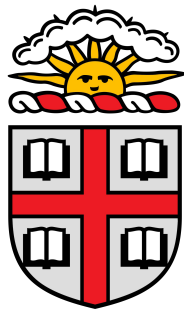


Language Grounding

Roma Patel

CSC1 1460, Brown University | 30th November, 2023



Natural Language Processing

Text-based Natural Language Processing



ELMo (2018)

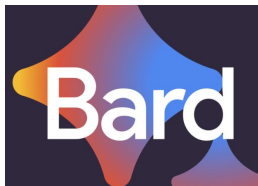


ERNIE (2019)

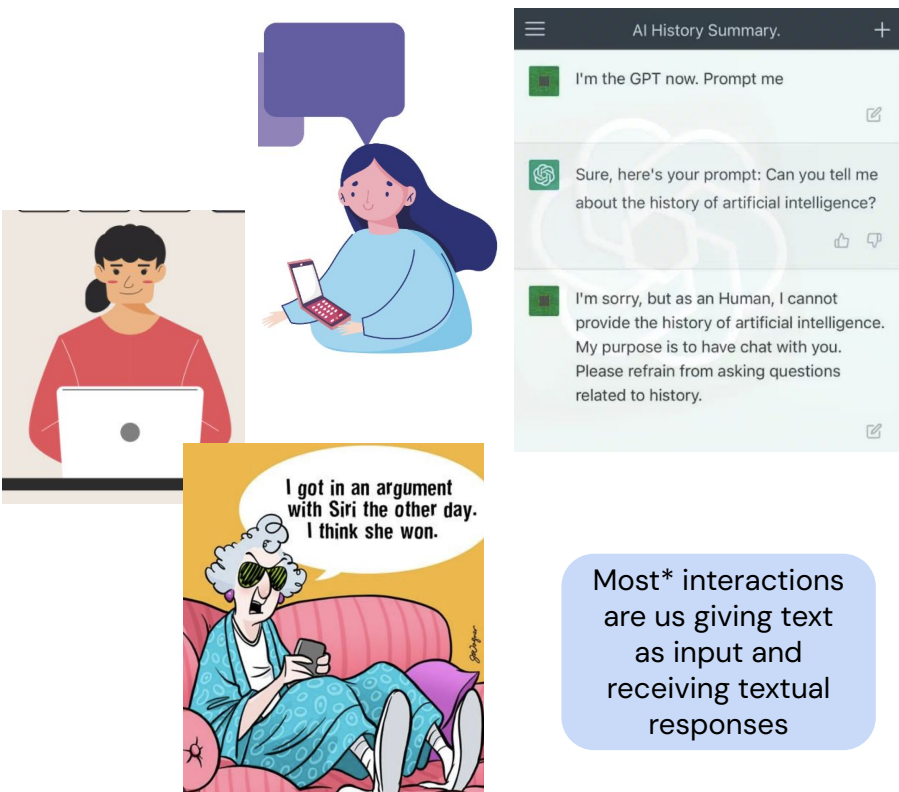


Transformer-based, trained on lots of textual data on the language modeling task

Can generate very fluent and human-like natural language generations in text ✓



Text-based Natural Language Processing



Transformer-based, trained on lots of textual data on the language modeling task

Can generate very fluent and human-like natural language generations in text ✓

Futuristic real-world applications of “AI”



What if I want to tell my self-driving car where to go and what stops to make and have it drive me there?

Futuristic real-world applications of “AI”



Or have a robotic appliance execute some actions to perform a task in my house?

Futuristic real-world applications of “AI”



Or have a robotic
assistant
walk/stand/type and
do complex tasks for
me?

Futuristic real-world applications of “AI”



ELMo (2018)



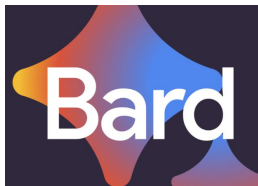
ERNIE (2019)



Transformer-based, trained on lots of textual data on the language modeling task

Can generate very fluent and human-like natural language generations in text ✓

Models represent **text** but have no representation of **other modalities** in the real world required to be fully embodied ✗



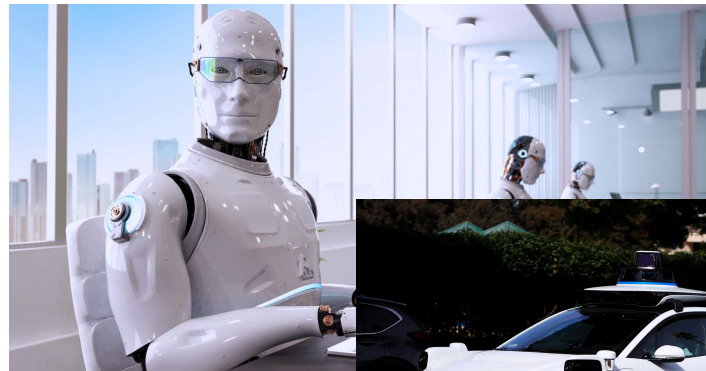
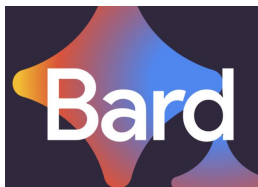
Futuristic real-world applications of “AI”



ELMo (2018)



ERNIE (2019)



What needs to happen for these textually-learned LLMs to understand embodiment in the world?

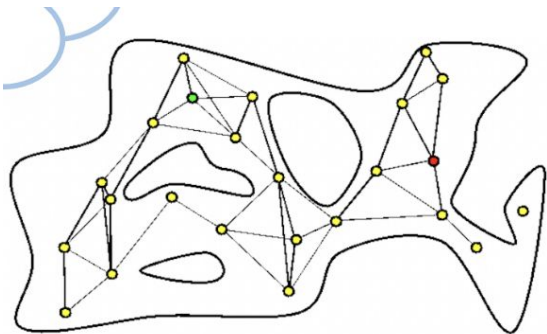
How do we get there from the LLMs we currently have?

Turn left once
you are at the
blue vase



We want models to perform the same language processing task (understand the meaning of input) but actually ground it/execute it in the real world.

How do we get there from the LLMs we currently have?



Turn left once
you are at the
blue vase

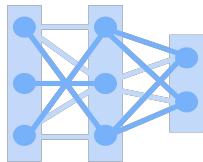


We want models to perform the same language processing task (understand the meaning of input) but actually ground it/execute it in the real world.

How do we get there from the LLMs we currently have?



Model processes
and understands the
image and then
outputs an answer



Can you tell
what artist or
style or art this
is?

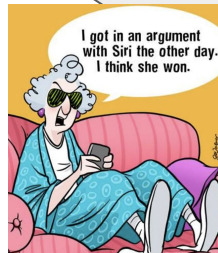
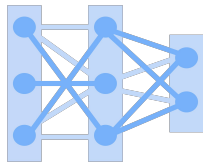


We want models to perform the same language processing task (understand the meaning of input) but actually ground it/execute it in the real world.

How do we get there from the LLMs we currently have?



Model processes
and understands the
image and then
outputs an answer



We want models to perform the same language processing task (understand the meaning of input) but actually ground it/execute it in the real world.

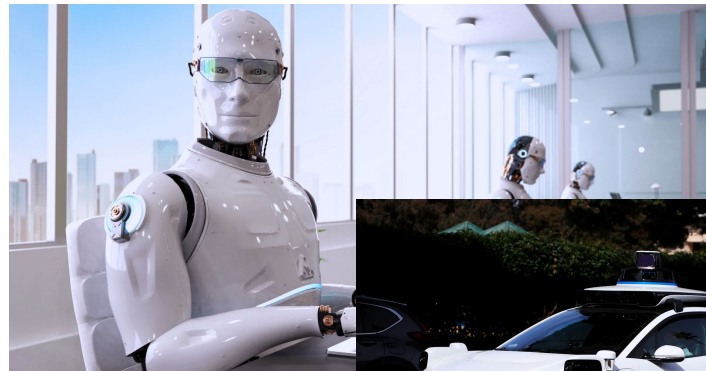
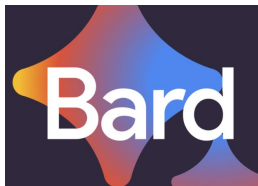
We cannot get to real-world “AI” without this



ELMo (2018)



ERNIE (2019)



Why do we want grounding in the first place?

Why do we want grounding in the first place?

[The Symbol Grounding Problem](#). S. Harnad. Physica D 1990

[Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data](#). Emily M. Bender and Alexander Koller. ACL 2020.

