COS 484: Natural Language Processing

L16: Language Grounding - 1

Spring 2022

Language representations

Contextualized Word Representations

- ELMo = **E**mbeddings from **L**anguage **Mo**dels

Deep contextualized word representations
https://arxiv.org › cs ▾
by ME Peters - 2018 - Cited by 1683 - Related articles
Deep contextualized word representations. ... Our **word** vectors are learned functions of the internal states of a **deep** bidirectional language model (biLM), which is pre-trained on a large text corpus.

- BERT = **B**idirectional **E**ncoder **R**epresentations from **T**ransformers

BERT: Pre-training of Deep Bidirectional Transformers for ...
https://arxiv.org › cs ▾
by J Devlin - 2018 - Cited by 2259 - Related articles
Oct 11, 2018 - Unlike recent language representation models, **BERT** is designed to pre-train deep ... As a result, the pre-trained **BERT** model can be fine-tuned with just one additional output ... Which authors of this **paper** are endorsers?

Symbol grounding problem

- Miller and Johnson-Laird (1976) — Language and Perception
- Harnad (1990) — Symbol grounding problem
  - How do we connect “symbols” to the world in the right way?

In a pure symbolic model the crucial connection between the symbols and their referents is missing; an autonomous symbol system, though amenable to a systematic semantic interpretation, is ungrounded. In a pure connectionist model, names are connected to objects through invariant patterns in their sensory projections, learned through exposure and feedback, but the crucial compositional property is missing; a network of names, though grounded, is not yet amenable to a full systematic semantic interpretation. In the hybrid system proposed here, there is no longer any autonomous symbolic level at all; instead, there is an intrinsically dedicated symbol system, its elementary symbols (names) connected to nonsymbolic representations that can pick out the objects to which they refer, via connectionist networks that extract the invariant features of their analog sensory projections.

- Neural networks (connectionism) help us connect symbolic reasoning to sensory inputs

(slide adapted from Greg Durrett)

Color test

- What color is this?

A) Blue B) Green C) Navy

Color test

- What color is this?

A) Pink B) Violet C) Purple

Color test

- What color is this?

