Grounded Semantics

Berkeley
N L P

Daniel Fried

with slides from Greg Durrett and Chris Potts

Language is Contextual

- Some problems depend on grounding into perceptual or physical environments:

"Add the tomatoes and mix"

"Take me to the shop on the corner"

- The world only looks like a database some of the time!
- Most of today: these kinds of problems

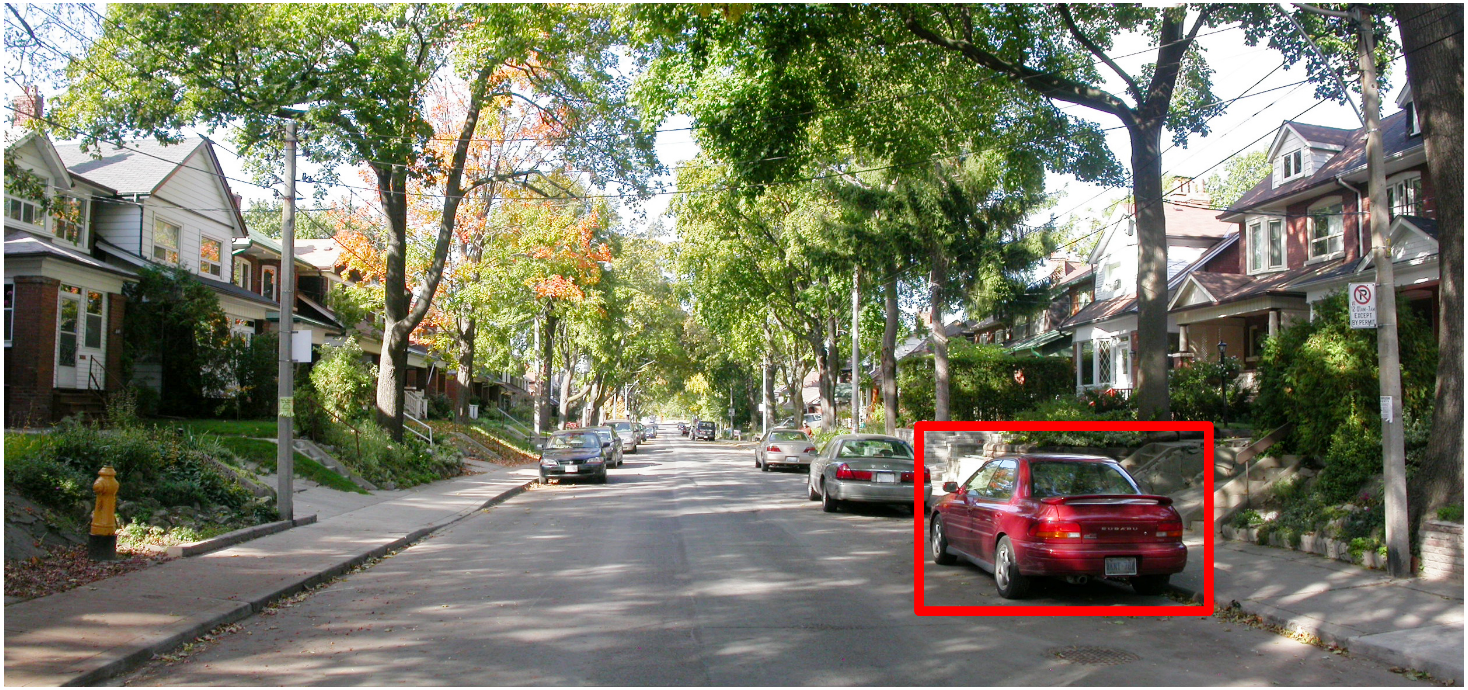
Grounded Semantics

What things in the world does language refer to?

“Stop at the second car”

Pragmatics

How does context influence interpretation and action?



"Stop at the car"

Language is Contextual

- Some problems depend on grounding indexicals, or references to context
- *Deixis*: “pointing or indicating”. Often demonstratives, pronouns, time and place adverbs
  - *I am speaking*
  - *We won* (a team I’m on; a team I support)
  - *He had rich taste* (walking through the Taj Mahal)

  - *I am here* (in my apartment; in this Zoom room)
  - *We are here* (pointing to a map)

  - *I’m in a class now*
  - *I’m in a graduate program now*
  - *I’m not here right now* (note on an office door)

Language is Contextual

- Some problems depend on grounding into speaker intents or goals:
  - “Can you pass me the salt”

    -> please pass me the salt
  - “Do you have any kombucha?” // “I have tea”

    -> I don’t have any kombucha
  - “The movie had a plot, and the actors spoke audibly”

    -> the movie wasn’t very good
  - “You’re fired!”

    -> *performative*, that changes the state of the world

- More on these in a future pragmatics lecture!

Language is Contextual

- Some knowledge seems easier to get with grounding:

Winograd schemas

*The large ball crashed right through the table because it was made of **steel**. What was made of steel?*

-> **ball**

*The large ball crashed right through the table because it was made of **styrofoam**. What was made of styrofoam?*

-> **table**

Winograd 1972; Levesque 2013; Wang et al. 2018

"blinking and breathing problem"

Word	*Teraword*	*Knext*	*Word*	*Teraword*	*Knext*
spoke	11,577,917	244,458	hugged	610,040	10,378
laughed	3,904,519	169,347	blinked	390,692	20,624
murdered	2,843,529	11,284	was late	368,922	31,168
inhaled	984,613	4,412	exhaled	168,985	3,490
breathed	725,034	34,912	was punctual	5,045	511

Table 1: Frequencies from [3] and the number of times Knext learns that *A person may* $\langle x \rangle$, including appropriate arguments, e.g., *A person may hug a person*. For *murder*, more frequently encountered in the passive, we include *be murdered*.

Gordon and Van Durme, 2013

Language is Contextual

- Children learn word meanings incredibly fast, from incredibly few data
    - Regularity and contrast in the input signal
    - Social cues
    - Inferring speaker intent
    - Regularities in the physical environment

Tomasello et al. 2005, Frank et al. 2012, Frank and Goodman 2014

Grounding

- (Some) possible things to ground into:
  - **Percepts**: *red* means this set of RGB values, *loud* means lots of decibels on our microphone, *soft* means these properties on our haptic sensor…
  - **High-level precepts**: *cat* means this type of pattern
  - **Effects on the world**: *go left* means the robot turns left, *speed up* means increasing actuation
  - **Effects on others**: polite language is correlated with longer forum discussions

Grounding

- (Some) key problems:
  - **Representation**: matching low-level percepts to high-level language (pixels vs *cat*)
  - **Alignment**: aligning parts of language and parts of the world
  - **Content Selection / Context**: what are the important parts of the environment to describe (for a generation system) or focus on (for interpretation)?
  - **Balance**: it's easy for multi-modal models to "cheat", rely on imperfect heuristics, or ignore important parts of the input
  - **Generalization**: to novel world contexts / combinations

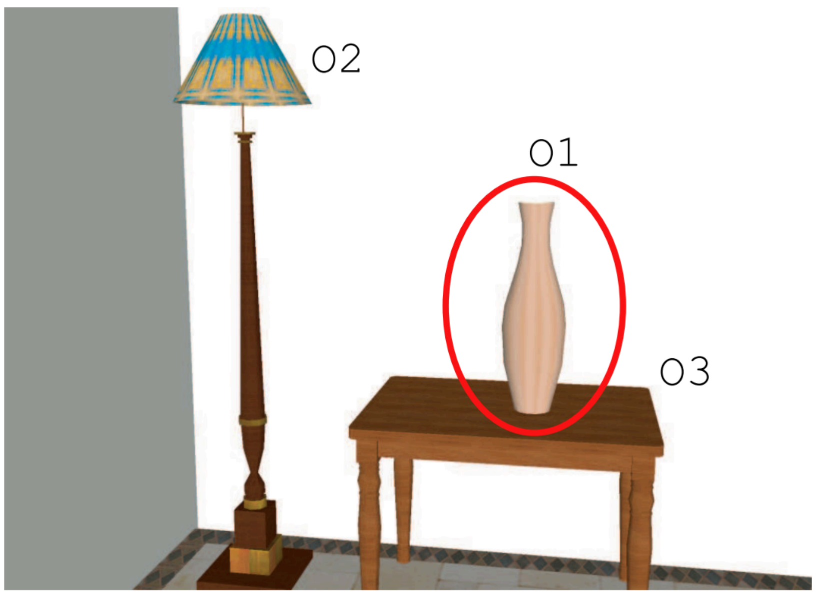
Grounding

- Today, survey:
  - Spatial relations
  - Image captioning
  - Visual question answering
  - Instruction following

Spatial Relations

Spatial Relations

Golland et al. (2010)



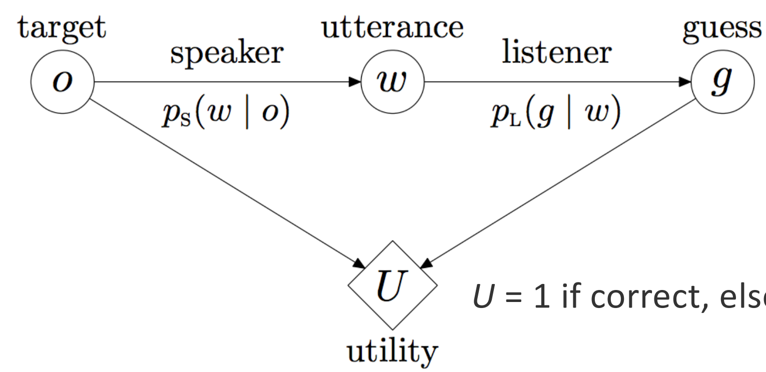
- How would you indicate O1 to someone with relation to the other two objects?
  (not calling it a vase, or describing its inherent properties)
- What about O2?
- Requires modeling listener — "right of O2" is insufficient though true

Spatial Relations

- Two models: a speaker, and a listener
- We can compute expected success:

$$\mathrm{EU}(\mathrm{S},\mathrm{L}) = \sum_{o,w,g} p(o) p_{\mathrm{S}}(w|o) p_{\mathrm{L}}(g|w) U(o,g)$$

Golland et al. (2010)