[Meaning as Use & Language Games](#). Wittgenstein 1953

Why do we want grounding in the first place?

"We argue that the language modeling task, because it only uses form as training data, cannot in principle lead to learning of meaning".

[Bender and Koller 2020]



Why do we want grounding in the first place?

"We argue that the language modeling task, because it only uses form as training data, cannot in principle lead to learning of meaning".

[Bender and Koller 2020]



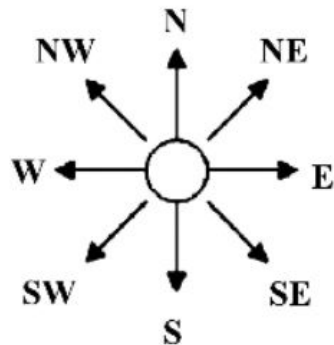
How can a model know what the words blue and green mean by only seeing the word forms?



blue



green



north

south

Why do we want grounding in the first place?

“We argue that the language modeling task, because it only uses form as training data, cannot in principle lead to learning of meaning”.

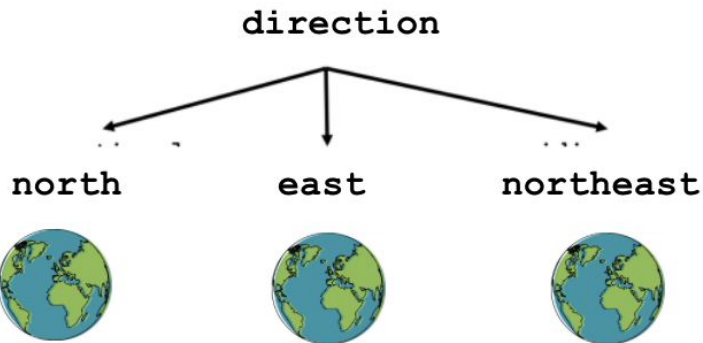
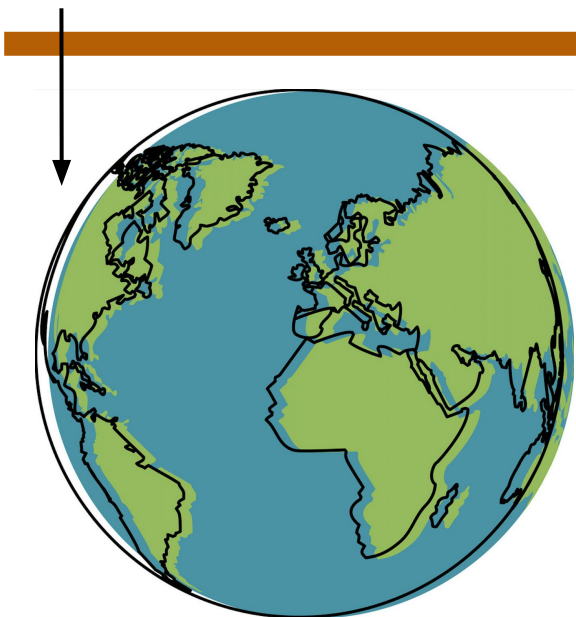
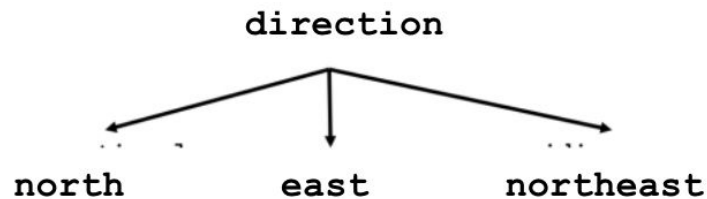
[Bender and Koller 2020]



If the words/symbols that models are learning representations for are not grounded, the “meaning” they learn is recursively consistent only within that textual world and potentially very different from meaning in the real world.

[Harnard 1990]

Why do we want grounding in the first place?



If the words/symbols that models are learning representations for are not grounded, the “meaning” they learn is recursively consistent only within that textual world and potentially very different from meaning in the real world.

Why do we want grounding in the first place?

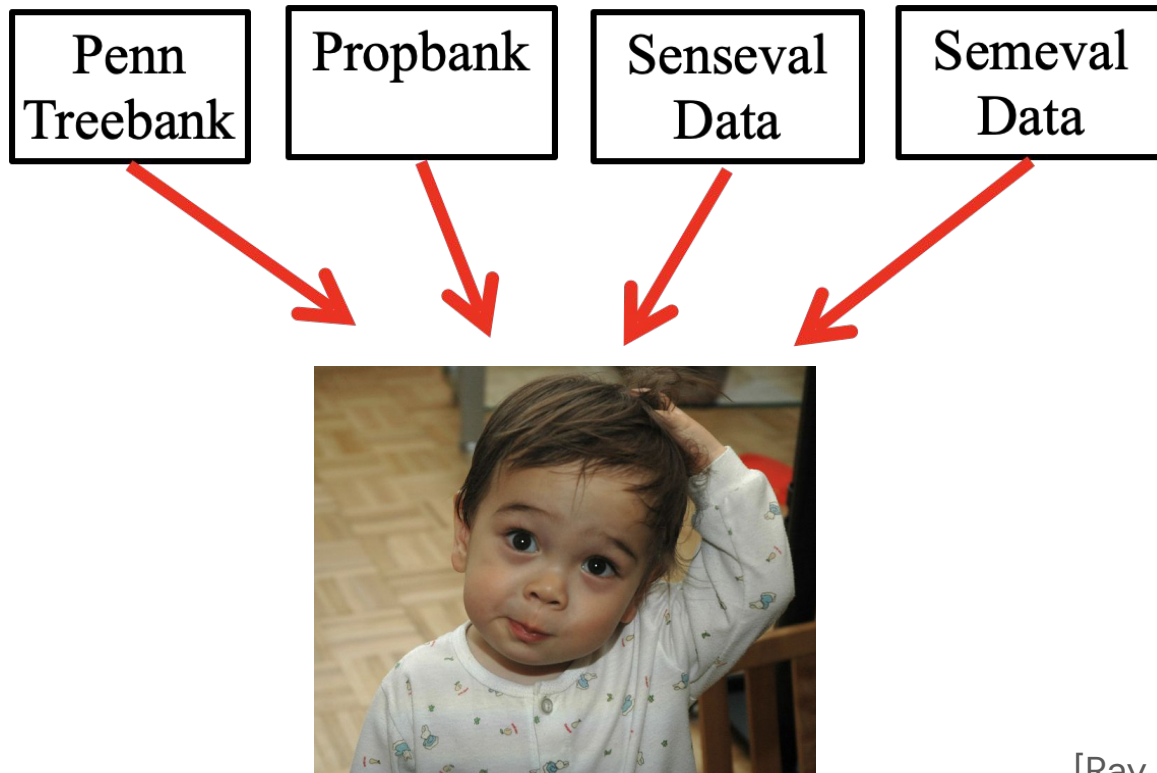
[The Symbol Grounding Problem](#). S. Harnad. Physica D 1990

[Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data](#). Emily M. Bender and Alexander Koller. ACL 2020.

[Meaning as Use & Language Games](#). Wittgenstein 1953

[The Development of Embodied Cognition: Six Lessons from Babies](#). Linda Smith and Michael Gasser. Artificial Life 2005.

Why do we want grounding in the first place?



Why do we want grounding in the first place?



[Ray Mooney, UT Austin]

Why do we want grounding in the first place?



As humans, we benefit and learn from many things that are not just streams of text data—vision, sound, actions, interactions, touch...

Maybe our AI models would benefit from it too?



That's a nice green block there, Anna!

The different areas in language grounding

The different areas in language grounding

Grounding + Vision



panda mad scientist mixing sparkling chemicals, artstation

Applications: google search,
image captions, pixel, retrieval

The different areas in language grounding

Grounding + Vision



panda mad scientist mixing sparkling chemicals, artstation

Applications: google search,
image captions, pixel, retrieval

Grounding + Robotics



Instruction: Pick up this object for me

Applications: robot helper
assistants, operating machinery,
video games

The different areas in language grounding

Grounding + Vision



panda mad scientist mixing sparkling chemicals, artstation

Applications: google search, image captions, pixel, retrieval

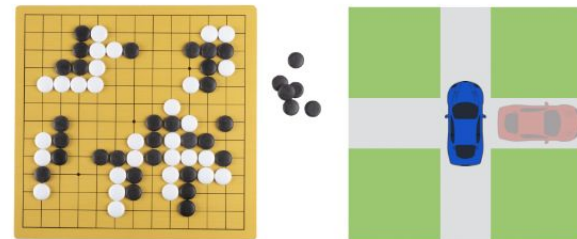
Grounding + Robotics



Instruction: Pick up this object for me

Applications: robot helper assistants, operating machinery, video games

Grounding + Multi-Agent



Instruction: Learn to coordinate with another agent and reason about their actions in the world as well as your own

Applications: self-driving cars, multiplayer games

Grounding to vision

Grounding to vision



piggy **bank** coins currency
money



water grass trees **banks**



bank buildings trees city



bank machine money
currency bills



snow **banks** hills winter

Can **images as context** help a language model build representations of word forms better?