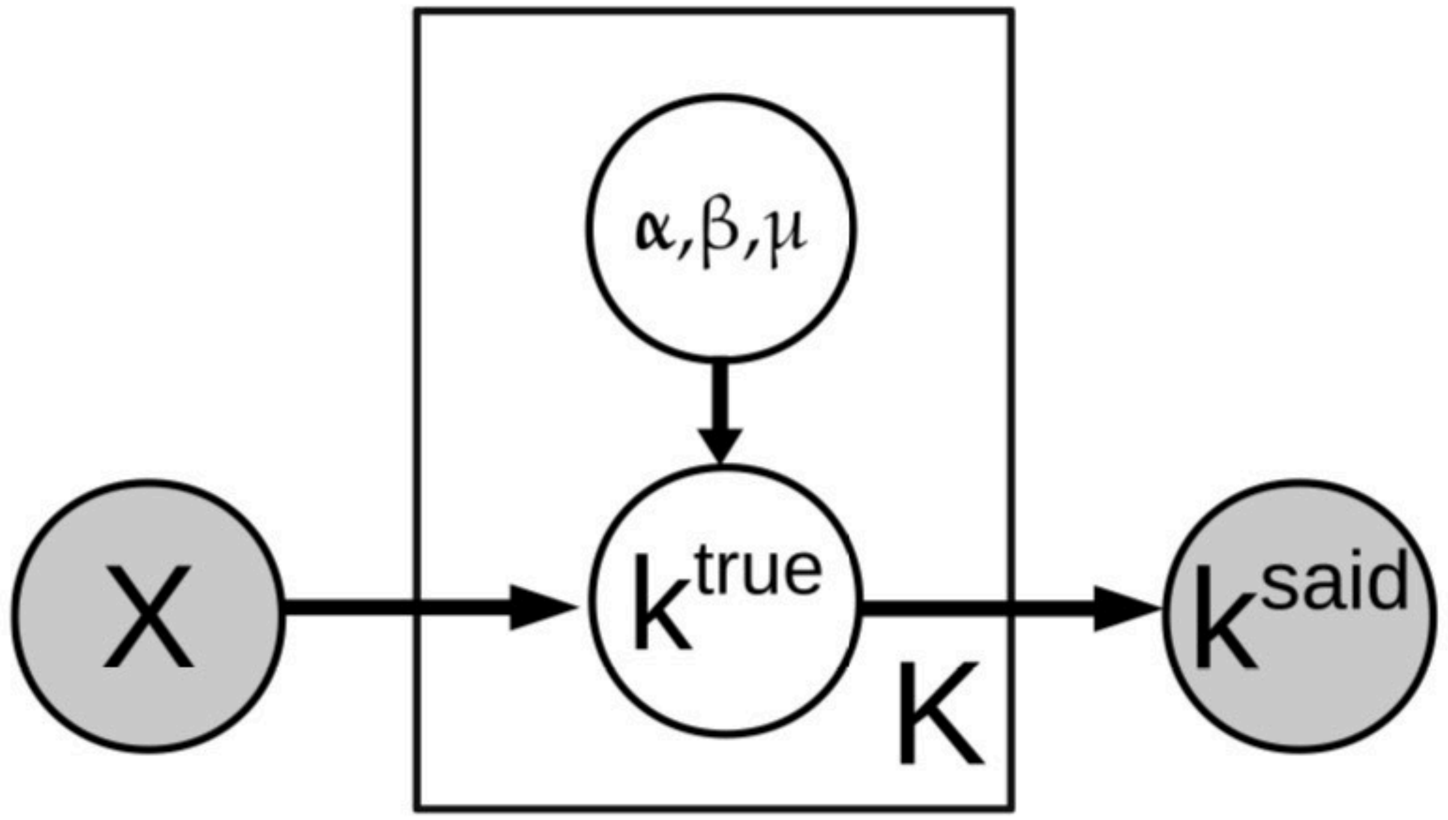
A) Lime B) Green C) Neon

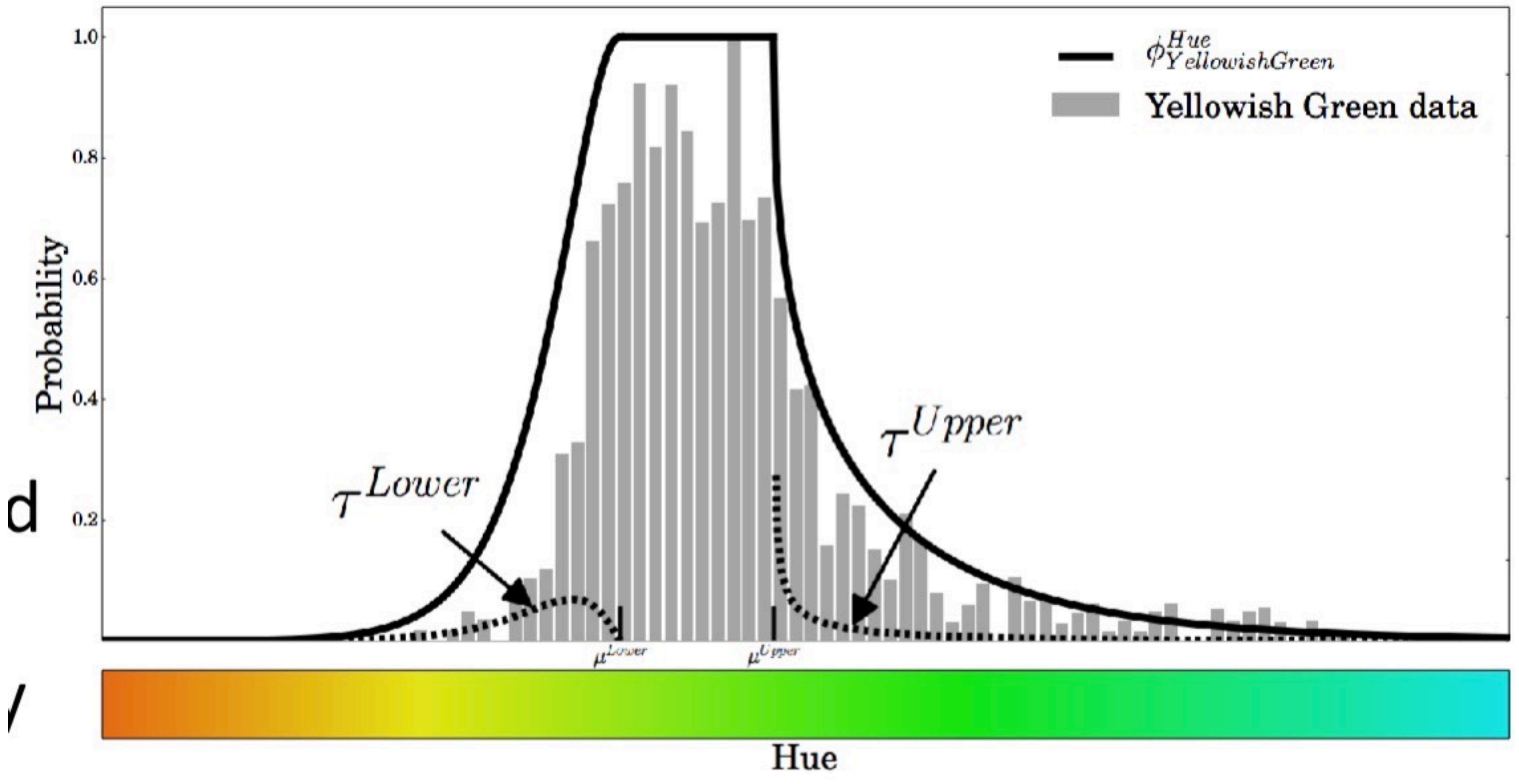
Color test

- What color is this?

Grounding color

- Bayesian model for grounded color semantics
- 829 color descriptions





(McMahan and Stone, 2014)

Gricean maxims

- Rules for cooperative, effective communication
- **Maxim of quantity:** Give as much information as needed, and no more
- **Maxim of quality:** Provide truthful information, supported by evidence
- **Maxim of relation:** Be relevant, say things pertinent to discussion
- **Maxim of manner:** Be clear, brief and orderly, avoid obscurity and ambiguity

Types of grounding

- **Perception**
  - Visual: *green* = [0,1,0] in RGB
  - Auditory: *loud* = >120 dB
  - Taste: sweet = >some threshold level of sensation on taste buds
  - High-level concepts:

cat

dog

Types of grounding

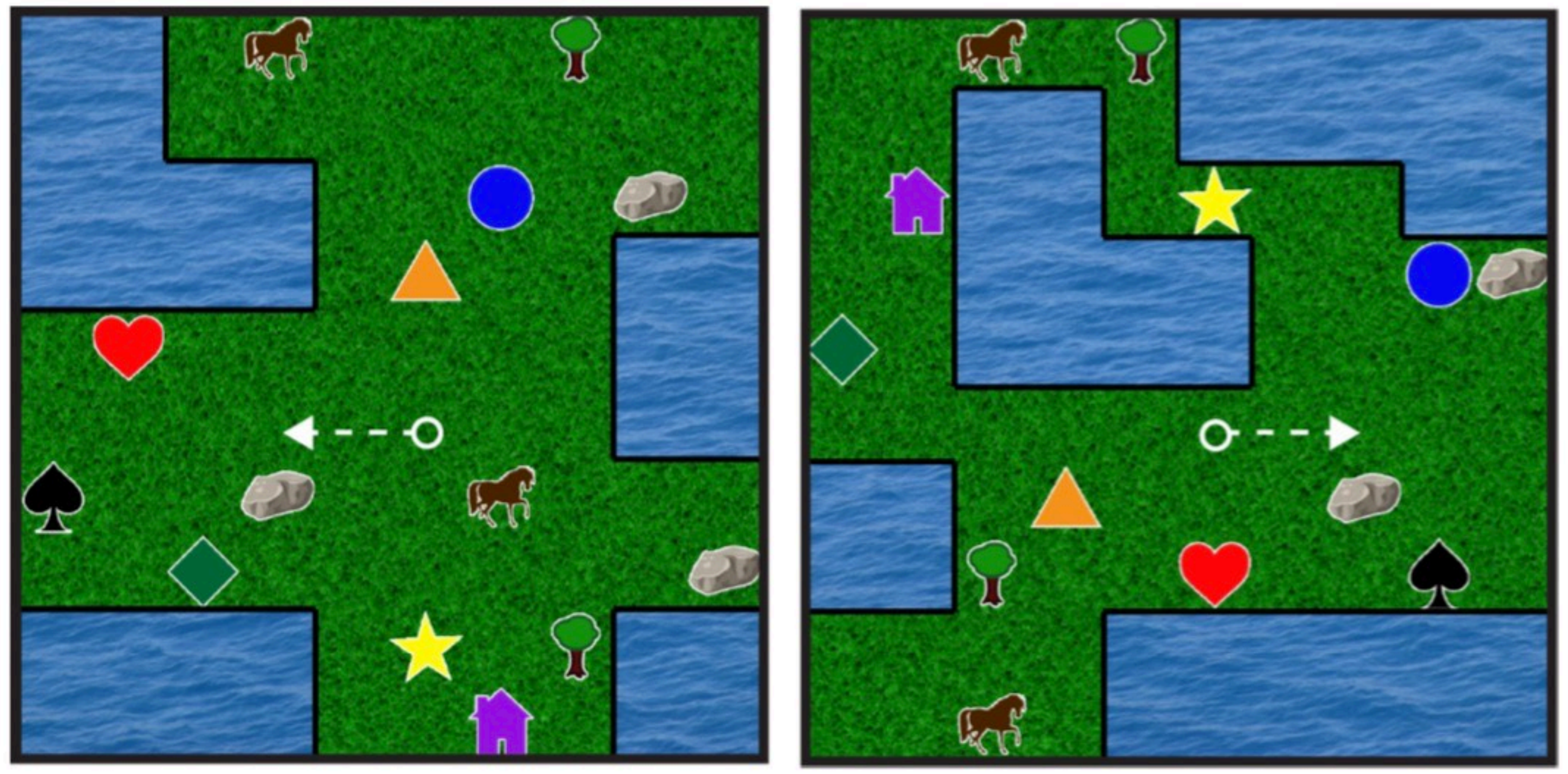
- **Temporal concepts**
  - *late evening* = after 6pm
  - *fast, slow* = describing rates of change
- **Actions**

running

eating

Types of grounding

- **Relations**
  - **Spatial:**
    - *left, on top of, in front of*
  - **Functional:**
    - *Jacket:* keeps people warm
    - *Mug*: holds water
  - **Size:**
    - Whales are *larger* than lions