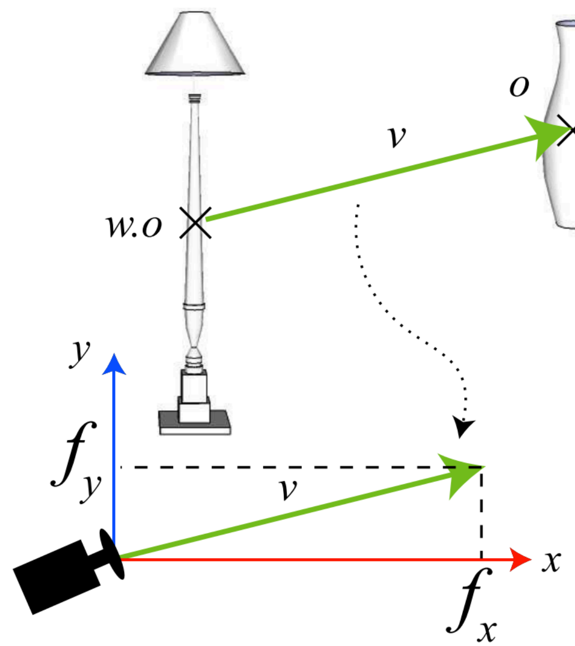
U = 1 if correct, else 0

- Modeled after cooperative principle of Grice (1975) : listeners should assume speakers are cooperative, and vice-versa
- For a fixed listener, we can solve for the optimal speaker, and vice-versa

Spatial Relations

- Listener model:
  - Objects are associated with coordinates (bounding boxes of their projections). Features map lexical items to distributions ("right" modifies the distribution over objects to focus on those with higher x coordinate)
  - Language -> spatial relations -> distribution over what object is intended

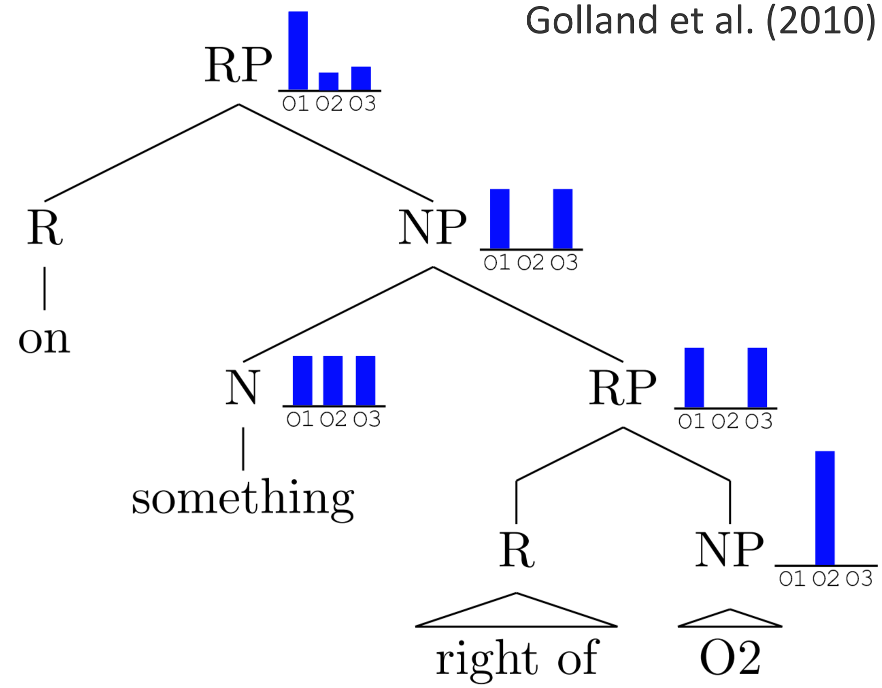
Golland et al. (2010)



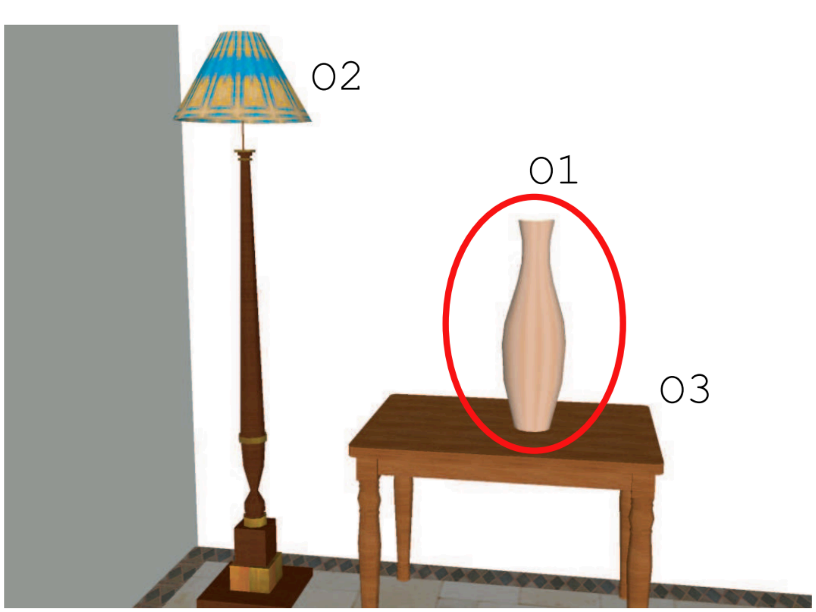
Spatial Relations

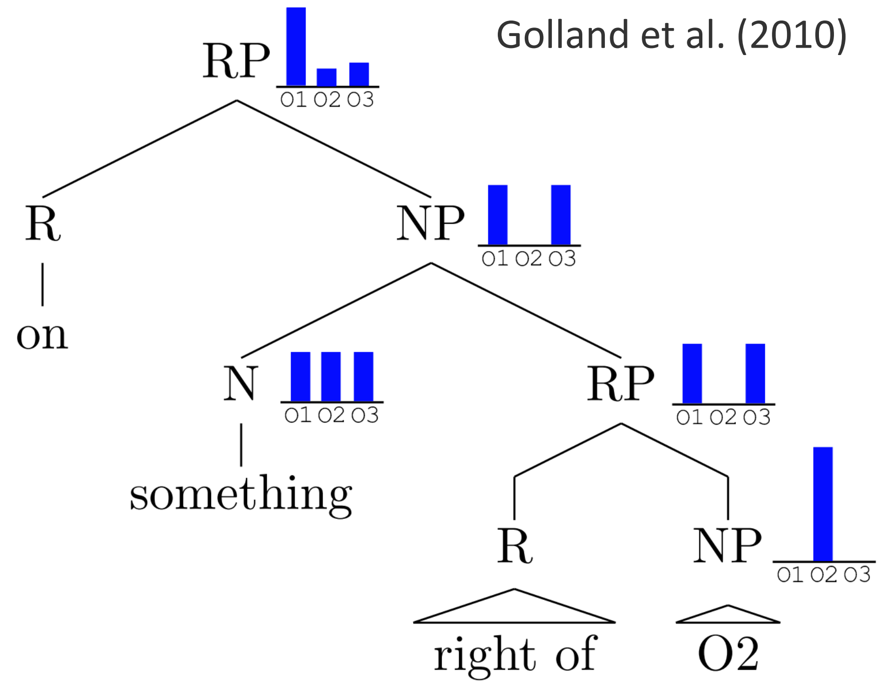
- Listener model:
  - Syntactic analysis of the particular expression gives structure
  - Rules (O2 = 100% prob of O2), features on words modify distributions as you go up the tree

Golland et al. (2010)



Spatial Relations





Golland et al. (2010)

- Put it all together: speaker will learn to say things that evoke the right interpretation
- Language is grounded in what the speaker understands about it

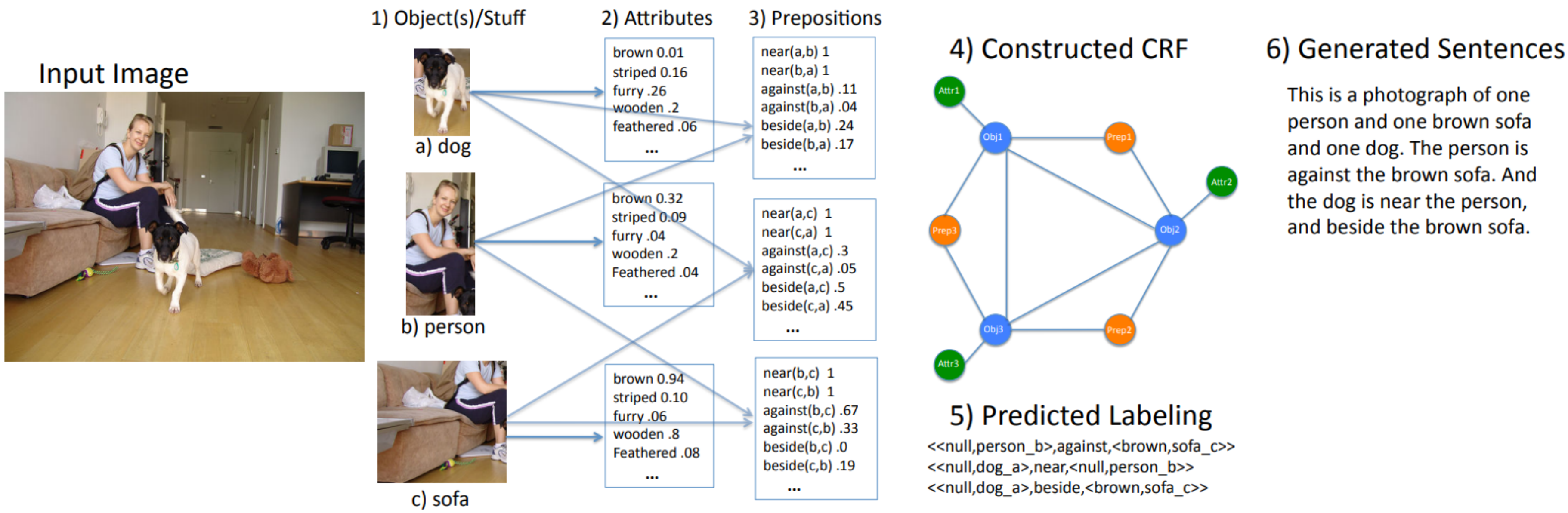
Image Captioning

How do we caption these images?

