COS 484/584: (Advanced) Natural Language Processing

# L20: Language Grounding

Spring 2021

# 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?



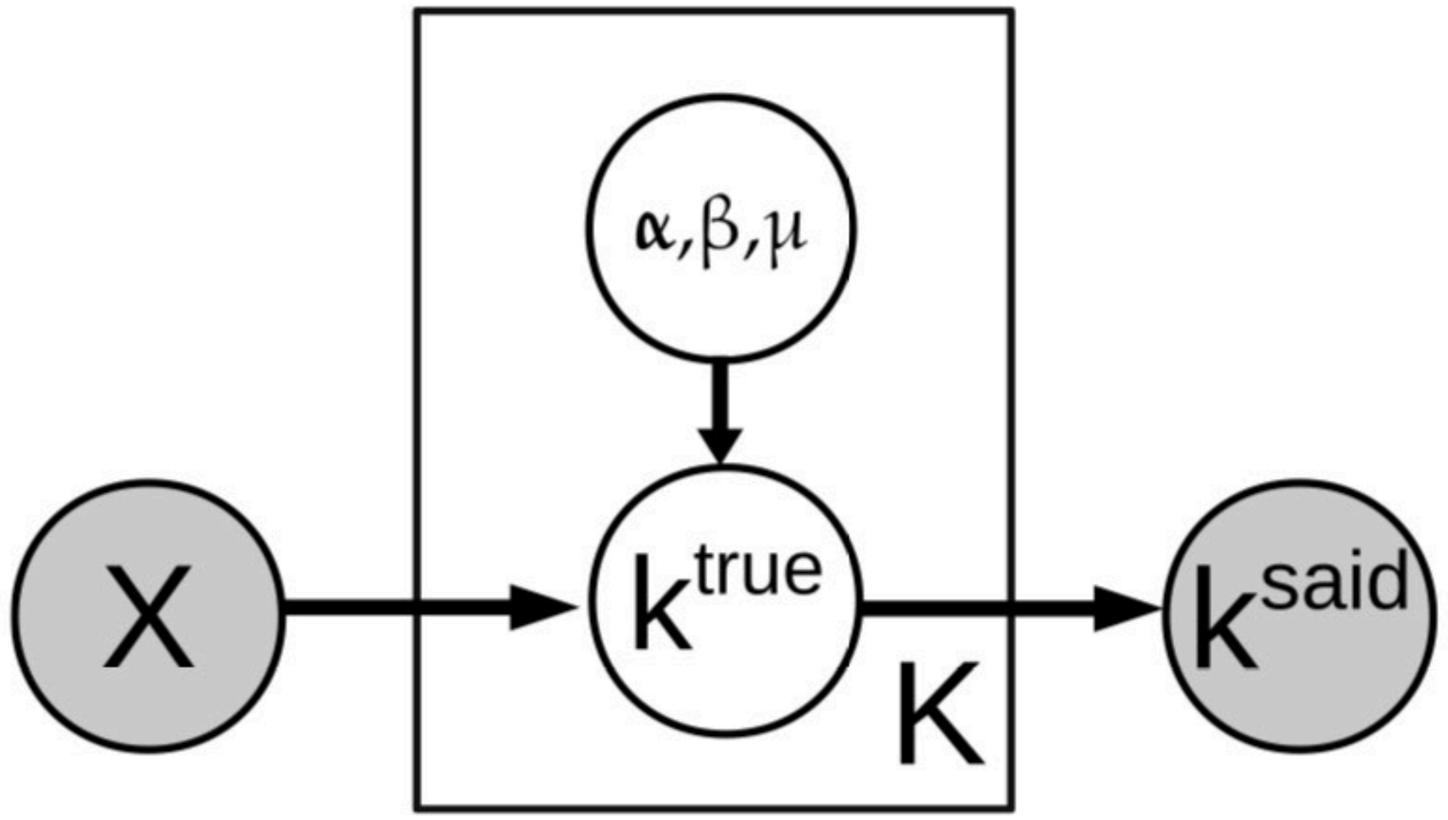
A) Lime B) Green C) Neon

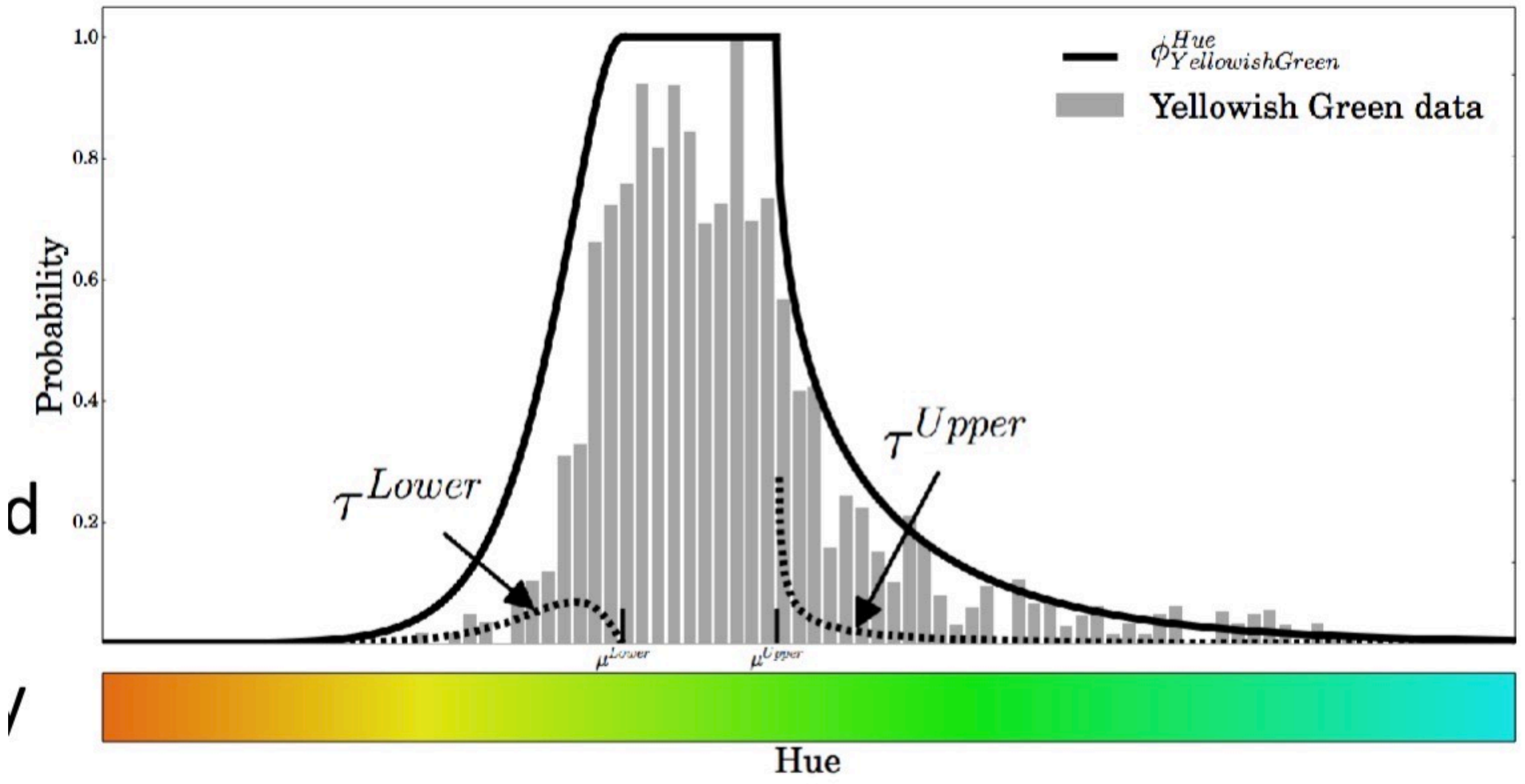
# Color test

- 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

# A chair

# A chair

green

armless

medium size

used to sit on

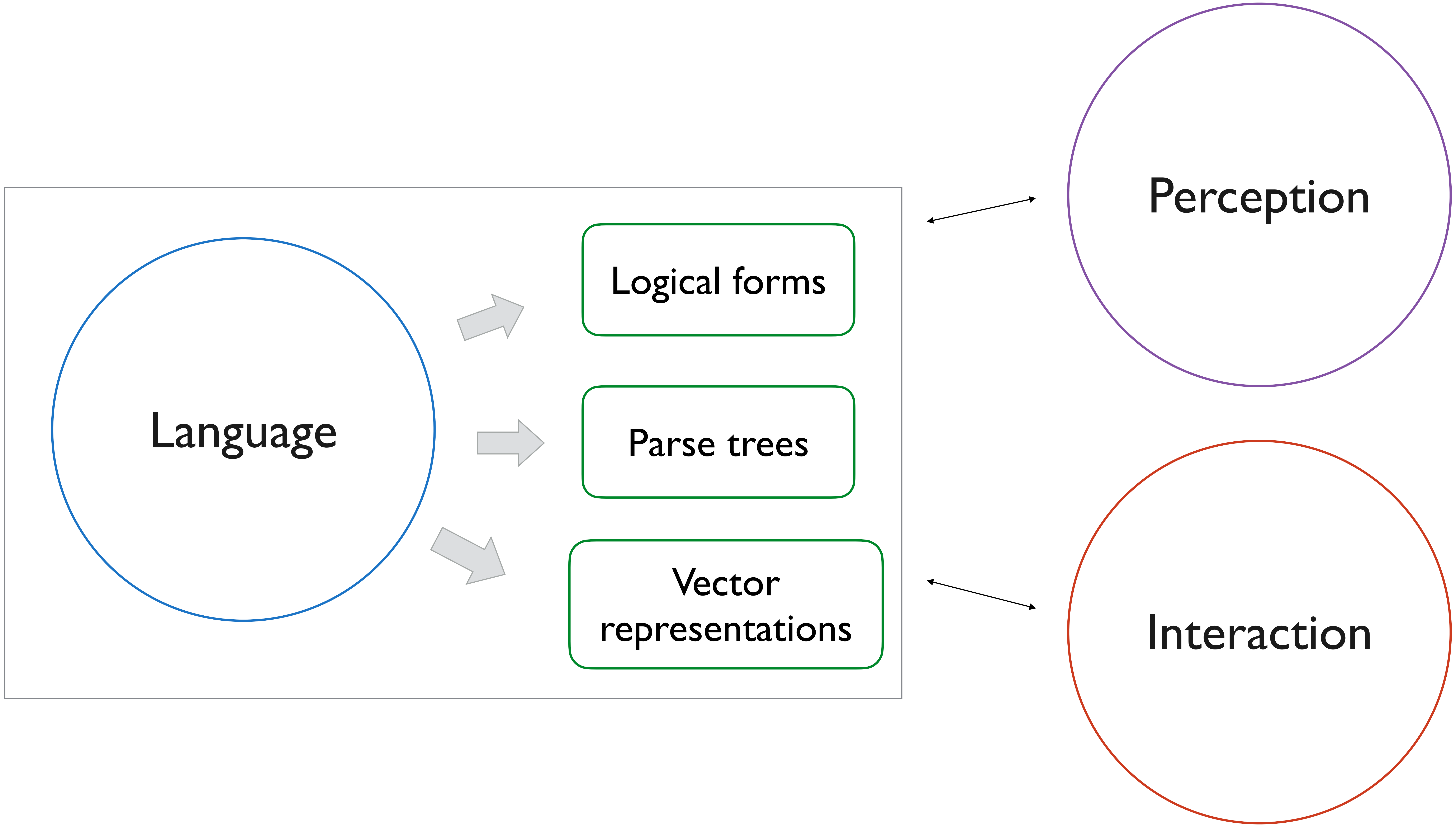
light

fragile

plush

Context is very important!