Reach the cell above the westernmost rock

A chair

A chair

green

armless

medium size

used to sit on

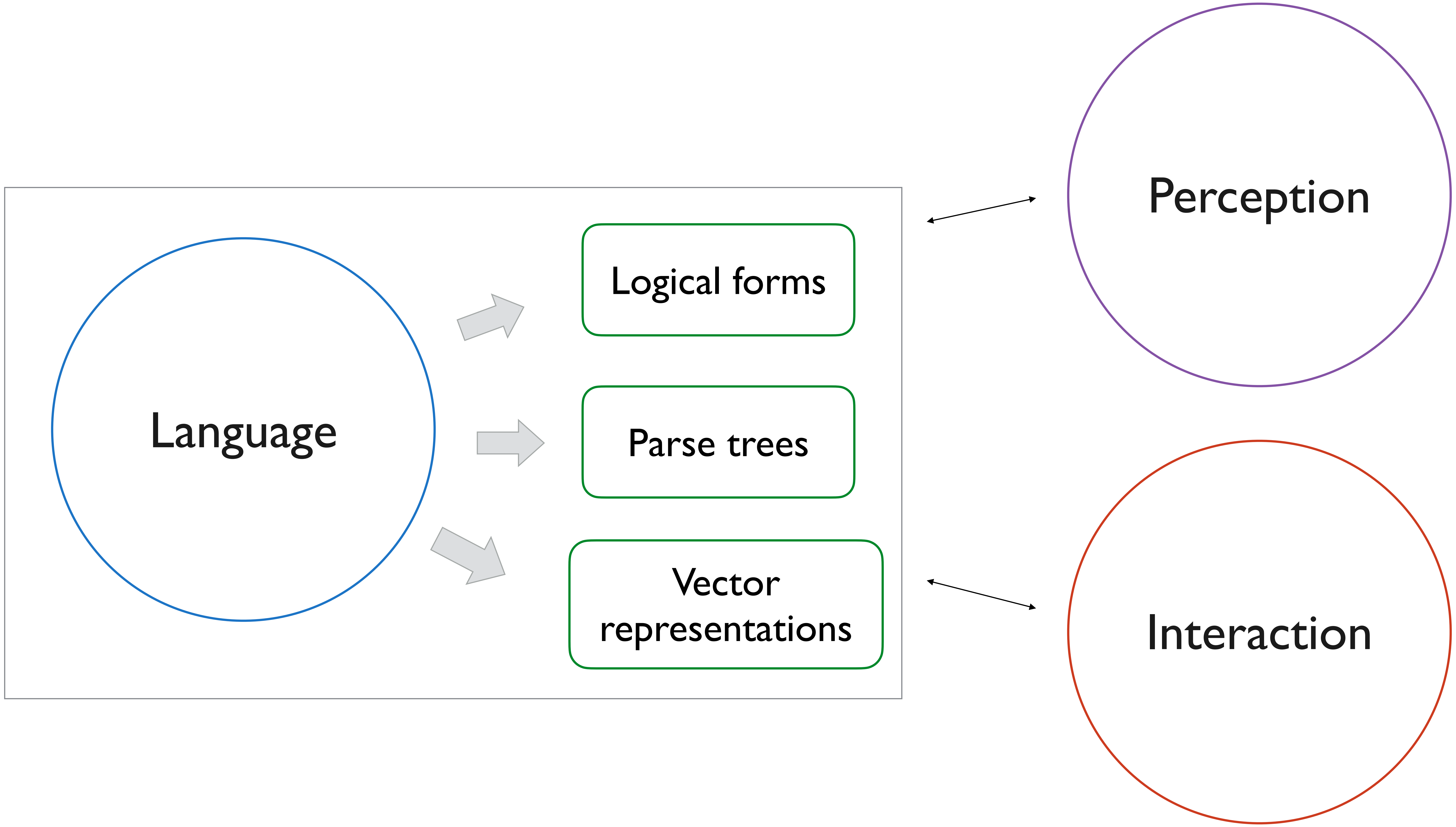
light

fragile

plush

Context is very important for understanding words!

Semantics does not exist in isolation



Some grounding tasks

- **Vision**
  - Captioning
  - Visual question answering (VQA)
  - Spatial reasoning
- **Interaction**
  - Instruction following
  - Text-based games

Image captioning

the girl is licking the spoon of batter

- Describe an image in a sentence

Image captioning

the girl is licking the spoon of batter

- Describe an image in a sentence
- Requires recognizing objects, attributes, relations in image
- Caption must be fluent

Applications?

(MS COCO, Chen et al., 2015)

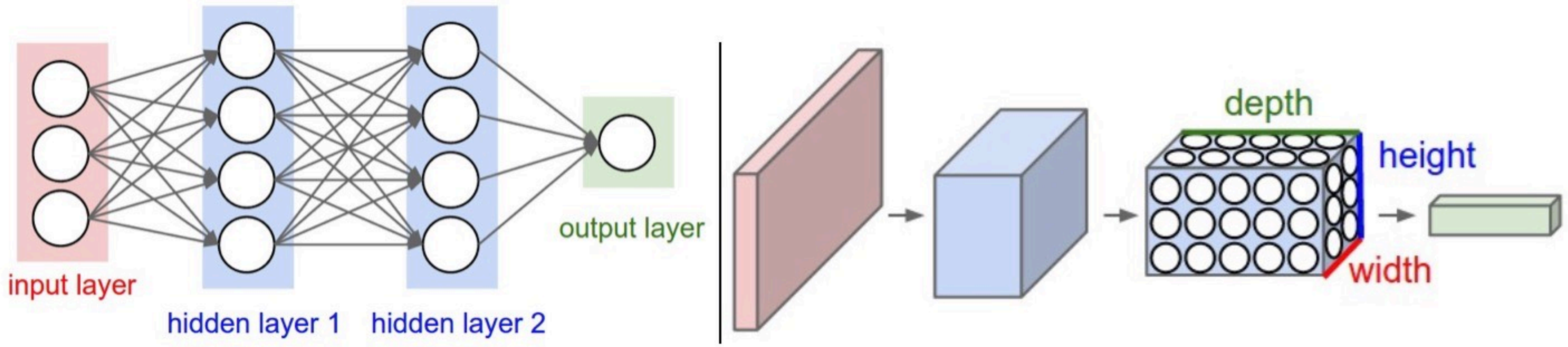
Captioning as multi-modal translation

French → Text encoder (e.g. RNN) → Text decoder (e.g. RNN) → *English*

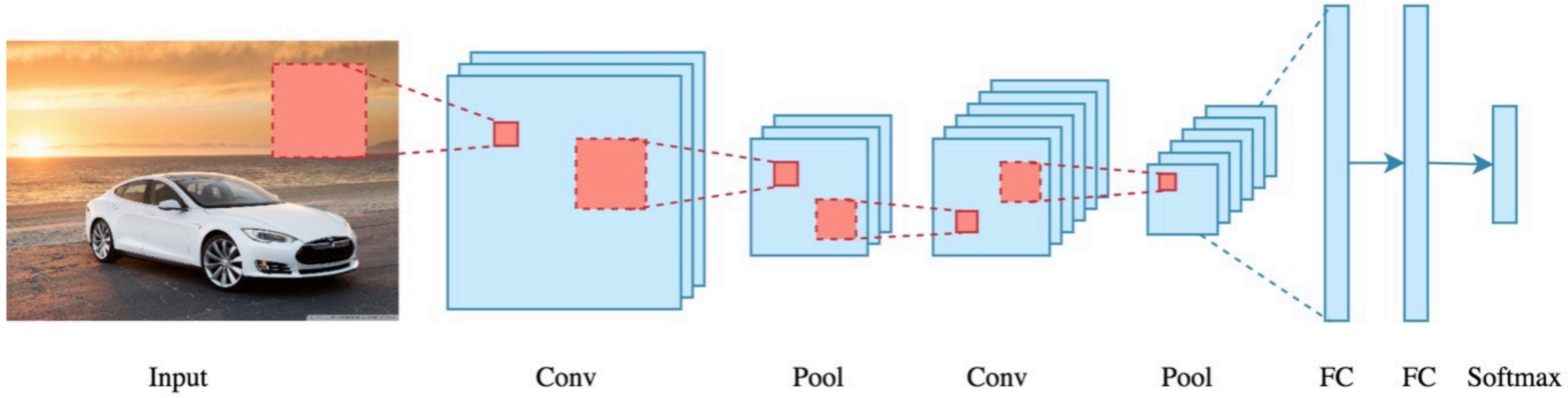
 → Image encoder → Text decoder (e.g. RNN) → *English*

(Donahue et al., 2015, Vinyals et al., 2015)