- Need to know what's going on in the images — objects, activities, etc.
- Choose what to talk about
- Generate fluid language

Pre-Neural Captioning: Objects and Relations

- *Baby Talk*, Kulkarni et al. (2011) [see also Farhadi et al. 2010, Mitchell et al. 2012, Kuznetsova et al. 2012]



- Detect objects using (non-neural) object detectors trained on a separate dataset
- Label objects, attributes, and relations. CRF with potentials from features on the object and attribute detections, spatial relations, and and text co-occurrence
- Convert labels to sentences using templates

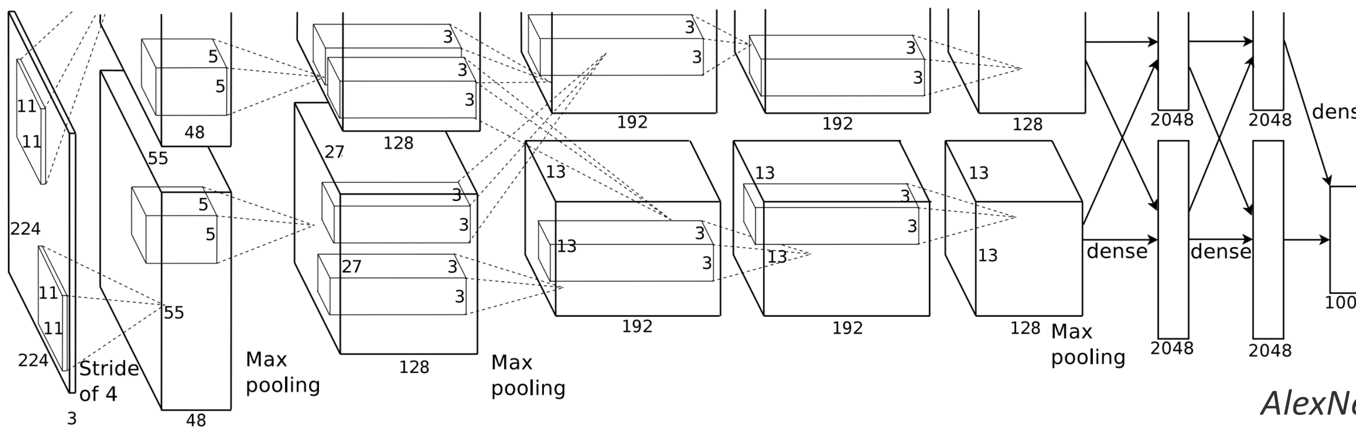
ImageNet models

- ImageNet dataset (Deng et al. 2009, Russakovsky et al. 2015)

  Object classification: single class for the image. 1.2M images, 1000 categories

  Object detection: bounding boxes and classes. 500K images, 200 categories

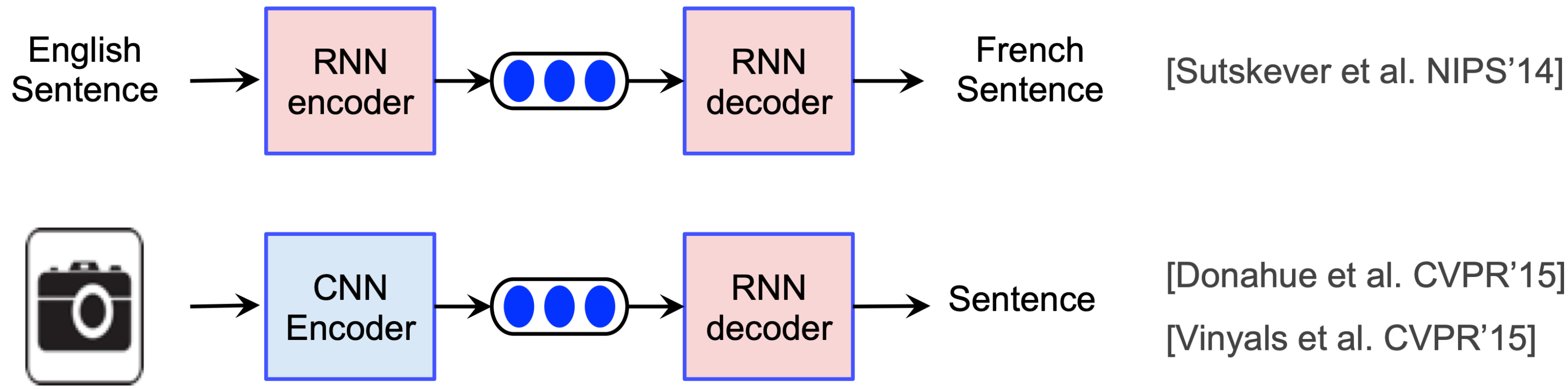
- 2012 ImageNet classification competition: drastic error reduction from deep CNNs



AlexNet, Krizhevsky et al. (2012)

- Last layer is just a linear transformation away from object detection — should capture high-level semantics of the image, especially what objects are in there

Neural Captioning: Encoder-Decoder



- Use a CNN encoder pre-trained for object classification (usually on ImageNet). Freeze the parameters.
- Generate captions using an LSTM conditioning on the CNN representation

What’s the grounding here?

food

a close up of a plate of ___

a dirt road

a couple of bears walking across ____

- What are the vectors really capturing?
  Objects, but maybe not deep relationships

Simple Baselines

- MRNN: take the last layer of the ImageNet-trained CNN, feed into RNN

- k-NN: use last layer of the CNN, find most similar train images based on cosine similarity with that vector. Obtain a consensus caption.

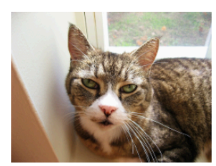
Devlin et al. (2015)

LM	PPLX	BLEU	METEOR
D-ME†	18.1	23.6	22.8
D-LSTM	14.3	22.4	22.6
MRNN	13.2	25.7	22.6
k-Nearest Neighbor	-	26.0	22.5
1-Nearest Neighbor	-	11.2	17.3

Table 1: Model performance on testval. †: From (Fang et al., 2015).

D-ME+DMSM	a plate with a sandwich and a cup of coffee
MRNN	a close up of a plate of food
D-ME+DMSM+MRNN	a plate of food and a cup of coffee
k-NN	a cup of coffee on a plate with a spoon

D-ME+DMSM	a black bear walking across a lush green forest
MRNN	a couple of bears walking across a dirt road
D-ME+DMSM+MRNN	a black bear walking through a wooded area
k-NN	a black bear that is walking in the woods



D-ME+DMSM	a gray and white cat sitting on top of it
MRNN	a cat sitting in front of a mirror
D-ME+DMSM+MRNN	a close up of a cat looking at the camera
k-NN	a cat sitting on top of a wooden table

Simple Baselines

System	Unique Captions	Seen In Training
Human	99.4%	4.8%
D-ME+DMSM	47.0%	30.0%
MRNN	33.1%	60.3%
D-ME+DMSM+MRNN	28.5%	61.3%
k-Nearest Neighbor	36.6%	100%

Table 6: Percentage unique (Unique Captions) and novel (Seen In Training) captions for testval images. For example, 28.5% unique means 5,776 unique strings were generated for all 20,244 images.

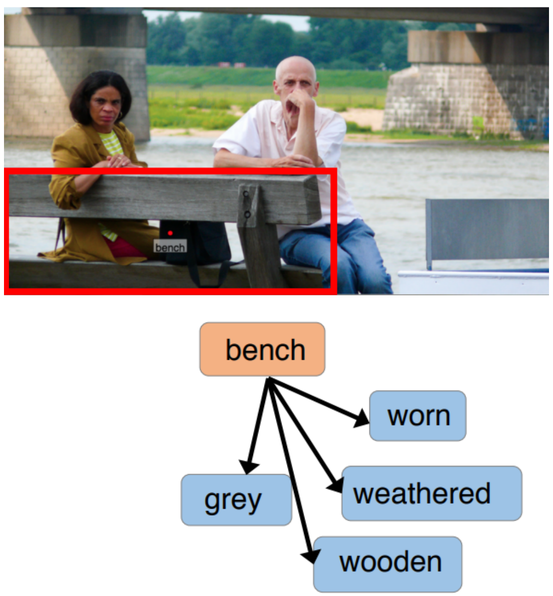
- Even from CNN+RNN methods (MRNN), relatively few unique captions even though it's not quite regurgitating the training

Devlin et al. (2015)

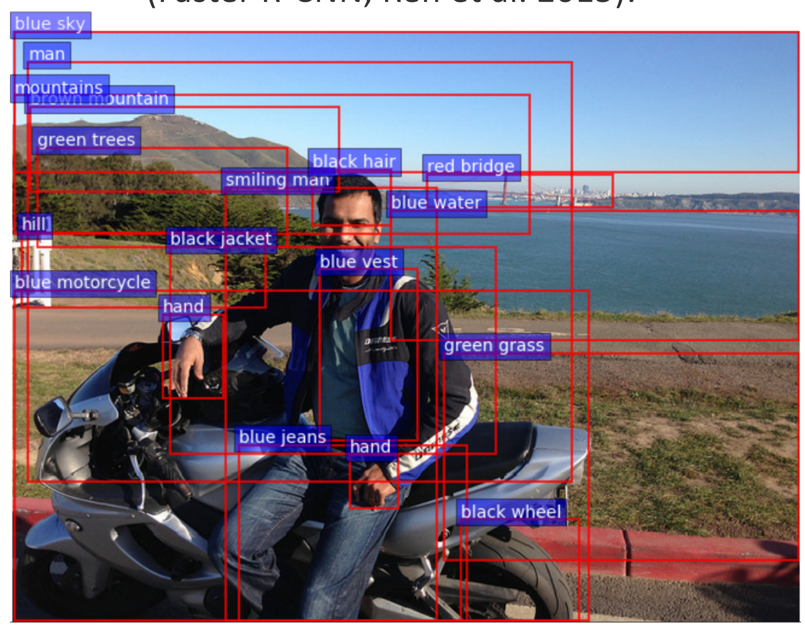
Neural Captioning: Object Detections

- Follow the pre-neural object-based systems: use features predictive of individual objects and their attributes

Training data
(Visual Genome, Krishna et al. 2015) :



Object and attribute detections
(Faster R-CNN, Ren et al. 2015):



Anderson et al. (2018)

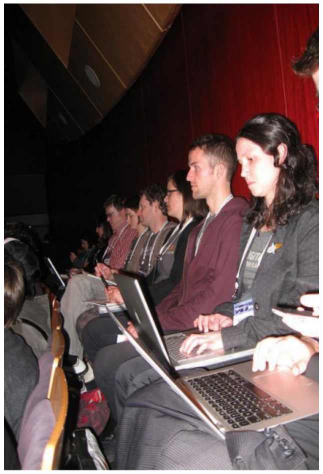
Neural Captioning: Object Detections

- Also add an attention mechanism: attend over the visual features from individual detected objects

Anderson et al. (2018)

Neural Hallucination

- Language model often overrides the visual context:



A group of people sitting around a ***table*** with laptops

A kitchen with a stove and a ***sink***

- Standard text overlap metrics (BLEU, METEOR) aren't sensitive to this!

