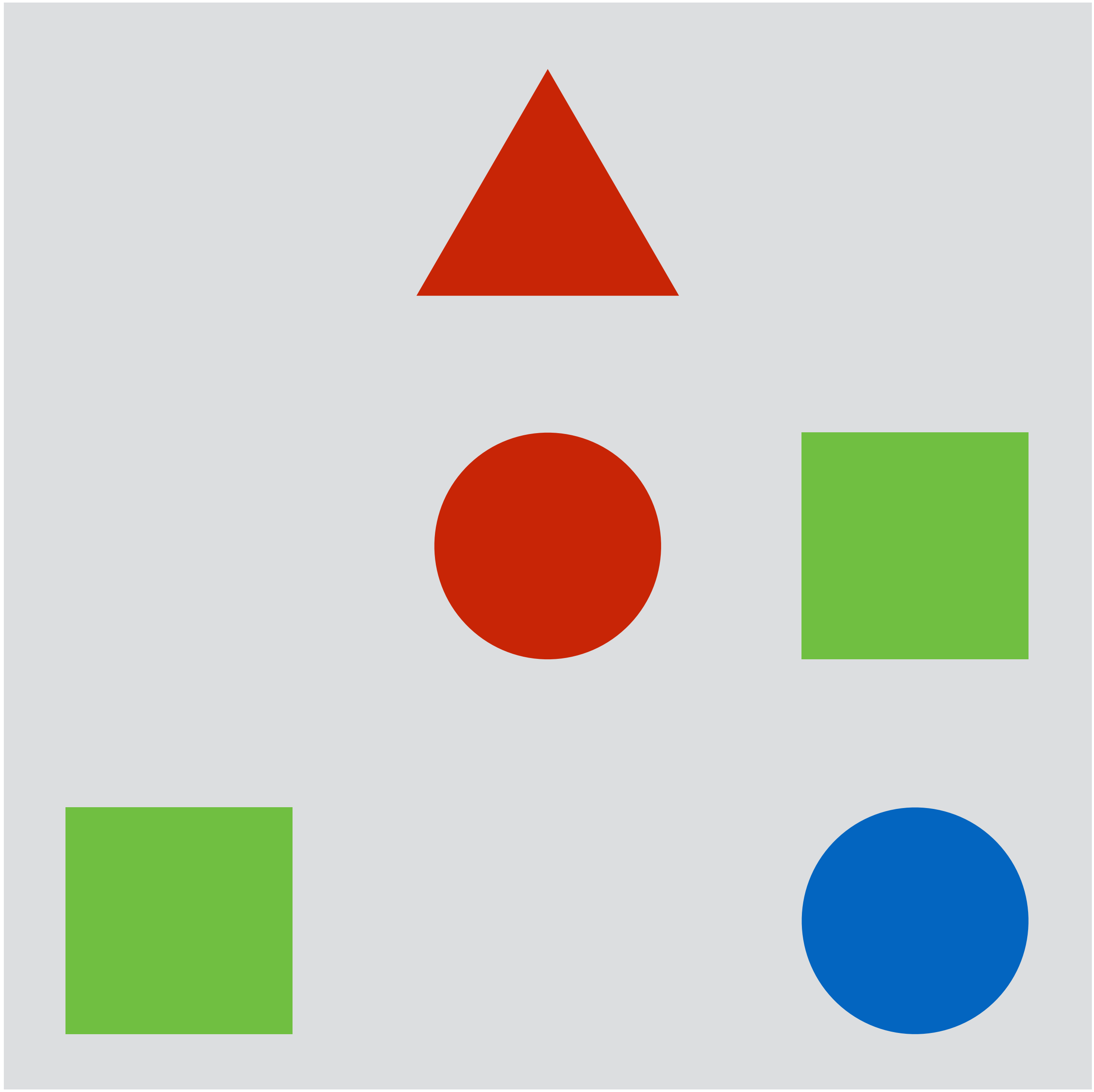
Representing meaning



Is there a red shape above a circle?

[Slides credit: Jacob Andreas]

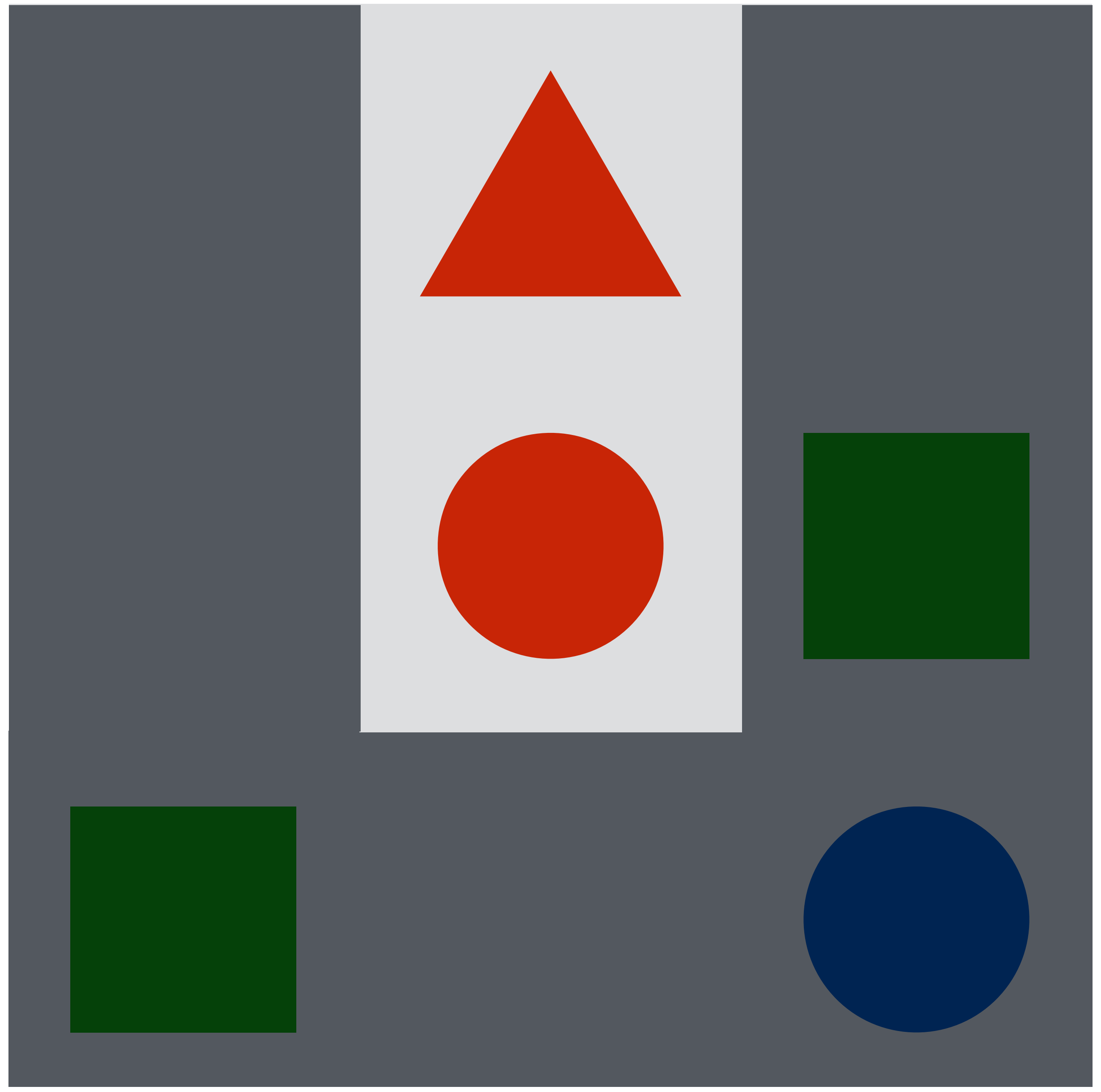
Representing meaning



Is there a red shape above a circle?

[Slides credit: Jacob Andreas]

