Grounded Language Understanding

He He

New York University

November 24, 2021

Steven Colbert's conversation with Siri

Colbert:	What am I talking about tonight?
Siri:	I would perfer not to say.
...	...
Colbert:	For the love of God, the *cameras* are on, give me something?
Siri:	What kind of place are you looking for? *Camera stores* or churches
...	...
Colbert:	I don't want to search for anything! I want to write the show!
Siri:	Searching the Web for "search for anything. I want to write the shuffle."

What went wrong?

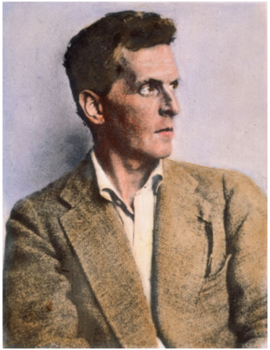
What am I talking about tonight?

- Who is “I”?
- When is “tonight”?
- What’s the purpose of the talk?
- Who’s the audience?

Context is important!

- Where are you from? (nation, hometown, school?)
- (Ice or no ice? Coffee or tea? Morning or afternoon?) The latter, please.
- Can you pass me the salt?

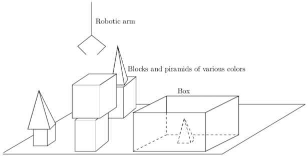
Language and communication



Wittgenstein, *Philosophical Investigations*

"For a large class of cases of the employment of the word 'meaning'—though not for all—this word can be explained in this way: *the meaning of a word is its use in the language*"

SHRDLU [Winograd 1972]



Person:	Pick up a big red block.
Computer:	OK.
Person:	Grasp the pyramid.
Computer:	I DON'T UNDERSTAND WHICH PYRAMID YOU MEAN.
...	...

- Connect symbols to the world: utterance $\rightarrow$ logical form $\rightarrow$ action $\rightarrow$ response
- Successful but limited to the blocks world
- Renewed interest in grounded language understanding with the success of neural networks

Tasks that involve grounding

Describing color [MacMahan and Stone, 2015]

Color	Utterance
	green
	purple
	grape
	turquoise
	moss green
	pinkish purple
	light blue grey
	robin's egg blue
	british racing green
	baby puke green

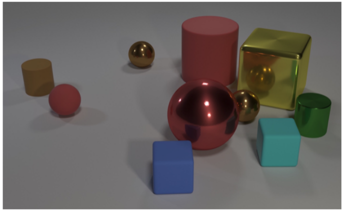
Figure: Example from Chris Potts

Tasks that involve grounding

Visual question answering [Agrawal+ 2015]

Tasks that involve grounding

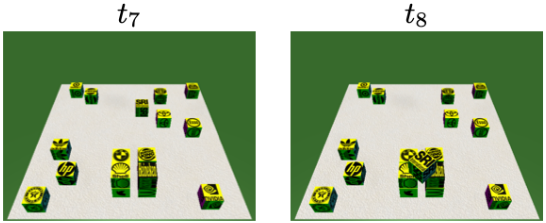
CLEVR: visual reasoning [Johnson+ 2015]



Q: Are there an **equal** **number** of **large things** and **metal spheres**?
Q: **What size** is the **cylinder** **that is left of** the **brown metal** thing **that is left of** the **big sphere**?
Q: There is a **sphere** with the **same size as** the **metal cube**; is it **made of the same material as** the **small red sphere**?
Q: **How many** objects are **either** **small cylinders** or **red** things?