Slide credit: Anja Rohrbach

Rohrbach & Hendricks et al. (2018)

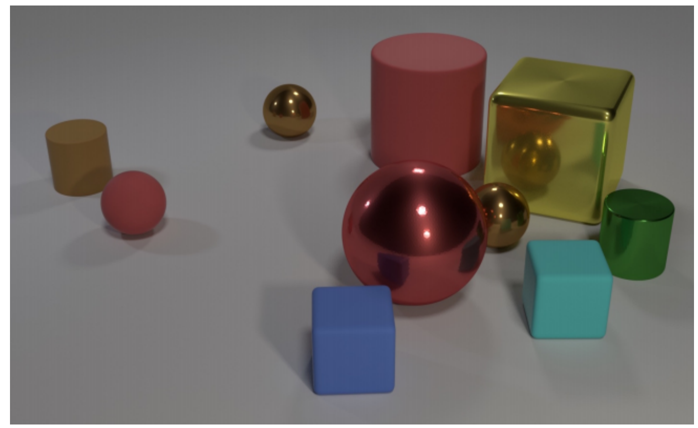
Visual Question Answering

Visual Question Answering

- Answer questions about images
- Frequently require compositional understanding of multiple objects or activities in the image

VQA: Agrawal et al. (2015)

Human-written questions

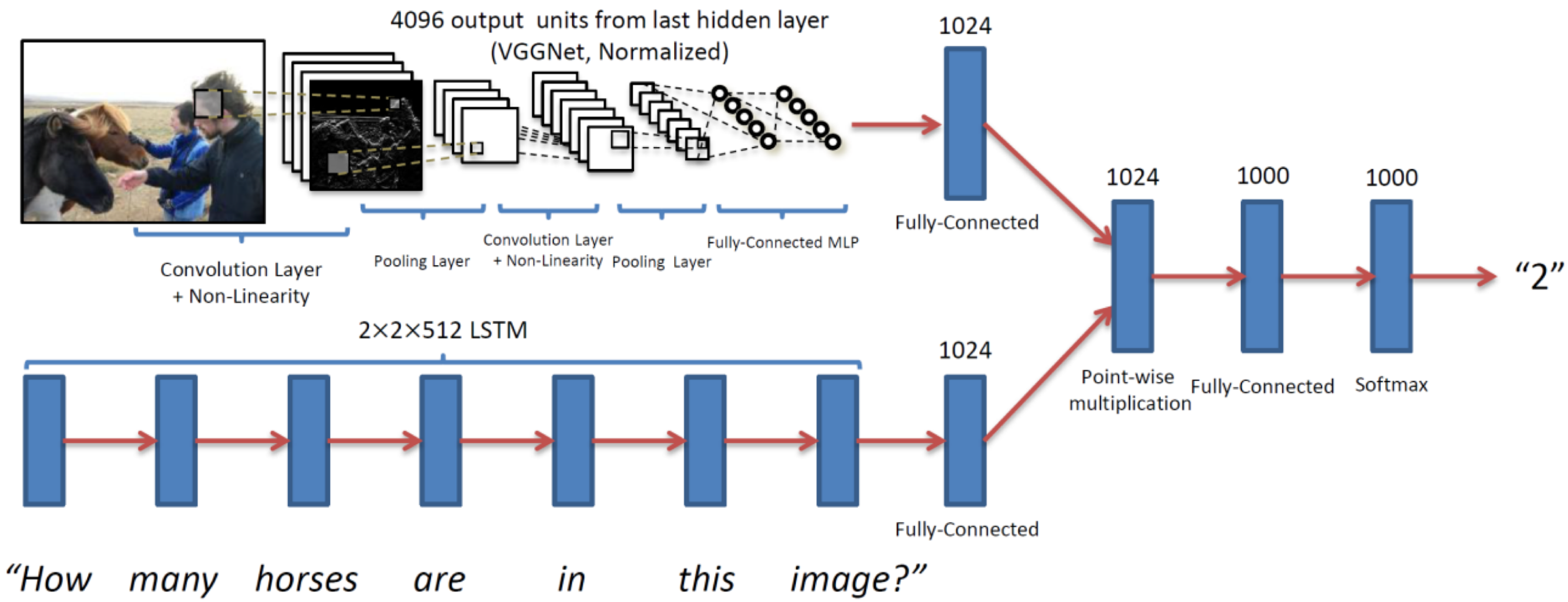


What size is the cylinder that is left of the brown metal thing that is left of the big sphere?

CLEVR: Johnson et al. (2017)

Synthetic, but allows careful control of complexity and generalization

Visual Question Answering



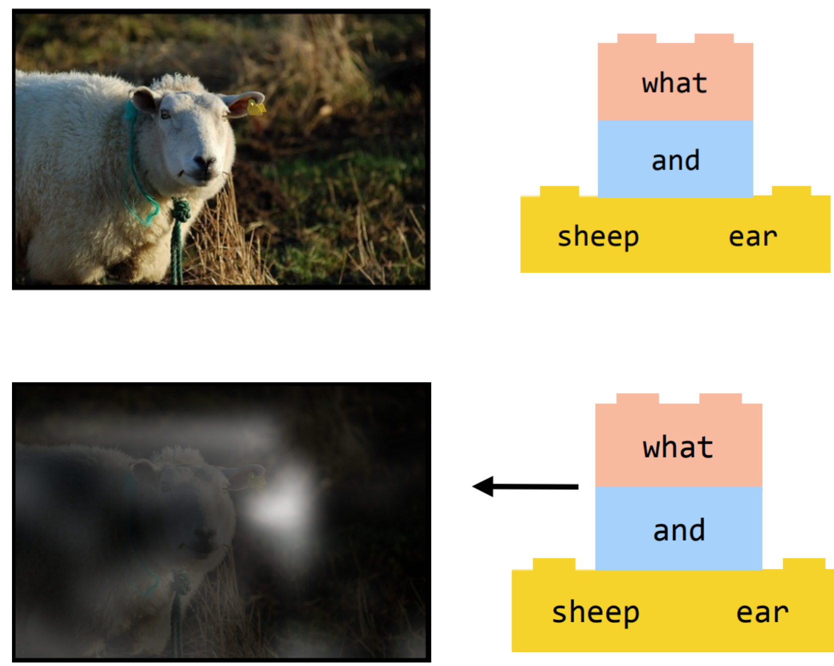
- Fuse modalities: pre-trained CNN processing of the image, RNN processing of the language
- What could go wrong here?

Agrawal et al. (2015)

Neural Module Networks

- Integrate compositional reasoning + image recognition
- Have neural network components like `find[sheep]`whose composition is governed by a parse of the question
- Like a semantic parser, with a learned execution function

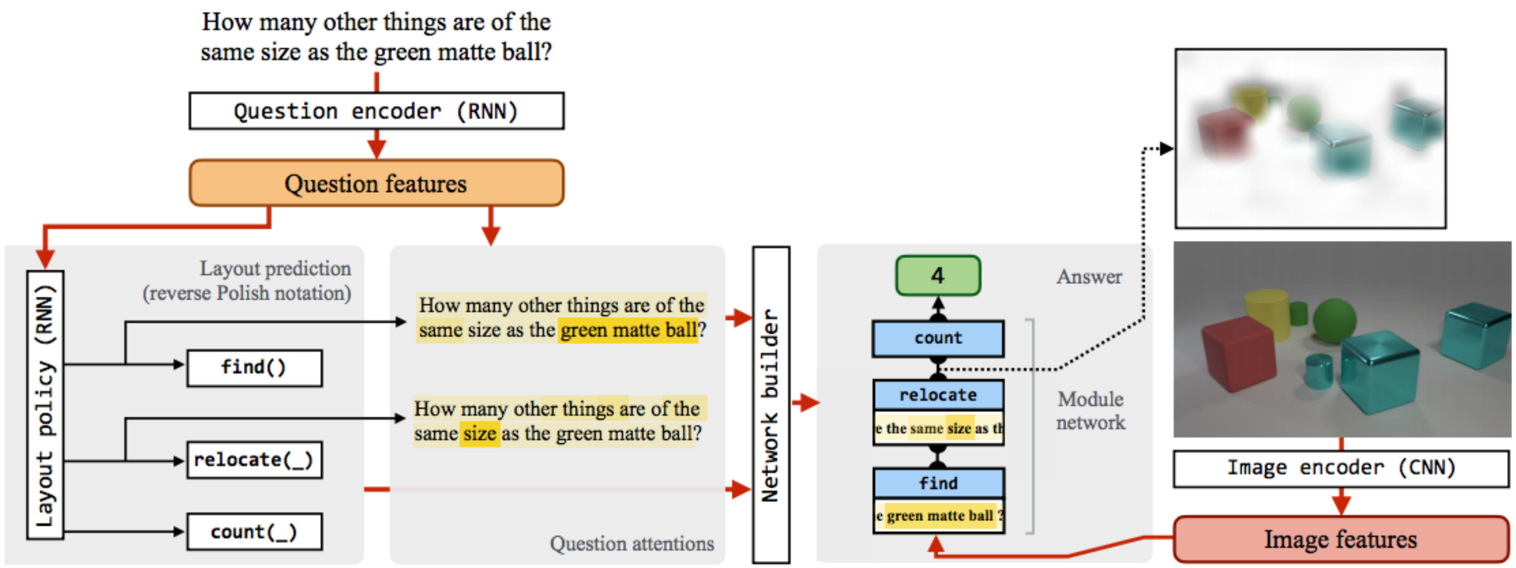
What is in the sheep's ear? => tag



Andreas et al. (2016), Hu et al. (2017)

Neural Module Networks

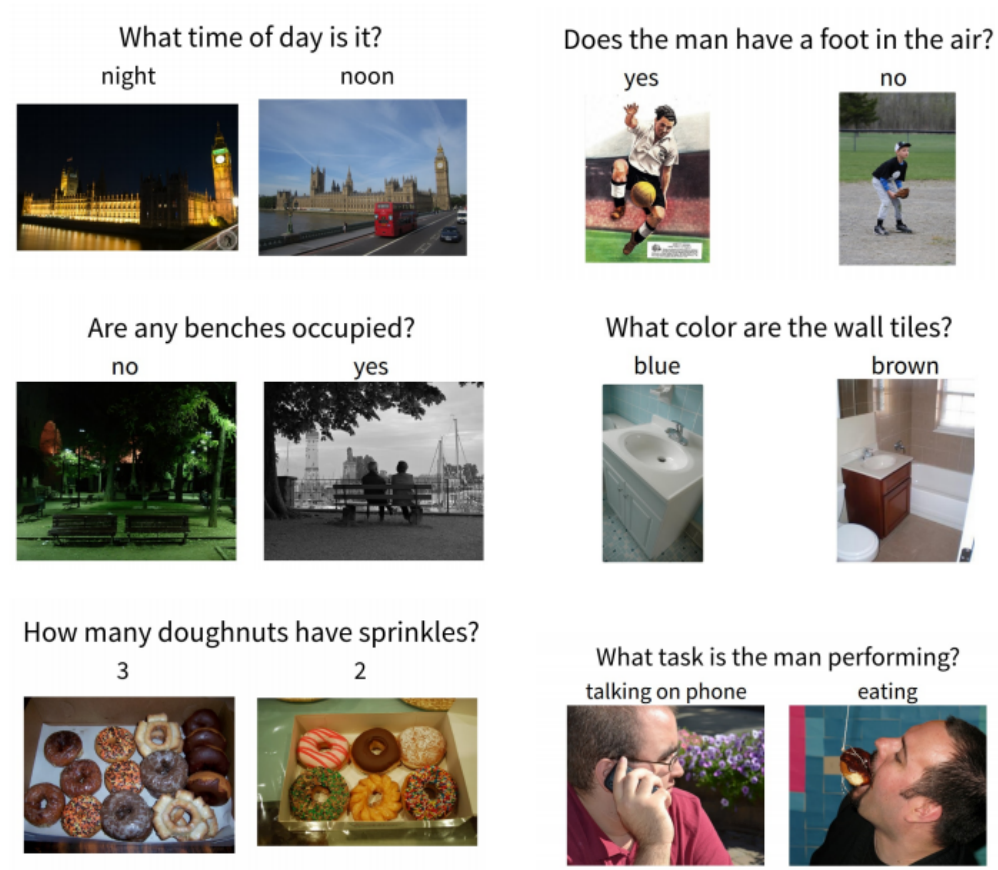
- Able to handle complex compositional reasoning, at least with simple visual inputs