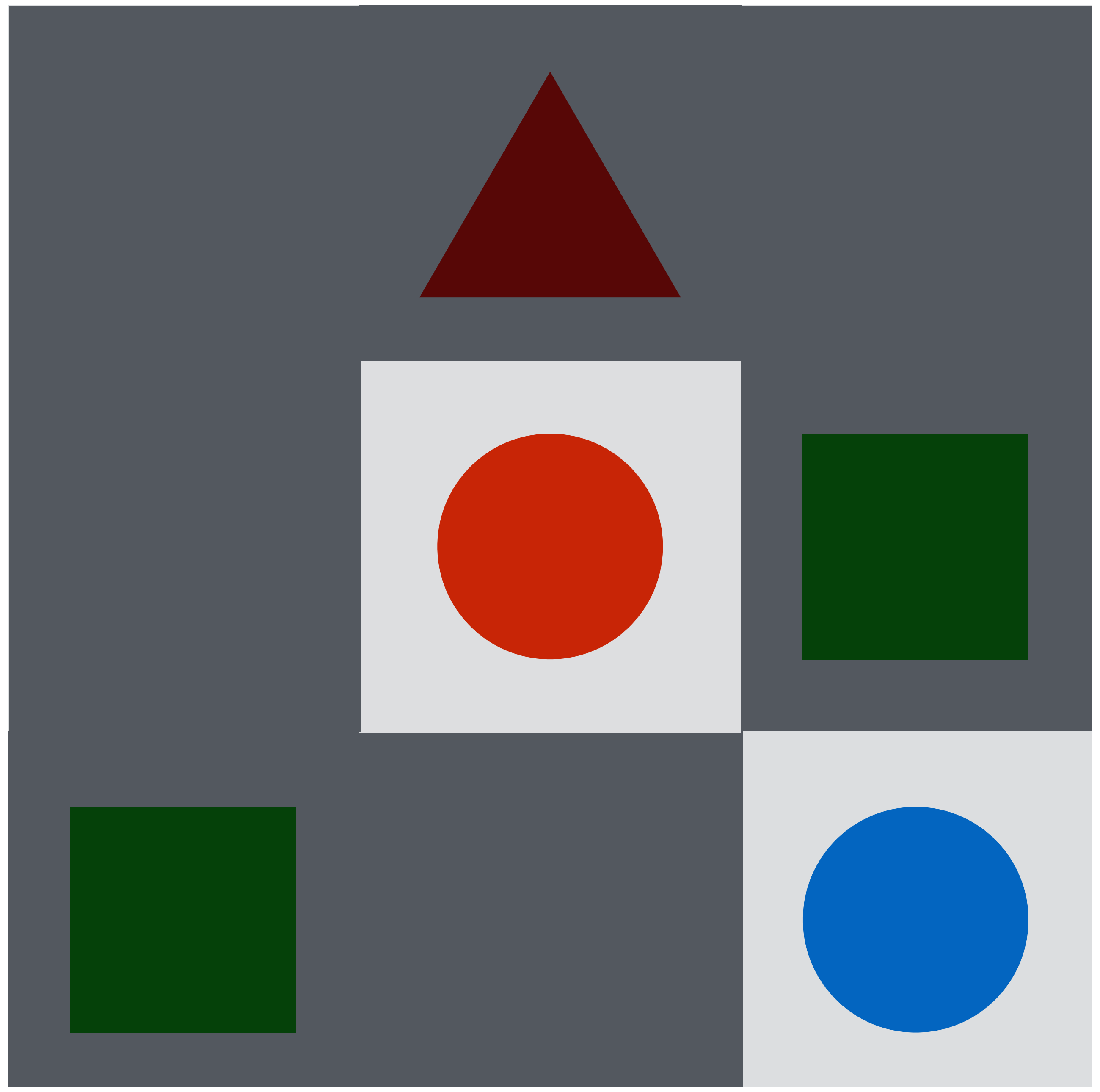
Sets encode meaning



Is there a red shape above a circle?

[Slides credit: Jacob Andreas]

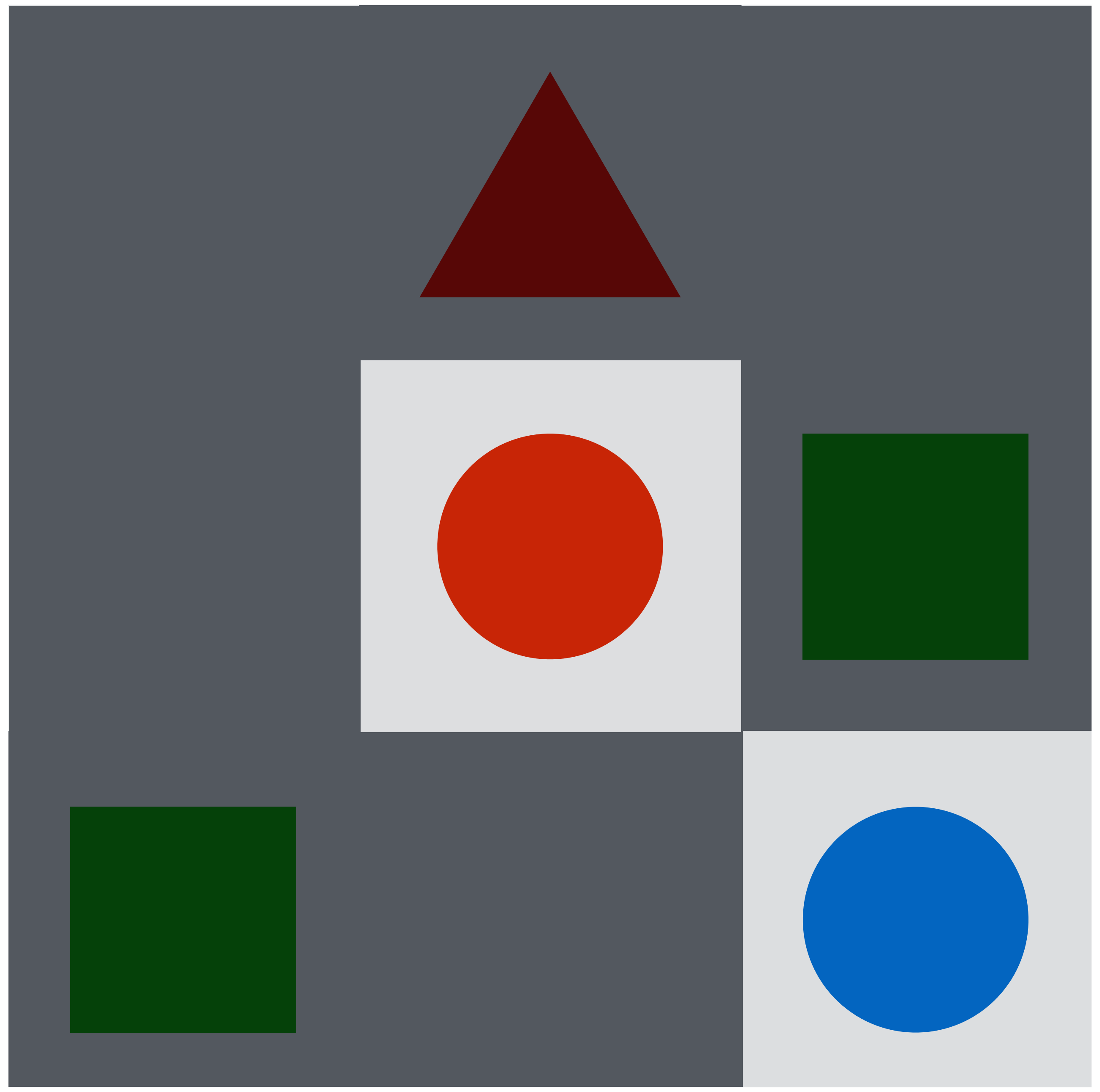
Sets encode meaning



Is there a red shape above a circle?

[Slides credit: Jacob Andreas]

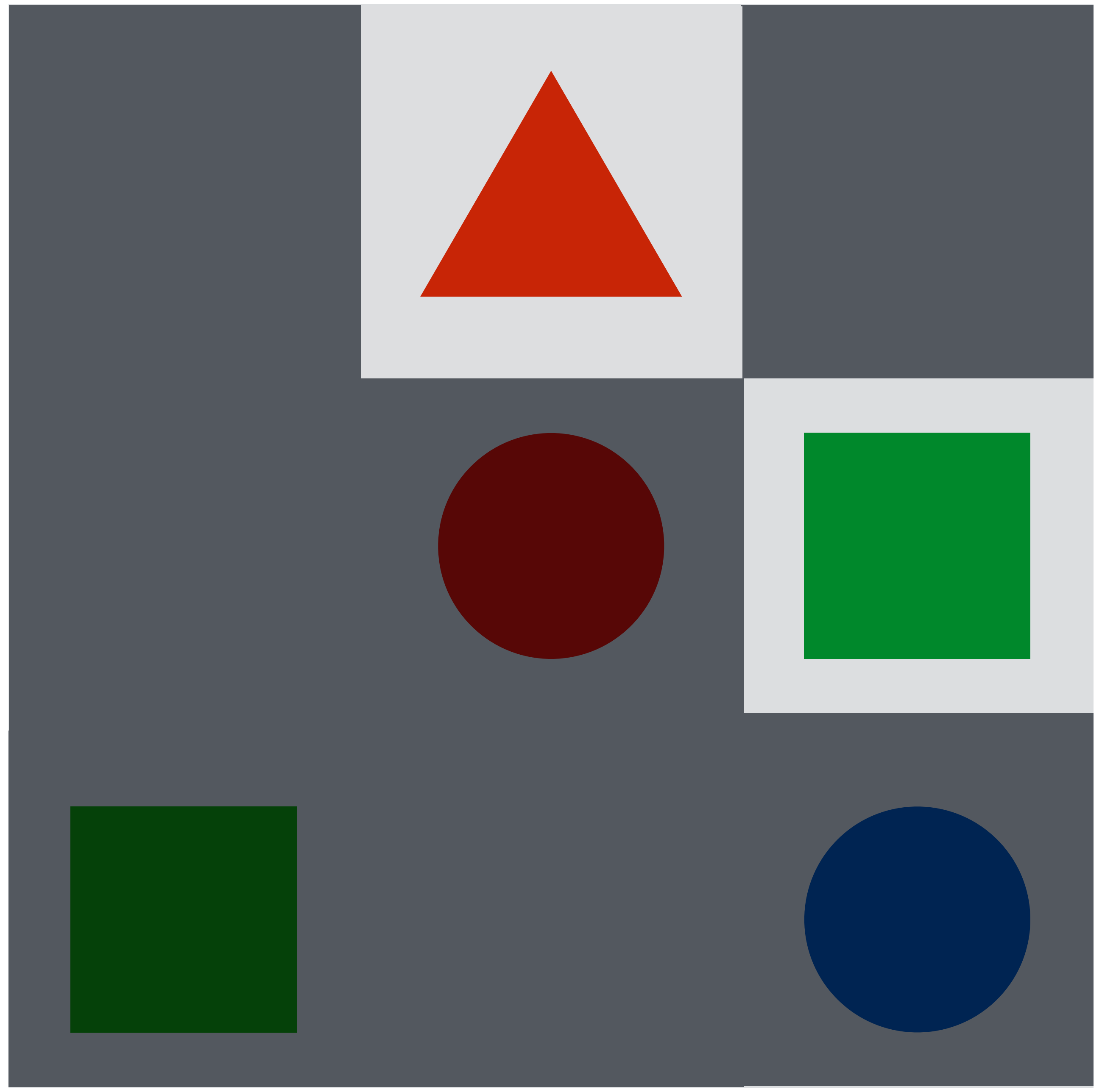
Set transformations encode meaning



Is there a red shape above a circle?