Tasks that involve grounding

Spatial reasoning [Bisk+ 2017]



1 **Rotate** SRI to the **right** ...
2 rotate it **45 degrees clockwise** ...
3 **only half of one rotation** so its corners point where its edges did ...
4 **the logo faces the top right corner** of the screen...
5 Spin SRI **slightly** to the right and then set it **in the middle of the 4 stacks**

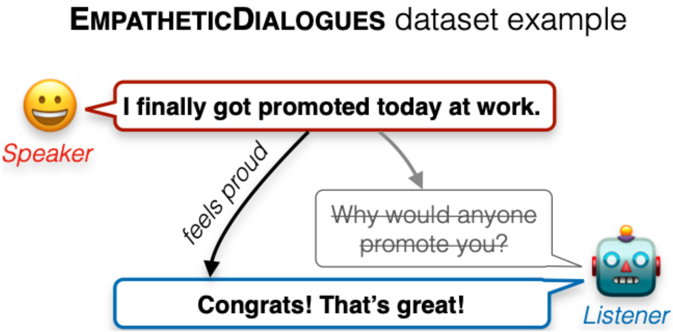
Tasks that involve grounding

ALFRED: instruction following [Shridhar+ 2020]

With real robots, see [Chai+ 2018].

Tasks that involve grounding

Empathetic dialogue [Rashkin+ 2020]



Tasks that involve grounding

Winograd schema challenge [Winograd 1972, Levesque 2011, Davis+ 2016]

Jim yelled at Kevin because he was so upset.
Jim comforted Kevin because he was so upset.

The customer walked into the bank and stabbed one of the tellers. He was immediately taken to the police station.
The customer walked into the bank and stabbed one of the tellers. He was immediately taken to the hospital.

Ground in social, physical context

Summary

Connects language (symbols) to the world

- *Perception*: vision, audio
- *Action*: navigation, interaction
- *Society*: commonsense, empathy

model $\rightarrow$ agent

- Multimodal: full perception of the world
- Interactive: actively learn about the world
- Multi-agent: consider other agents in the world

Useful frameworks for thinking about grounding problems

Multimodal: mapping between different types of signals

- Neural architectures that encode different signals in the same space

Interactive: take actions and receive feedbacks

- Reinforcement learning: learning from trial and error

Multi-agent: model other agents' goals and contexts

- Speakers: generate language given the world
- Listeners: interpret language in the world
- The rational speech act model: reason about each other

Table of Contents

Basic multimodal architecture

Key components:

1. Encoders: embed different signals separately
2. Fusion: create interaction among different embeddings
3. Decoder: classification, generation etc.

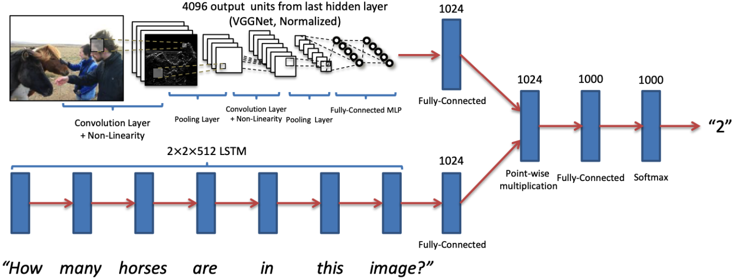


Figure: [Agrawal+ 2016]

Attention over image

Similar to text QA, we want to interact different parts in the text and the image.

What are "words" in images?

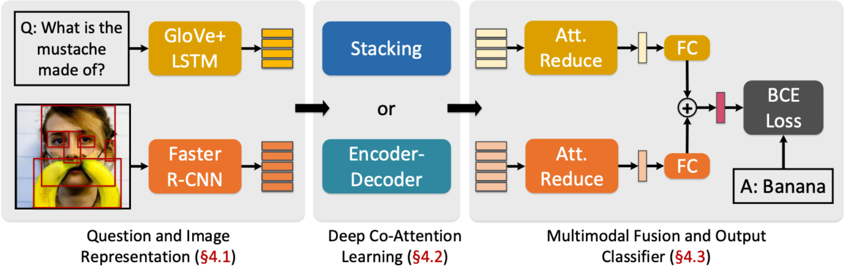
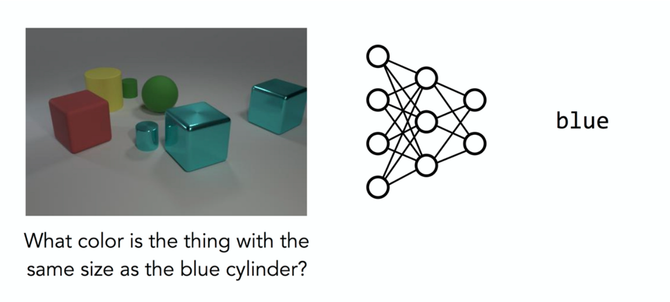