# 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

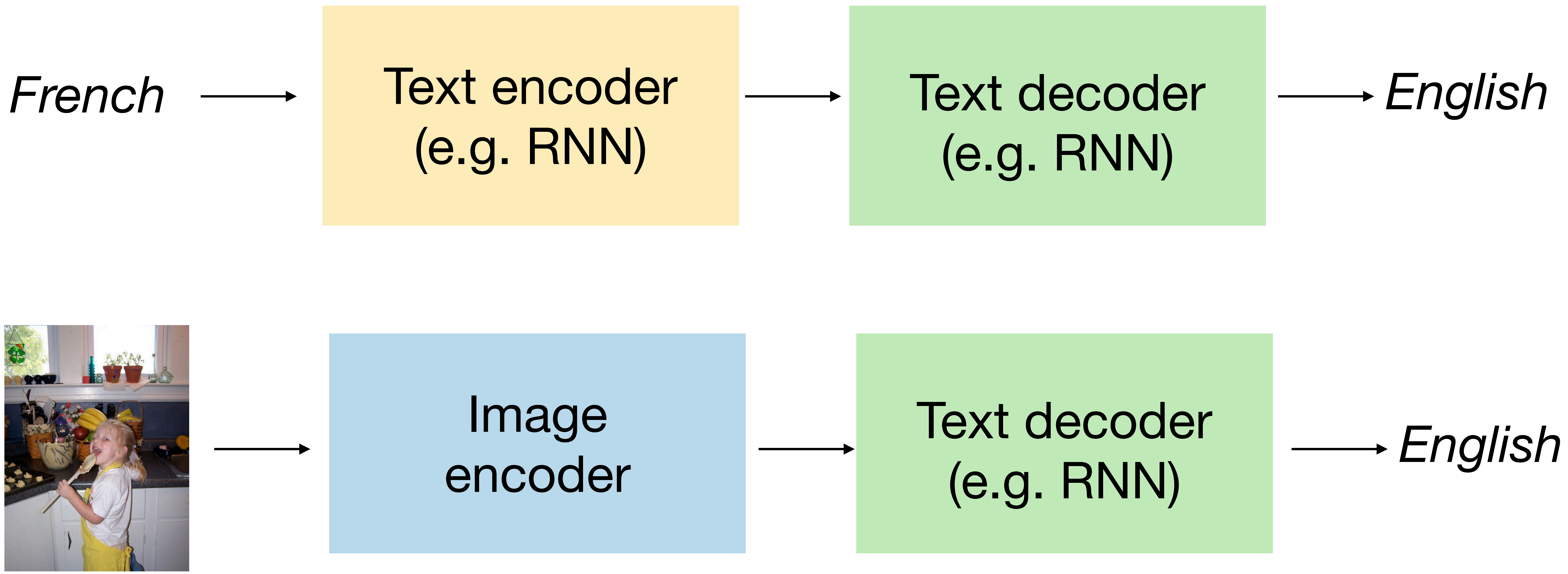
the girl is licking the spoon of batter

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

Applications?

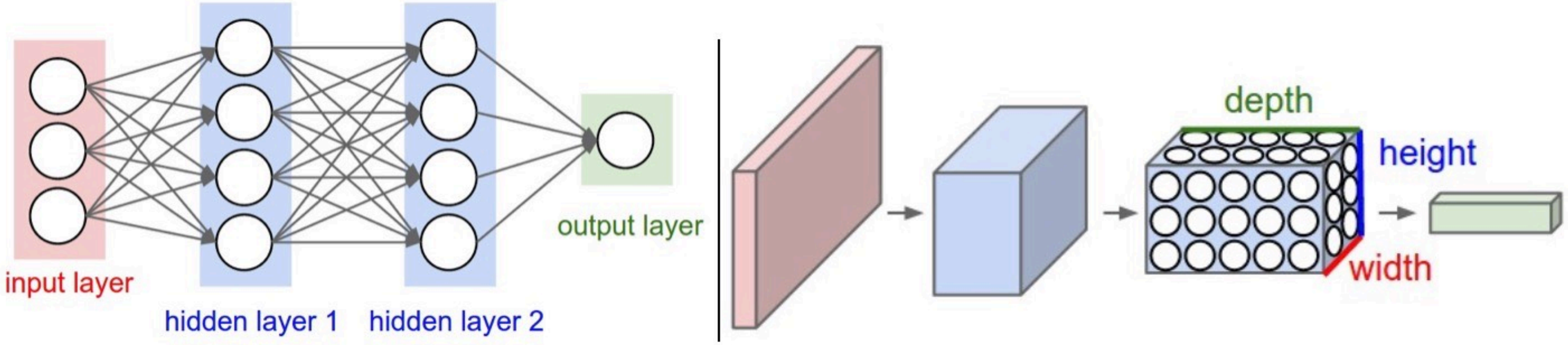
*(MS COCO, Chen et al., 2015)*

# Captioning as multi-modal translation

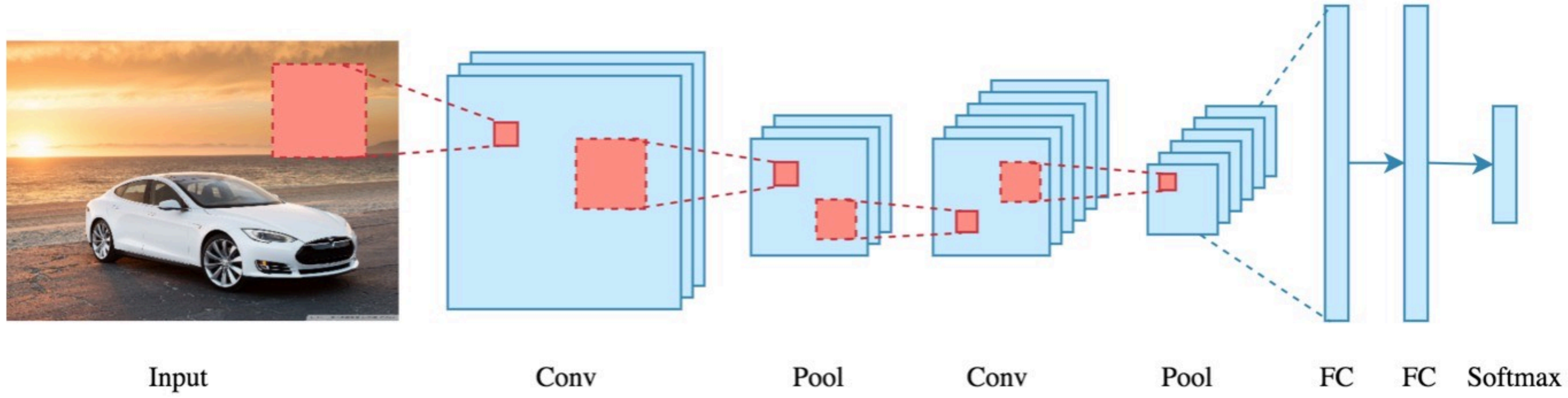



(Donahue et al., 2015, Vinyals et al., 2015)

# Convolutional Neural Networks



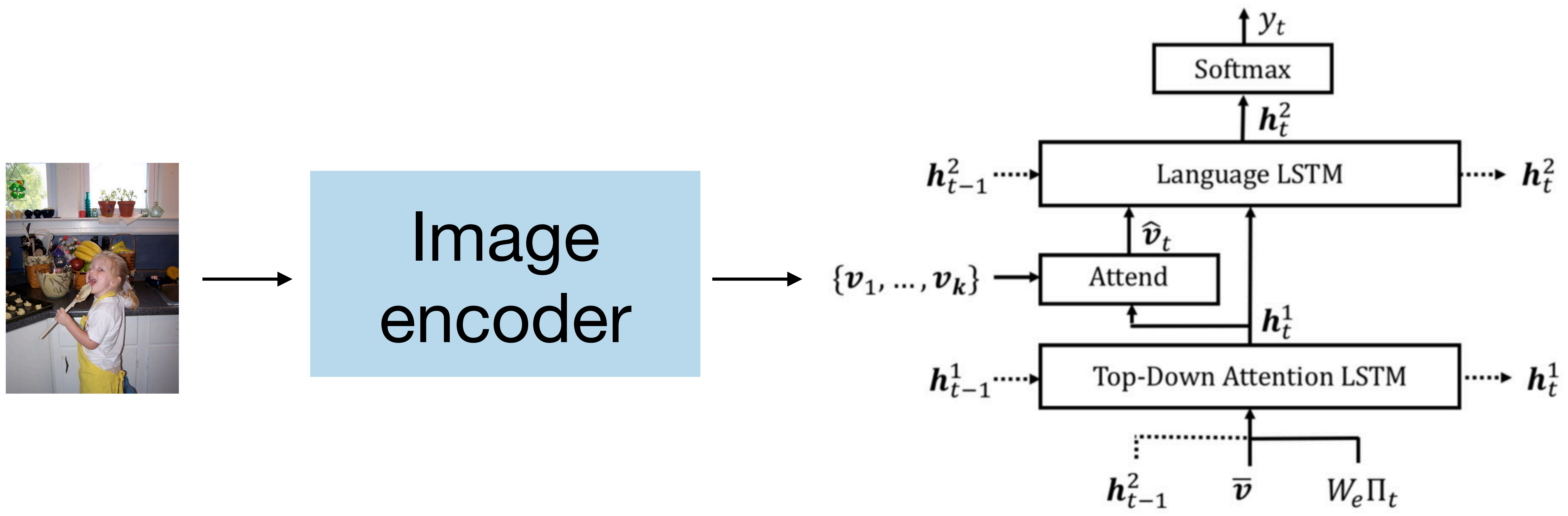

Left: A regular 3-layer Neural Network. Right: A ConvNet arranges its neurons in three dimensions (width, height, depth), as visualized in one of the layers. Every layer of a ConvNet transforms the 3D input volume to a 3D output volume of neuron activations. In this example, the red input layer holds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