[Slides credit: Jacob Andreas]

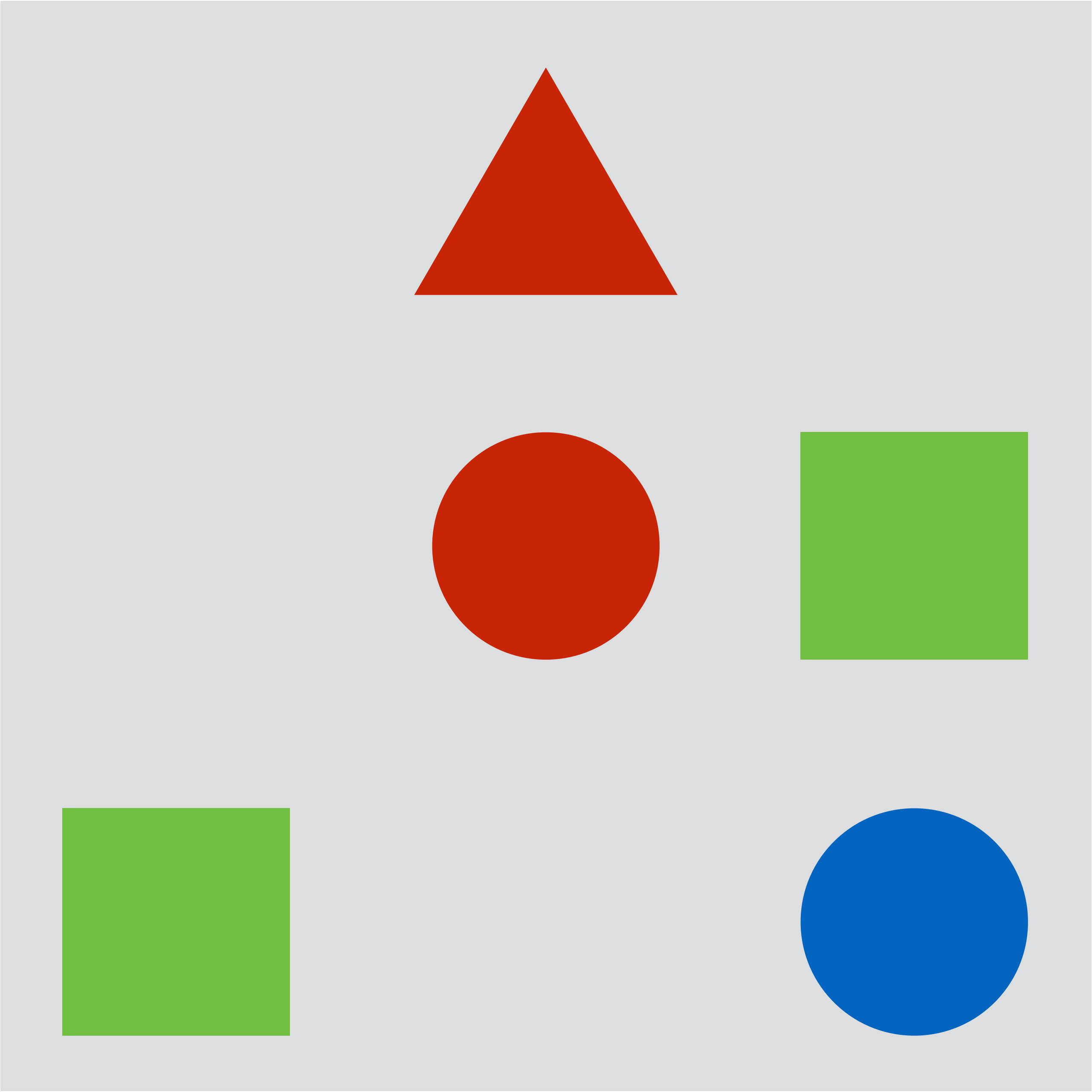
Set transformations encode meaning



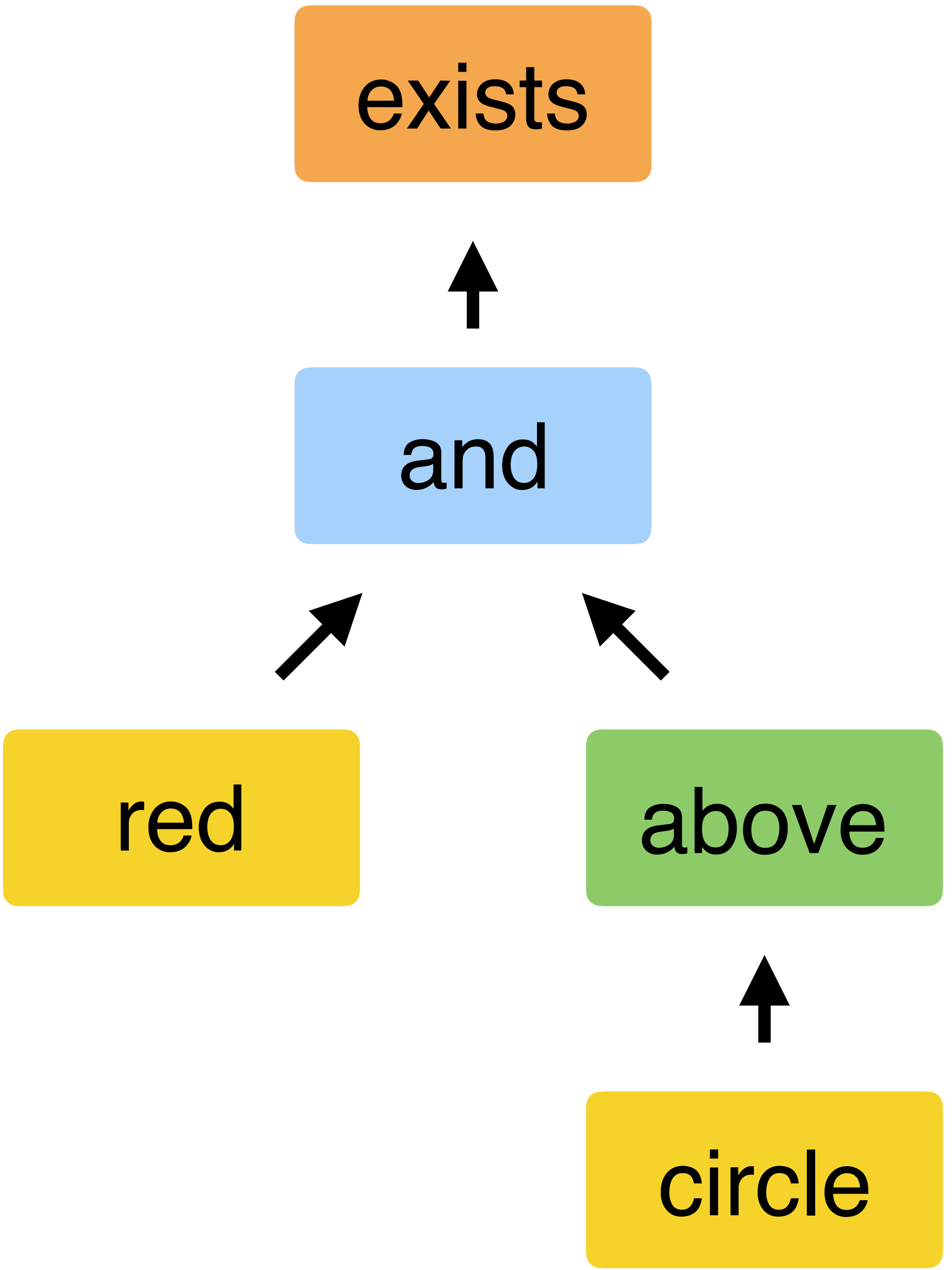
Is there a red shape above a circle?