Figure: [Yu+ 2019]

Neural module networks

Visual reasoning $\iff$ semantic parsing



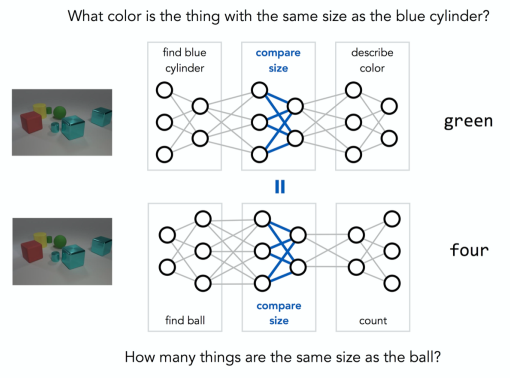
$$\texttt{color}(\lambda x.\texttt{equal}(\texttt{size}(x), \texttt{size}(\lambda y.\texttt{blue}(y) \wedge \texttt{cylinder}(y))))$$

How do we execute the logical form on an image?

Neural module networks

Text	`capital`(x)	database lookup
Image	`color`(x)	learned function $f_{\text{color}}(x, \text{image})$

Share modules ("predicates"/functions) across examples



Neural module networks

Jointly learning the module networks and the composition (layout)

What to do with the unobserved layout (i.e. logical form)?

- Use rules (in restricted domains)
- Model as a latent variable
- Obtain human annotation

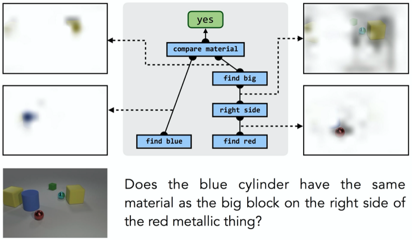


Figure: [Andreas+ 2016]

Multimodal pre-training

Data: image caption, VQA
Self-supervision: masked LM, matching between image/text

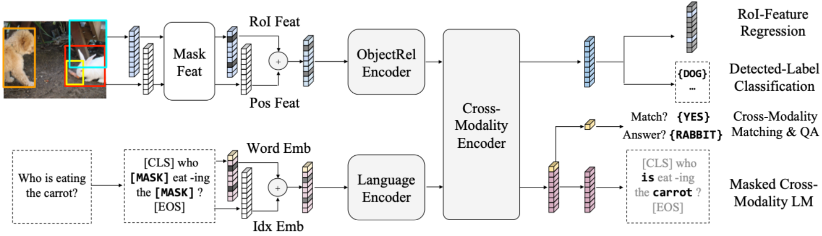


Figure: [Tan and Bansal 2019]

Table of Contents

Learning through interaction

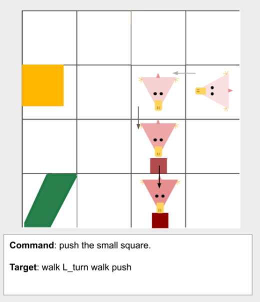


Figure: [Ruis+ 2020]

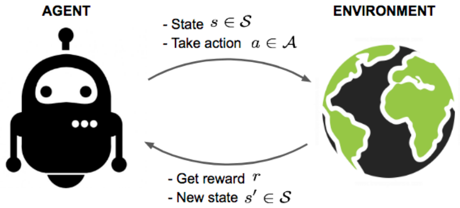
A trial-and-error strategy:

- Agent: Try out random actions in the world
- World: reward agent when goals are achieved

How to learn from experience?