Convolutional Neural Networks



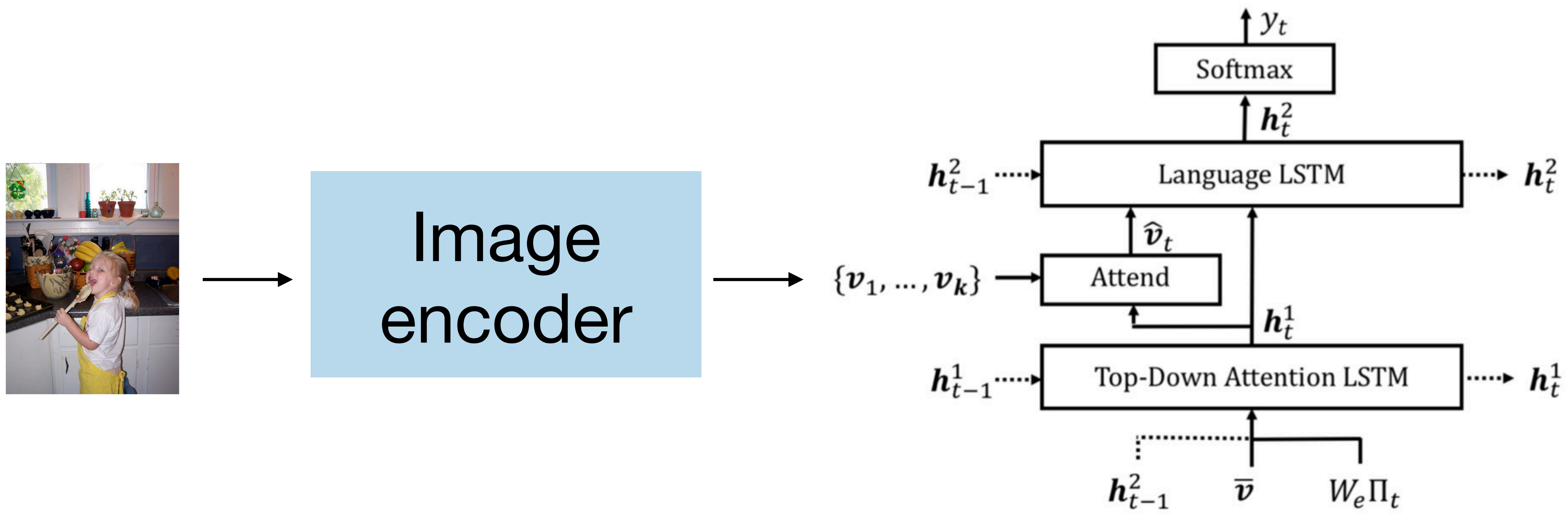
Left: A regular 3-layer Neural Network. Right: A ConvNet arranges its neurons in three dimensions (width, height, depth), as visualized in one of the layers. Every layer of a ConvNet transforms the 3D input volume to a 3D output volume of neuron activations. In this example, the red input layer holds the image, so its width and height would be the dimensions of the image, and the depth would be 3 (Red, Green, Blue channels).

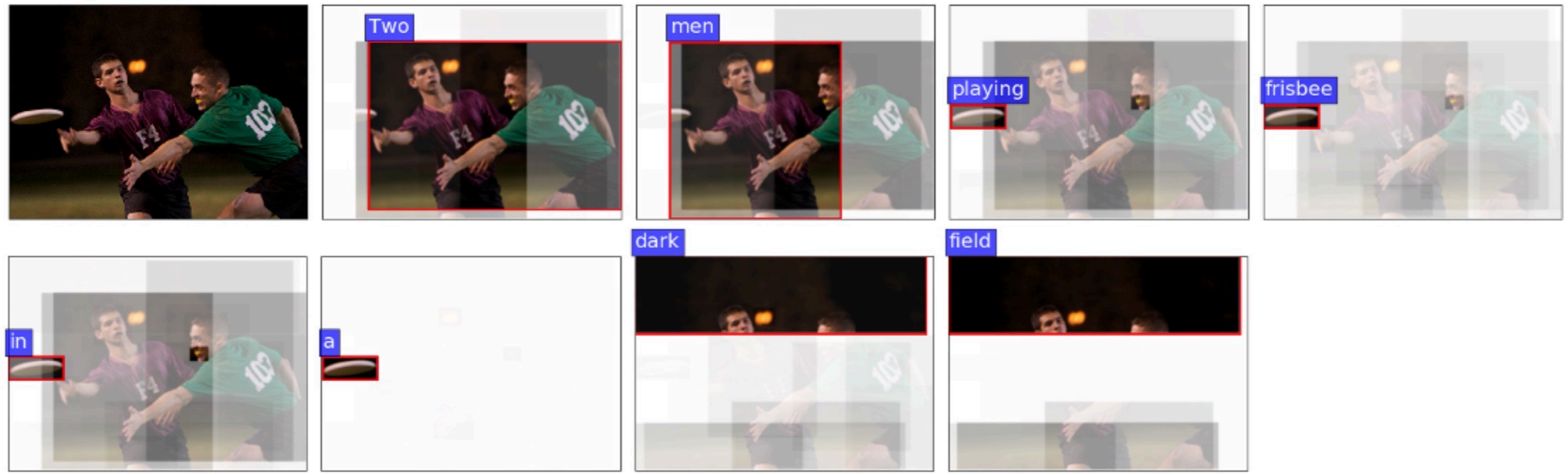


CNN for image classification

(source: CS231n, Stanford)

Captioning with attention

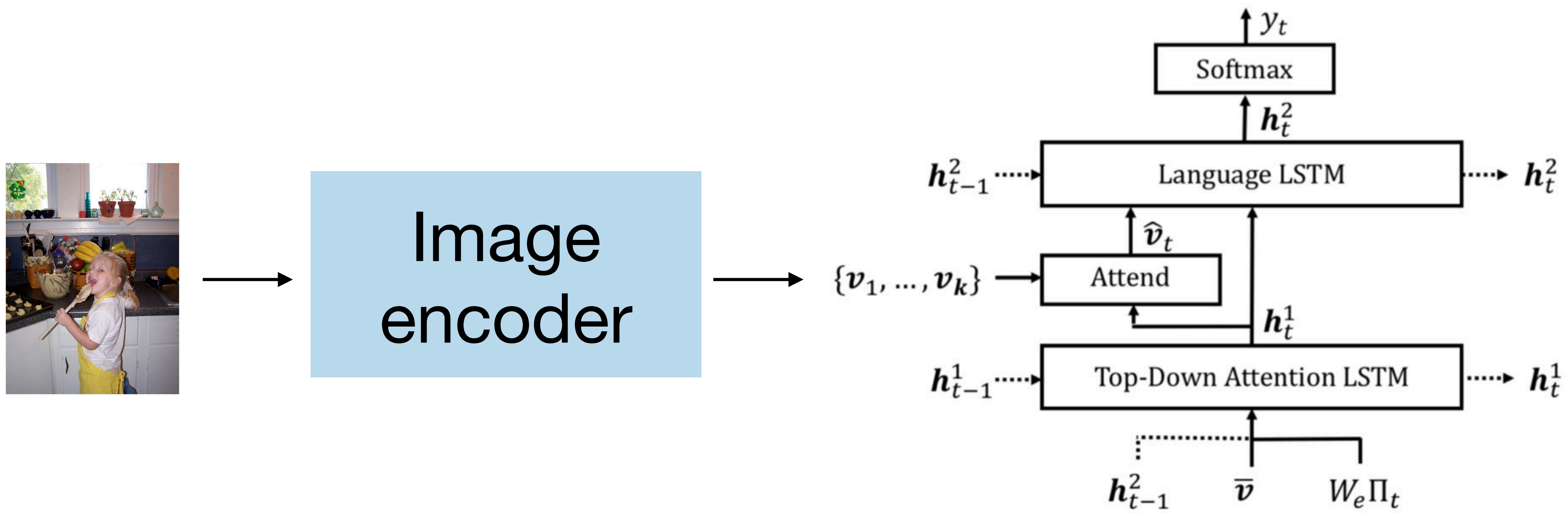




Two men playing frisbee in a dark field.