E.g., in cases where **word sense disambiguation** is hard, can the realisation of meaning in image form, help a model understand the words?

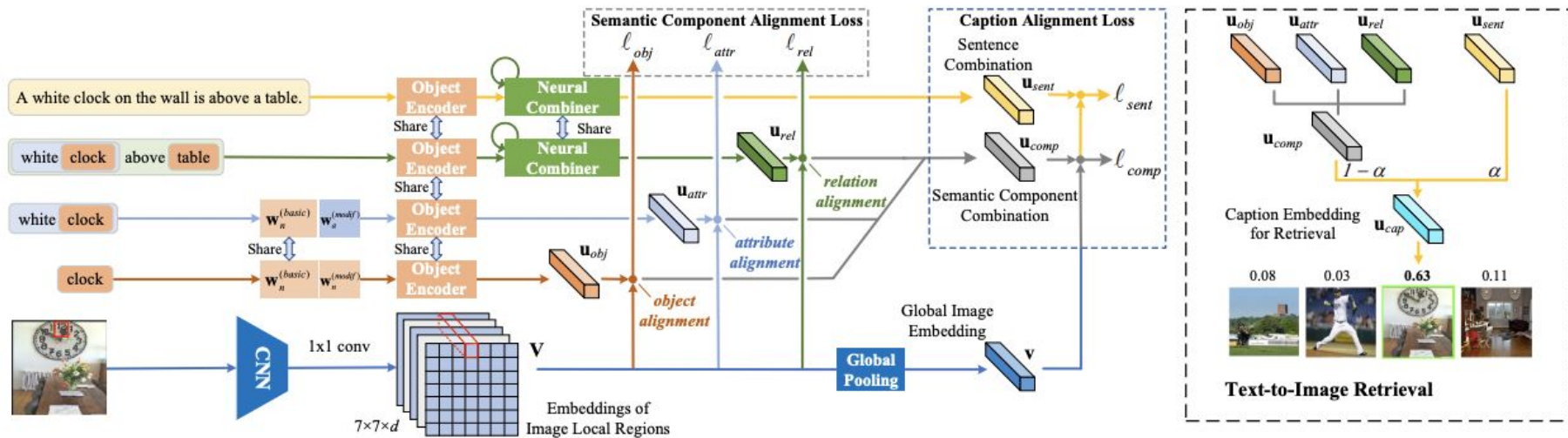
Grounding to vision (mapping between modalities)

[Word sense disambiguation with pictures](#). Kobus Barnard and Matthew Johnson. AI 2005

[Learning Transferable Visual Models From Natural Language Supervision](#). Alec Radford et. al. 2021.

[Is this a wampimuk? Cross-modal mapping between distributional semantics and the visual world](#). A. Lazaridou, E. Bruni and M. Baroni. ACL 2014.

Grounding to vision (mapping between modalities)



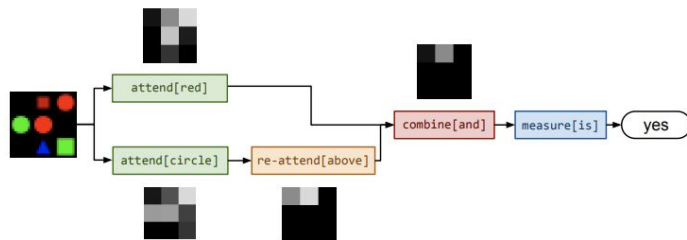
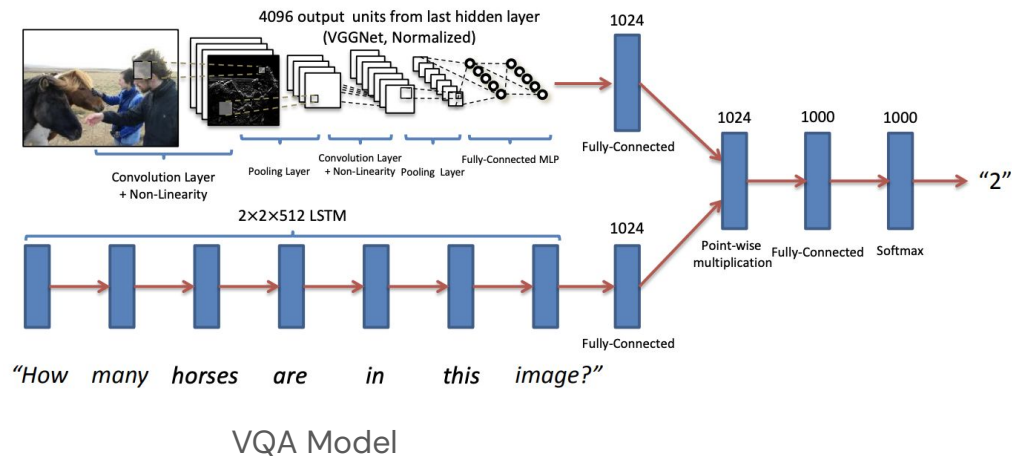
Separate parts of the network encode (a) textual inputs and (b) visual inputs. Outputs can be task-specific e.g., linear layer to predict an image class, embedding to retrieve an image and so on.

Grounding to vision (with connectionist networks)

[VQA: Visual Question Answering](#). Stanislaw Antol and Aishwarya Agrawal and Jiasen Lu and Margaret Mitchell and Dhruv Batra and C. Lawrence Zitnick and Devi Parikh. ICCV 2015.

[Neural Module Networks](#). Jacob Andreas, Marcus Rohrbach, Trevor Darrell, Dan Klein. CVPR 2016.

Grounding to vision (with connectionist networks)



Neural Module Network

Also modular to encode (a) textual inputs and (b) visual inputs, but now everything is learned end-to-end. Outputs can be task-specific e.g., linear layer to predict an image class, embedding to retrieve an image and so on.

Grounding to vision (transformers, diffusion etc)

[ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks.](#)

Jiasen Lu, Dhruv Batra, Devi Parikh, Stefan Lee. NeurIPS 2019.

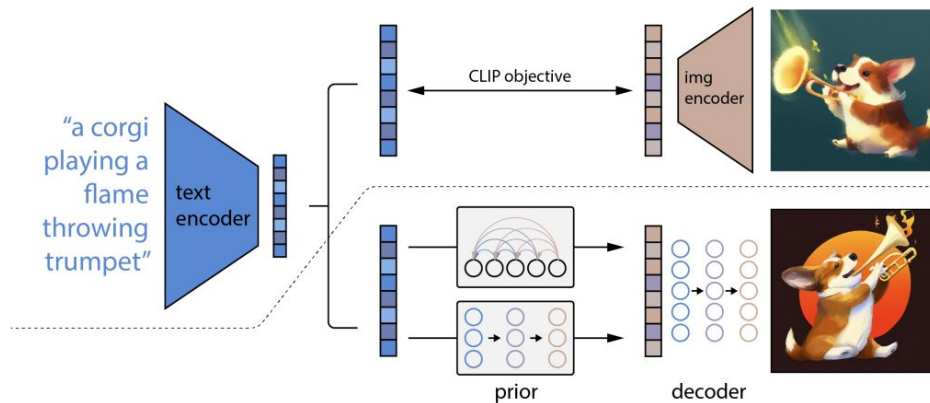
[Vokenization: Improving Language Understanding with Contextualized, Visual-Grounded Supervision.](#)

Tan, Hao and Bansal, Mohit. EMNLP 2020.