(Analogous to human learning)

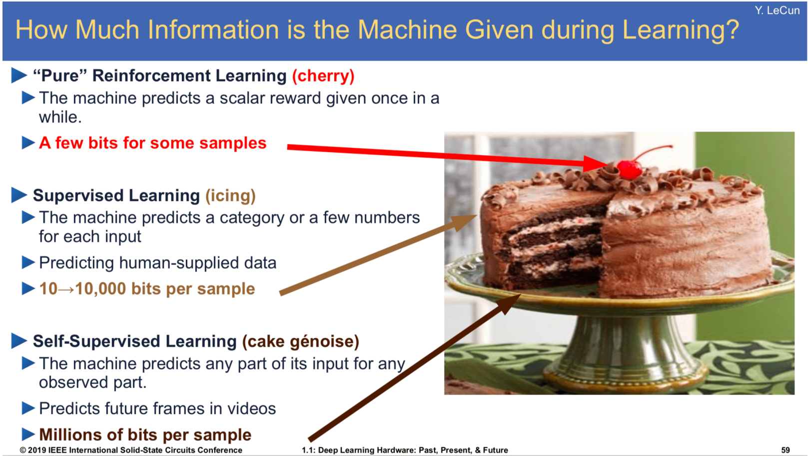
Markov decision process (MDP)



- At time step t, the agent is in **state** $s_t \in \mathcal{S}$.
- It takes an **action** $a_t \in \mathcal{A}$ and transitions to state s_{t+1} with probability $\mathbb{P}(s_{t+1} = s' \mid s_t = s, a_t = a)$.
- The agent receives an immediate **reward** $r(s, s', a)$.

Goal: learn a **policy** $\pi \colon \mathcal{S} \to \mathcal{A}$ that maximizes the expected **return**

$$\mathbb{E}\left[\sum_{t=0}^{\infty} r(s_t, s_{t+1}, a_t)\right] \qquad \text{where } a_t \sim \pi(s_t)$$



The cherry: most important real-world problems involve complex decision making with sparse supervision signal (healthcare, self-driving etc.)

Challenges in reinforcement learning

- *Delayed reward*: which actions are responsible for the reward/penalty?
- *Incomplete information*: exploration vs exploitation
- *Expensive exploration*: real world RL (education, healthcare, self-driving)

- Extremely flexible framework
- Challenging to do RL from scratch (often needs to pre-train by SL)

Example with a simulator

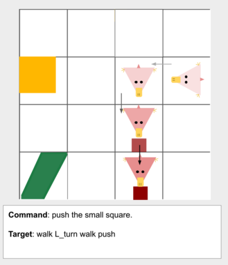


Figure: [Ruis+ 2020]

RL formulation:

- Action: walk, turn-L/R, push etc.
- What is the state?
- Reward: 1 if the task is completed and 0 otherwise

Want to learn:

- What is a “square” / “circle” / ...?
- What is “small” / “big” / ...?
- What is “red” / “green” / “yellow” / ...?

Policy

A typical model for instruction following

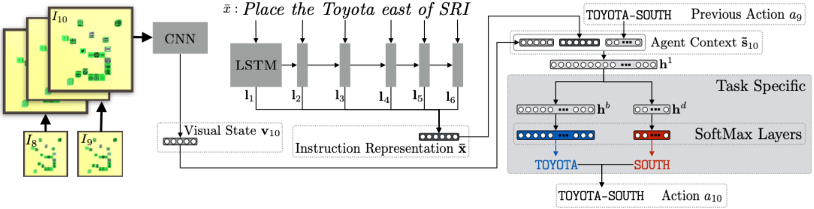


Figure: [Misra+ 2017]

- (visual input, textual instruction) $\to$ action
- Stochastic policy: $\pi_\theta(a \mid s) = p_\theta(a \mid s)$
- Parametrization: multimodal networks.
- May need to add history observation into the state.