(Anderson et al., 2018)

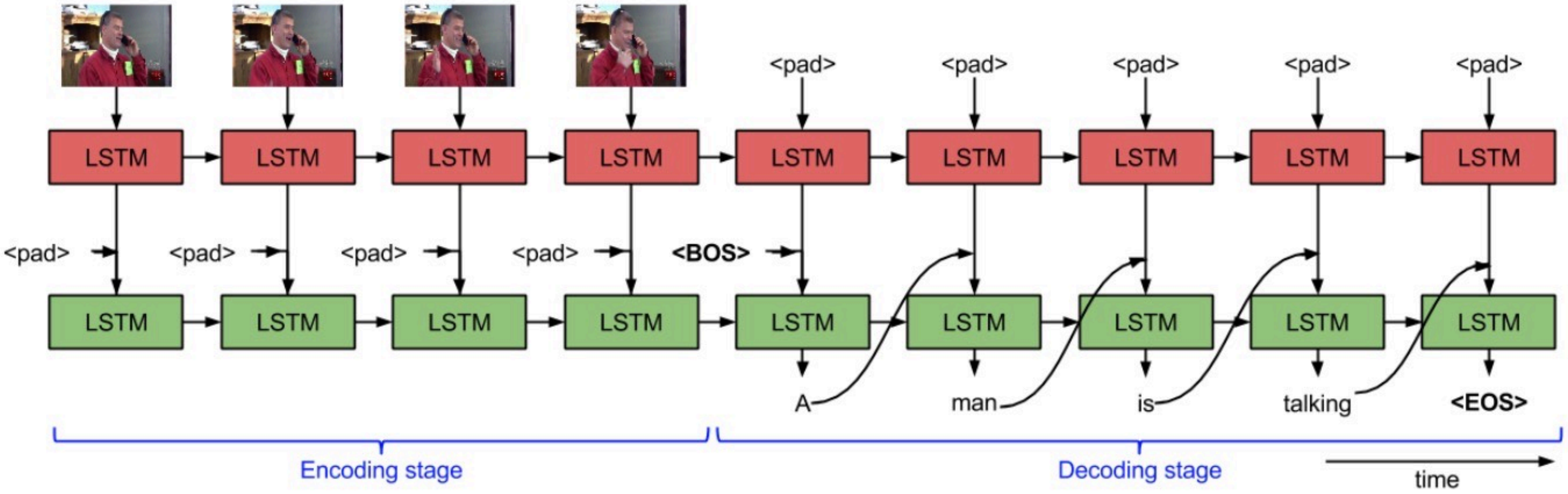
Captioning with attention



	BLEU-1		BLEU-2		BLEU-3		BLEU-4		METEOR		ROUGE-L		CIDEr		SPICE	
	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40
Review Net [48]	72.0	90.0	55.0	81.2	41.4	70.5	31.3	59.7	25.6	34.7	53.3	68.6	96.5	96.9	18.5	64.9
Adaptive [27]	74.8	92.0	58.4	84.5	44.4	74.4	33.6	63.7	26.4	35.9	55.0	70.5	104.2	105.9	19.7	67.3
PG-BCMR [24]	75.4	-	59.1	-	44.5	-	33.2	-	25.7	-	55	-	101.3	-	-	-
SCST:Att2all [34]	78.1	93.7	61.9	86.0	47.0	75.9	35.2	64.5	27.0	35.5	56.3	70.7	114.7	116.7	20.7	68.9
LSTM-A_3 [49]	78.7	93.7	62.7	86.7	47.6	76.5	35.6	65.2	27	35.4	56.4	70.5	116	118	-	-
Ours: Up-Down	**80.2**	**95.2**	**64.1**	**88.8**	**49.1**	**79.4**	**36.9**	**68.5**	**27.6**	**36.7**	**57.1**	**72.4**	**117.9**	**120.5**	**21.5**	**71.5**

(Anderson et al., 2018)

Video captioning



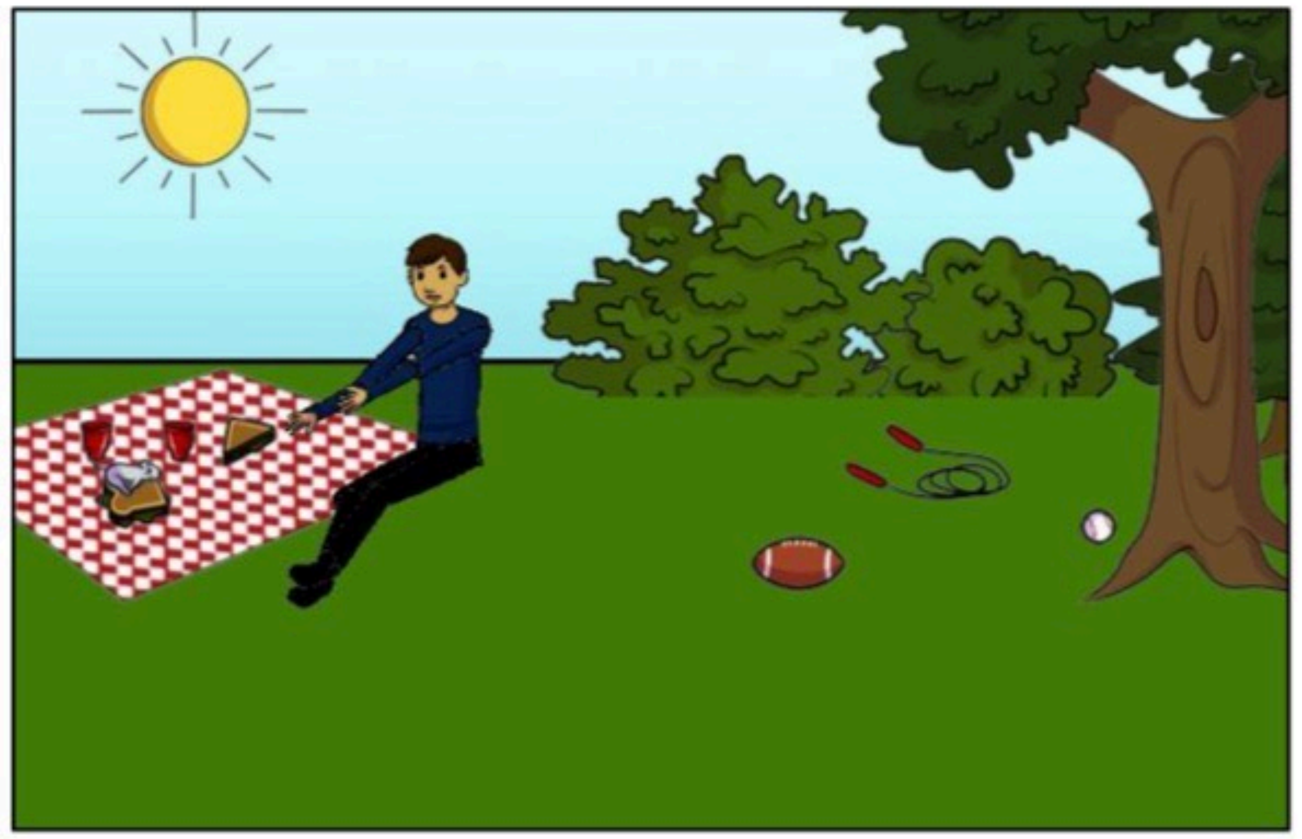
An overview of the S2VT video to text architecture.