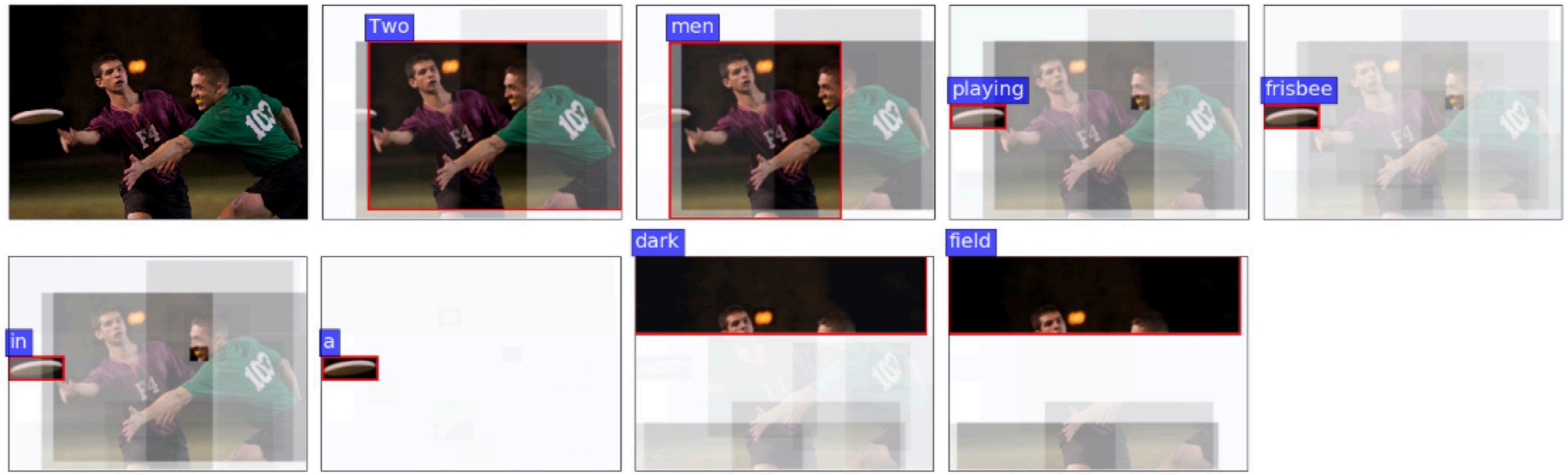



CNN for image classification

*(source: CS231n, Stanford)*

# Captioning with attention

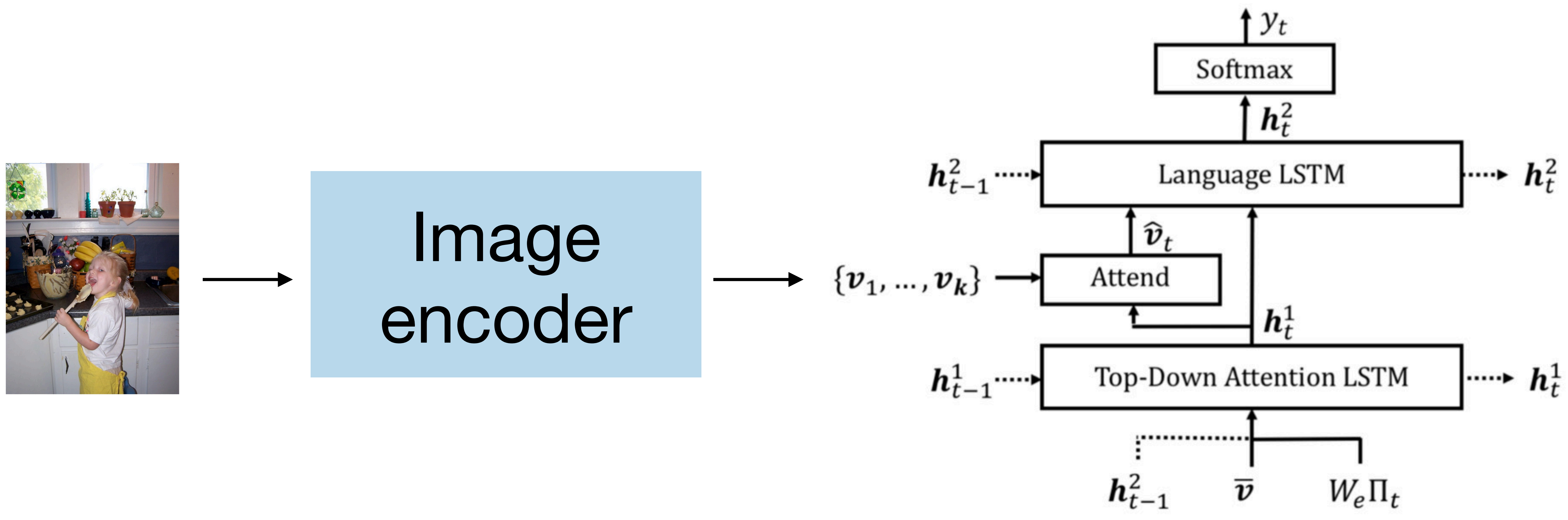



Two men playing frisbee in a dark field.

*(Anderson et al., 2018)*

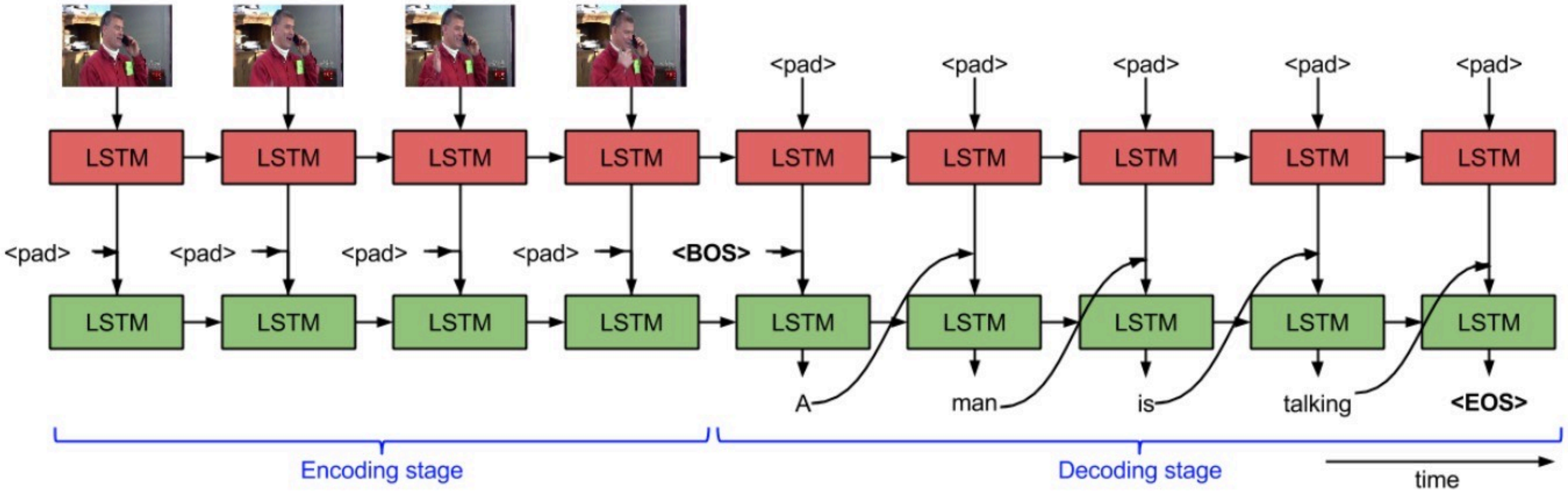
# Captioning with attention




| | 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



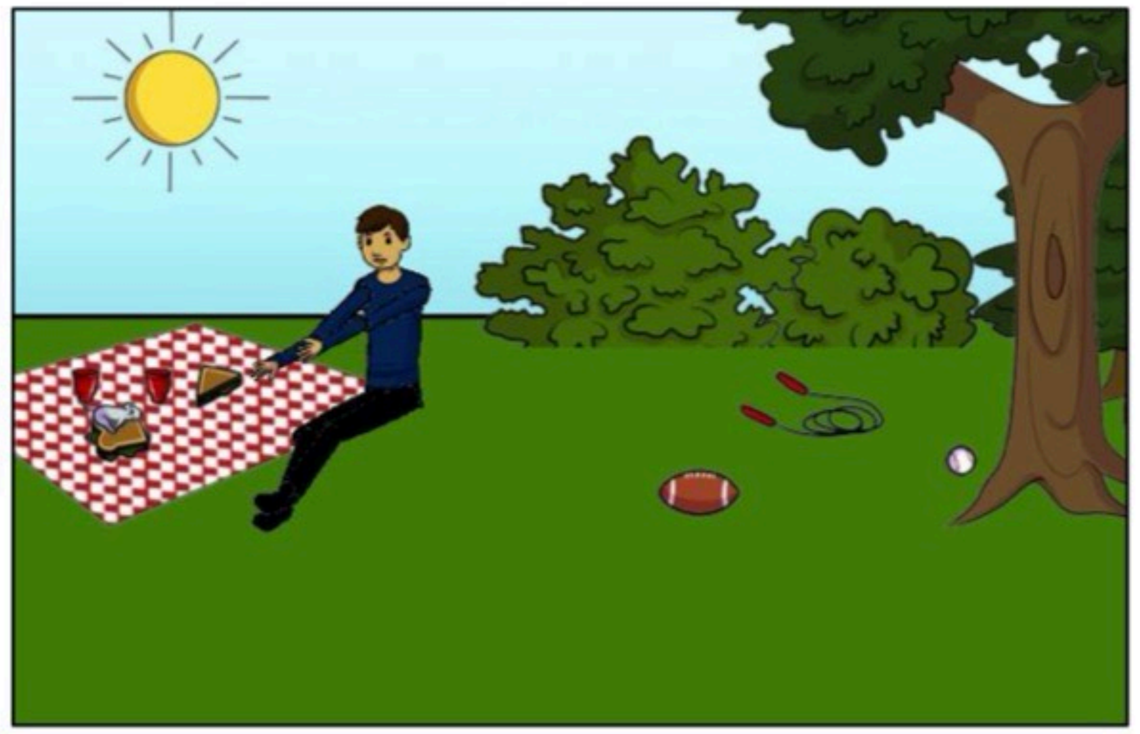

An overview of the S2VT video to text architecture.

(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?

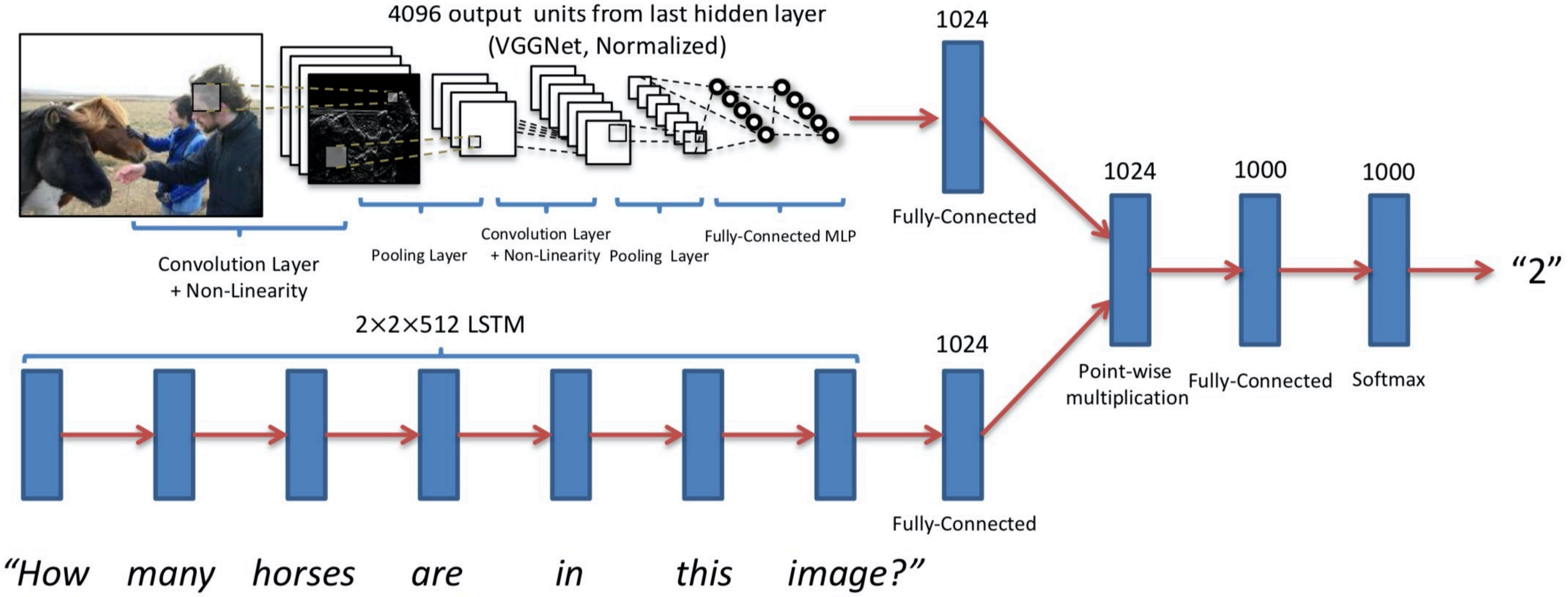

Is this person expecting company?
What is just under the tree?