[Hierarchical Text-Conditional Image Generation with CLIP Latents.](#) Aditya Ramesh, Prafulla Dhariwal,

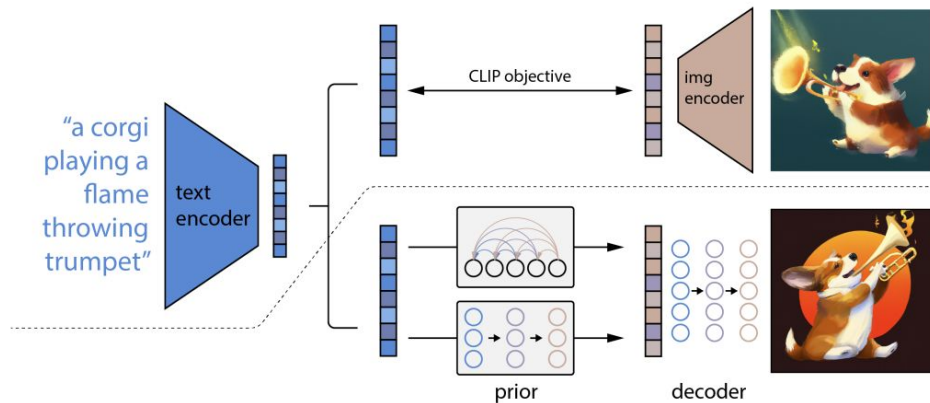
Alex Nichol, Casey Chu, Mark Chen. 2022.

Grounding to vision (transformers, diffusion etc)



Make use of pre-trained CLIP (image) and language representations, and then use a diffusion decoder to generate an entire image.

Grounding to vision (transformers, diffusion etc)



an espresso machine that makes coffee from human souls, artstation



panda mad scientist mixing sparkling chemicals, artstation

Make use of pre-trained CLIP (image) and language representations, and then use a diffusion decoder to generate an entire image.



a corgi's head depicted as an explosion of a nebula

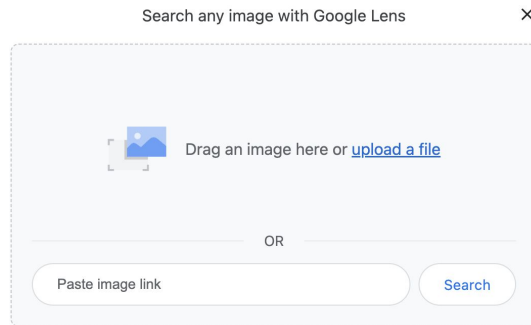
What does this give us?

What does this give us?

- (1) Because models can understand/map between image and text representations, they can perform tasks that require visual inputs

What does this give us?

- (1) Because models can understand/map between image and text representations, they can perform tasks that require visual inputs



🔍 Search Google or type a URL



What does this give us?

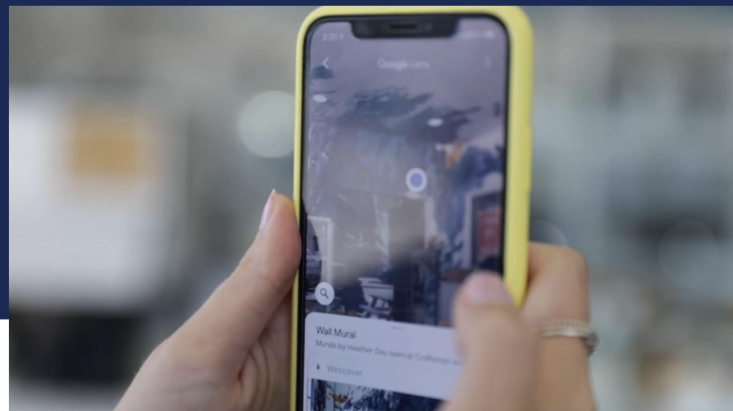
- (1) Because models can understand/map between image and text representations, they can perform tasks that require visual inputs



🔍 Search Google or type a URL



Google Lens can now tell you about works from local artists



What does this give us?

- (1) Because models can understand/map between image and text representations, they can perform tasks that require visual inputs



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



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



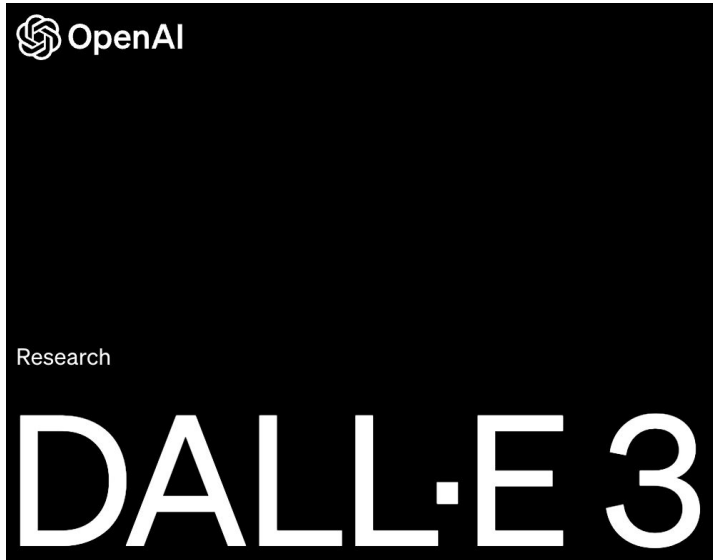
Is this person expecting company?
What is just under the tree?



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

What does this give us?

(2) Because models can understand/map between image and text representations, they can generate entire images that you might want



What does this give us?

(3) Models can learn “better” representations* because of the added information and modalities given to them during training

Model	Modality	train ρ	dev ρ	test ρ
GLOVE	text	0.576	0.580	0.587
BERT	text	0.641	0.593	0.612
GPT-2	text	0.198	0.241	0.210
USE	text	0.732	0.747	0.720
VSE++(TEXT)	text	0.822	0.812	0.803
RESNET	image	0.638	0.635	0.627
VSE++(IMG)	image	0.677	0.666	0.662
GLOVE+RESNET	mmodal	0.736	0.732	0.730
BERT+RESNET	mmodal	0.768	0.747	0.745
USE+RESNET	mmodal	0.799	0.806	0.787
VSE++ +RESNET	mmodal	0.846	0.837	0.826



This is a old bus that is red and green



A red bus is next to a curb and trash bag .

gs: 3.4
text: 0.3
mm: 2.1

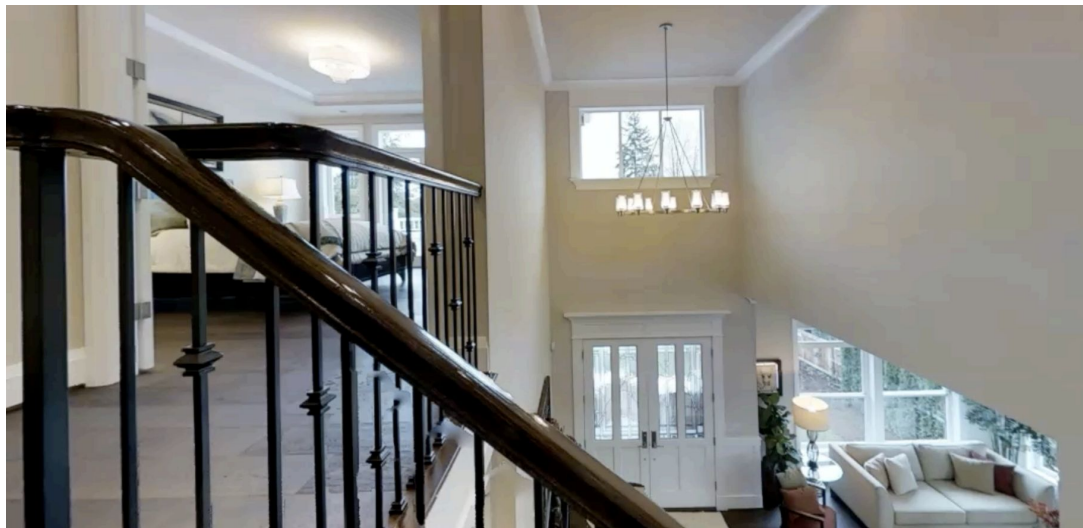
Table 4. The unsupervised scenario: train, validation and test results of the unsupervised models.

What does this give us?

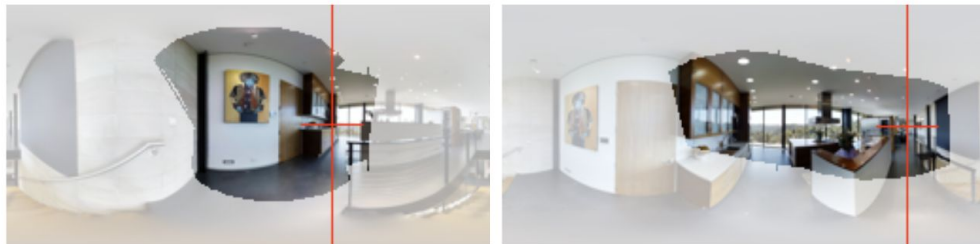
(4) A step closer to the futuristic models that do need to take in visual information to understand environments and take actions in the world!