[Slides credit: Jacob Andreas]

Sentence meanings are computations

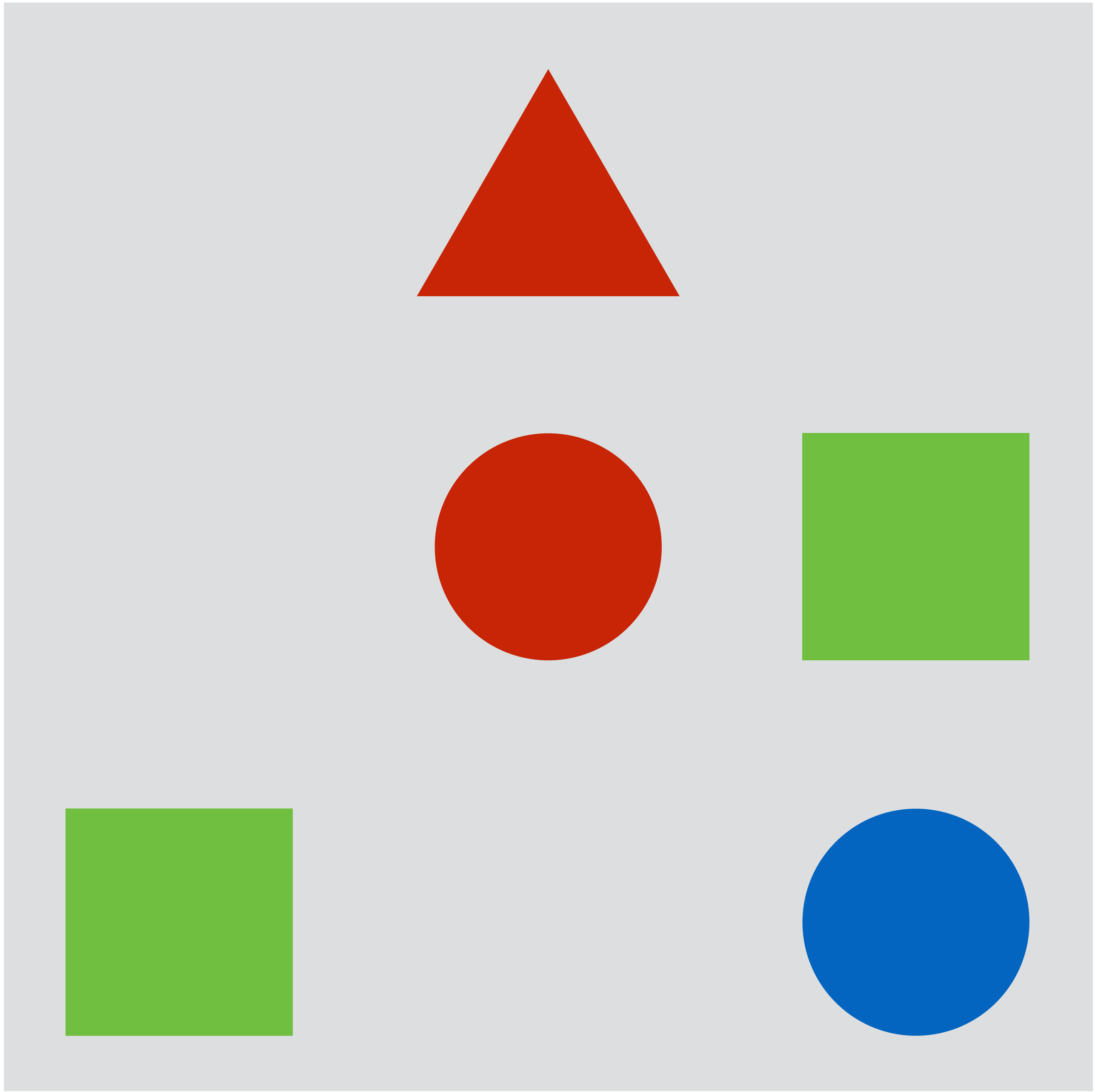


Is there a red shape above a circle?

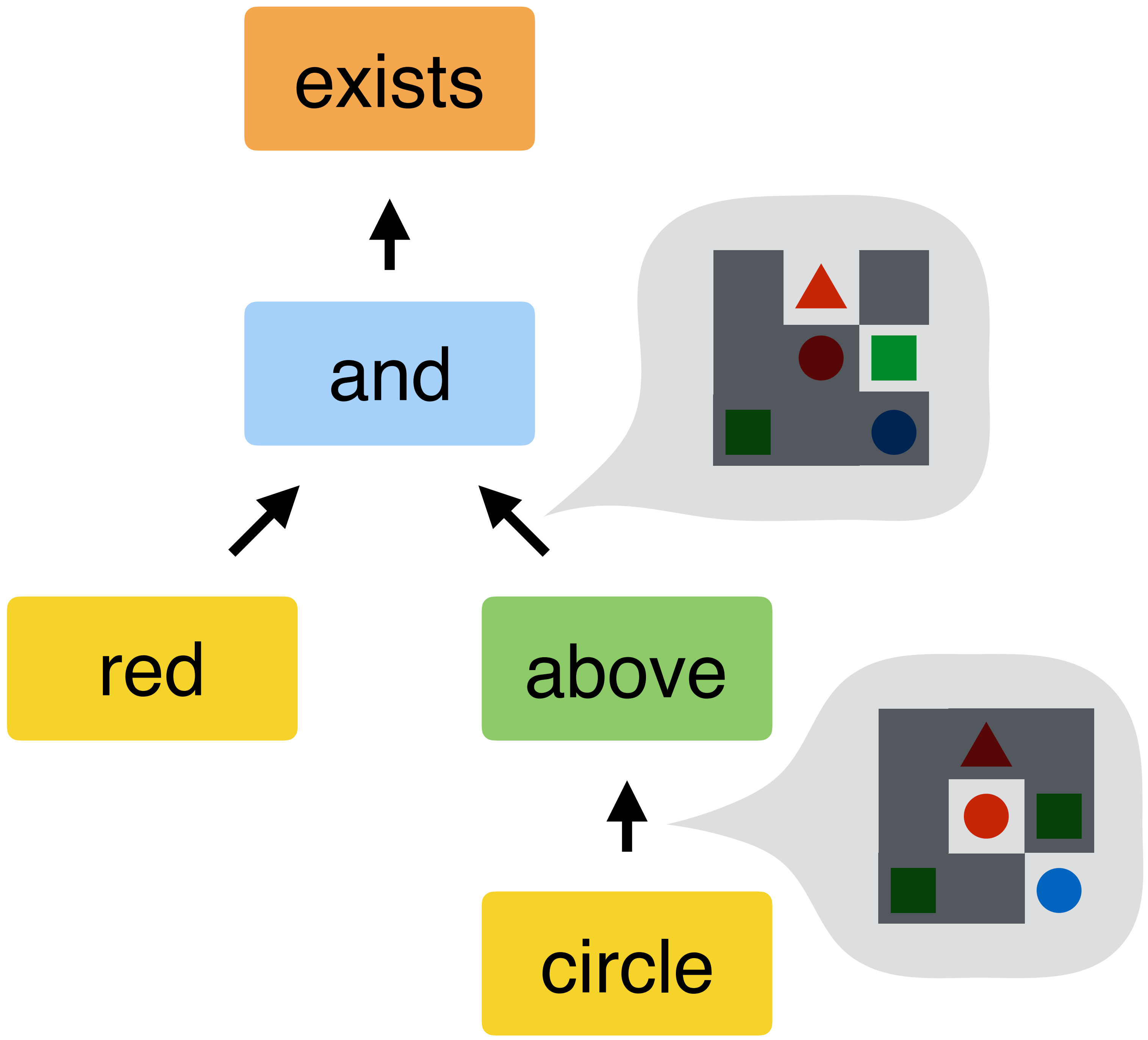


[Slides credit: Jacob Andreas]

Sentence meanings are computations

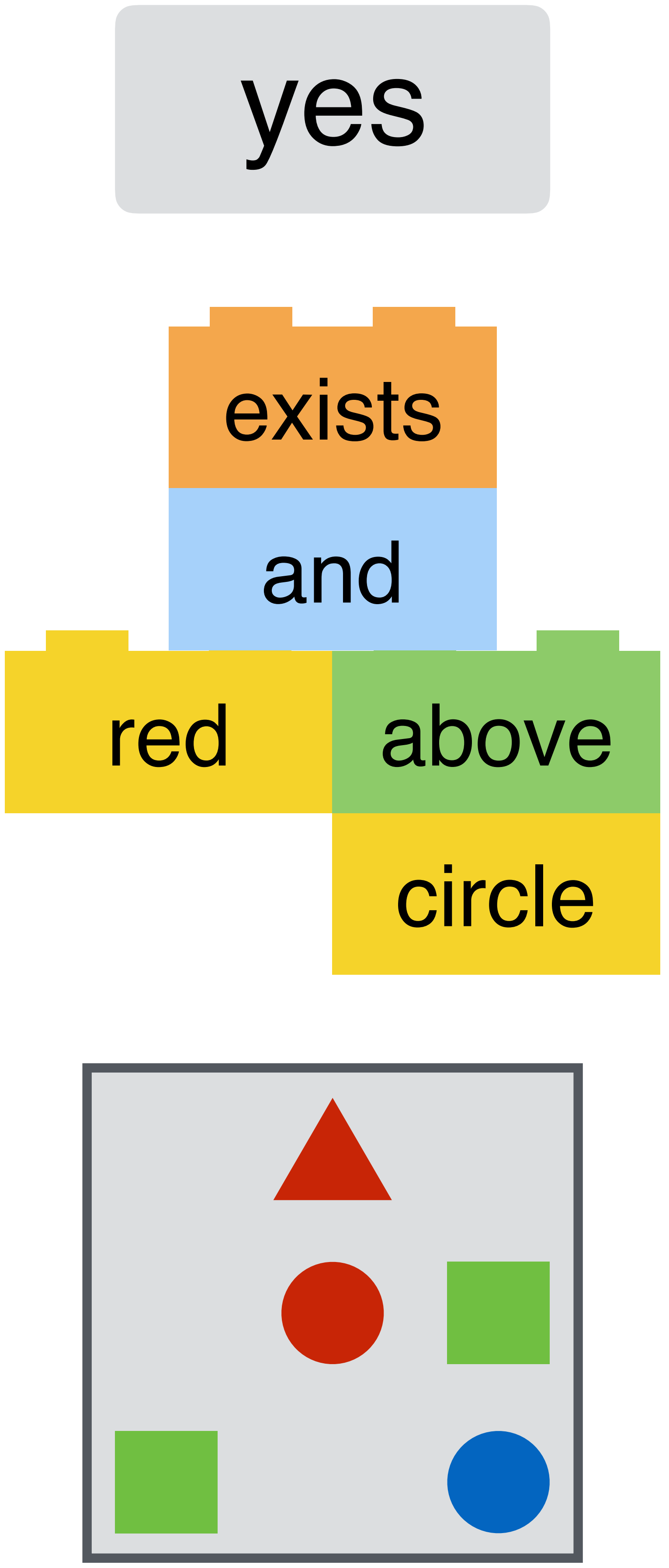


Is there a red shape above a circle?