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

- Require *multi-modal* knowledge and reasoning
- Well-defined *evaluation metric* (accuracy)

*(Agrawal et al., 2015)*

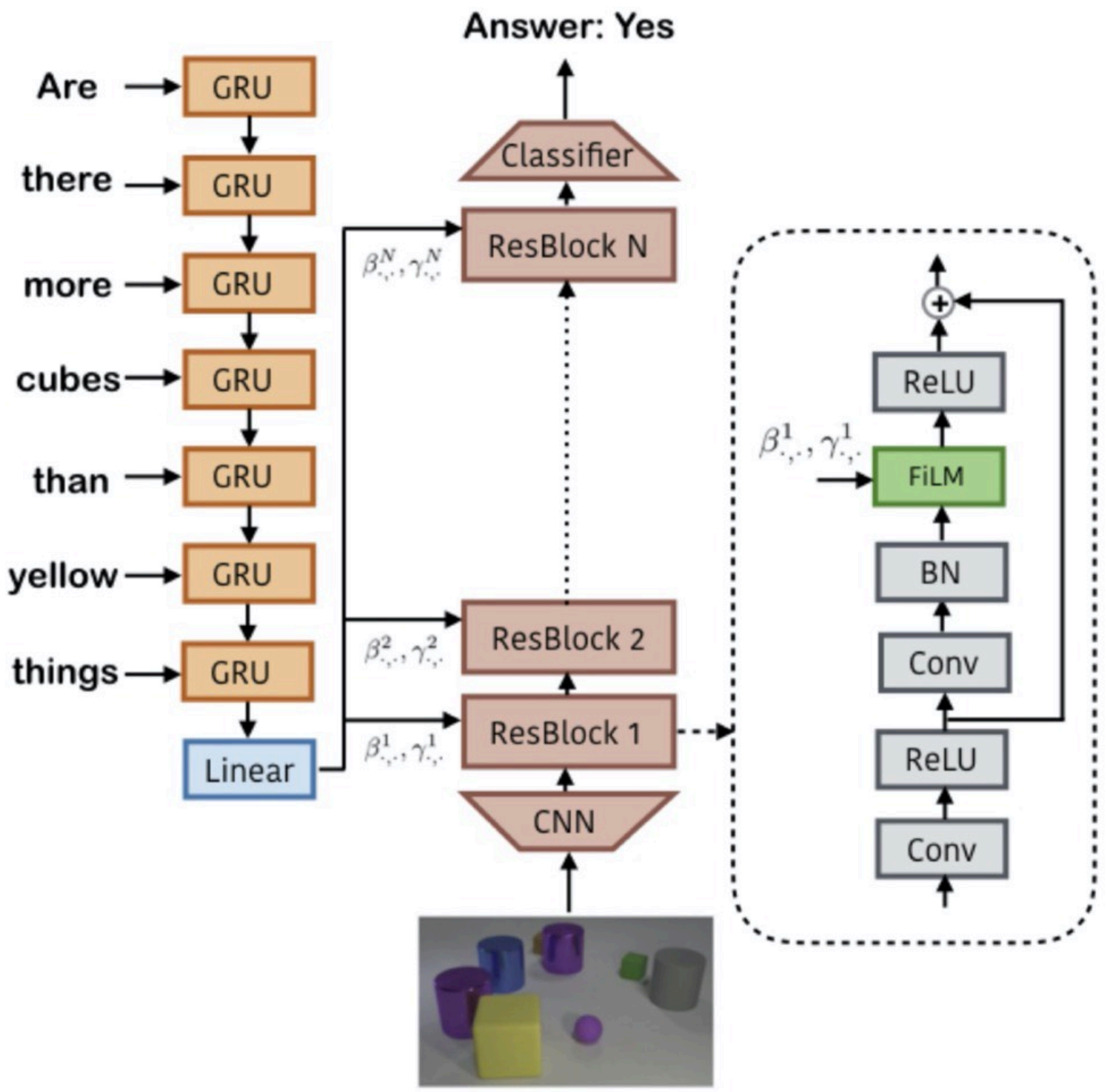
# Visual Question Answering




Any issues?

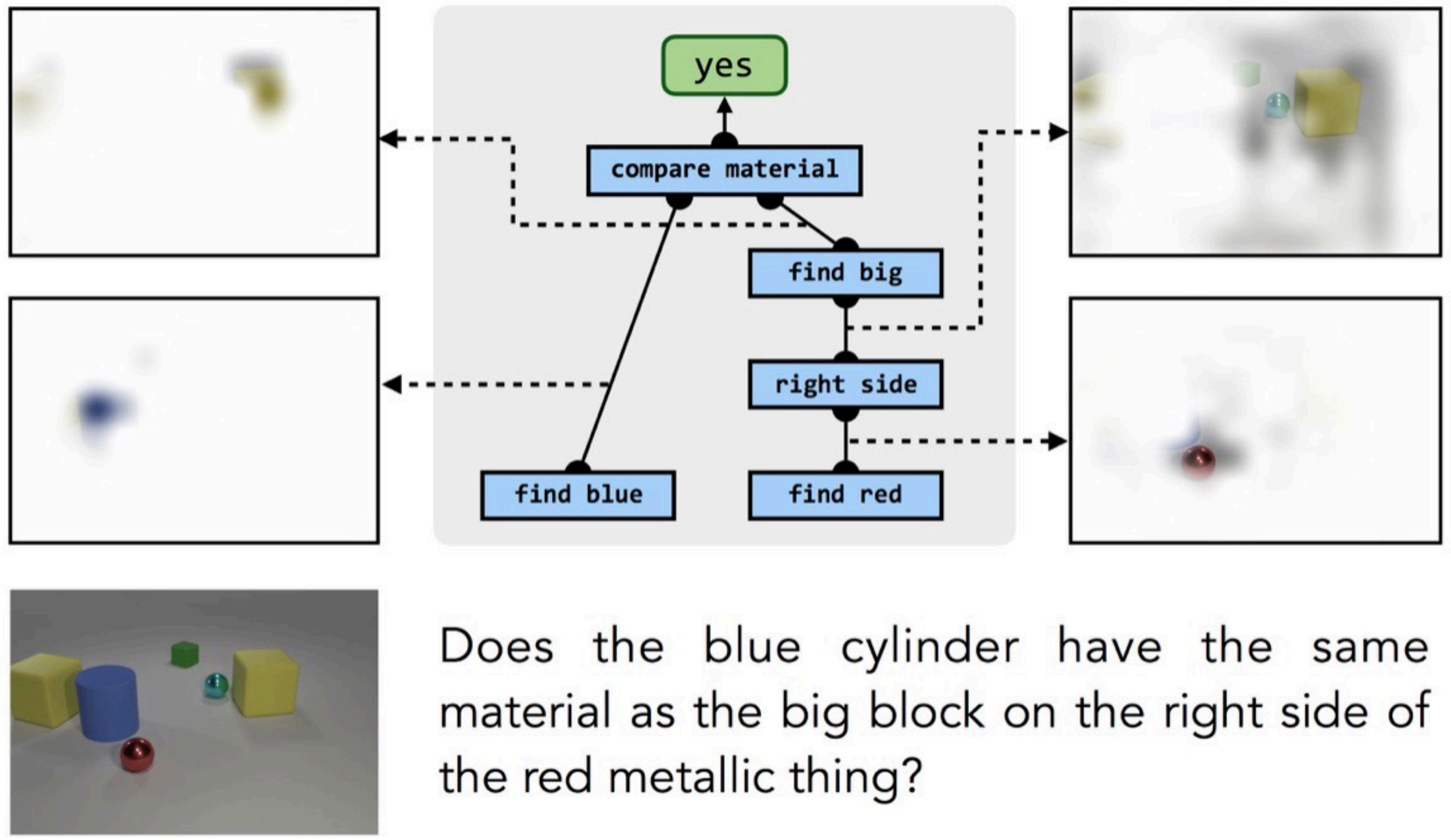
(Agrawal et al., 2015)