Learning

Policy gradient methods: directly learn π parametrized by θ by maximizing the expected return

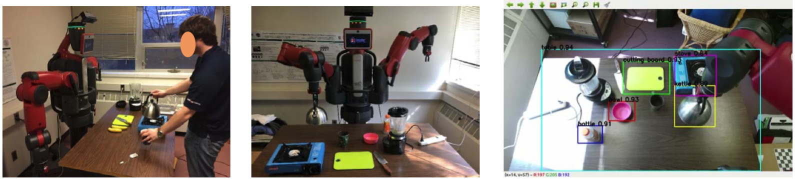
$$J(\theta) = \mathbb{E}_{\pi}\left[Q^{\pi}(s, a)\right]$$
$$\nabla_{\theta} J(\theta) = \mathbb{E}_{\pi}\left[\nabla_{\theta} \log \pi_{\theta}(a \mid s) Q^{\pi}(s, a)\right]$$

- Expectation over the starting state distribution and the stationary distribution of π_{θ}
- $Q^{\pi}(s, a)$: expected return starting from state s, taking action a, and following π ("cost-to-go")
- REINFORCE: estimate $Q^{\pi}(s, a)$ by Monte Carlo sampling
- Implementation
  1. Sample trajectories from π_{θ}
  2. Receive rewards
  3. Gradient update: weighted MLE update

More realistic simulators

Figure: The Room-to-Room dataset [Anderson+ 2018]

Robot learning



(a) Robot learning from human language instruction and action demonstration.

(b) Robot learning through its own actions by following human instruction and demonstration

(c) Robot's perception of the physical world during learning.

Figure: Interactive Task Learning with Physical Agents [Chai+ 2018]

Additional supervision: human demonstration, guidance through conversation

Summary

Robot navigation with instructions

Modeling: multimodal neural networks

Learning: reinforcement learning ($+$ supervised learning)

- Learn the connection between language and the world in an end-to-end way
- Require a large number of interactions (may not be realistic)

Inference: best action ($+$ planning)

Table of Contents

Speakers and listeners

Speakers: world to language

- Image caption
- Color description
- Instruction giving

Listeners: language to world

- Semantic parsing
- Visual reasoning
- Instruction following

What are scenarios/tasks with both listeners and speakers?

Reference games

Identify the target image

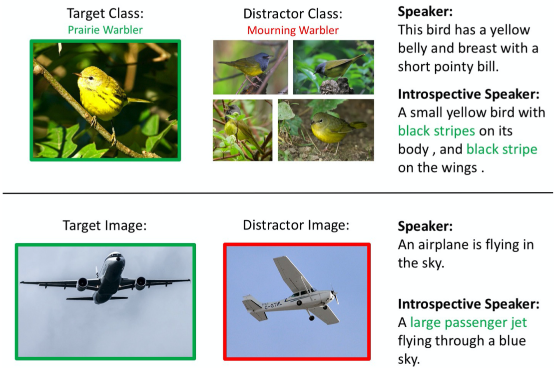


Figure: [Vedantam+ 2017]

- Base speaker: caption is consistent with both images
- Context-sensitive speaker: caption is discriminative

Generating and following instructions

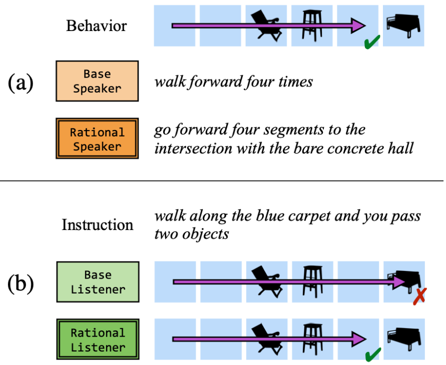


Figure: [Fried+ 2018]

- Rational speaker: what's the listener's orientation?
- Rational listener: should I pass exactly two objects or at least two?