Andreas et al. (2016), Hu et al. (2017)

Visual Question Answering

- In many cases, language as a prior is pretty good!
  - “Do you see a…” = yes (87% of the time)
  - “How many…” = 2 (39%)
  - “What sport…” = tennis (41%)
- When only the question is available, baseline models are super-human!
- Balanced VQA: reduce these regularities by having pairs of images with different answers



Goyal et al. (2017)

Challenge Datasets

- NLVR2: Difficult comparative reasoning; balanced dataset construction; human-written

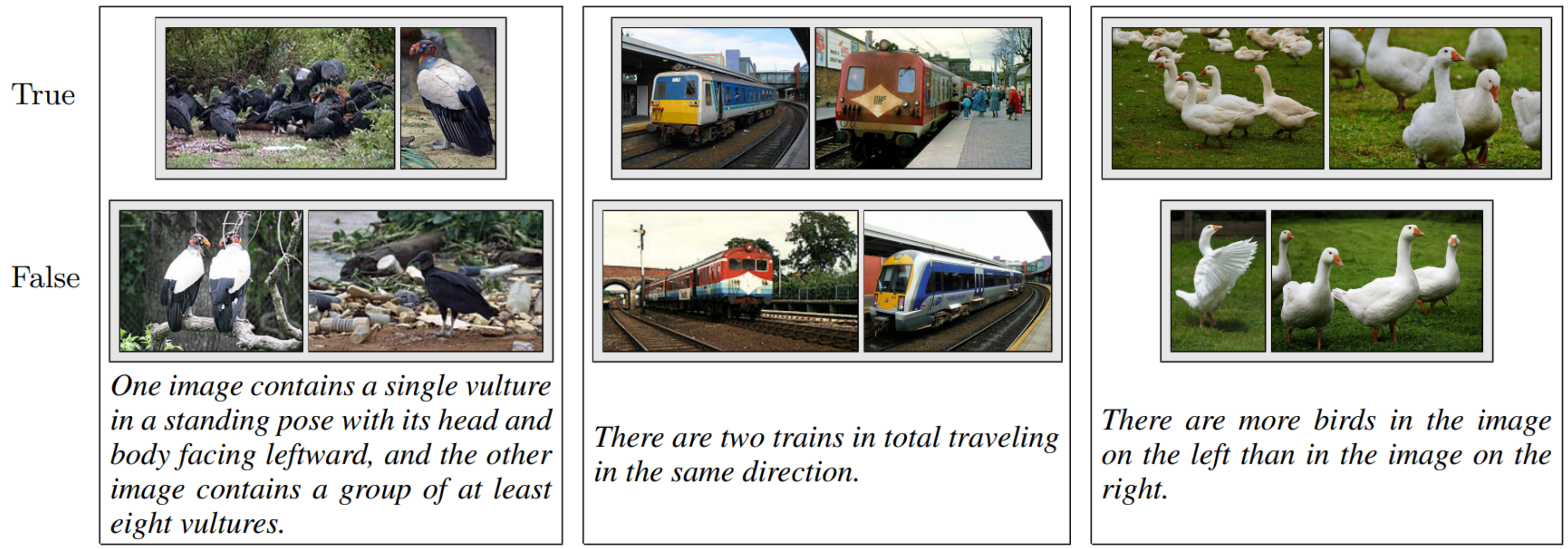


Table 3: Six examples with three different sentences from NLVR2. For each sentence, we show two examples using different image-pairs, each with a different label.

Majority class baseline: 50%
Current best system: 80%
Human performance: 96%

Suhr & Zhou et al., 2019

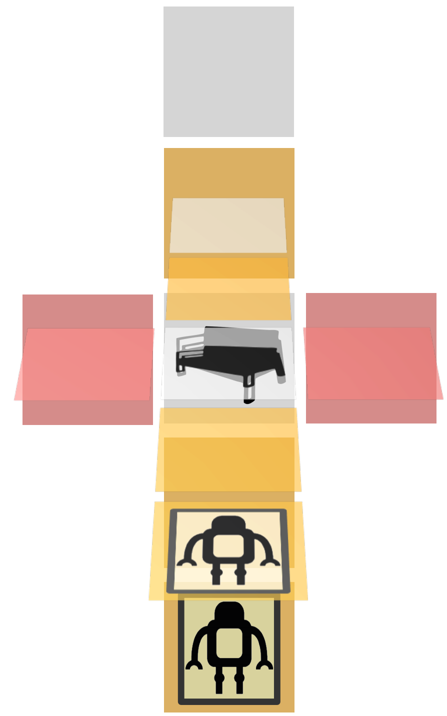
Instruction Following

Instruction Following

- SAIL dataset: navigational instructions in synthetic grid worlds, with furniture and patterns

MacMahon et al., 2006; Chen and Mooney, 2011

Human annotator view

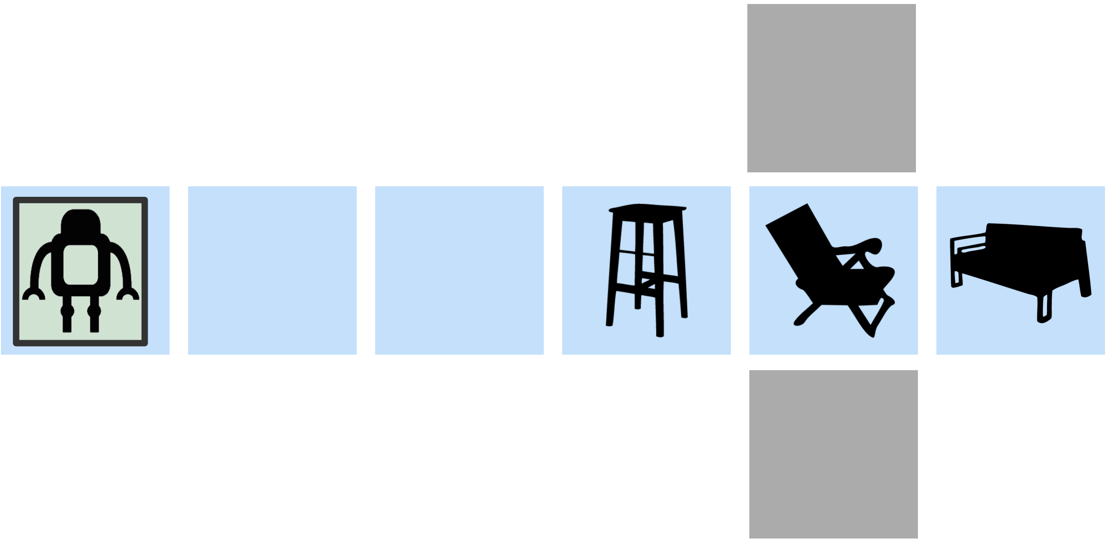


System view

Instruction Following

Input instruction: *go to the chair. turn left and go forward to the fish painting. head to the right until you get to a coat rack*

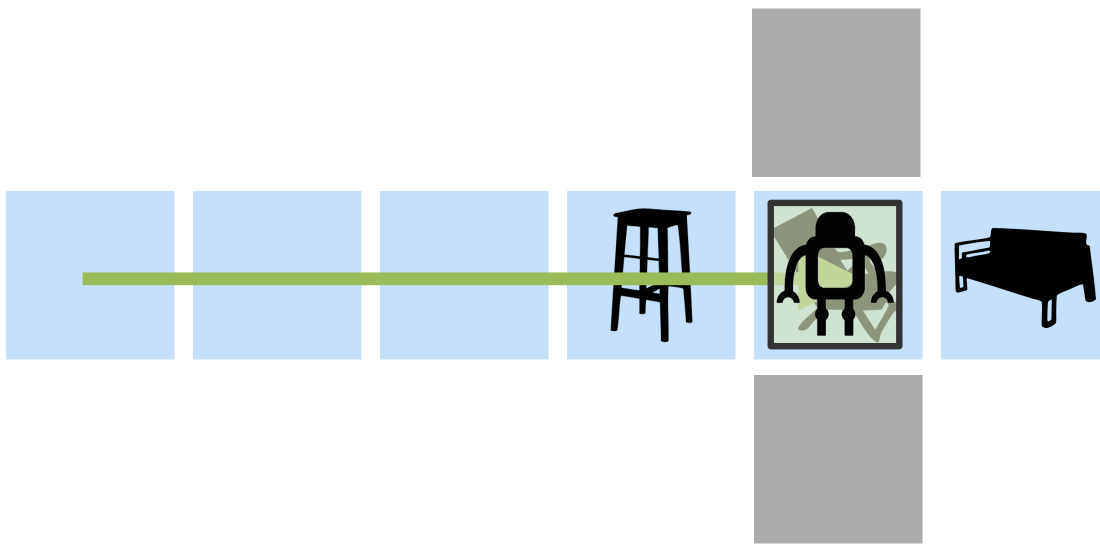
Output actions:



Instruction Following

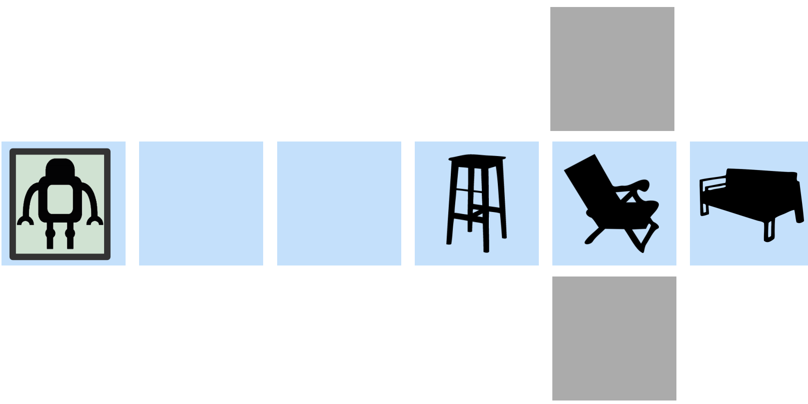
Input instruction: *go to the chair. turn left and go forward to the fish painting. head to the right until you get to a coat rack*