# Better multimodal reasoning




FiLM

*(Perez et al., 2017)*




Neural module networks

*(Andreas et al., 2016)*

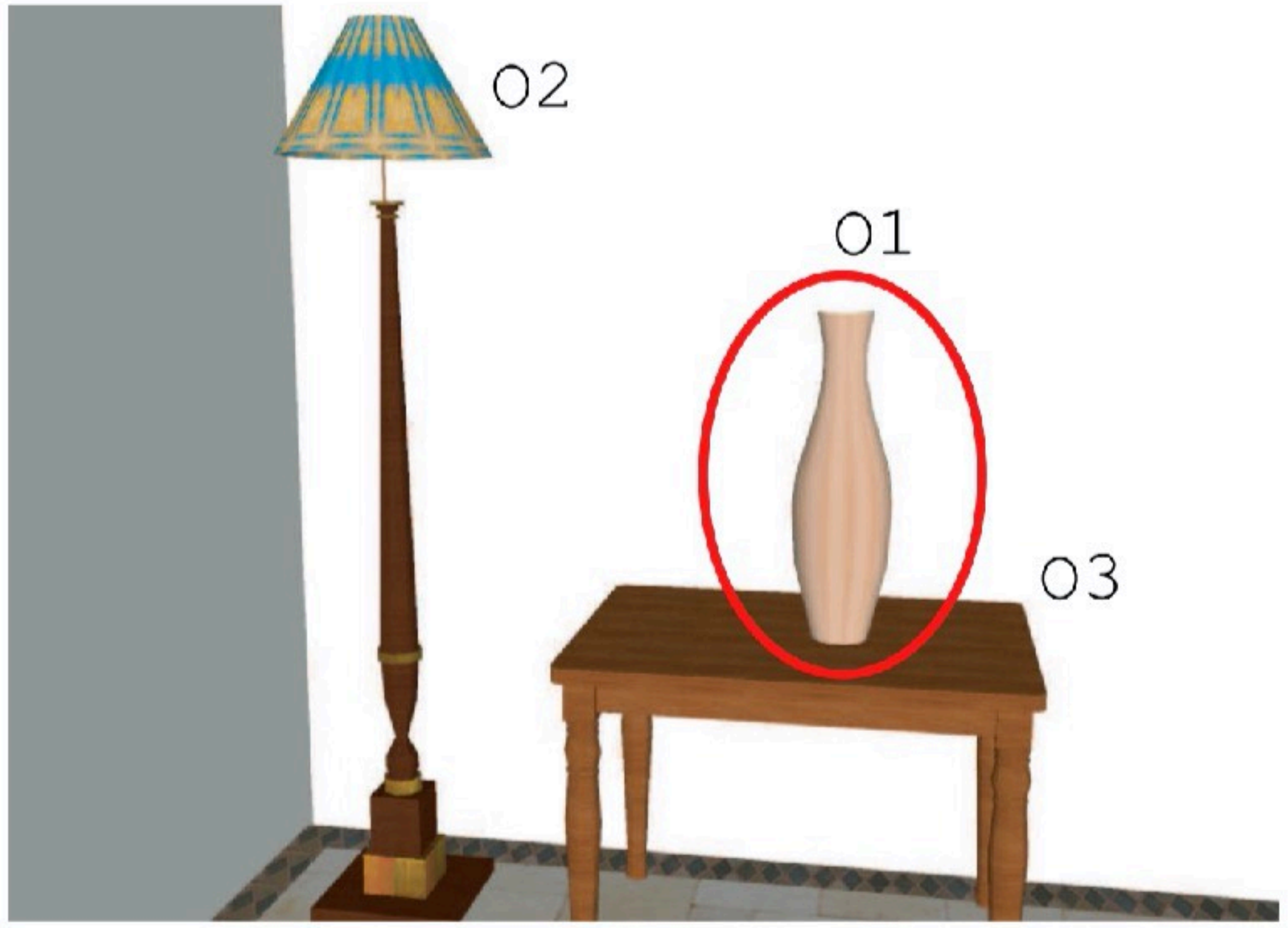
# Visual Question Answering

- 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)*

*(slide adapted from Greg Durrett)*

# Spatial Relations

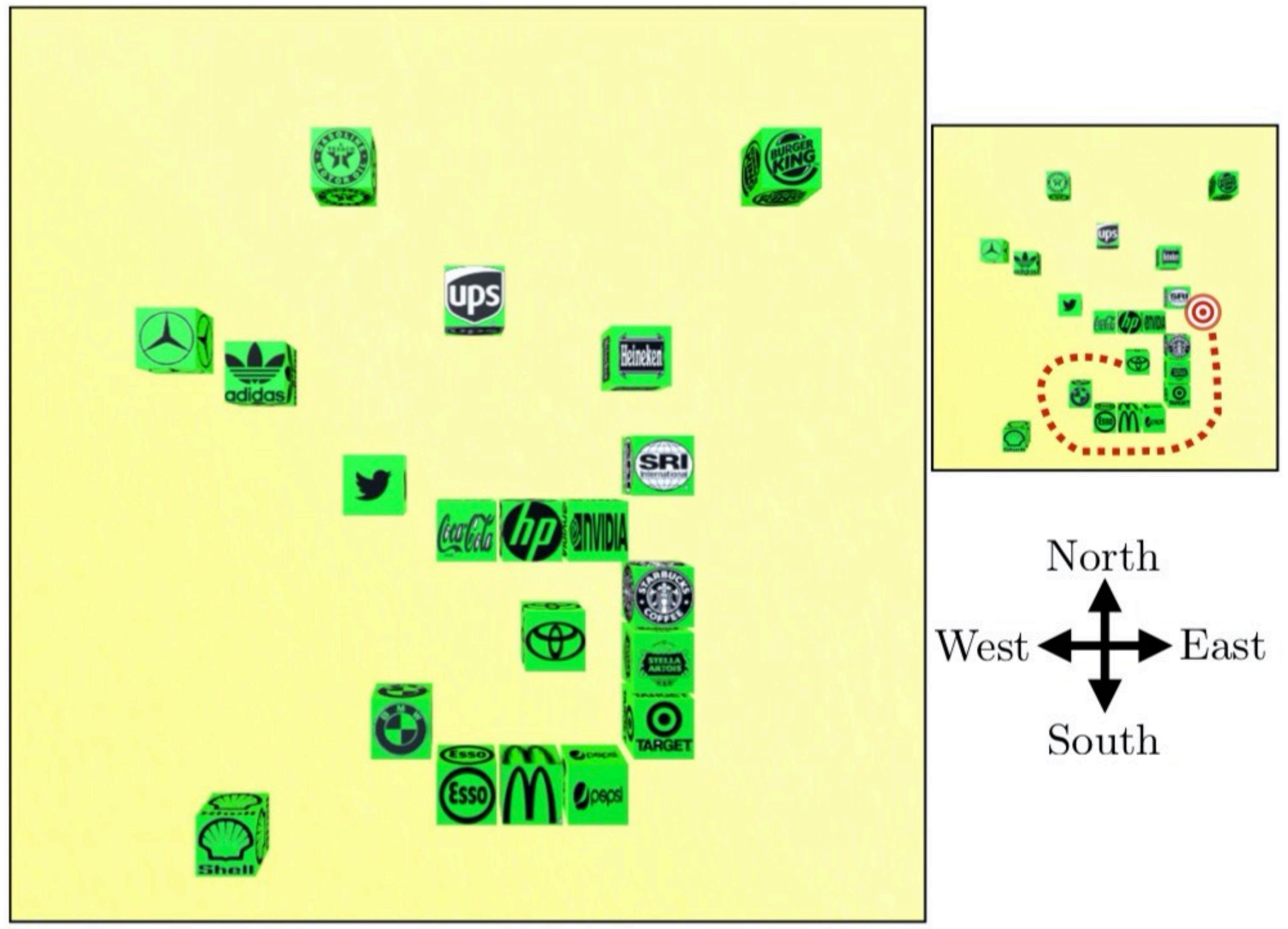



Golland et al. (2010)

- 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

*(slide adapted from Greg Durrett)*

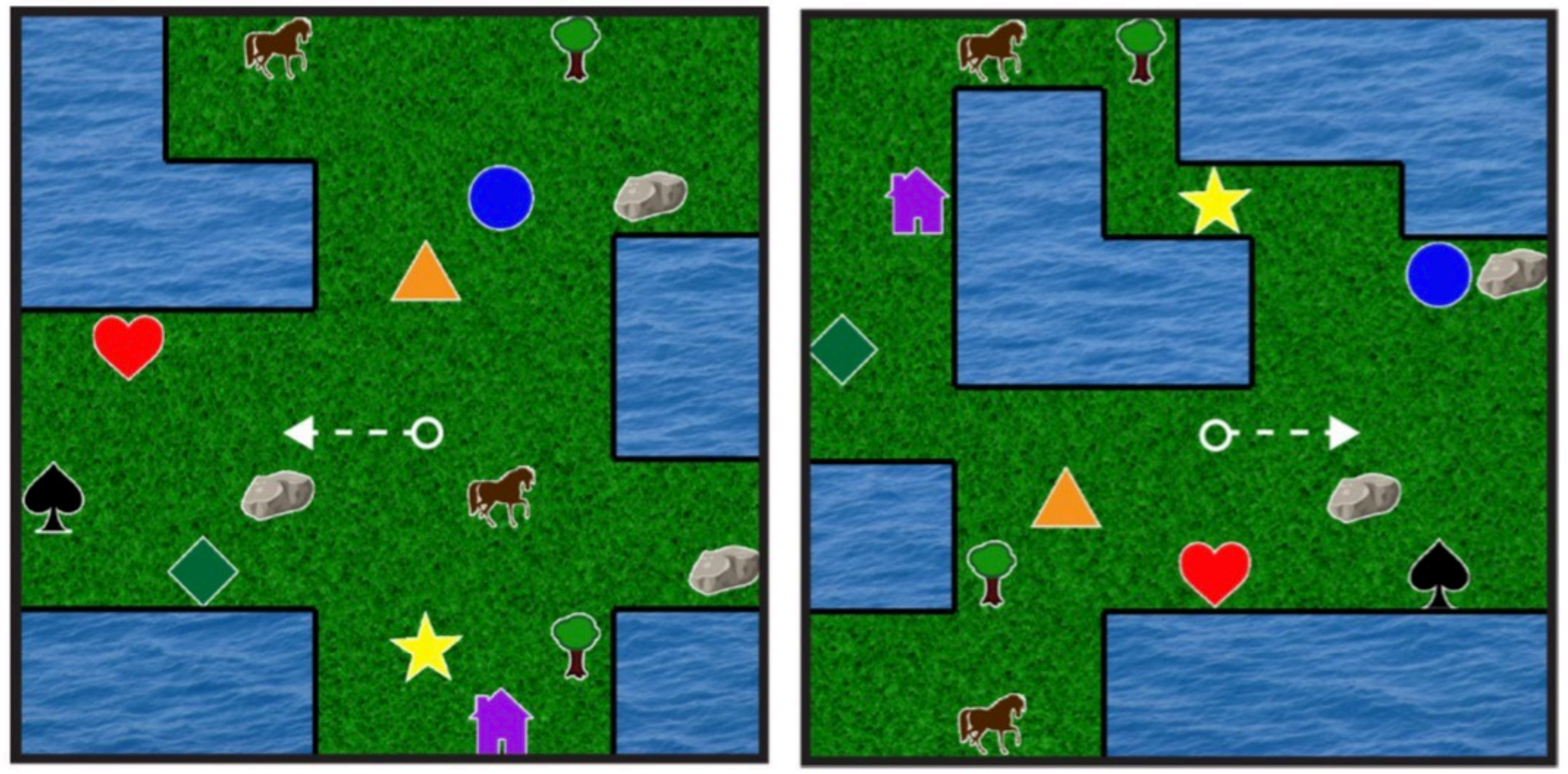
# Spatial Reasoning




| |
|---|
| *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)*

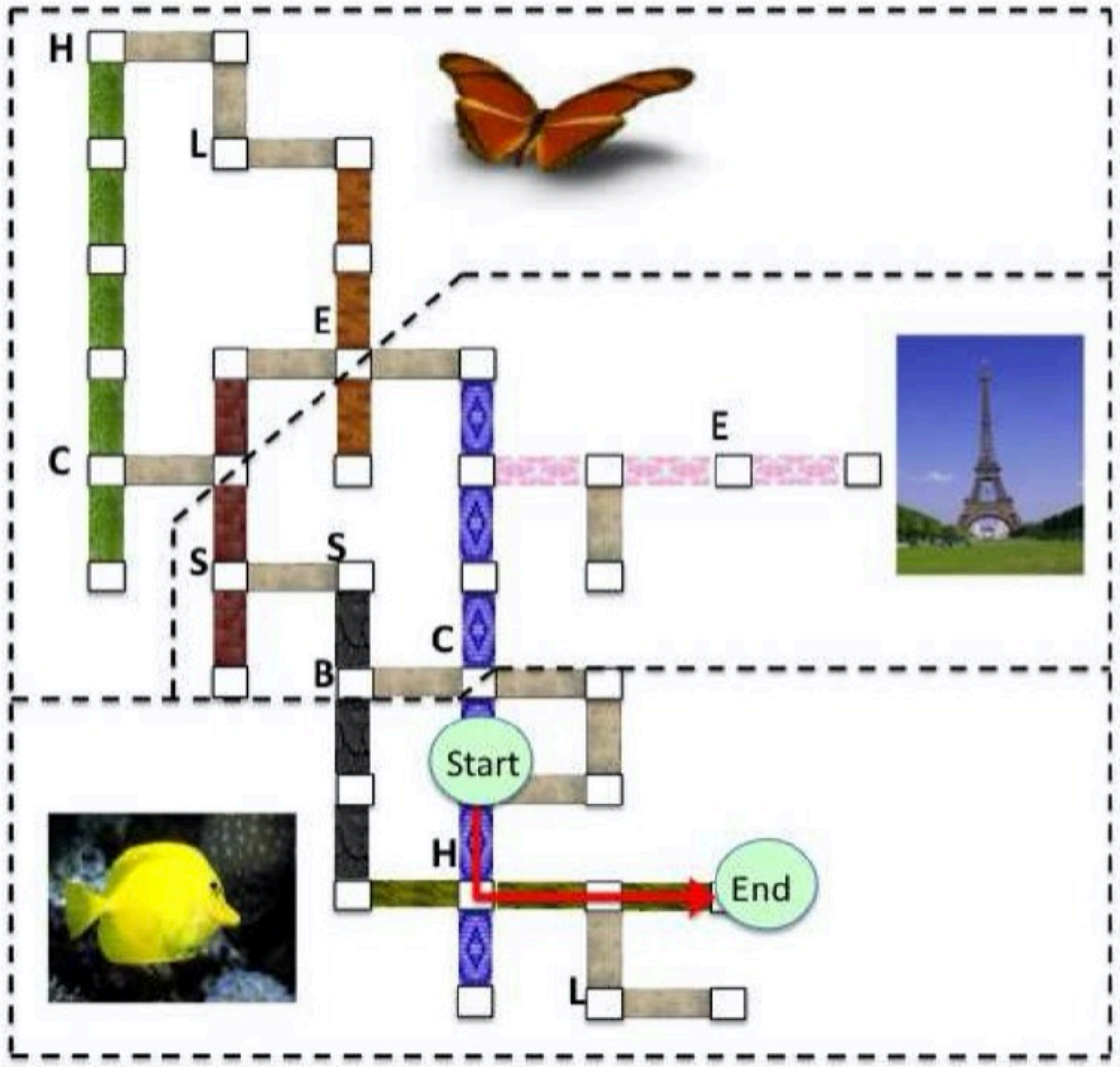


*Reach the cell above the westernmost rock*

Autonomous navigation

*(Janner et al., 2017)*

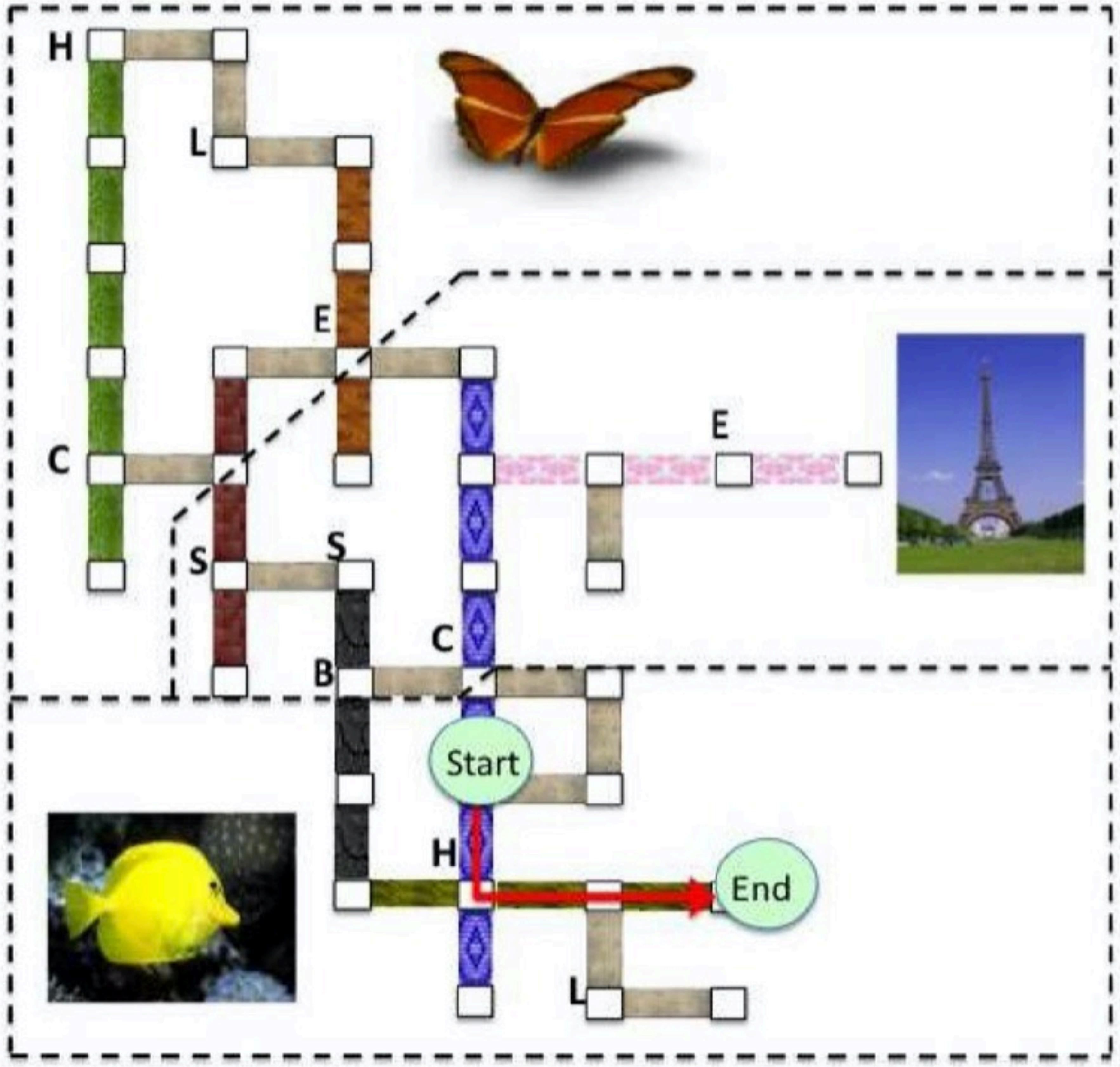
# Instruction Following




- Want to be able to follow instructions in a virtual environment
- "Go along the blue hall, then turn left away from the fish painting and walk to the end of the hallway"