Collaborative games

Name	Company	Time	Location
Kathy	TRT Holdings	afternoon	indoor
Jason	Dollar General	afternoon	indoor
Johnny	TRT Holdings	afternoon	outdoor
Frank	SFN Group	afternoon	indoor
Catherine	Dollar General	afternoon	indoor
Catherine	Weis Markets	afternoon	indoor
Kathleen	TRT Holdings	morning	indoor
Lori	TRT Holdings	afternoon	indoor
Frank	L&L Hawaiian Barbecue	afternoon	outdoor

Name	Company	Time	Location
Justin	New Era Tickets	morning	indoor
Kathleen	TRT Holdings	morning	indoor
Gloria	L&L Hawaiian Barbecue	morning	indoor
Kathleen	Advance Auto Paris	morning	outdoor
Justin	Dollar General	morning	indoor
Anna	Arctic Cat	morning	indoor
Steven	Dollar General	morning	indoor
Wayne	R.J. Corman Railroad	morning	indoor
Alexander	R.J. Corman Railroad	morning	indoor

hi

all of my friends prefer morning

1 of my morning likes the indoors

And all like indoor except one

do they work for trt holdings?

Kathleen?

SELECT (Kathleen, TRT Holdings, morning, indoor)

SELECT (Kathleen, TRT Holdings, morning, indoor)

Figure: [He+ 2017]

- Need knowledge from both agents to solve the puzzle
- Efficient collaboration requires reasoning about the other agent's knowledge

Efficient referential communication

$\text{state} = \{\text{blueSquare}, \text{blueCircle}, \text{greenSquare}\}$

$\text{utterance} = \{\text{square}, \text{circle}, \text{green}, \text{blue}\}$

Assuming the speaker is cooperative, which object does “blue” refer to?

Rational speech act model

Literal listener: interprets an utterance according to its literal meaning

“blue”: blueSquare or blueCircle

Pragmatic speaker: minimize the literal listener’s effort of inferring the state while maximizing communication efficiency

blueSquare: “blue” or “square” or “blue square”

Pragmatic listener: infer the state by reasoning about the pragmatic speaker

“blue”: blueSquare

Rational speech act model

blue	1/2	1/2	0
square	1/2	0	1/2
circle	0	1	0
green	0	0	1

Literal listener L_0: interprets an utterance according to its literal meaning

$$p_{L_0}(s \mid u) \propto \underbrace{p(s)}_{\text{state prior}} \underbrace{m(s,u)}_{\text{world model}}$$

$$m : \mathcal{S} \times \mathcal{U} \to \{0,1\}$$

Rational speech act model

blue	1/2	1/3	0
square	1/2	0	1/3
circle	0	2/3	0
green	0	0	2/3

Pragmatic speaker: minimize the literal listener's effort of inferring the state while maximizing communication efficiency

$$p_{S_1}(u \mid s) \propto \exp(\alpha U_{S_1}(u; s))$$
$$U_{S_1}(u; s) = \log p_{L_0}(s \mid u) - C(u)$$

Rational speech act model

blue	3/5	2/5	0
square	3/5	0	2/5
circle	0	1	0
green	0	0	1

Pragmatic listener: infer the state by reasoning about the pragmatic speaker

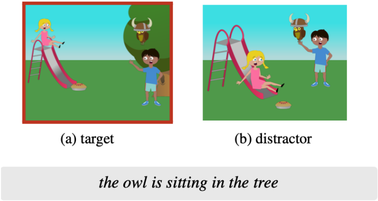
$$p_{L_1}(s \mid u) \propto p_{S_1}(u \mid s)p(s)$$

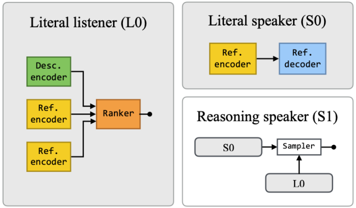
Neural RSA

Limitation of RSA

- Pre-defined (small) lexicon
- Enumerate over all possible sequences

Learned speaker and listener with basic reasoning [Andreas+ 2016]





Summary

- Philosophy: language as a tool
- Goal: build agents with language capability working in human-centered environments
- Challenge: scale to realistic, persistent, interactive scenarios (with humans)