Output actions:



Instruction Following

- Several successful approaches using semantic parsing
  (Chen and Mooney 2011; Artzi and Zettlemoyer 2013; Artzi et al. 2014)



$$
\begin{array}{cccc}
\text{go} & \text{to} & \text{the} & \text{chair} \\
\hline
S & AP/NP & NP/N & N \\
\lambda a.move(a) & \lambda x.\lambda a.to(a,x) & \lambda f.\iota x.f(x) & \lambda x.chair(x)
\end{array}
$$

$$\frac{NP/N : \lambda f.\iota x.f(x) \quad N : \lambda x.chair(x)}{NP : \iota x.chair(x)} >$$

$$\frac{AP/NP : \lambda x.\lambda a.to(a,x) \quad NP : \iota x.chair(x)}{AP : \lambda a.to(a, \iota x.chair(x))} <$$

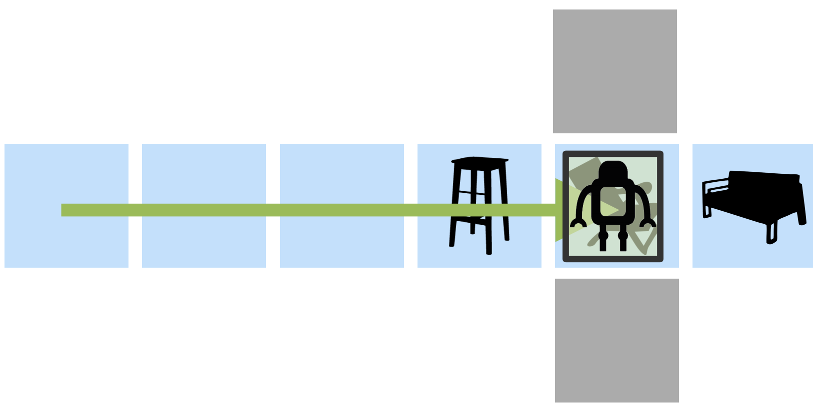
$$\frac{AP : \lambda a.to(a, \iota x.chair(x))}{S\backslash S : \lambda f.\lambda a.f(a) \wedge to(a, \iota x.chair(x))}$$

$$\frac{S : \lambda a.move(a) \quad S\backslash S : \lambda f.\lambda a.f(a) \wedge to(a, \iota x.chair(x))}{S : \lambda a.move(a) \wedge to(a, \iota x.chair(x))} <$$

examples from Yoav Artzi

Instruction Following

- Several successful approaches using semantic parsing
  (Chen and Mooney 2011; Artzi and Zettlemoyer 2013; Artzi et al. 2014)



go to the chair

$\lambda a.move(a) \wedge to(a, \iota x.chair(x))$

move until you reach the chair

$\lambda a.move(a) \wedge$
$post(a, intersect(\iota x.chair(x), you))$

- Logical forms denote action sequences, often using post-conditions
- Learn from action sequences paired with instructions

examples from Yoav Artzi

Instruction Following

- This is a sequence-to-sequence task, right?

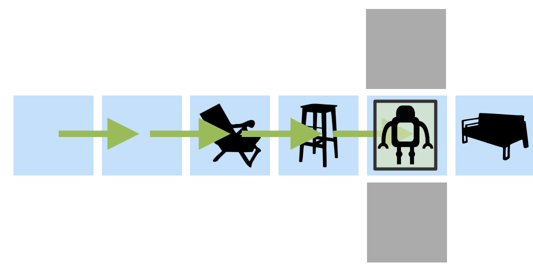
Inputs

go forward to the grey hallway

Instruction

Listener

Outputs

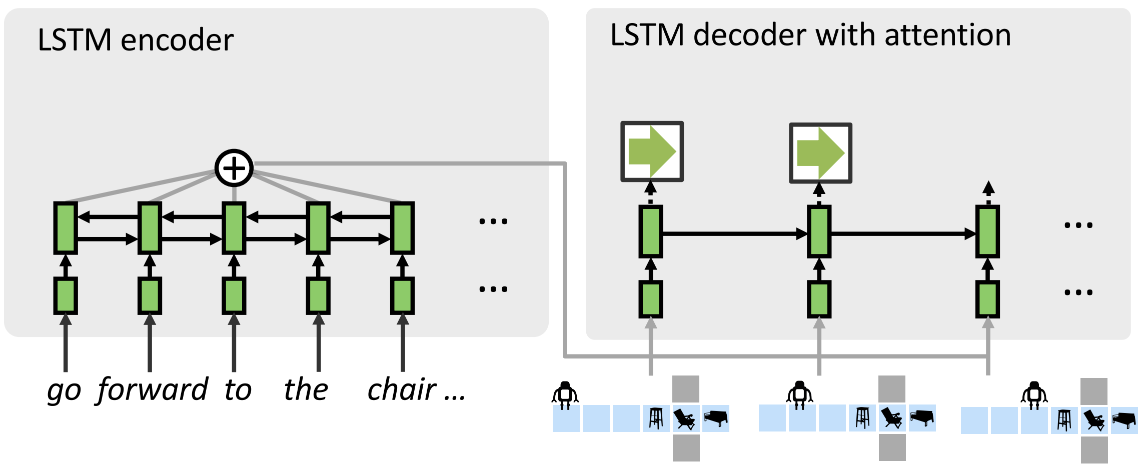


Actions in context

Neural Instruction Following

- Encoder-decoder setup with attention to the instruction

- Decoder takes as input embeddings for all the (symbolic) world features the agent can see



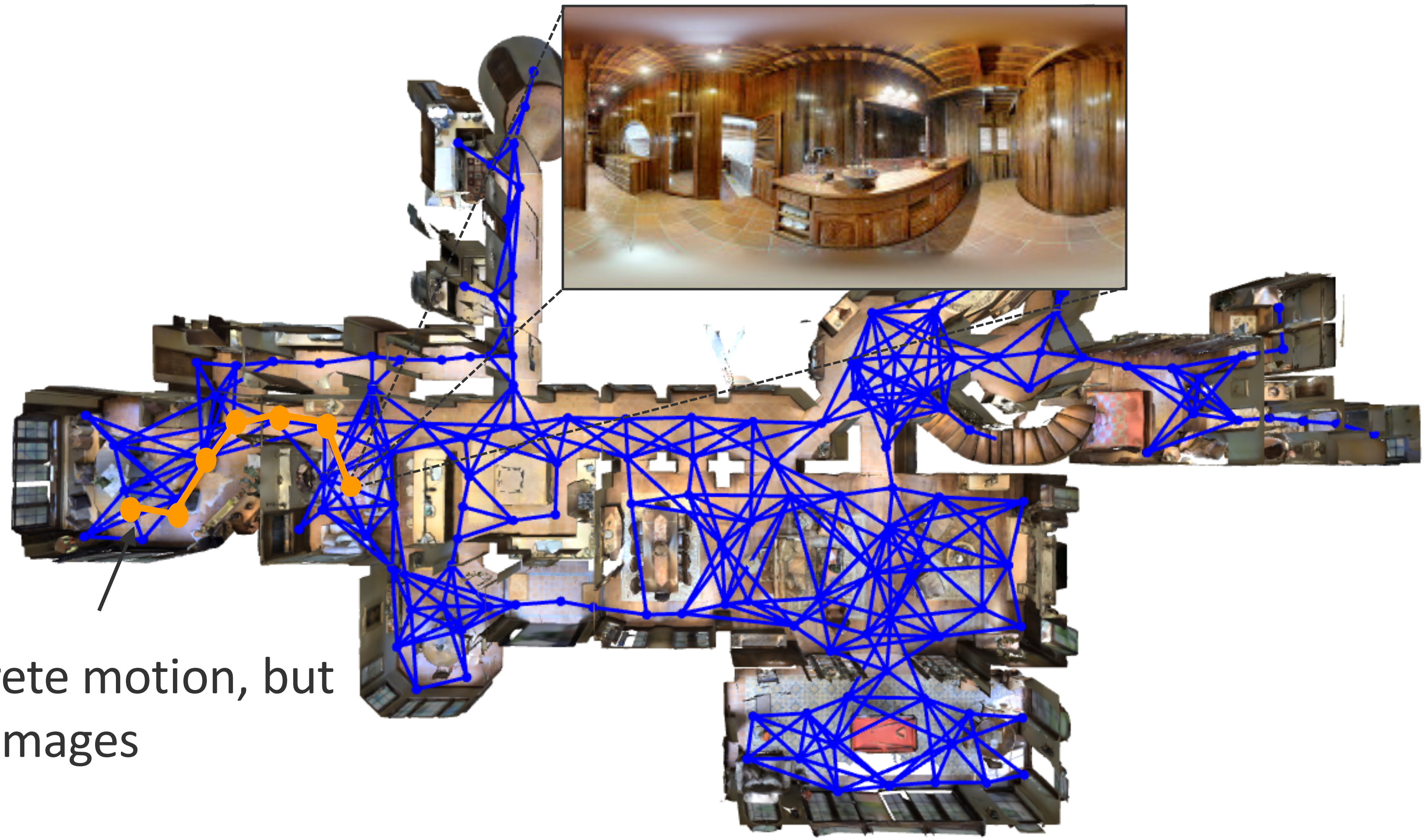
- Almost as good as the best semantic parsing approach

Mei et al. (2016)