*(MacMahon et al., 2006)*

# Instruction Following




**Instruction:** "Go away from the lamp to the intersection of the red brick and wood"

Basic: Turn ( ) ,
Travel ( steps: 1 )

Landmarks: Turn ( ) ,
Verify ( left: WALL , back: LAMP , back: HATRACK , front: BRICK HALL) ,
Travel ( steps: 1 ) ,
Verify ( side: WOOD HALL )

- Train semantic parser on (utterance, action) pairs
- Language is grounded in actions in the world

*(Chen and Mooney, 2011)*

# Grounding semantics in control applications

1. Use feedback from control application to understand language

Walk across the bridge


*Reward +1*

*Alleviate dependence on large scale annotation*

2. Use language to improve performance in control applications

*Score: 7*

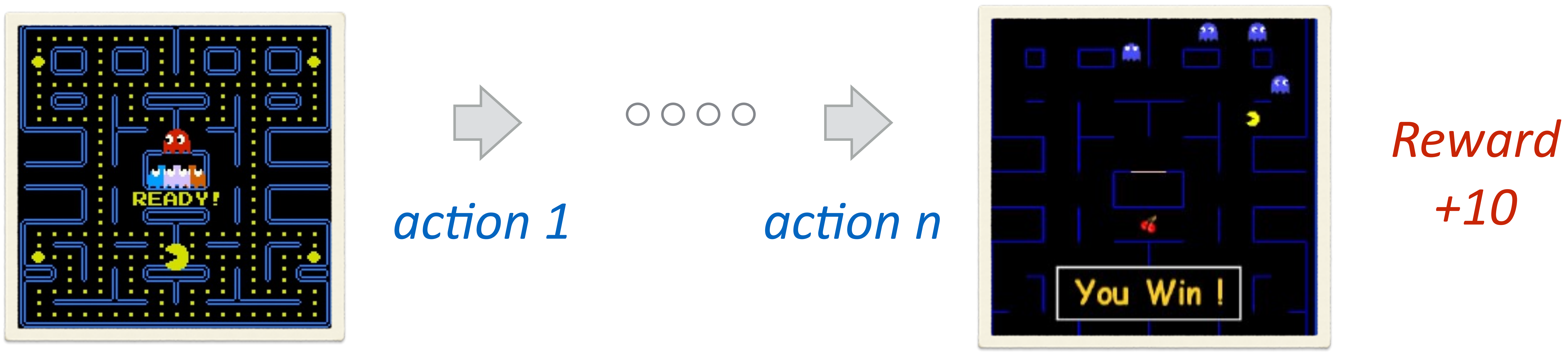
+

1. **Ghosts** chase and try to kill you
2. Collect all the **pellets**
3. …

*Score: 107*

# Reinforcement Learning

- Delayed feedback



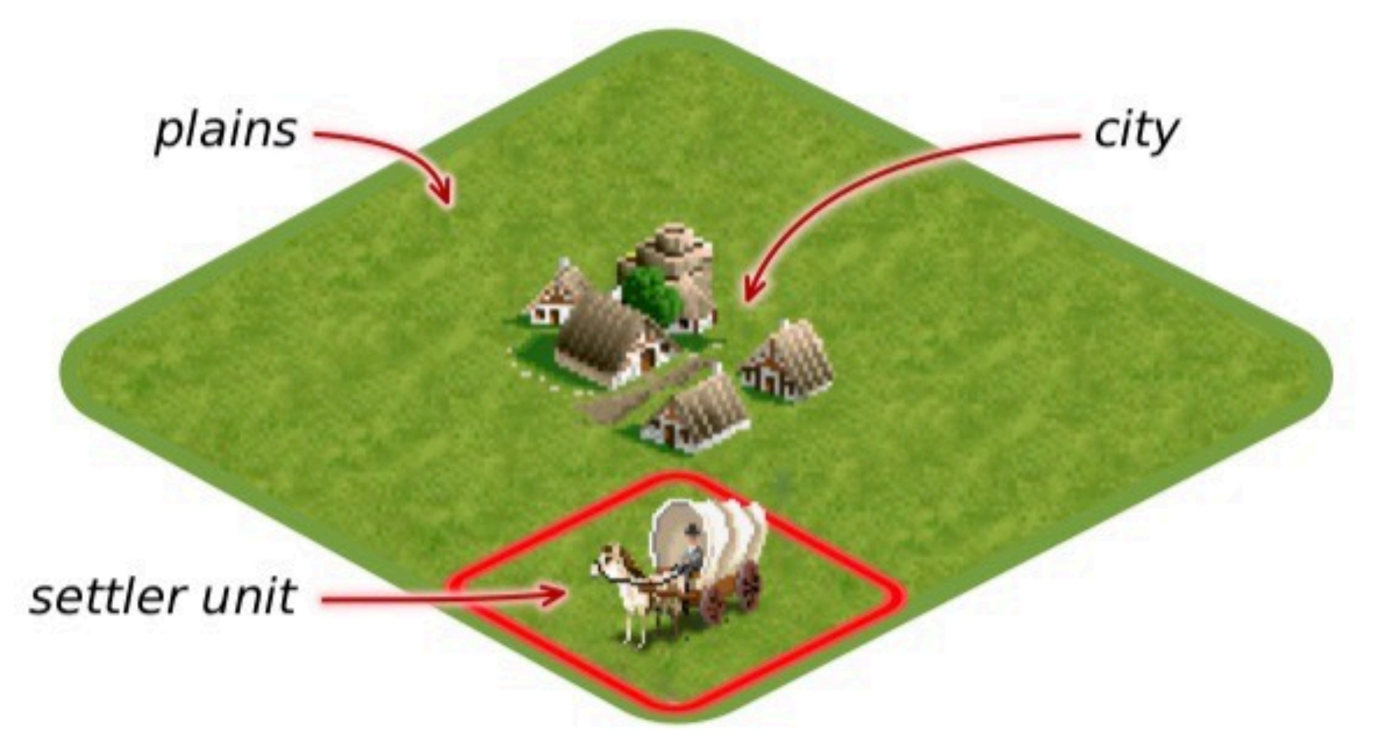

⇒ *How to perform credit assignment for individual actions*

- Large number of possible action sequences

⇒ *Need for effective exploration*

Improved language understanding translates
to improved task performance

# Playing Civilization by reading game manuals




Relevant text: *"Use settlers to irrigate land near your city"*

Predicted action words: *"irrigate", "settler"*

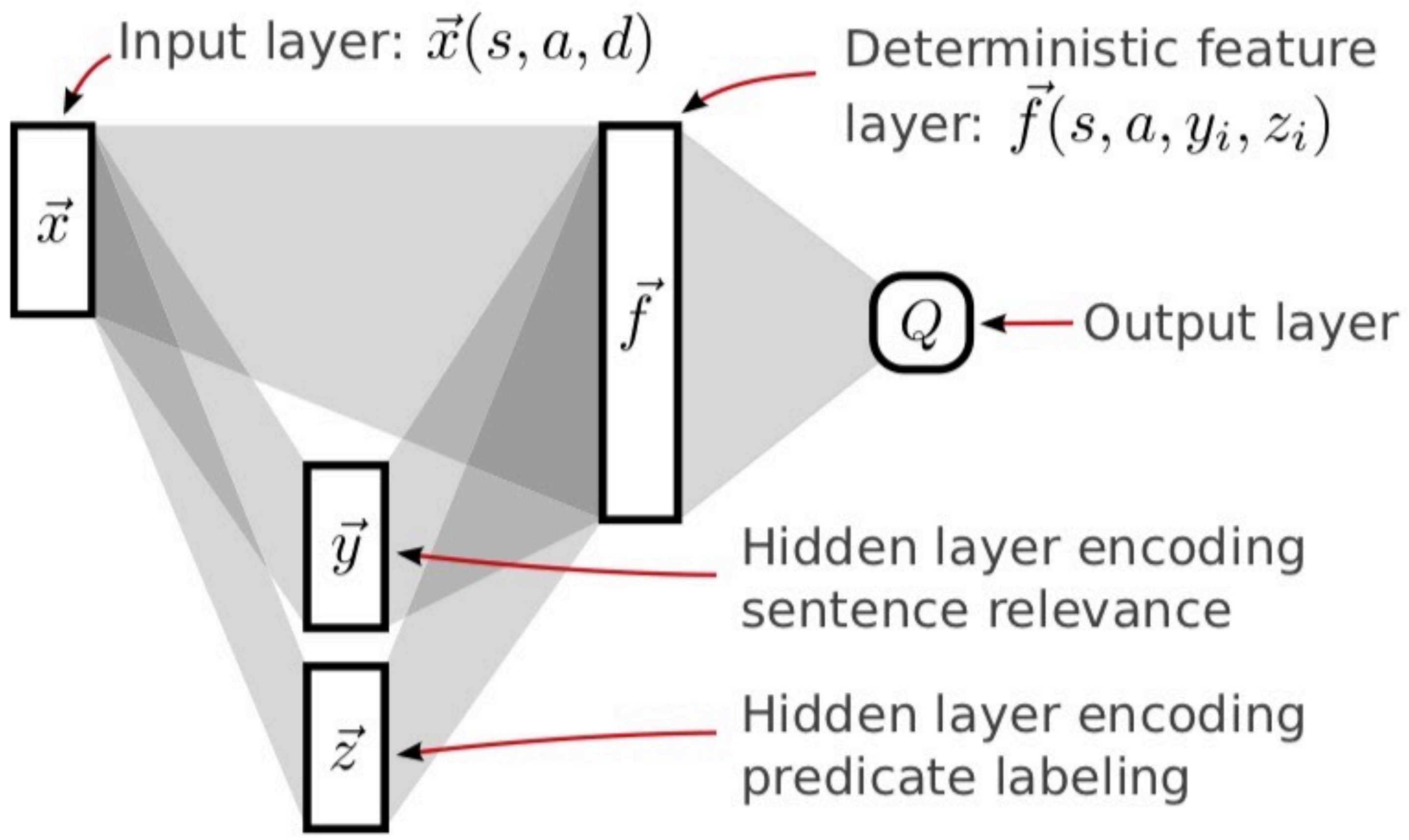
Predicted state words: *"land", "near", "city"*

Settlers unit, candidate action 1: ***irrigate***

Features:
action = **irrigate** and action-word = "*irrigate*"
action = **irrigate** and state-word = "*land*"
action = **irrigate** and terrain = plains
action = **irrigate** and unit-type = settler
state-word = "*city*" and near-city = true

Settlers unit, candidate action 2: ***build-city***

Features:
action = **build-city** and action-word = "*irrigate*"
action = **build-city** and state-word = "*land*"
action = **build-city** and terrain = plains
action = **build-city** and unit-type = settler
state-word = "*city*" and near-city = true