(Venugopalan et al., 2015)

Visual Question Answering

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

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



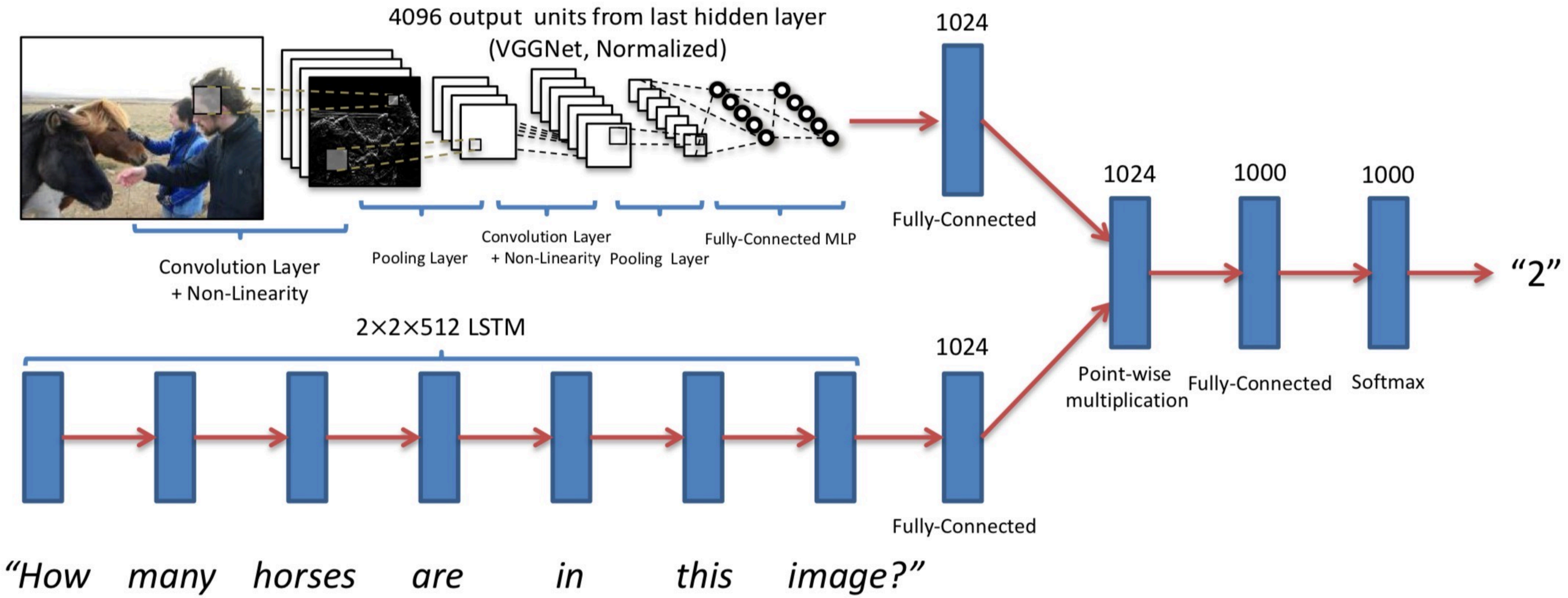
Is this person expecting company?
What is just under the tree?

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

- Answer questions about an image
- Require *multi-modal* knowledge and reasoning
- Well-defined *evaluation metric* (accuracy)

(Agrawal et al., 2015)