Vision-and-Language Navigation

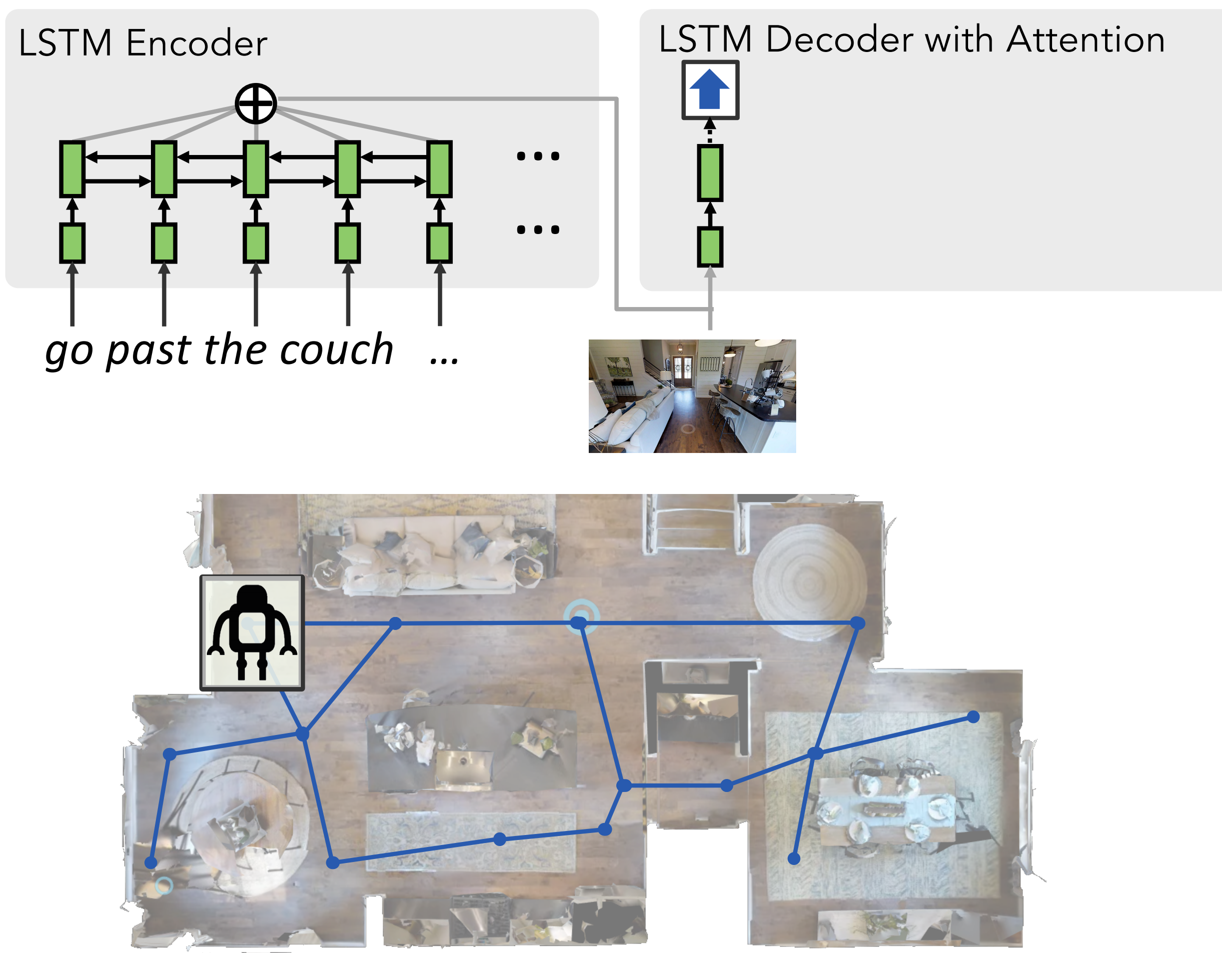
Turn left and take a right at the table. Take a left at the painting and then take your first right. Wait next to the exercise equipment.

Anderson et al. (2018)



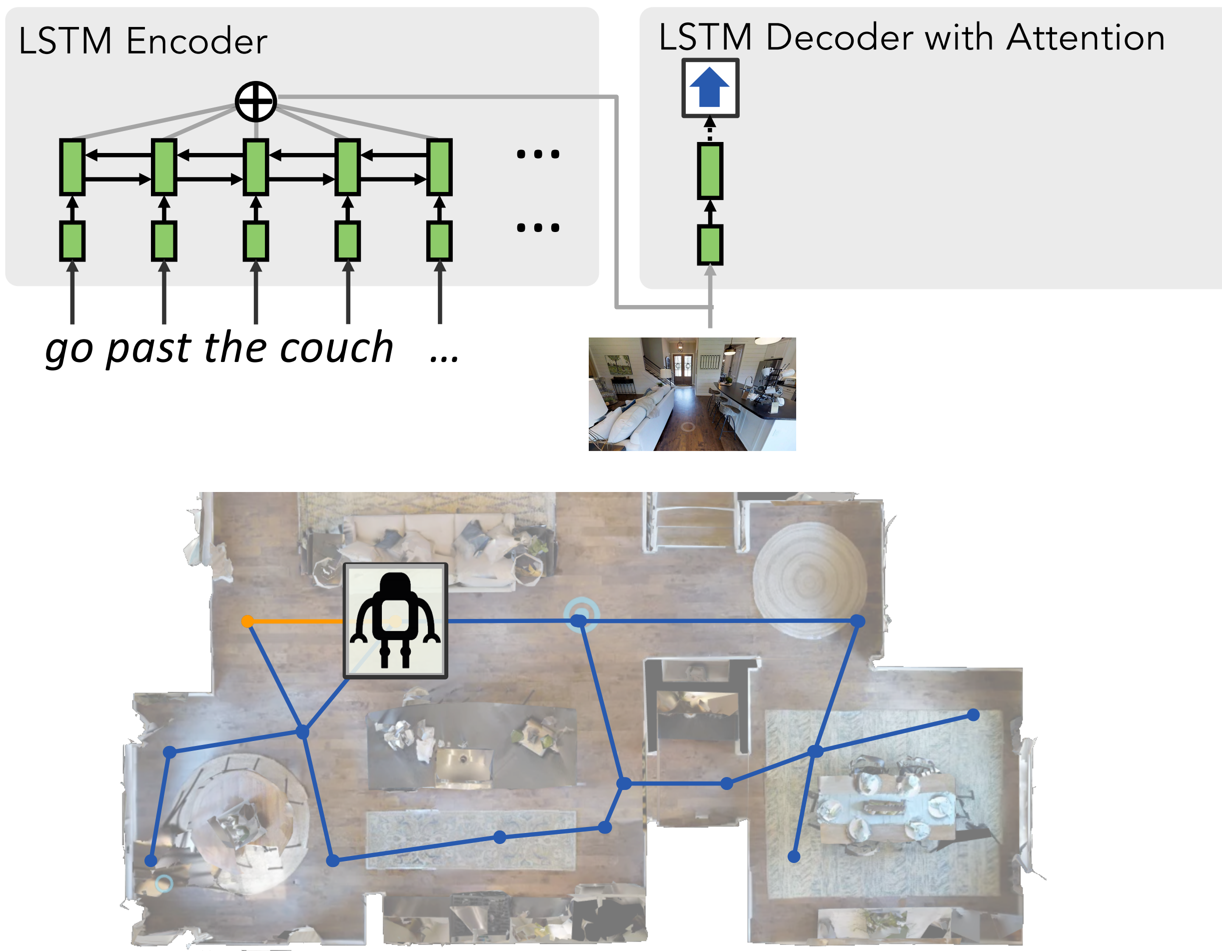
Discrete motion, but real images

Vision-and-Language Navigation



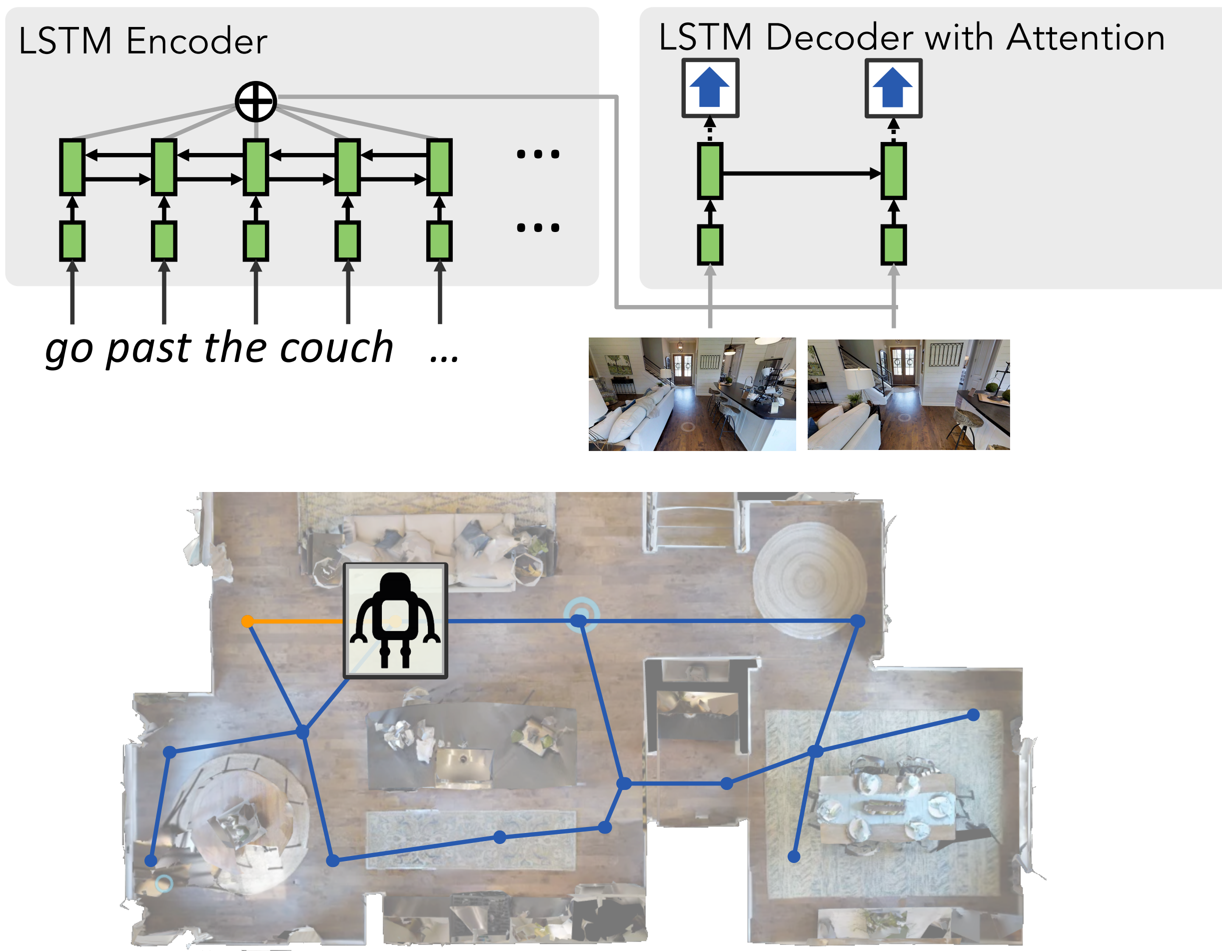
Anderson et al. (2018)