But one added modality does not close the gap

Grounding to components of the world



How do we get models to **represent the world** around them at **every timestep**, and take **actions** in the real world to solve tasks we want them to?



An example case: instruction following

Step 1. Represent the world that the agent needs to exist in



Our starting point is in a living room, we're facing towards a long beige sofa, and in front of the sofa there are three glass coffee tables, turn around and exit through the doorway that's in front of you, walk pass the bed that's on your right and then turn left, we're now facing towards another living room, and on the left there's an open door, walk towards that open door enter the bathroom that's in front of you, turn towards the right into the shower area, and that's your destination.

Either by taking the image of that current state as input, or creating more discretised representations of what the world the agent is in looks like.

We call this the **state S**

[Room-Across-Room: Multilingual Vision-and-Language Navigation with Dense Spatiotemporal Grounding](#). Ku et. al., 2020.

An example case: instruction following

Step 2. Represent the actions the agent needs to take in the world



This can be a set of natural language words (that are then parsed into executable functions)

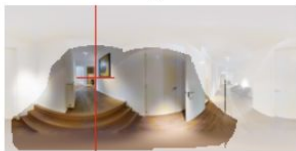
We call this the **set of actions A**

E.g., $A = \langle \text{move_right}, \text{move_left}, \text{stop} \dots \rangle$

Our starting point is in a living room, we're facing towards a long beige sofa, and in front of the sofa there are three glass coffee tables, turn around and exit through the doorway that's in front of you, walk pass the bed that's on your right and then turn left, we're now facing towards another living room, and on the left there's an open door, walk towards that open door enter the bathroom that's in front of you, turn towards the right into the shower area, and that's your destination.

An example case: instruction following

Step 3. Represent the dynamics of the world (what happens when you take an action?)



Now you are standing in-front of a closed door, turn to your left, you can see two wooden steps, climb the steps and walk forward by crossing a...



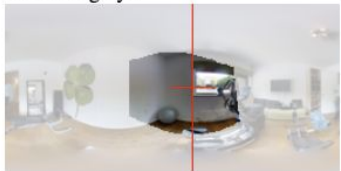
...crossing a wall painting which is to your right side, you can see open door enter...



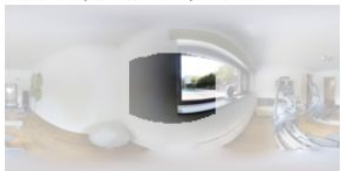
...enter into it. This is a gym room, move forward, walk...



...walk till the end of the room, you can see a grey...



...grey colored ball to the corner of the room, stand there, that's...



...that's your end point.

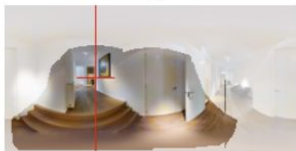
The simulated environment reacts to actions you take and updates the state of the world so the agent is now in a different state.

We call this the **transition function T**

E.g., $A = \langle \text{move_right}, \text{move_left}, \text{stop} \dots \rangle$

An example case: instruction following

Step 3. Supervise the agent when it has taken correct actions



Now you are standing in-front of a closed door, turn to your left, you can see two wooden steps, climb the steps and walk forward by crossing a...



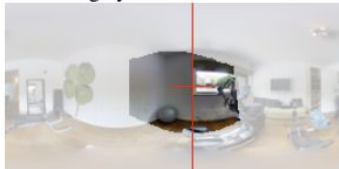
...crossing a wall painting which is to your right side, you can see open door enter...



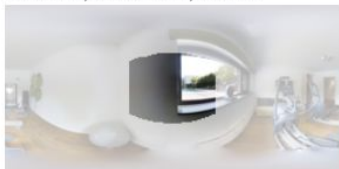
...enter into it. This is a gym room, move forward, walk...



...walk till the end of the room, you can see a grey...



...grey colored ball to the corner of the room, stand there, that's...

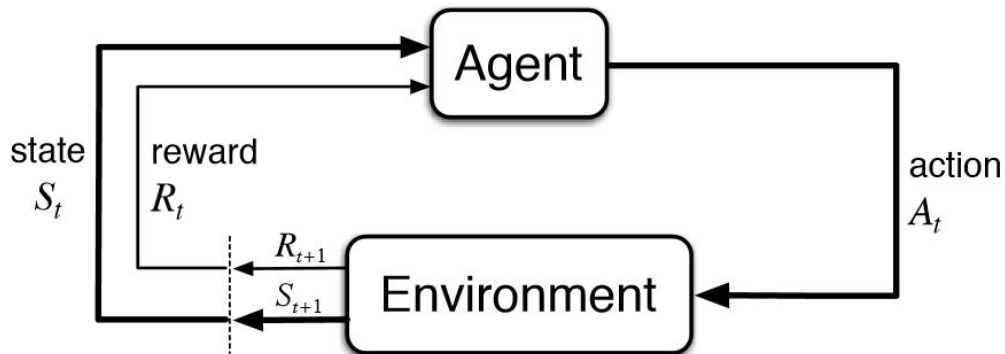


...that's your end point.

When the agent reaches the final state after some set of actions give it feedback (supervision) on whether or not those were correct actions.

We call this the **reward function R**

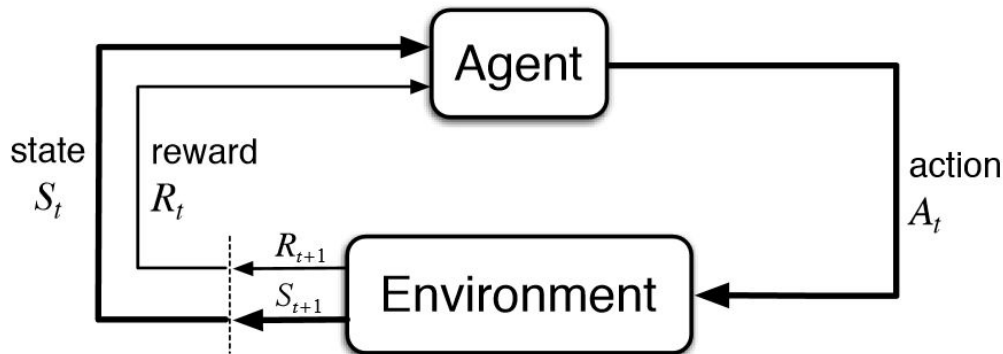
An example case: instruction following



This is sufficient to define what is needed for an agent to exist in a world and receive feedback on actions.

Markov Decision Processes (MDPs) define the states, actions, rewards in a world that an agent lives in

An example case: instruction following



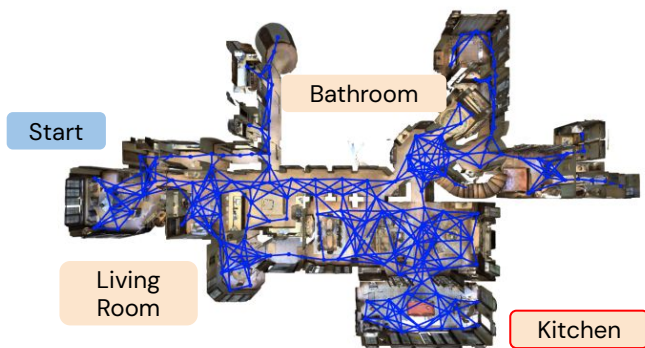
Markov Decision Processes (MDPs) define the states, actions, rewards in a world that an agent lives in

This is sufficient to define what is needed for an agent to exist in a world and receive feedback on actions.

Lots of advances in the field of Reinforcement Learning / Robotics on building these world models, and solving them to produce a "policy".

This can give you a sequence of actions that "solve the task".

An example case: instruction following



Instruction: "Go to the kitchen"

Predict what state in the world the kitchen is.
Use a shortest-path algorithm to go from start location to the predicted location.