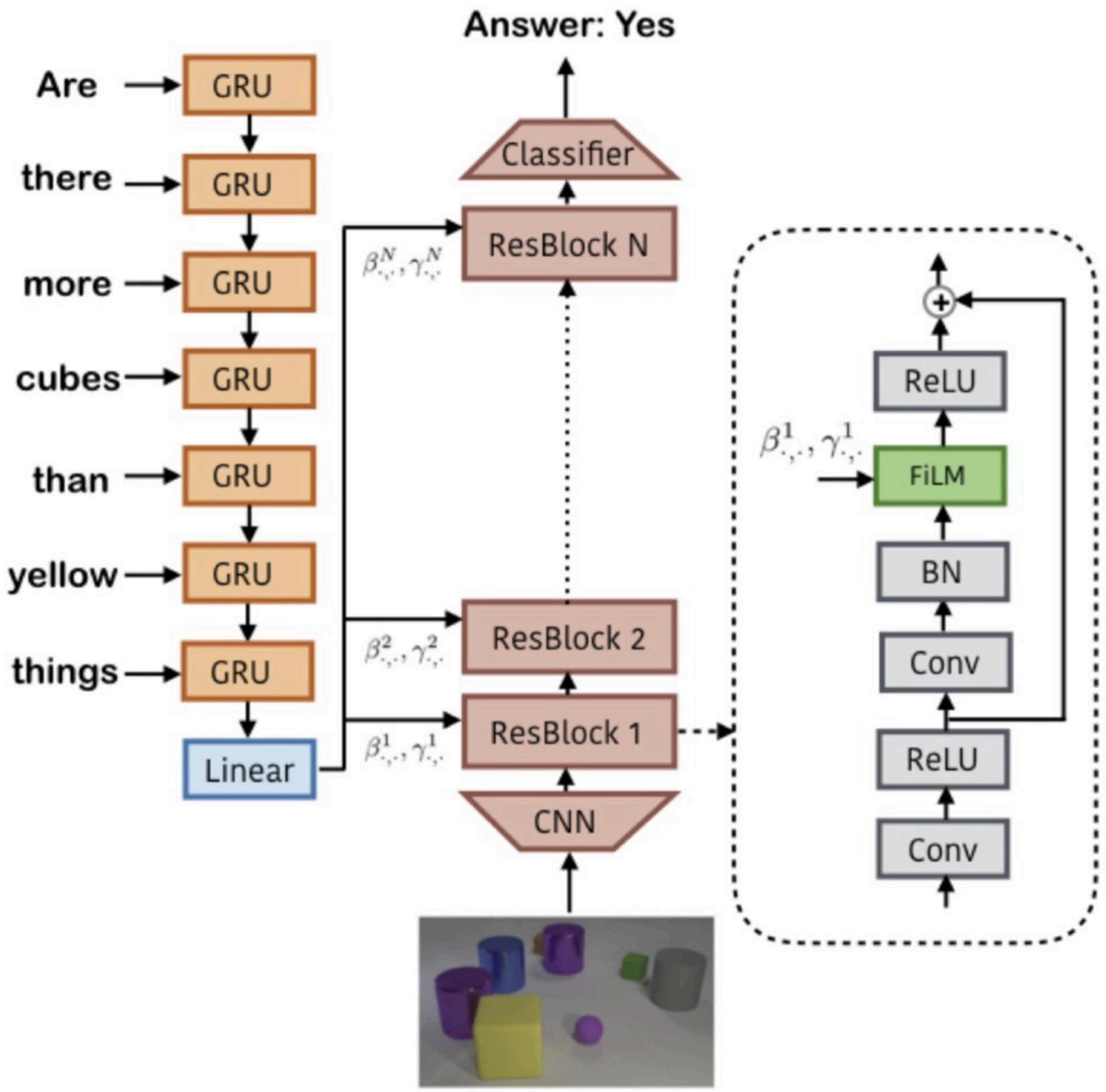
Visual Question Answering



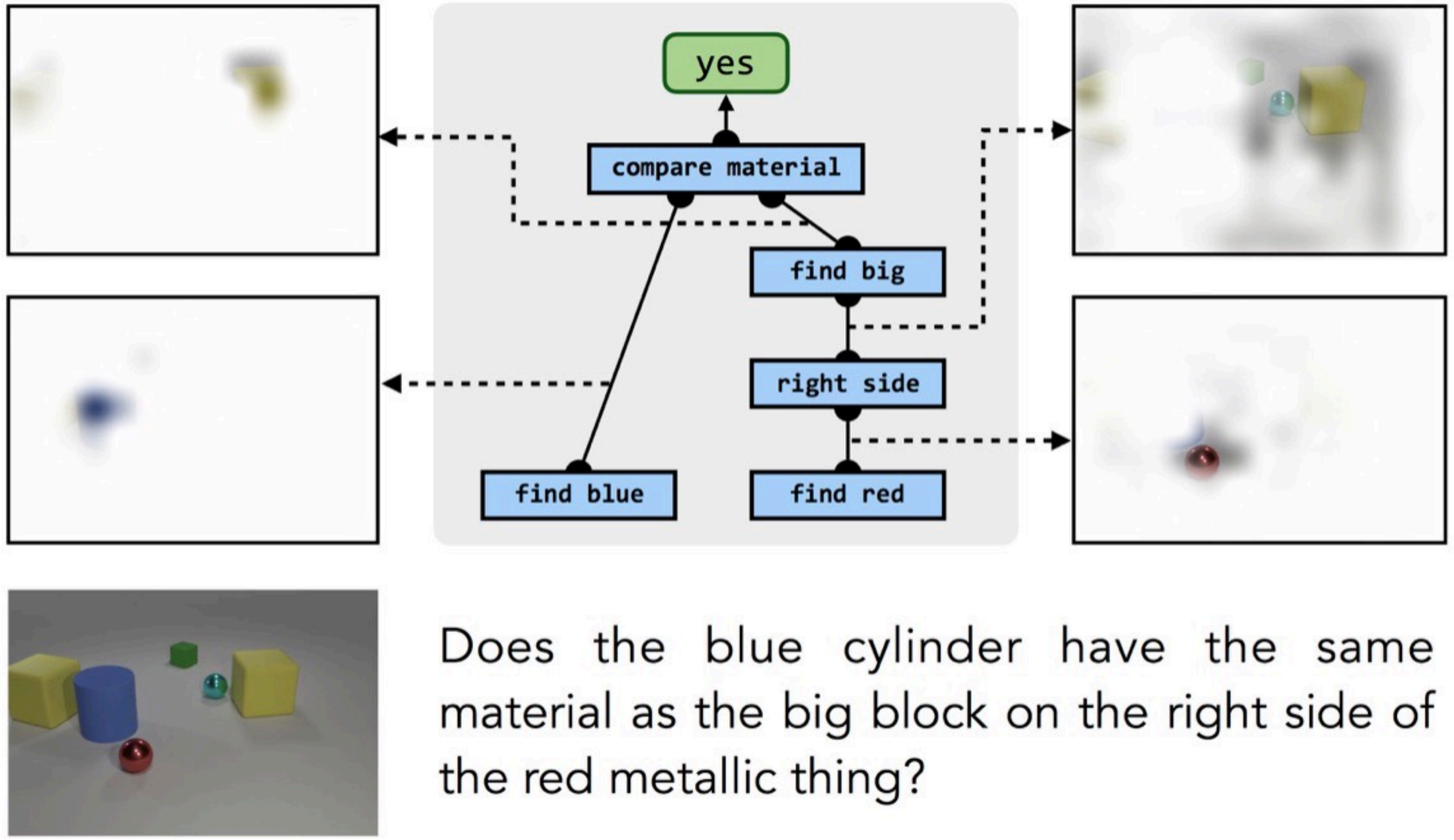
Any issues?

(Agrawal et al., 2015)

Better multimodal reasoning



FiLM



Does the blue cylinder have the same material as the big block on the right side of the red metallic thing?

Neural module networks

(Perez et al., 2017)

(Andreas et al., 2016)

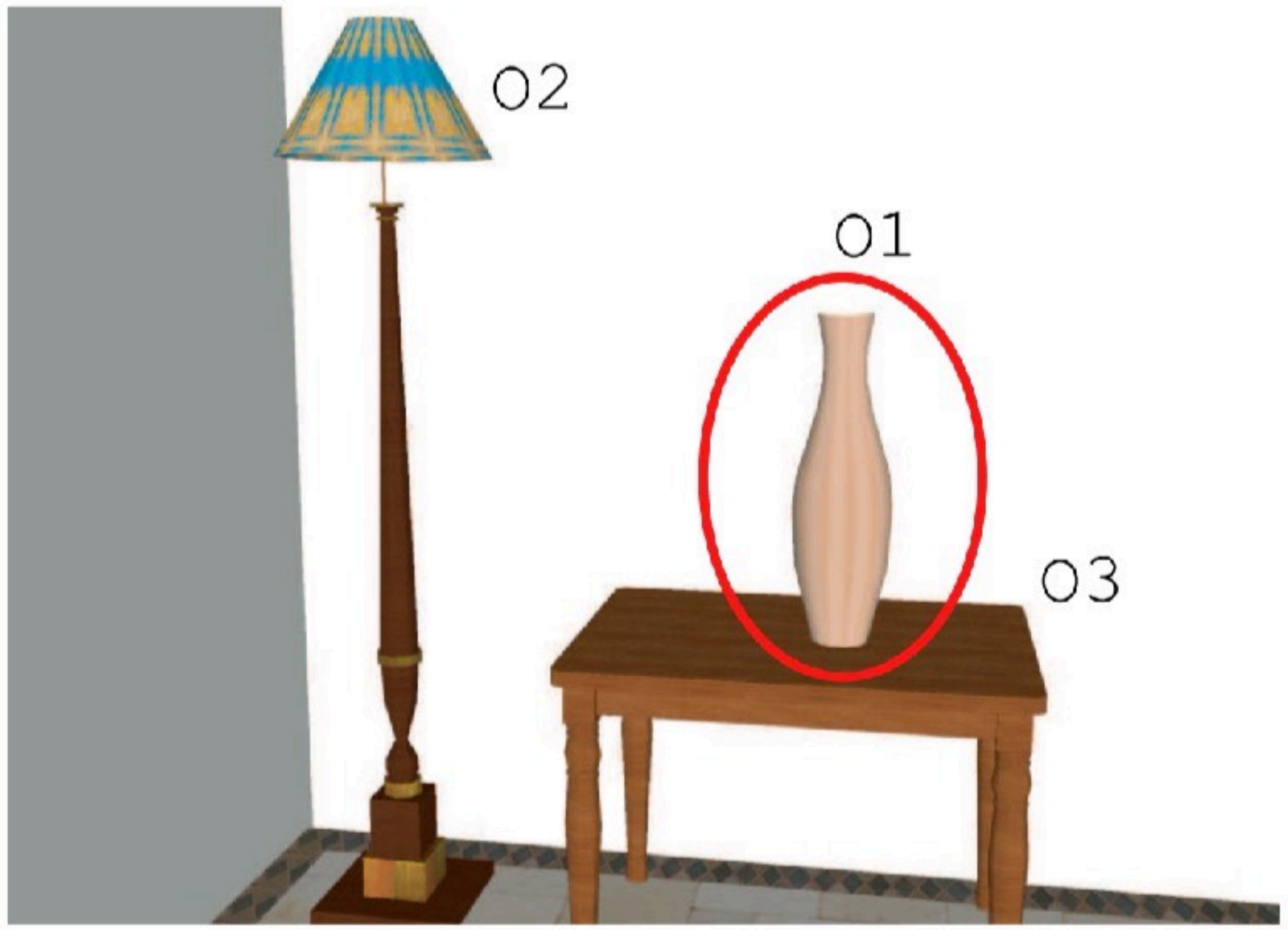
Visual Question Answering

- On deeper examination:
  - Just using language is a pretty good prior!
  - “Do you see a ..” = yes (87% of the time)
  - “How many…” = 2 (39%)
  - “What sport …” = tennis (41%)

Balanced VQA

(Goyal et al., 2017)