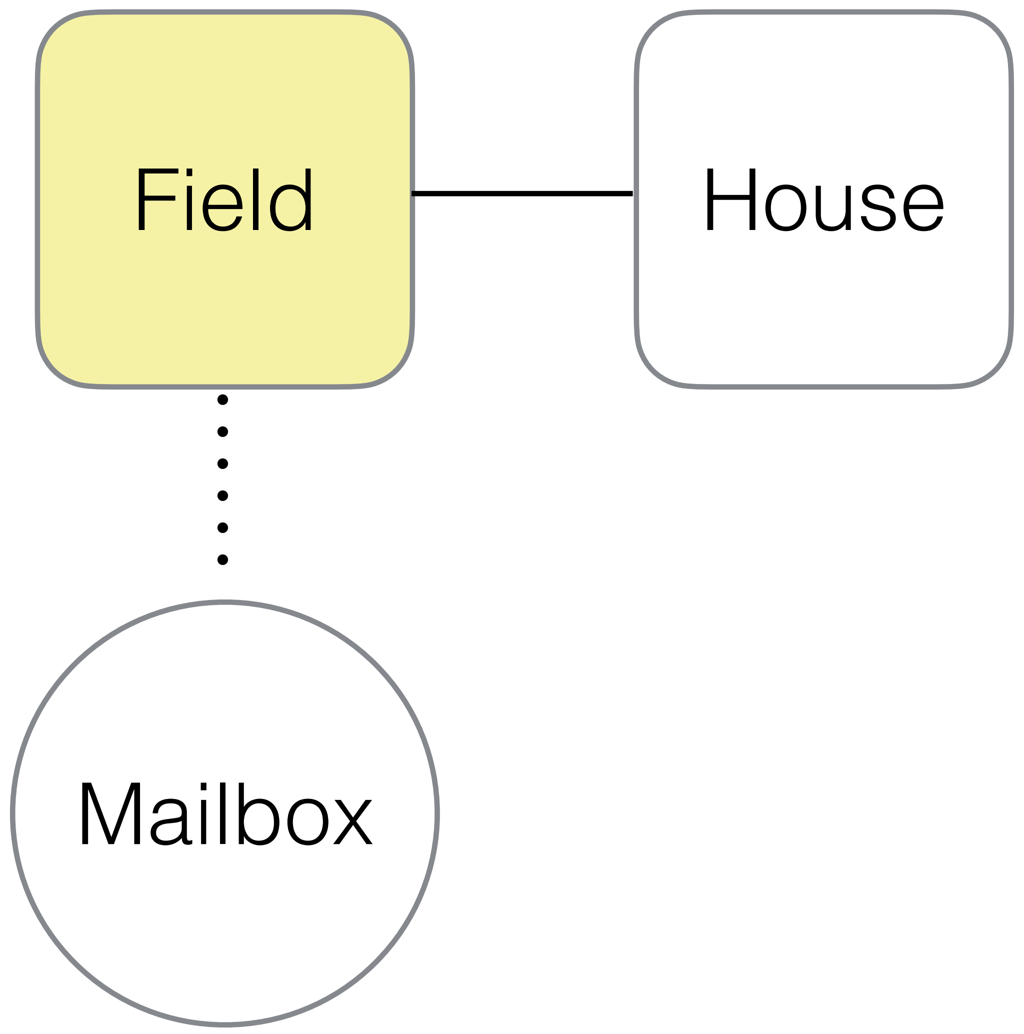

Neural network for policy

| Method | % Win | % Loss | Std. Err. |
|---|---|---|---|
| Random | 0 | 100 | — |
| Built-in AI | 0 | 0 | — |
| Game only | 17.3 | 5.3 | ± 2.7 |
| Latent variable | 26.1 | 3.7 | ± 3.1 |
| **Full model** | **53.7** | 5.9 | ± 3.5 |
| Randomized text | 40.3 | 4.3 | ± 3.4 |

*(Branavan et al., 2012)*

# Text-based games

You are standing in an open field west of a white house, with a boarded front door. There is a small mailbox here.



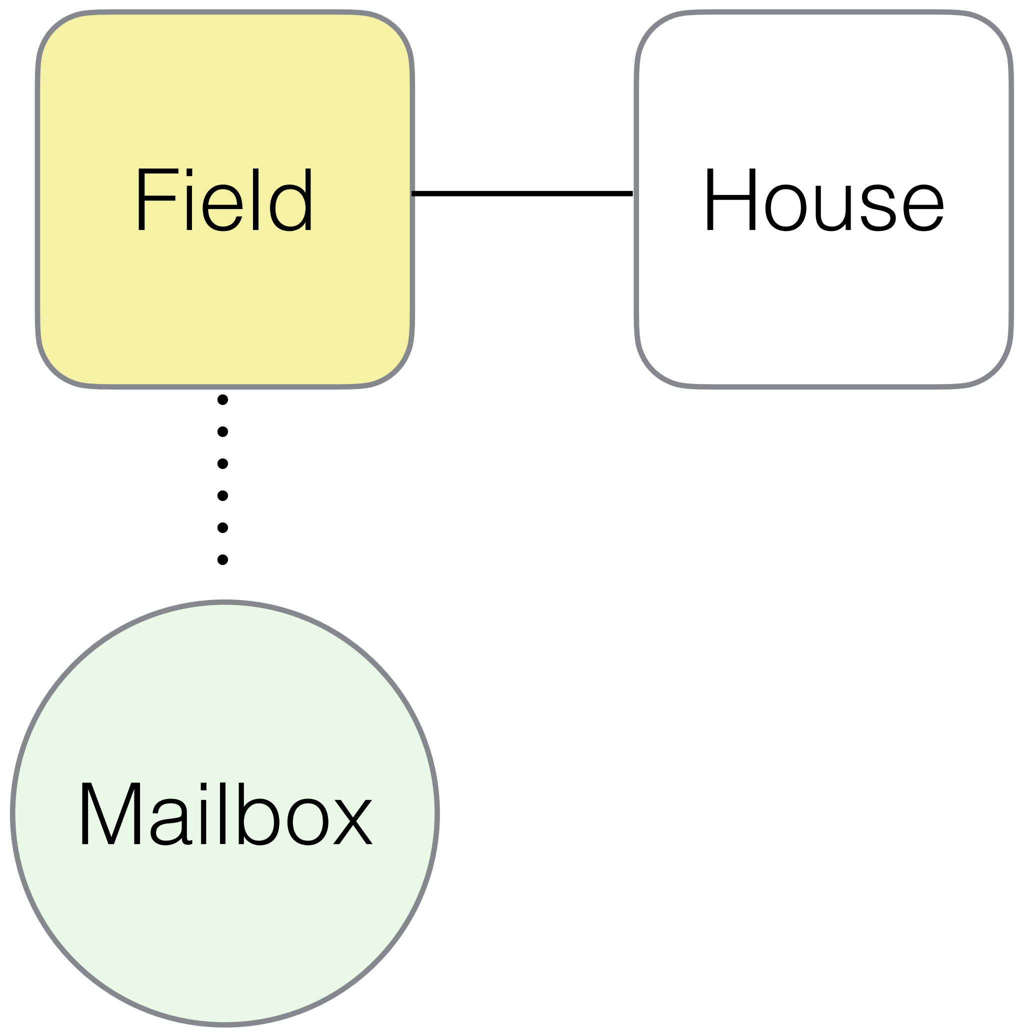

- open mailbox
- go east
- search field

*Underlying game state (h1)*

*(Narasimhan et al., 2015)*

# Text-based games

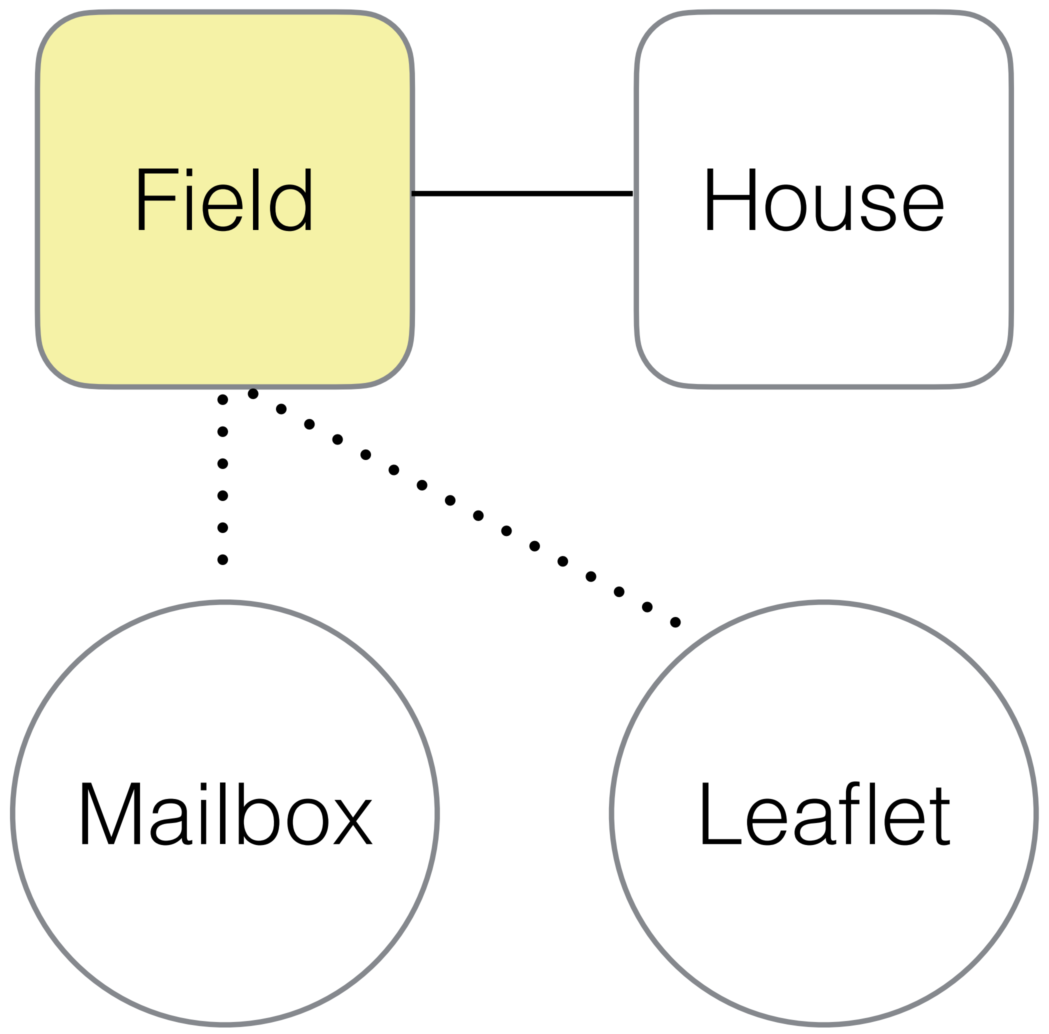
You are standing in an open field west of a white house, with a boarded front door. There is a small mailbox here.




✦ open mailbox

# Text-based games

Opening the mailbox reveals a leaflet.




*Underlying game state (h2)*

*No symbolic representation available*

You are standing in an open field west of a white house, with a boarded front door. There is a small mailbox here.

~~*Location:* Field~~

~~*Wind level:* 3~~

~~*Time:* 12pm~~

*Varying text descriptions*

You are standing in an open field west of a white house, with a boarded front door. There is a small mailbox here.

You are in an open field next to a white house. The house's front door is boarded shut. You see a small mailbox here.

# Opportunity

Grounded language learning

You are standing in an open field west of a white house, with a boarded front door. There is a small mailbox here.