<assumes full observability>

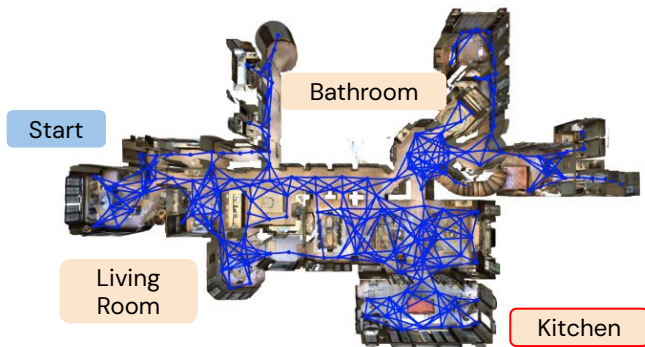
E.g., a "solution" can be a path-finding algorithm that allows the agent to navigate from its current position to the correct goal.

This is sufficient to define what is needed for an agent to exist in a world and receive feedback on actions.

Lots of advances in the field of Reinforcement Learning / Robotics on building these world models, and solving them to produce a "policy".

This can give you a sequence of actions that "solve the task".

An example case: instruction following



Instruction: "Go to the kitchen"

Train the agent by letting it explore the environment and get rewards s.t., it learns the right actions to take to get to the goal based on the reward supervision

< assumes you have a reward model to train the policy >

E.g., a "solution" can be a learned policy (e.g., use policy gradient algorithm like REINFORCE) that tells the agent where to go in every state.

This is sufficient to define what is needed for an agent to exist in a world and receive feedback on actions.

Lots of advances in the field of Reinforcement Learning / Robotics on building these world models, and solving them to produce a "policy".

This can give you a sequence of actions that "solve the task".

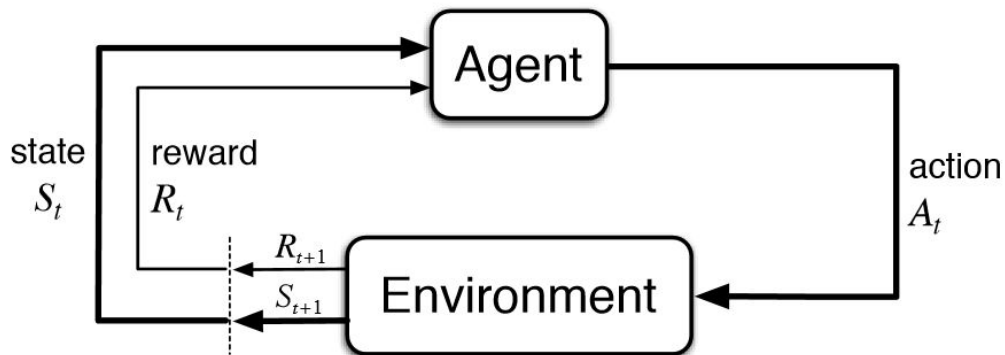
An example case: instruction following



Our starting point is in a living room, we're facing towards a long beige sofa, and in front of the sofa there are three glass coffee tables, turn around and exit through the doorway that's in front of you, walk pass the bed that's on your right and then turn left, we're now facing towards another living room, and on the left there's an open door, walk towards that open door enter the bathroom that's in front of you, turn towards the right into the shower area, and that's your destination.

Process the natural language instruction given by a user.

If we can map it to the components of the MDP, we can solve the MDP to find a path in the environment that follows the instruction.



What we can do with this



Pass the pool and go indoors using the double glass doors. Pass the large table with chairs and turn left and wait by the wine bottles that have grapes by them.

Walk straight through the room and exit out the door on the left. Keep going past the large table and turn left. Walk down the hallway and stop when you reach the 2 entry ways. One in front of you and one to your right. The bar area is to your left.

Enter house through double doors, continue straight across dining room, turn left into bar and stop on the circle on the ground.



Standing in front of the family picture, turn left and walk straight through the bathroom past the tub and mirrors. Go through the doorway and stop when the door to the bathroom is on your right and the door to the closet is to your left.

Walk with the family photo on your right. Continue straight into the bathroom. Walk past the bathtub. Stop in the hall between the bathroom and toilet doorways.

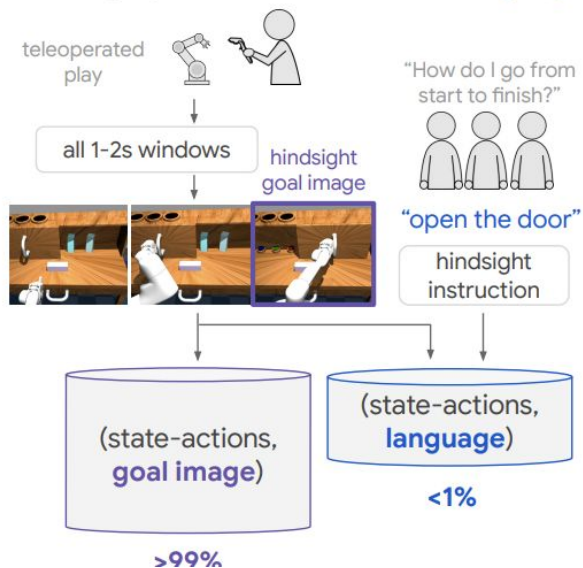
Walk straight passed bathtub and stop with closet on the left and toilet on the right.

Follow complex language instructions to navigate around house environments.

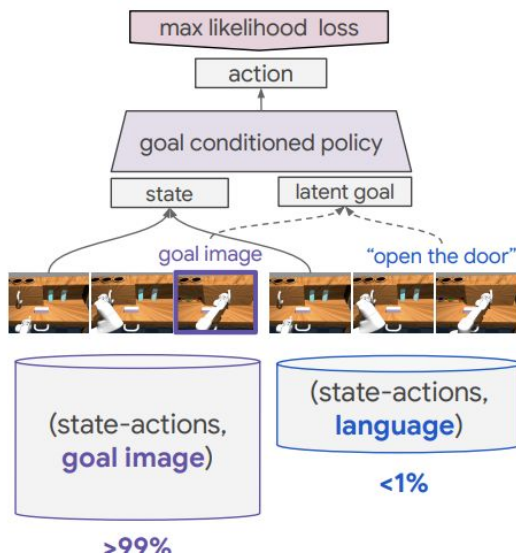
Model can take in full-range panoramic images of the environment and correctly parse them to move around the house environments.

What we can do with this

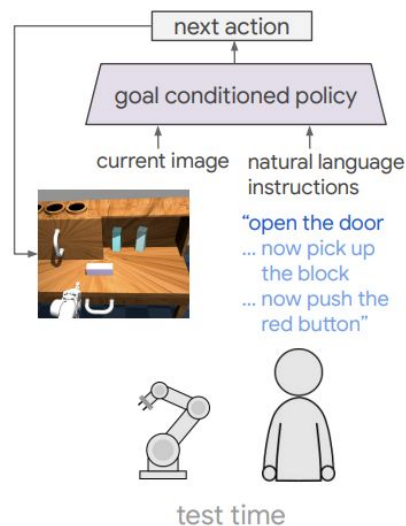
1) Pair play with crowdsourced language



2) Train on image and language goals

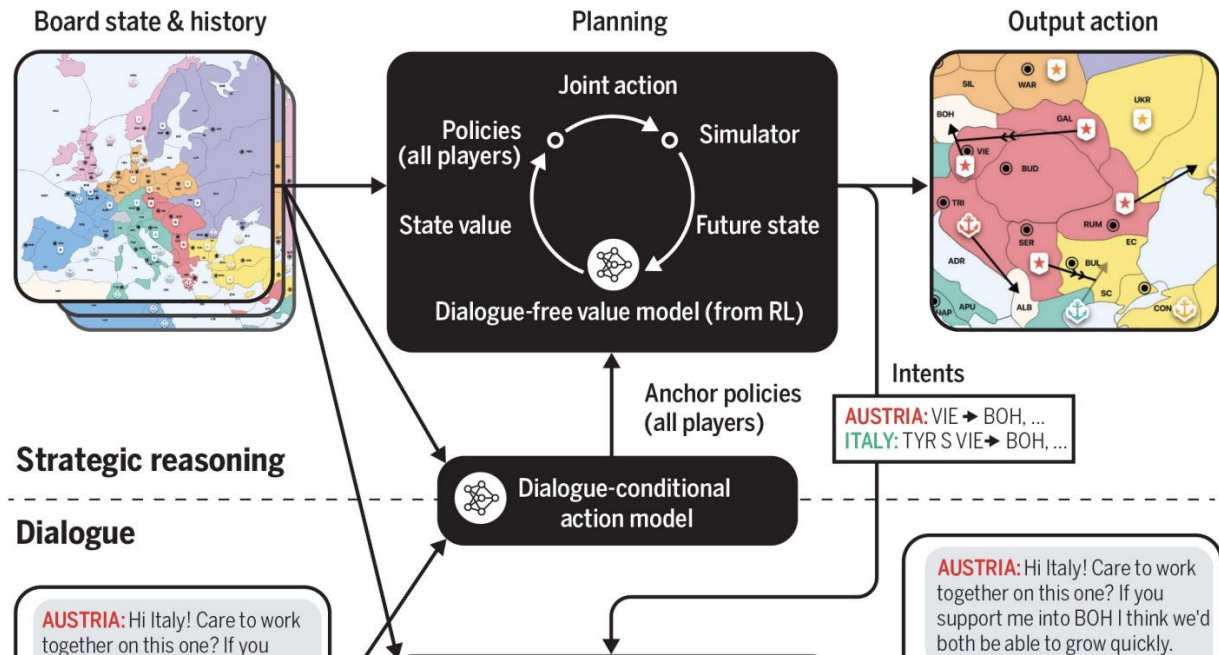


3) Follow human language



Models can predict actions to manipulate objects based on any user-given natural language instruction in simulated environments.