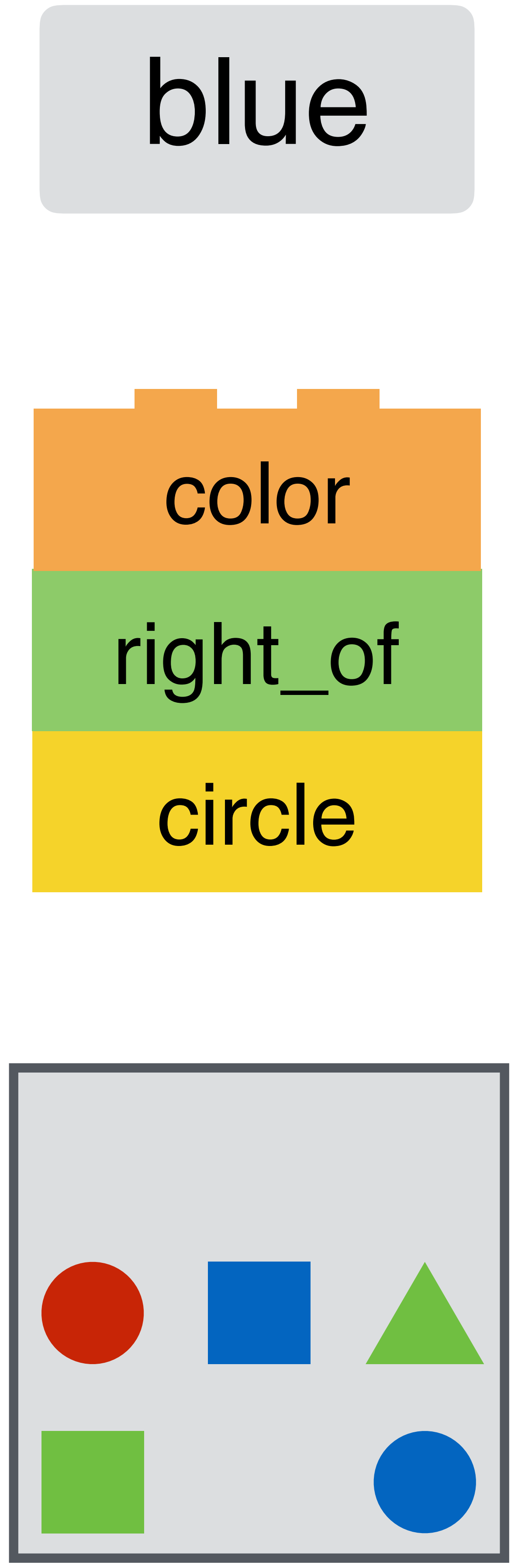


[Slides credit: Jacob Andreas]

Learning

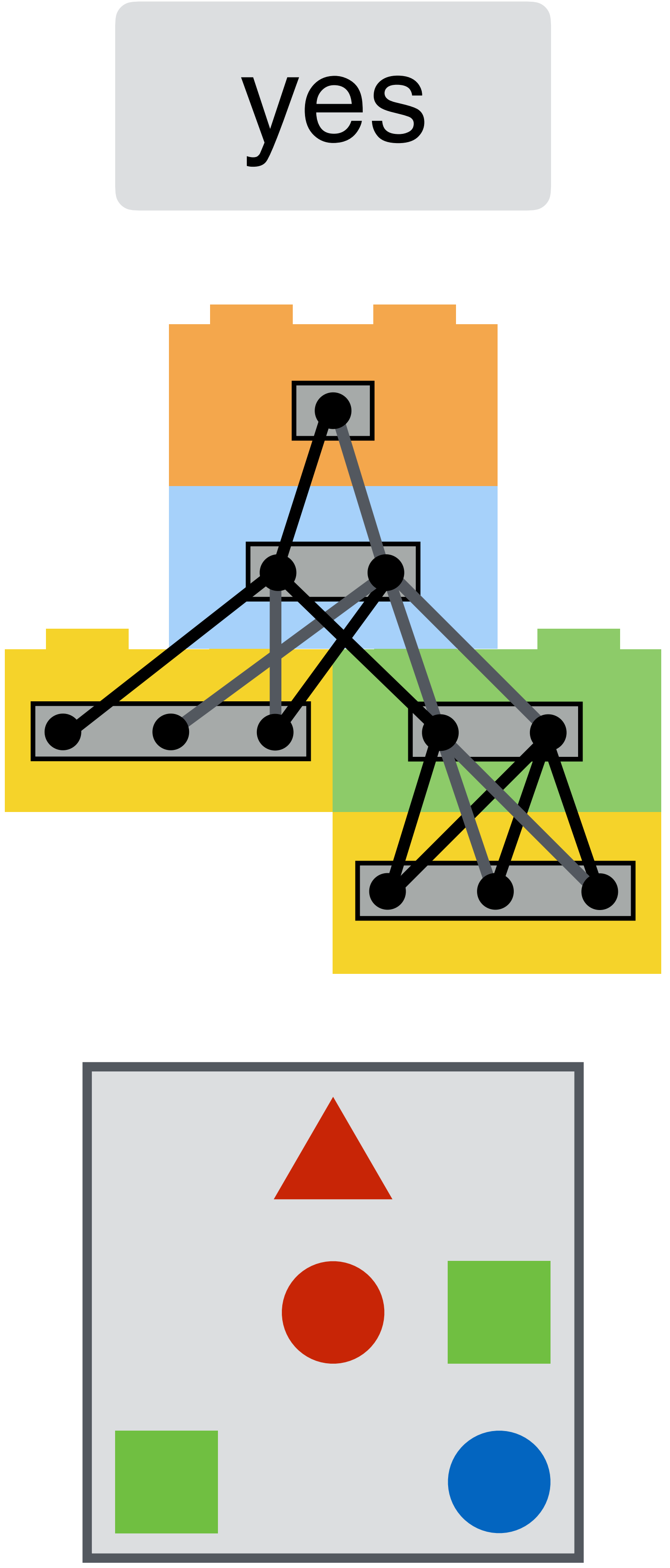


Is there a red shape above a circle?

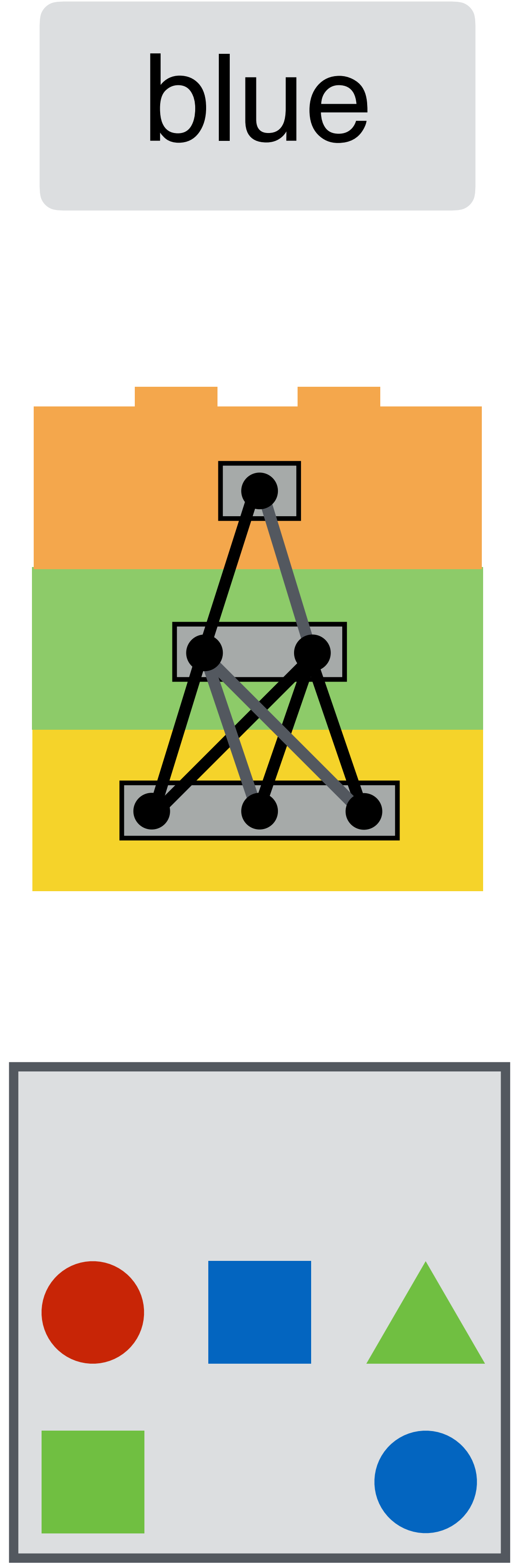


What color is the shape right of a circle?