Vision-and-Language Navigation



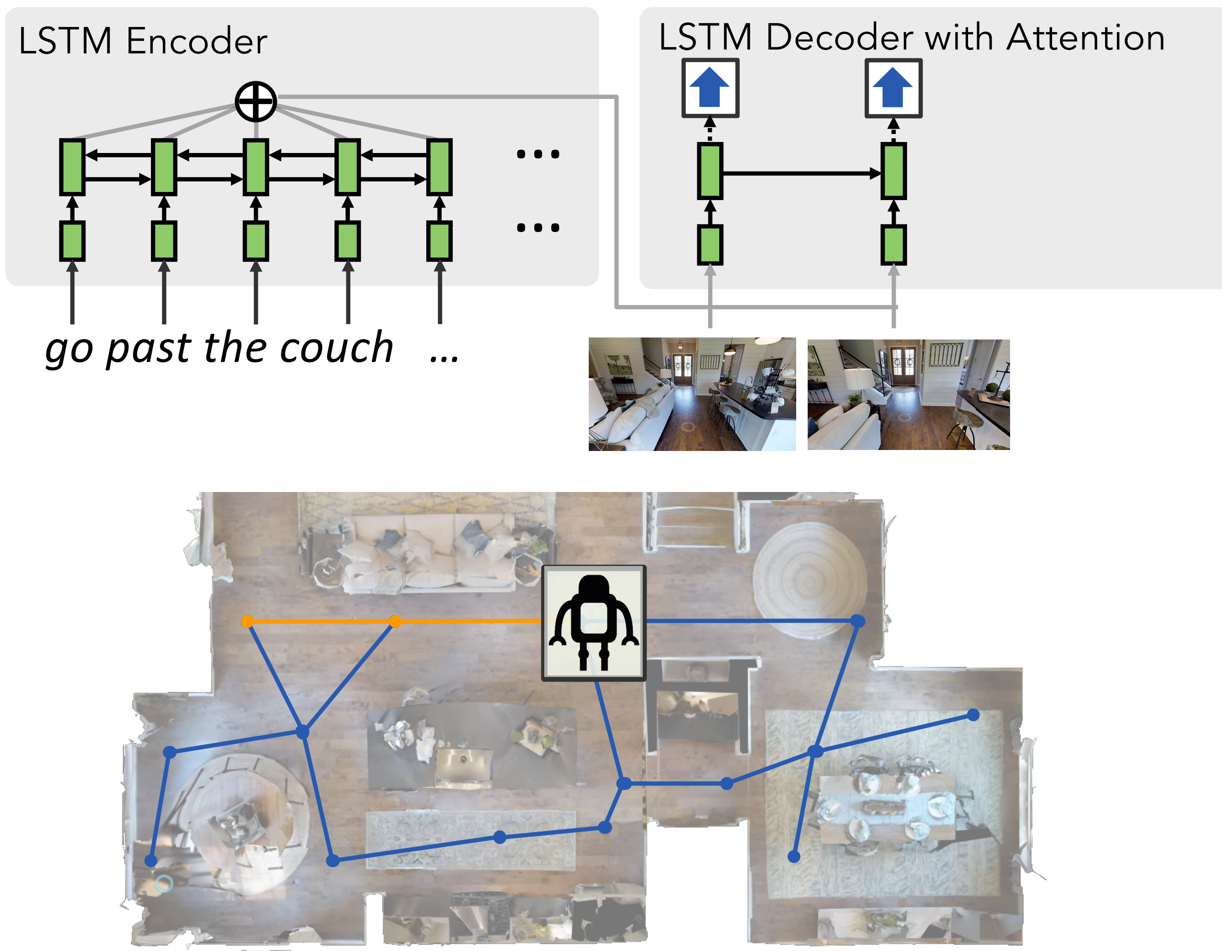
Anderson et al. (2018)

Vision-and-Language Navigation



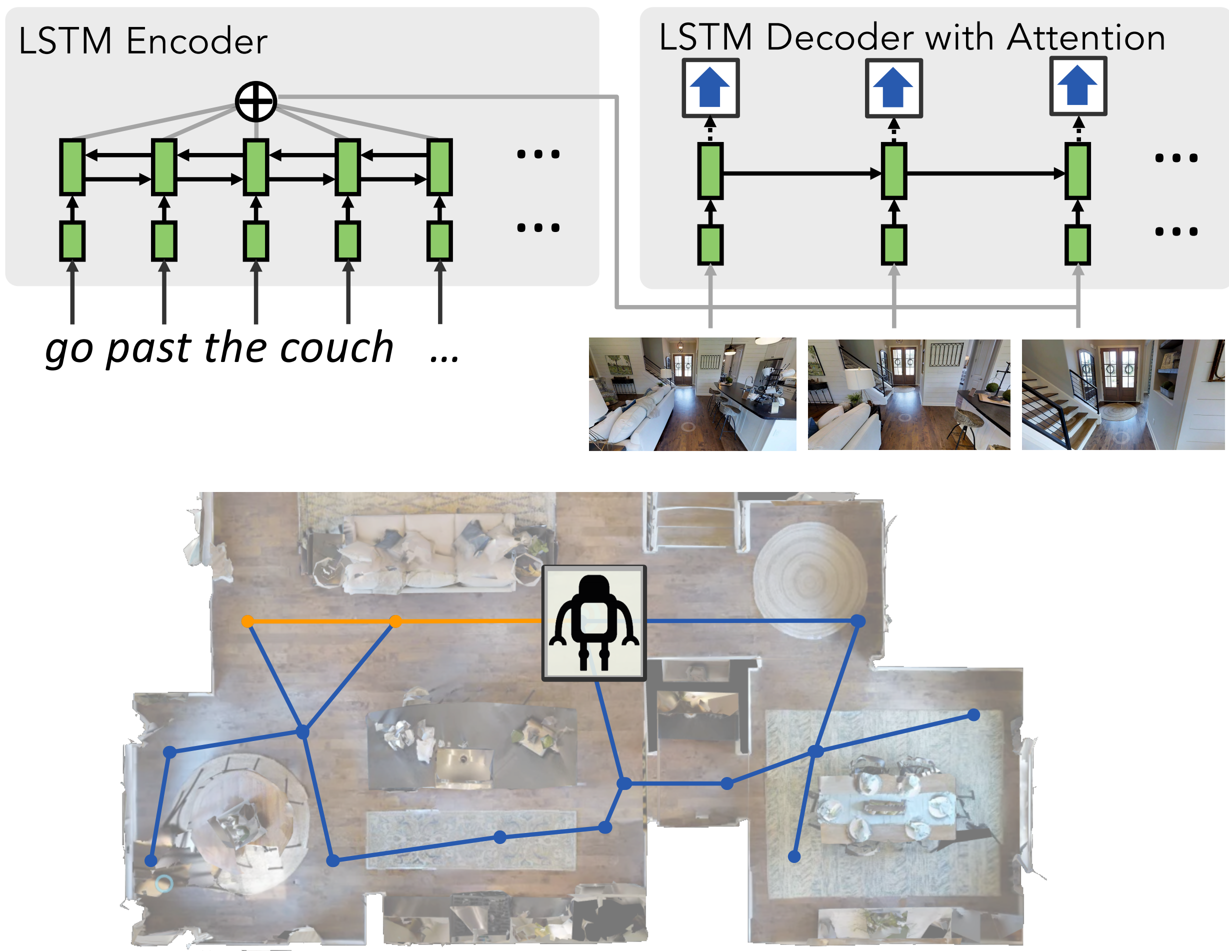
Anderson et al. (2018)

Vision-and-Language Navigation



Anderson et al. (2018)

Vision-and-Language Navigation



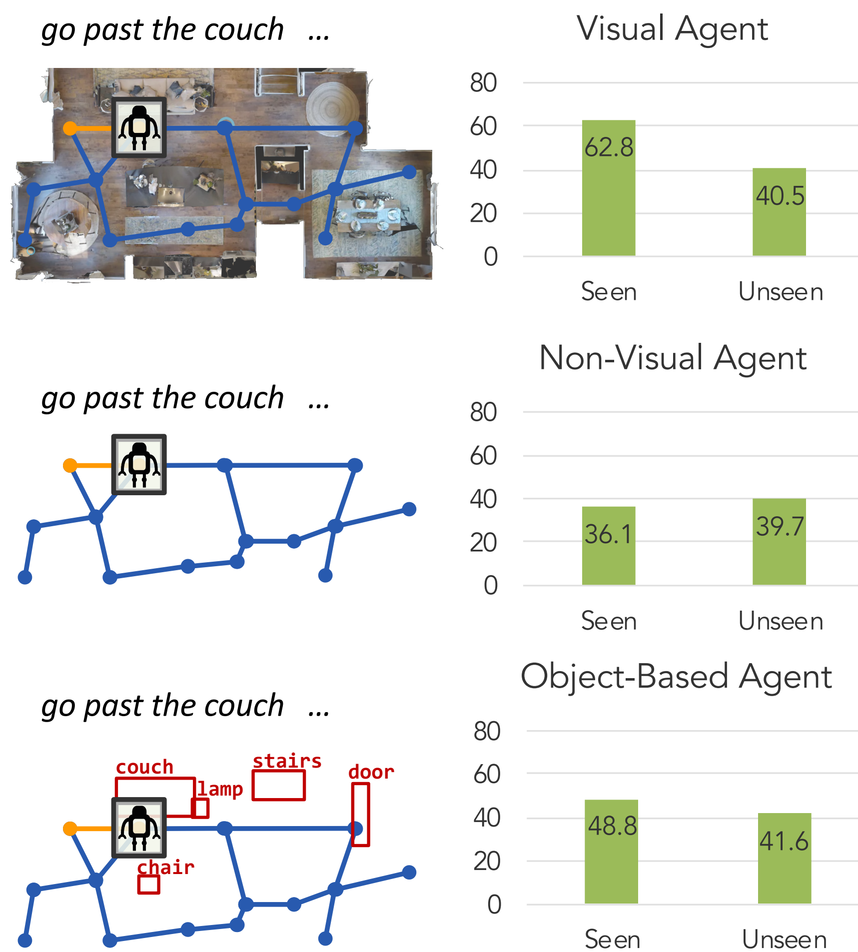
Anderson et al. (2018)

Walk past hall table. Walk into bedroom. Make left at table clock. Wait at bathroom door threshold.

Fried, Hu, Cirik et al. (2018)

Vision-and-Language Navigation

- Best current models: 72% accuracy; humans: 86%
- But, what are the models actually grounding into?
- Some combination of:
  - generalizable representations
  - environments seen in training
  - biases in the routes themselves



Gordon et al. 2018, Hu et al. 2019

Challenge Tasks

Touchdown

Chen et al. 2019, Mehta et al. 2020

Turn and go with the flow of traffic. At the first traffic light turn left. Go past the next two traffic lights …

- Long, complex routes through NYC’s StreetView graph, with associated imagery
- SOTA model: 5% accuracy. Human: 92%

Challenge Tasks

ALFRED Shridhar et al. 2020

- Interact with objects in a household setting
- Long time horizons, non-reversible state changes
- Baseline model: 1% accuracy. Human: 91%

Takeaways

- Lots of problems where natural language has to be interpreted in an environment and can be understood in the context of that environment
- Neural models make it easier to fuse representations from multiple modalities (but they sometimes learn to cheat)
- Symbolic methods guided by linguistic structure; neural systems with learned representations; some work productively combines both