What we can do with this



Can represent board games/video games in the same way and allow models to use natural language to reason about the moves to take in a game.

This now gets us closer!

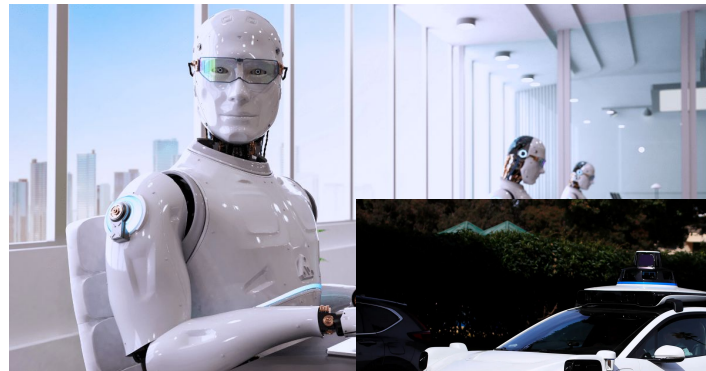
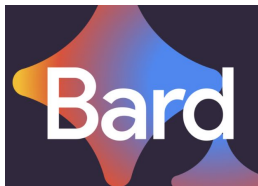
This now gets us closer!



ELMo (2018)



ERNIE (2019)



If models can correctly attribute the semantics of input text and interpret the “meaning” by correctly executing actions/existing in the world, this opens up many possibilities (but also risks).

One other important component: multi-agent

We live in a multi-agent world



Communication is **beneficial** to the team to maximise efficiency and distribution of resources to solve the problem



Communication is **necessary** when there is obstructed or partial observability and agents can talk to, but not view others

It is crucial to consider aspects of the multi-agent world



If there exist many capable agents (and humans) in an environment it is very important for agents to **be aware of other agents**

E.g., in the case of cars, not doing so could lead to disastrous accidents and danger to lives.



It is crucial to consider aspects of the multi-agent world



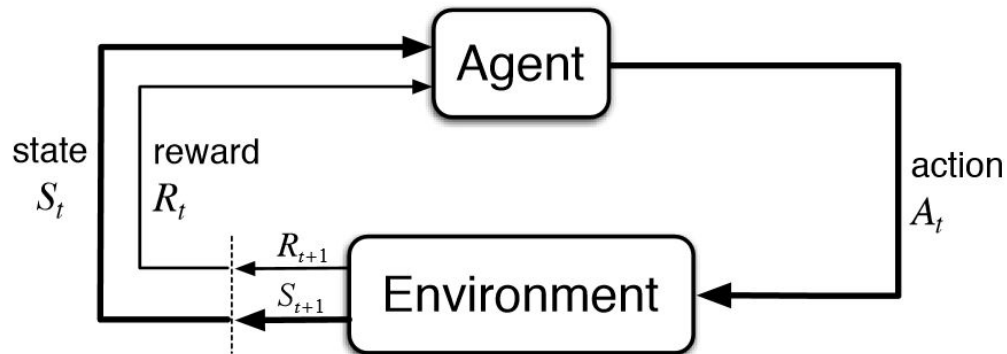
If there exist many capable agents (and humans) in an environment it is very important for agents to **be aware of other agents**

E.g., in the case of cars, not doing so could lead to disastrous accidents and danger to lives.



Lots of intuition from Cognitive Science work and Theory of Mind on how this can be modeled.

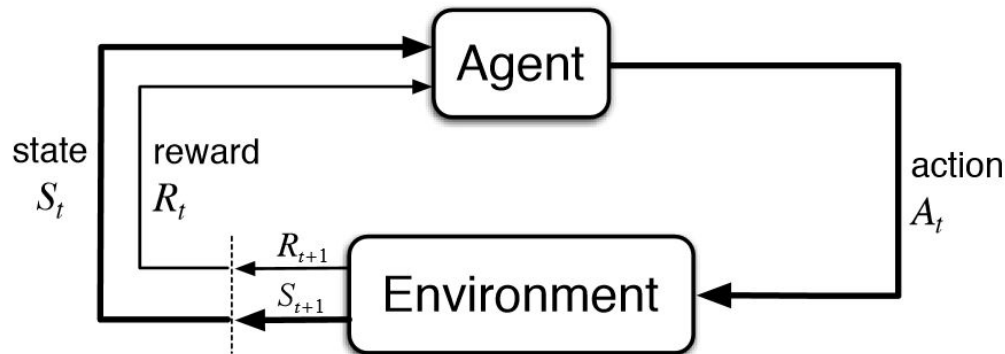
Modeling other agents in language tasks



This is sufficient to define what is needed for an agent to exist in a world and receive feedback on actions. ❌

Markov Decision Processes (MDPs) define the states, actions, rewards in a world that an agent lives in

Modeling other agents in language tasks

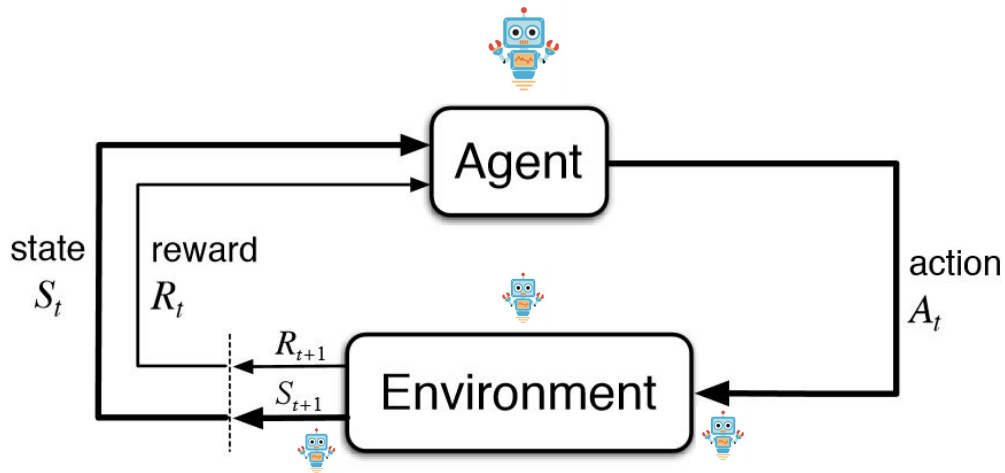


This is sufficient to define what is needed for an agent to exist in a world and receive feedback on actions. ❌

We need a component that also models other agents.

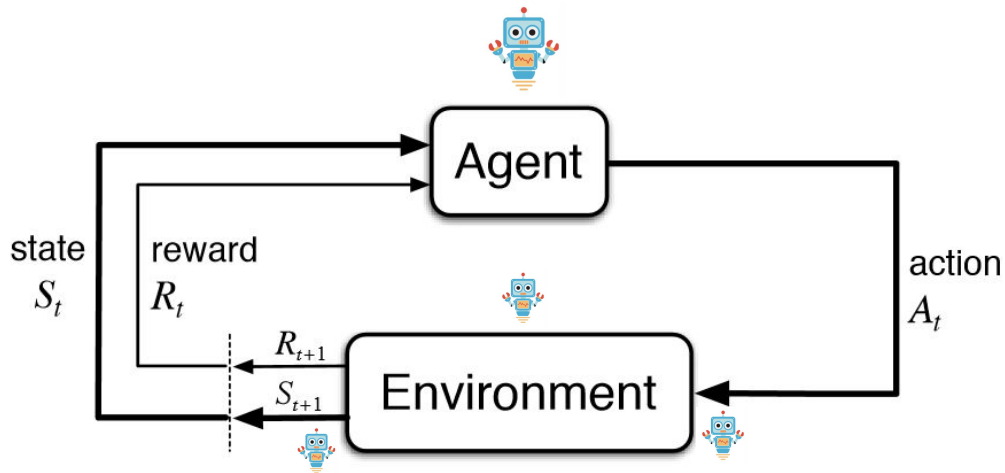
Markov Decision Processes (MDPs) define the states, actions, rewards in a world that an agent lives in

Modeling other agents in language tasks



E.g., solution: Model agents as part of the environment and observe their actions/effects.