Learning

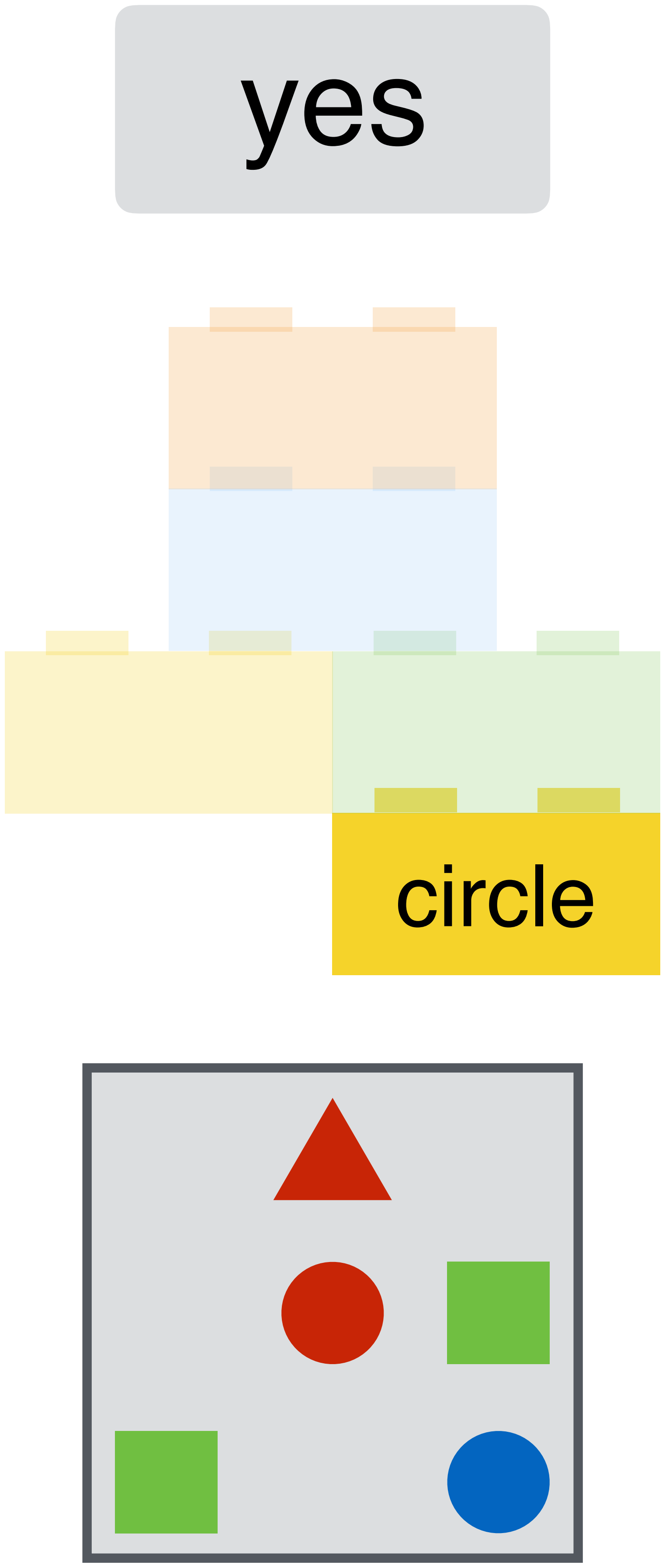


Is there a red shape above a circle?

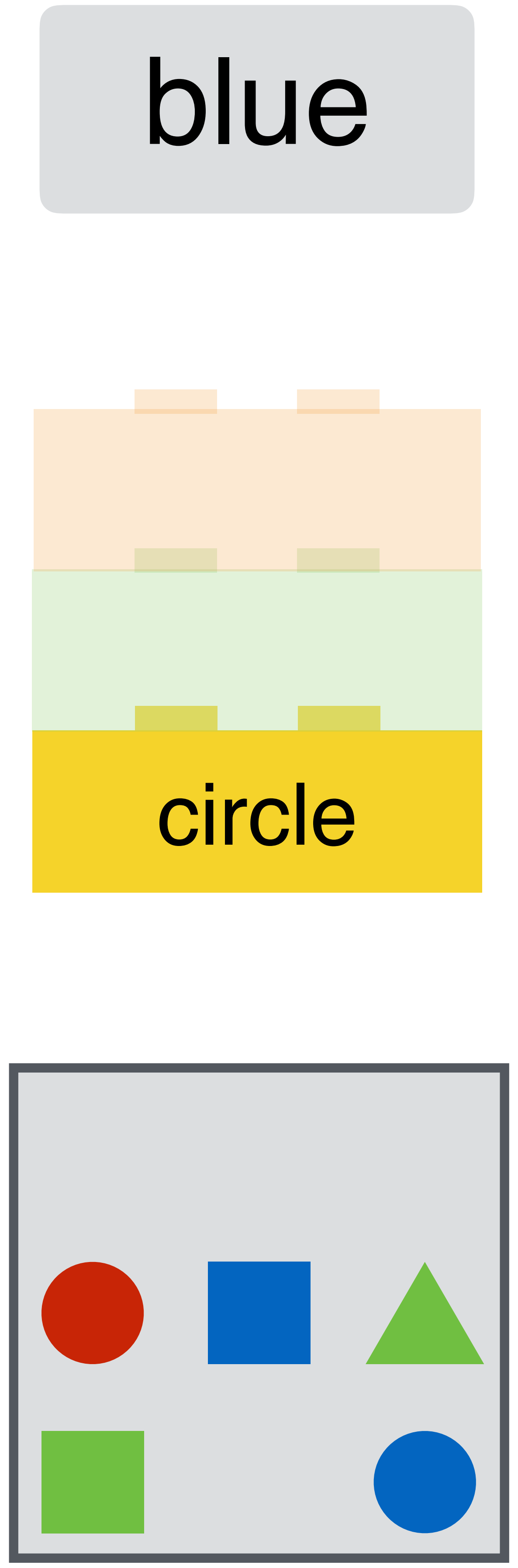


What color is the shape right of a circle?

Parameter tying

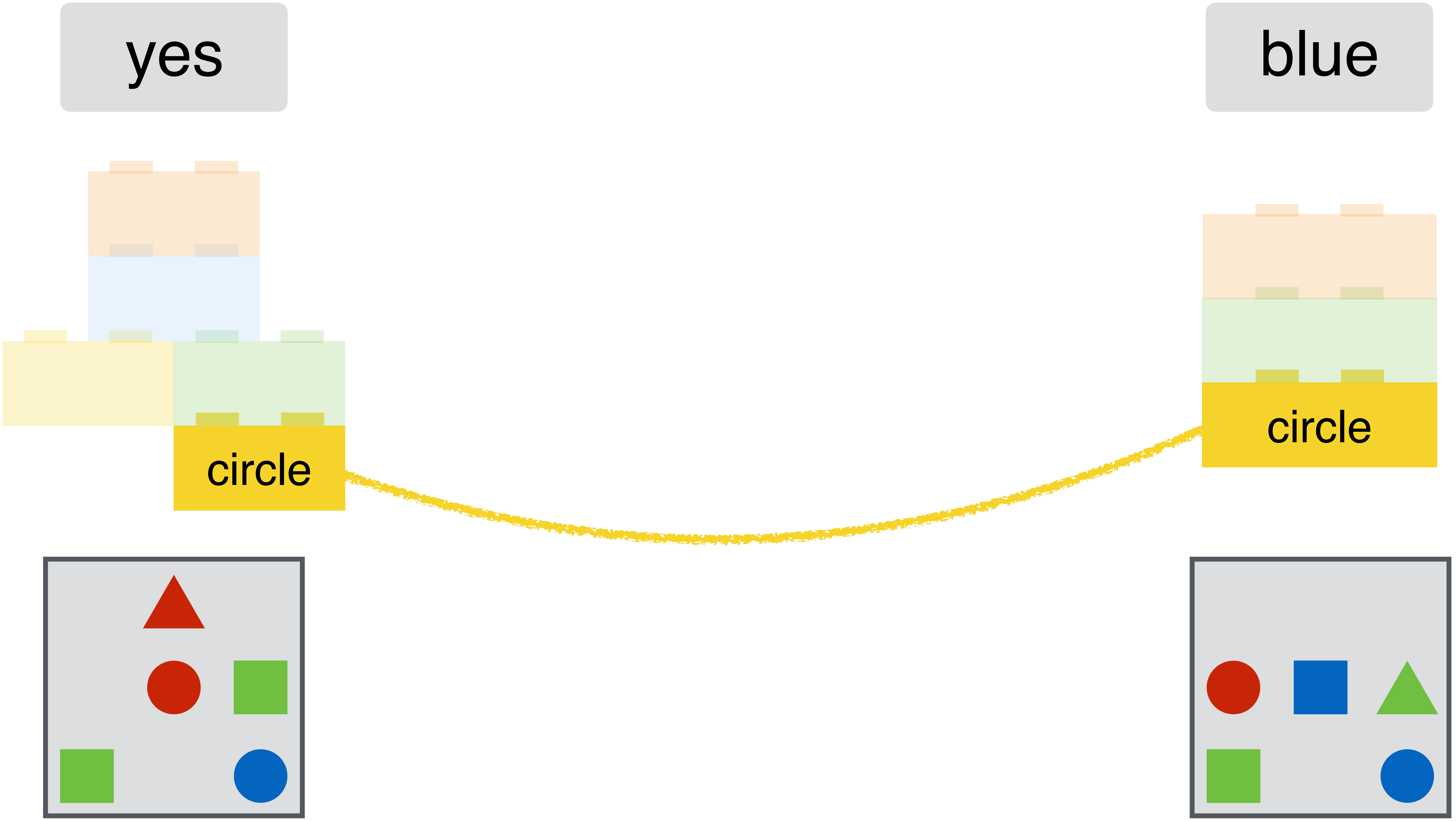


Is there a red shape above a circle?