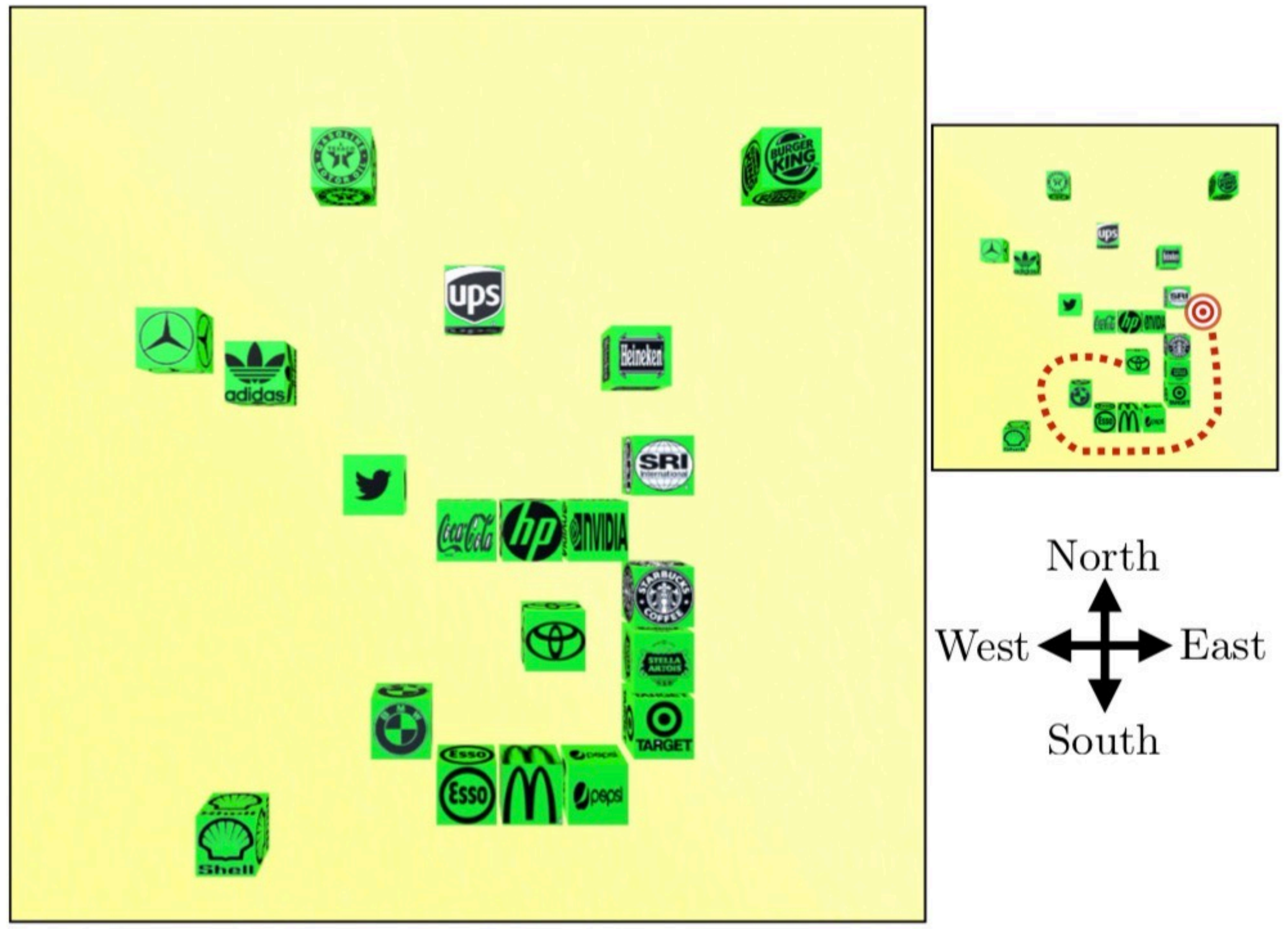
Spatial Relations



Golland et al. (2010)

- How would you indicate O1 to someone with relation to the other two objects? (not calling it a vase, or describing its inherent properties)
- What about O2?
- Requires modeling listener — "right of O2" is insufficient though true

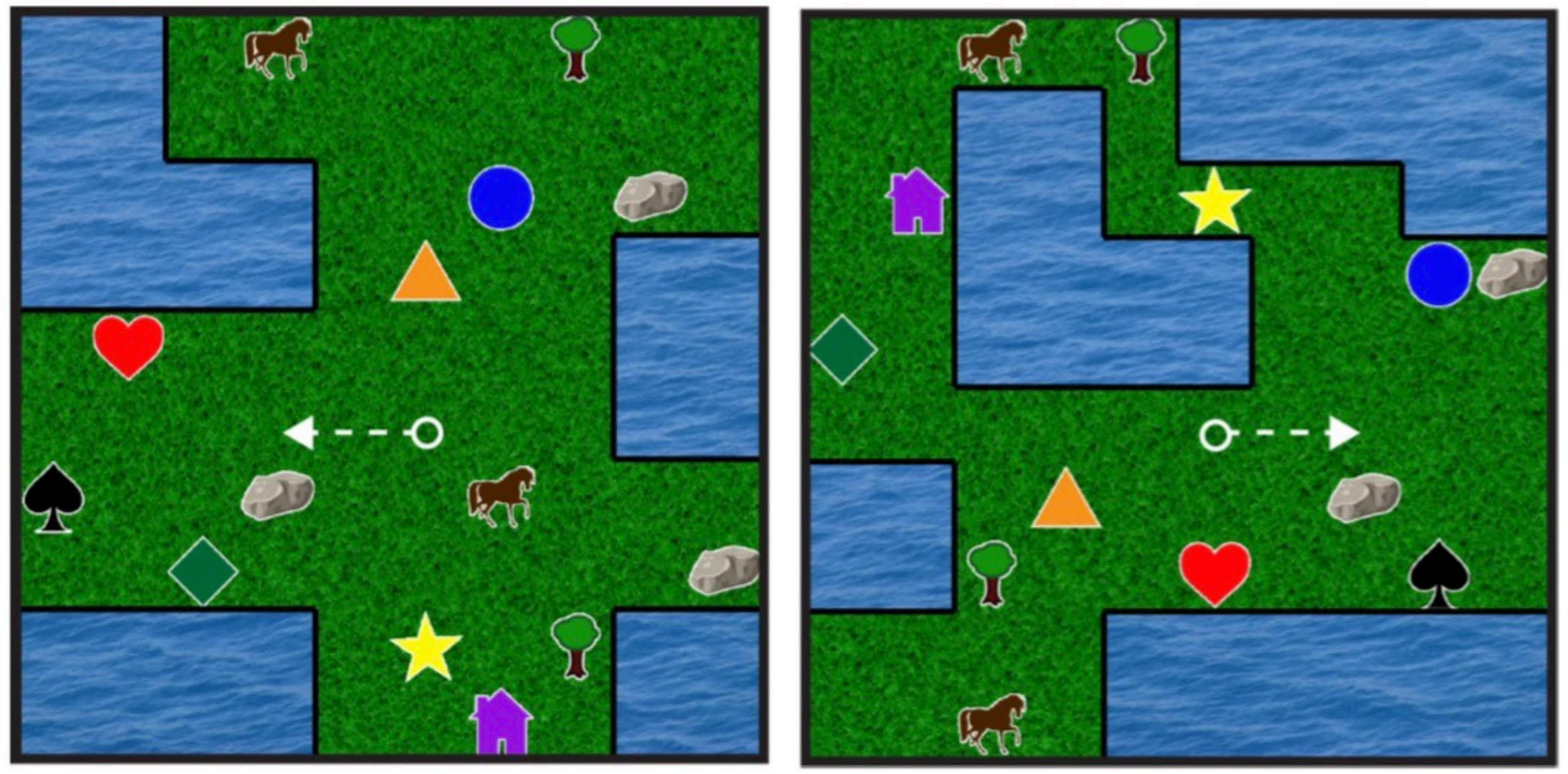
Spatial Reasoning



Put the Toyota block in the same row as the SRI block, in the first open space to the right of the SRI block
Move Toyota to the immediate right of SRI, evenly aligned and slightly separated
Move the Toyota block around the pile and place it just to the right of the SRI block
Place Toyota block just to the right of The SRI Block
Toyota, right side of SRI

Robotic Manipulation

(Bisk et al., 2016, Misra et al., 2017)



Reach the cell above the westernmost rock

Autonomous navigation

(Janner et al., 2017)