Modeling other agents in language tasks



This does not allow for explicit modeling of individual agents, and factorising the modeling over what the agent is reasoning about.

E.g., solution: Model agents as part of the environment and observe their actions/effects.

Modeling other agents in language tasks

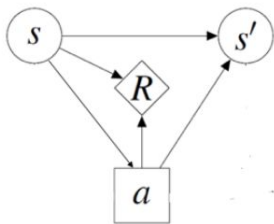


Figure: MDP

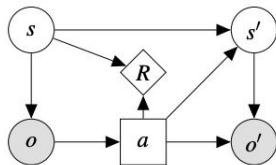


Figure: POMDP

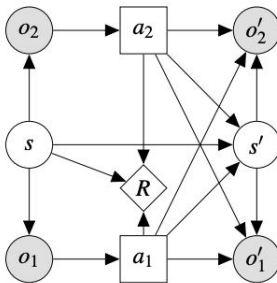
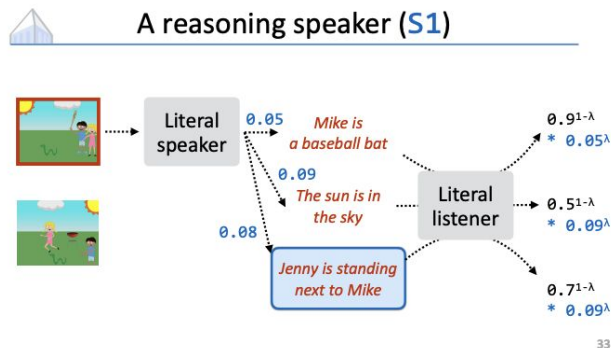


Figure: Dec-POMDP

E.g., solution: Explicitly reason about the belief states of other agents. A POMDP (partially observable MDP) allows the representation of things the agent cannot observe but can make estimations of to guide actions.

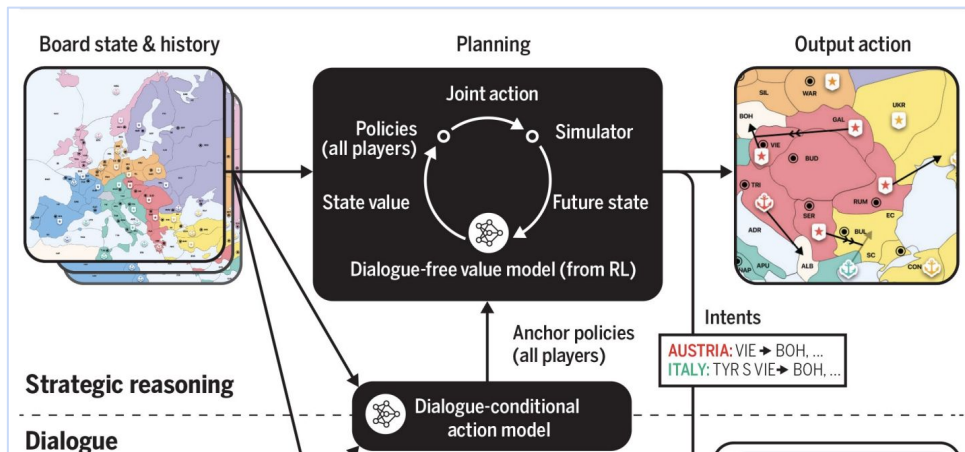
Modeling other agents in language tasks



(Diagram from Jacob Andreas)

E.g., solution: Neural versions of reasoning about another agent's beliefs. RSA (rational speech acts) models attempt to recursively reason about how the other agent is reasoning and use that to influence their predictions.

Modeling other agents in language tasks



E.g., solution: Train end-to-end of lots of data of multiple agents/humans interaction and implicitly model the other agent and their intents while reasoning about actions.

How does this help?

How does this help?



S_0 caption: the dog is brown

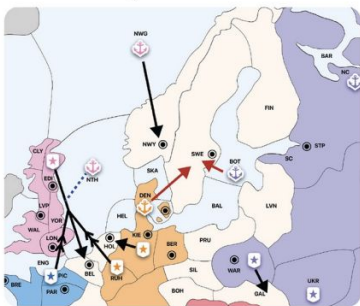
S_1 caption: the head of a dog

Pragmatic reasoning (taking into account the other agent's reasoning strategy) allows for better and more reasonable language model generations

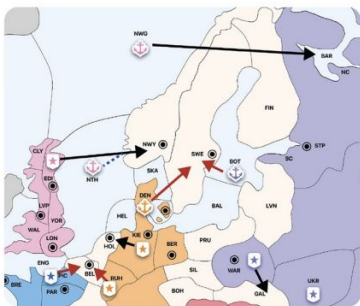
How does this help?

Depicted intents:

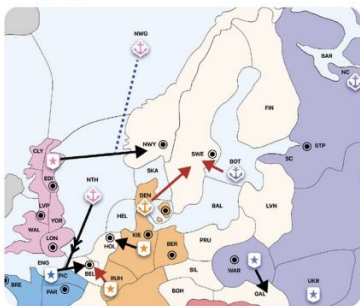
England conveys an army to Belgium with the support of France and Germany while taking Norway in a manner friendly to Russia.



England conveys an army to Norway and also aggressively moves to the Barents Sea against Russia.



England conveys an army to Norway and supports France to gain Belgium, preventing Germany from taking Belgium.



Can play multiplayer games with such models to a reasonable degree of proficiency

Example generations from the dialogue model with these actions as the conditioned intent:

ENG → FRA Mind supporting Edi - Bel?

ENG → GER Do you wanna support my convoy to Bel? With Italy going aggressive France will fall quickly and we can make gains off of both Russia and France.

ENG → FRA What's the plan for Bel?

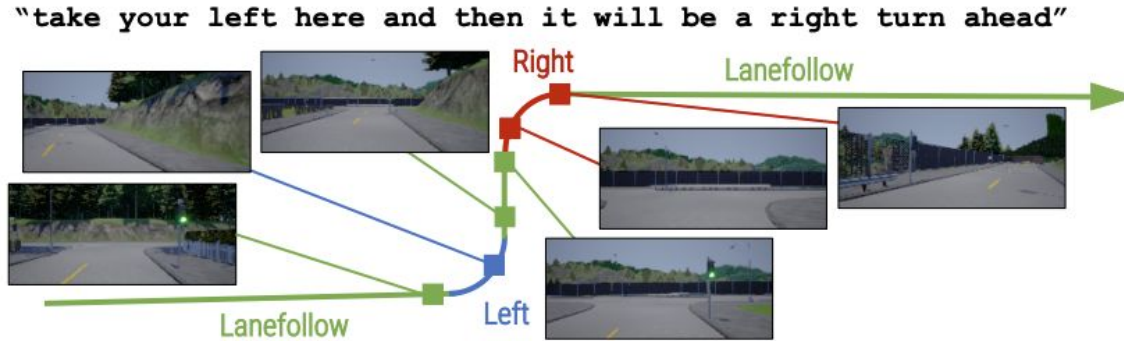
ENG → GER I will be conveying my army to Norway and moving to Barrens Sea.

ENG → RUS Hey, I moved to Edi to keep my options open. Does you have a problem with me

ENG → FRA Let me know if you want me to support you to Bel, otherwise I'll probably poke Hol.

ENG → GER Looks like you'll get three builds unless France bounces you! Are you gonna bounce Russia or not?

How does this help?



Advance in "conditional driving" applications

In summary

In summary

1. Lots of ***limitations of text-only language models*** that prevent them from being useful in many applications
2. There's also lots of speculation on whether LLMs ***can*** learn language correctly if they are not grounded
3. We've covered a few different ways of attempting to ground these language models to allow them to take in the necessary modalities (vision, worlds) to exist in the real world to solve tasks
4. Draws on lots of insights from cognitive science (theory of mind), computer vision (processing images), robotics and RL (representing the world) and multi-agent RL (representing agents)