✦ open mailbox

# Opportunity

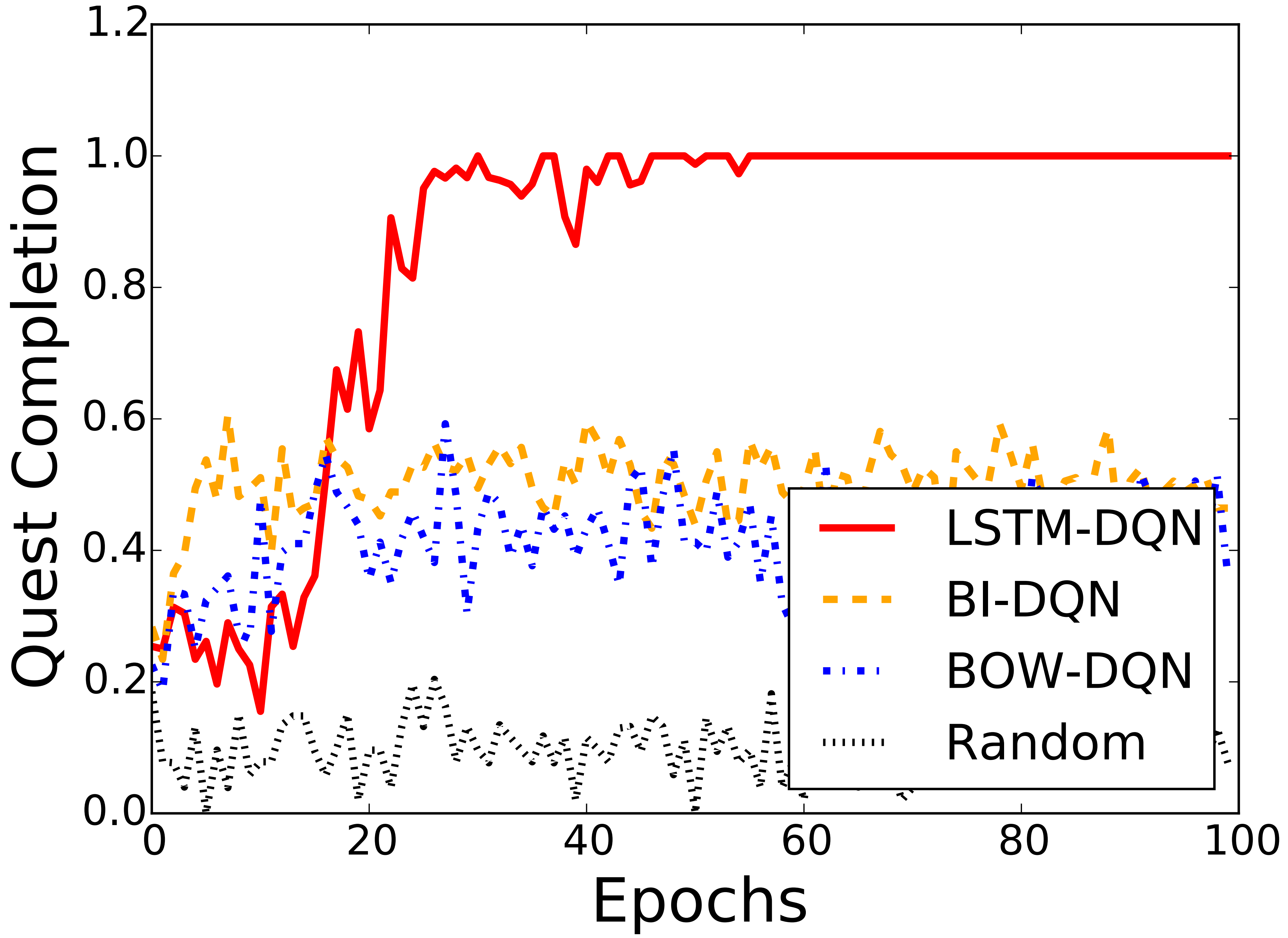
Grounded language learning

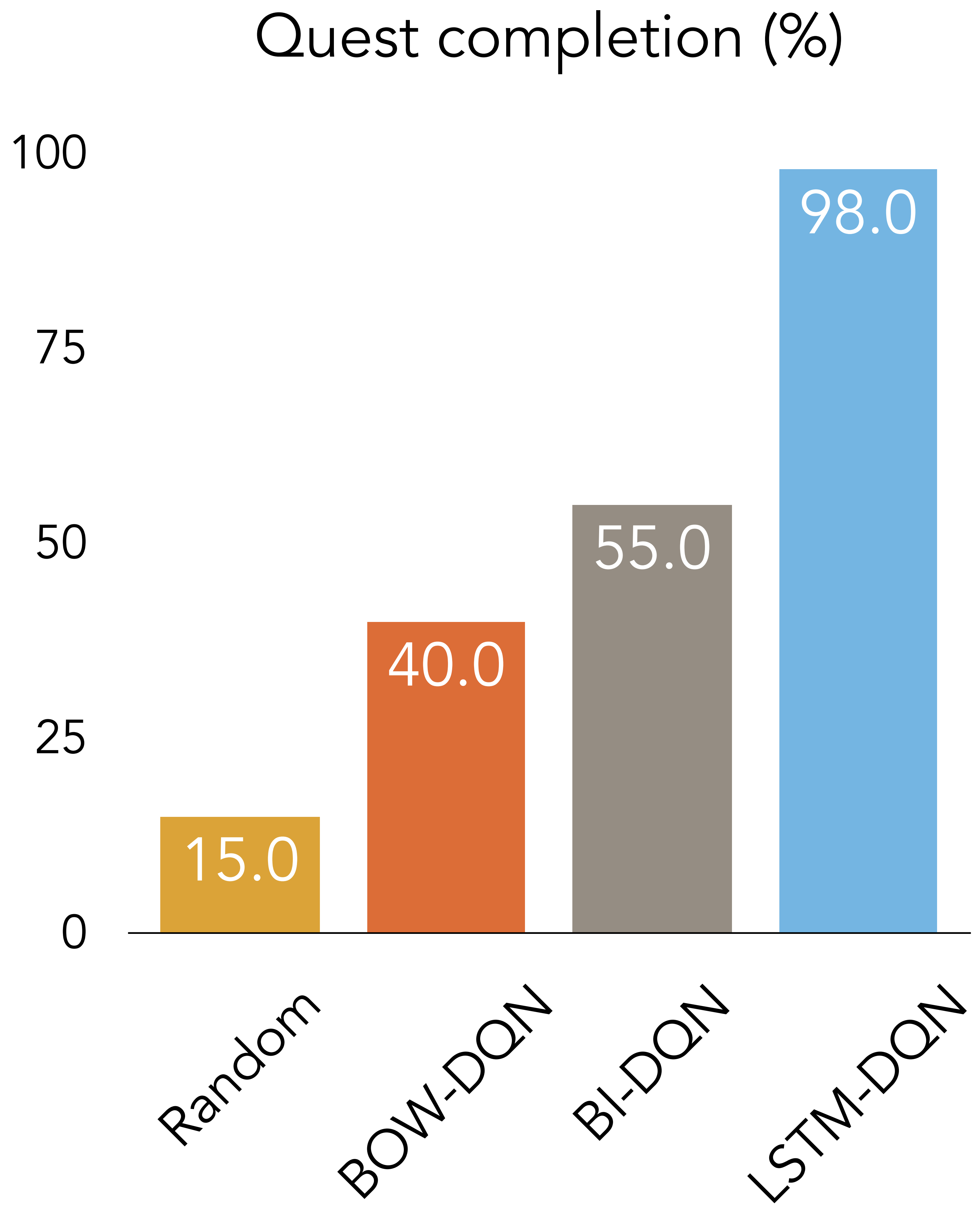
In-game rewards provide unstructured feedback

+10 gold

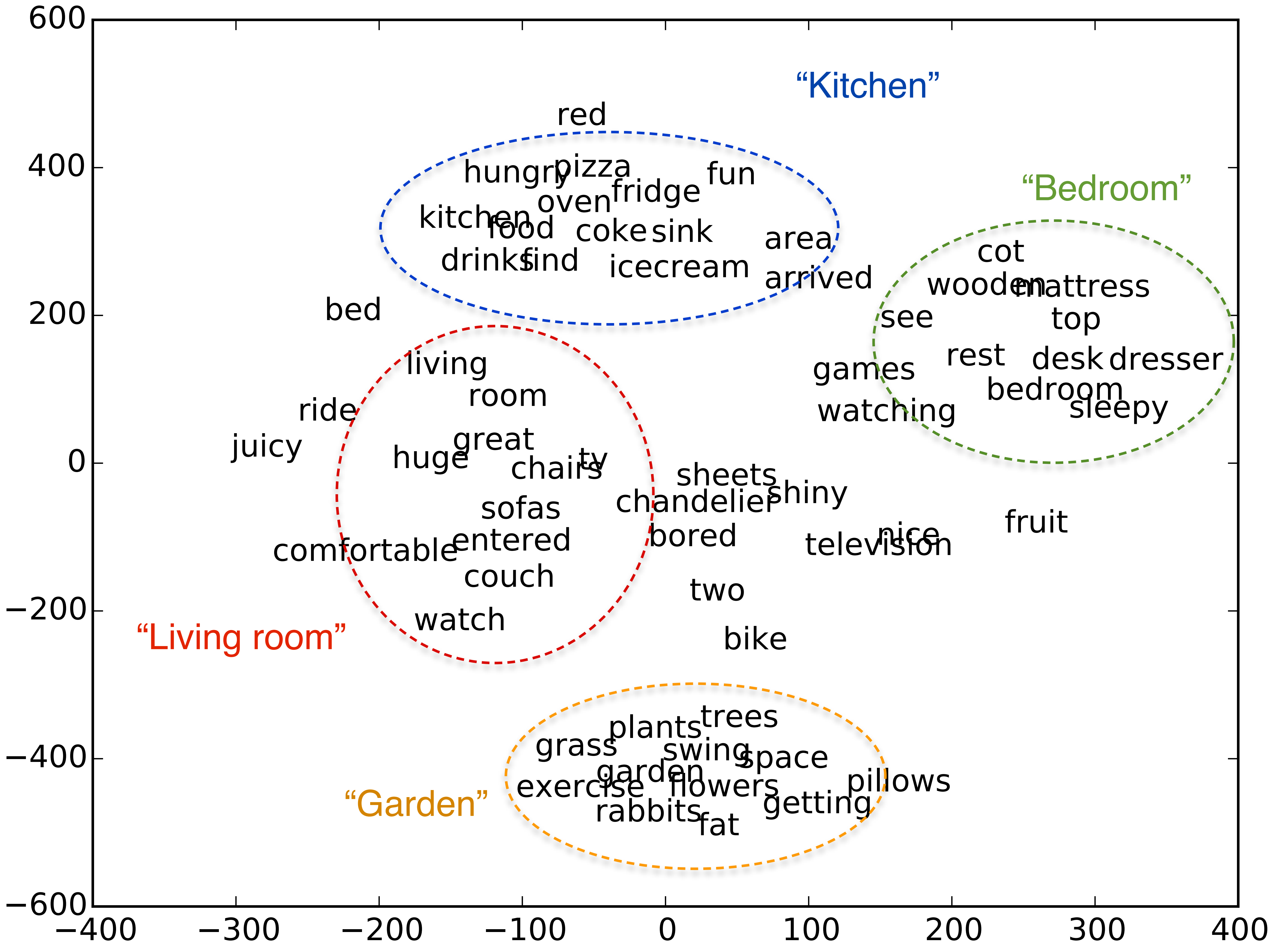
+5 health

# Results


Quest Completion
1.2
1.0
0.8
0.6
0.4
0.2
0.0
0
20
40
60
80
100
Epochs
LSTM-DQN
BI-DQN
BOW-DQN
Random


Quest completion (%)
100
75
50
25
0
15.0
40.0
55.0
98.0
Random
BOW-DQN
BI-DQN
LSTM-DQN

# Visualizing Learnt Representations




t-SNE visualization of vectors learnt by agent

# Contextual Action Language Model (CALM)

- **Want:** Generate sensible action commands
- **Idea:** Train a *single* language model to generate action candidates for any game
  - Actions are subsequently reranked by an RL agent using game-specific rewards

*Observation:* You are in the living room. There is a doorway to the east, a wooden door with strange gothic lettering to the west, which appears to be nailed shut, a trophy case, and a large oriental rug in the center of the room. You are carrying: A brass lantern . . .

**Random Actions**:
close door, north a, eat troll with egg, . . .
**CALM (n-gram) Actions**:
enter room, leave room, lock room, open door, close door, knock on door, . . .
**CALM (GPT-2) Actions**:
east, open case, get rug, turn on lantern, move rug, unlock case with key, . . .

*Next Observation*: With a great effort, the rug is moved to one side of the room, revealing the dusty cover of a closed trap door...

(Yao et al., 2020)

# Semantics does not exist in isolation