What color is the shape right of a circle?

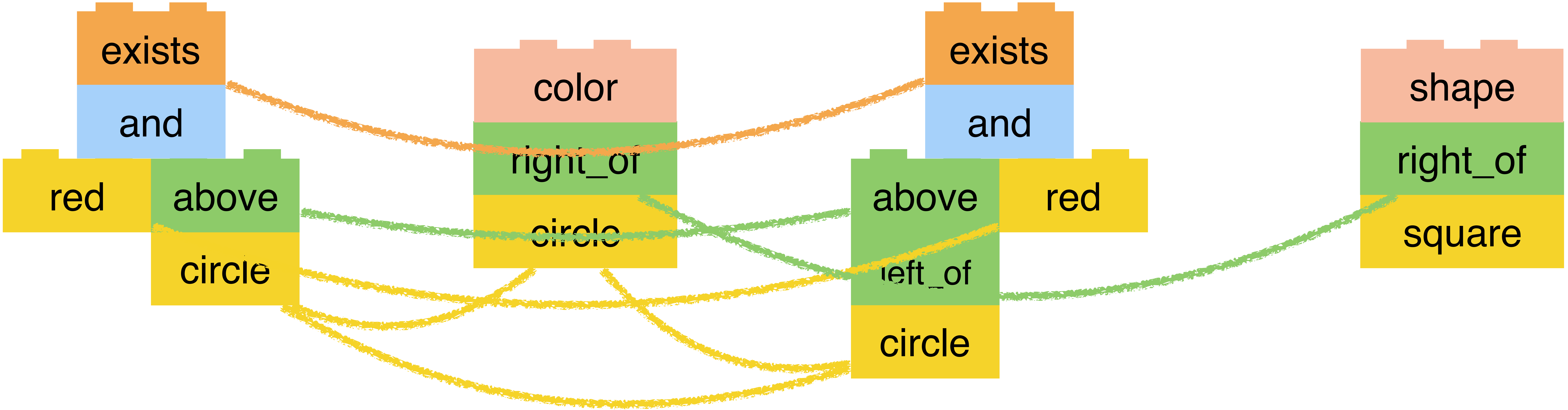
Parameter tying



Is there a red shape above a circle?

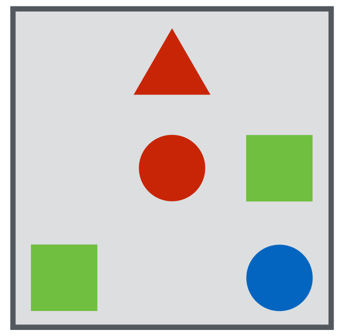
What color is the shape right of a circle?

Extreme parameter tying



[Slides credit: Jacob Andreas]

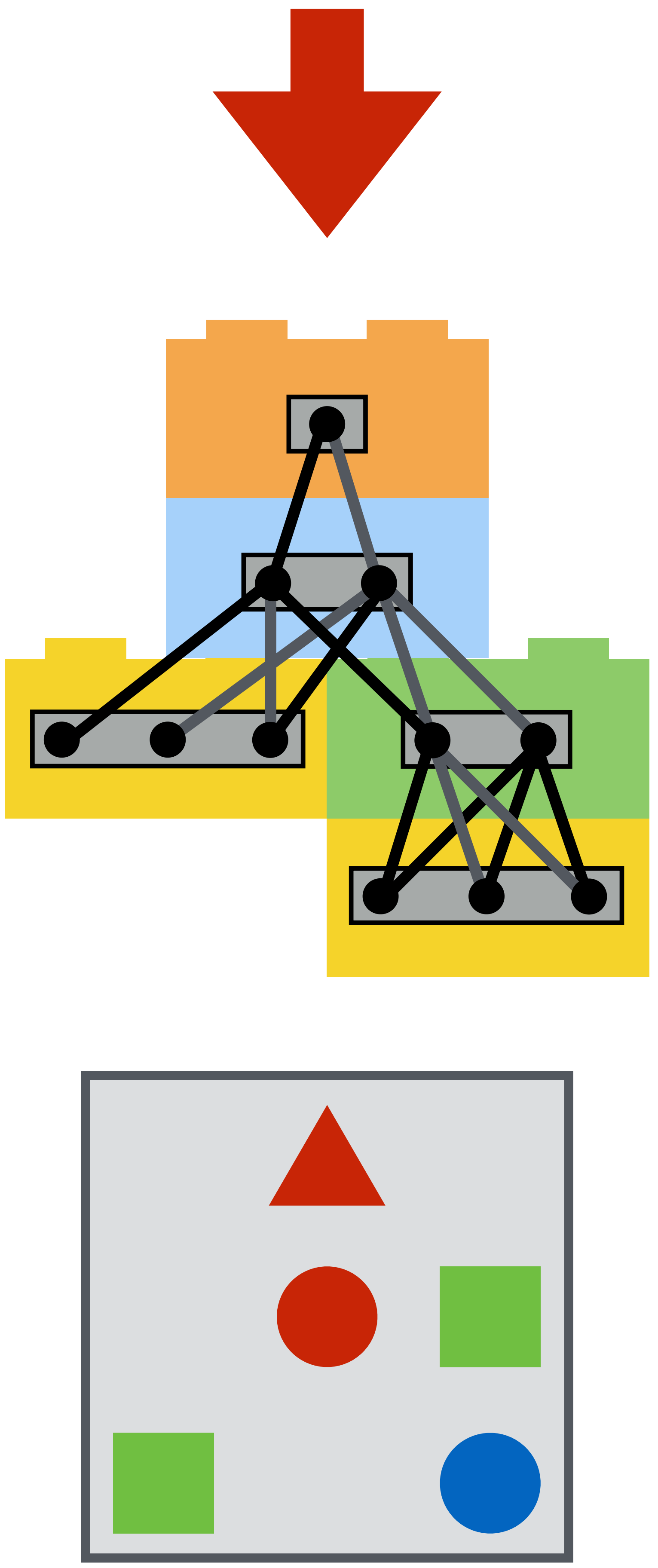
Learning with fixed layouts is easy!

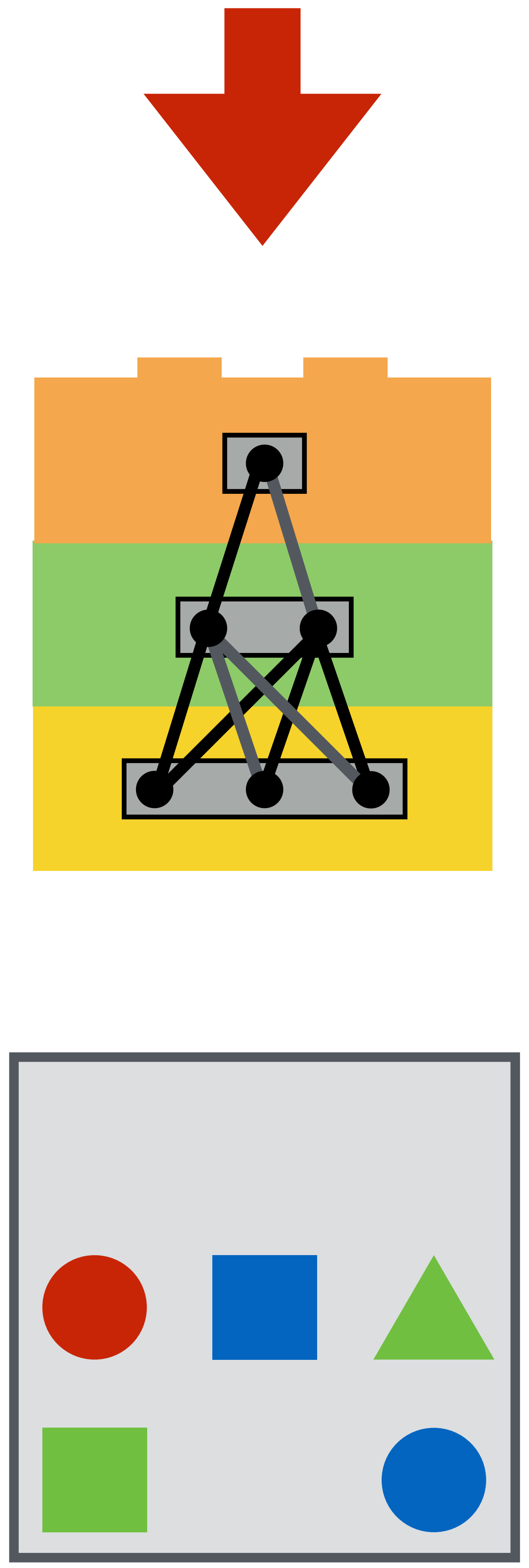
$\underset{W}{\arg\max} \sum p($ yes $|$ ,  $; W)$

(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

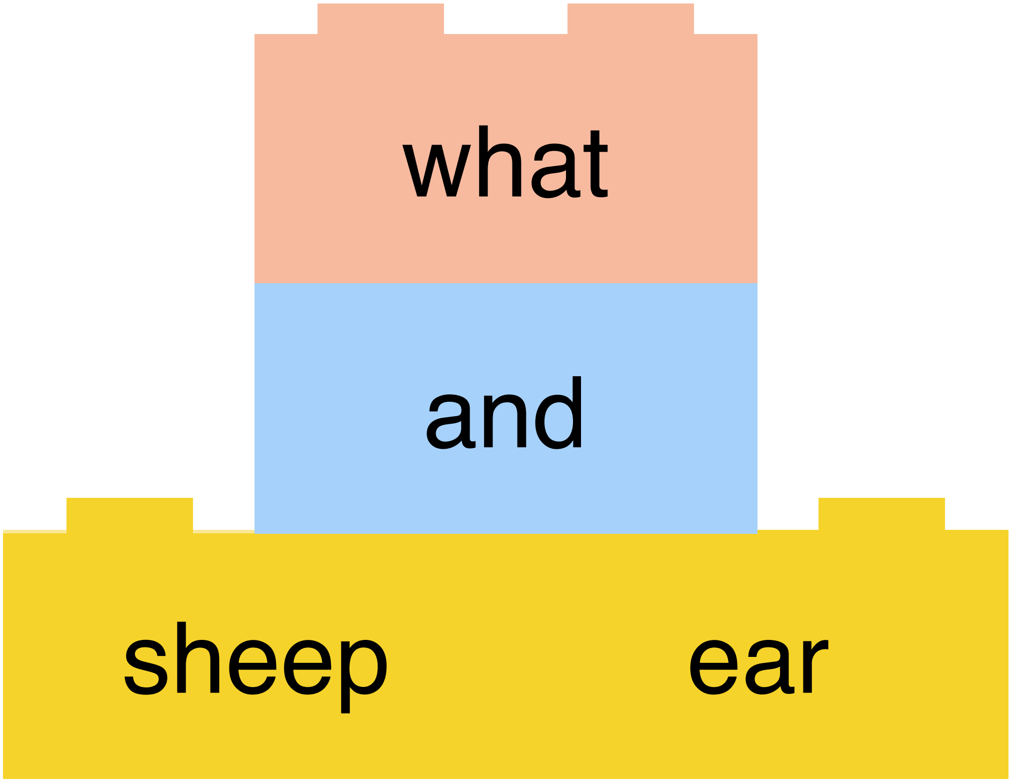




[Slides credit: Jacob Andreas]

Experiments: VQA Dataset

What is in the sheep's ear?



tag

[Slides credit: Jacob Andreas]

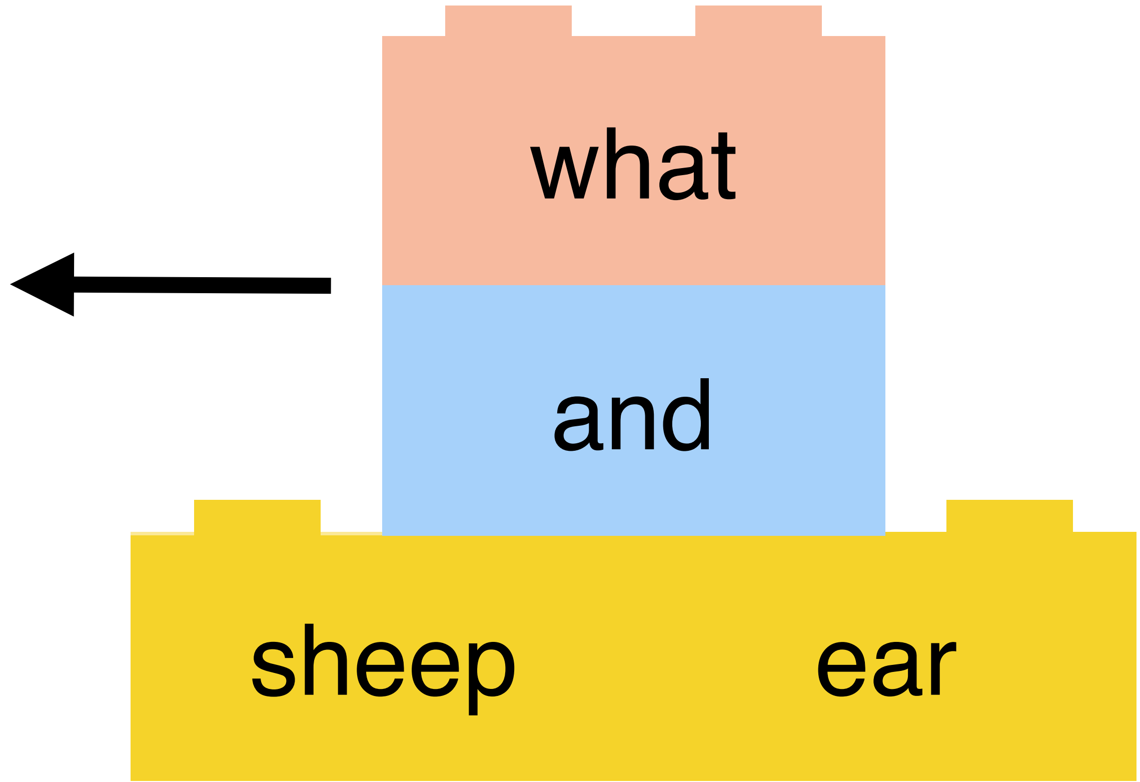
Experiments: VQA Dataset

What i
sheep

tag

Experiments: VQA Dataset

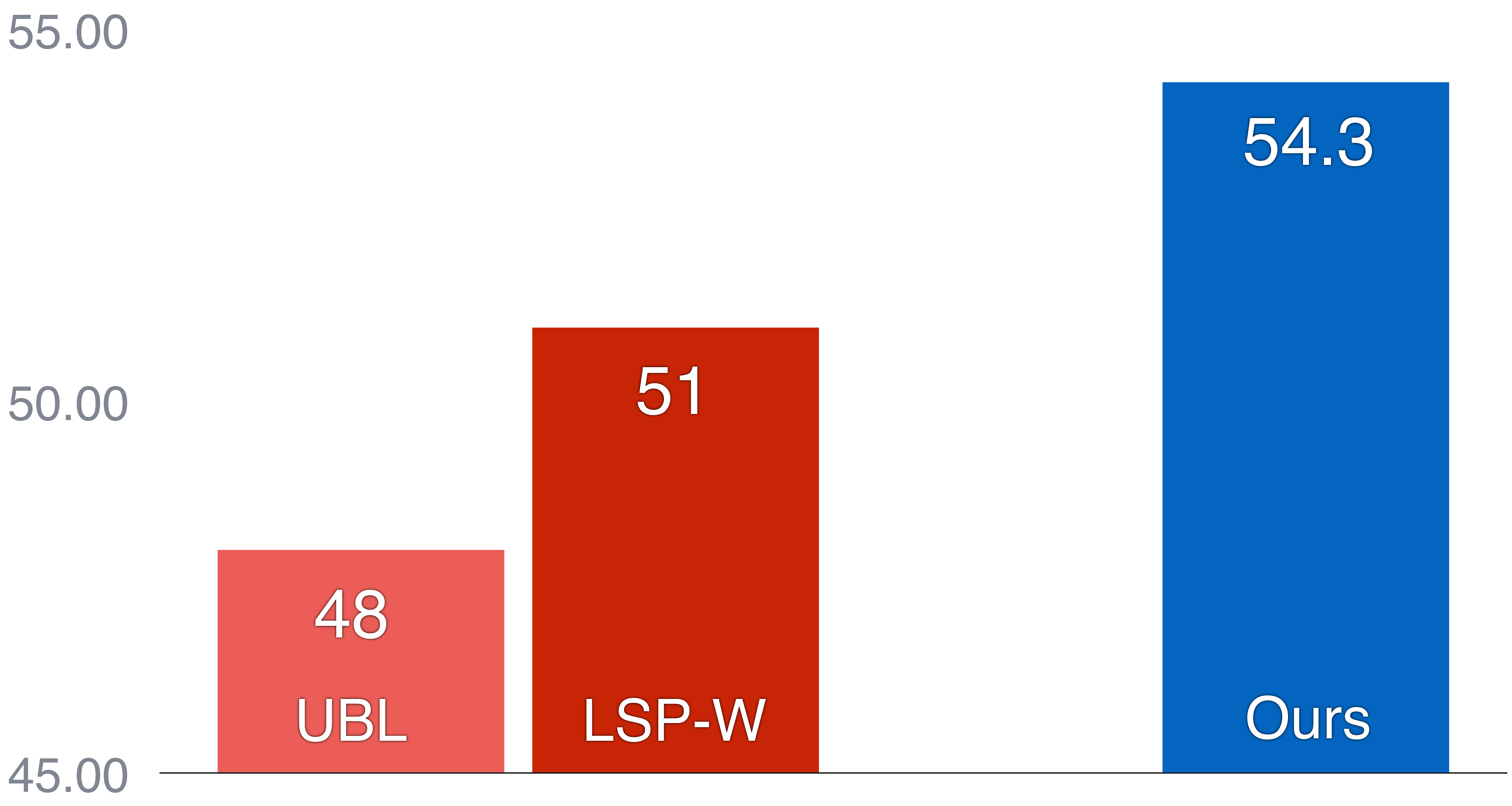
What is in the sheep's ear?



tag

[Slides credit: Jacob Andreas]

Experiments: GeoQA dataset



[Kwiatkowski et al. 2010, Krishnamurthy & Kollar 2013]