Learning Language by Grounding Language

Karl Moritz Hermann
DeepMind

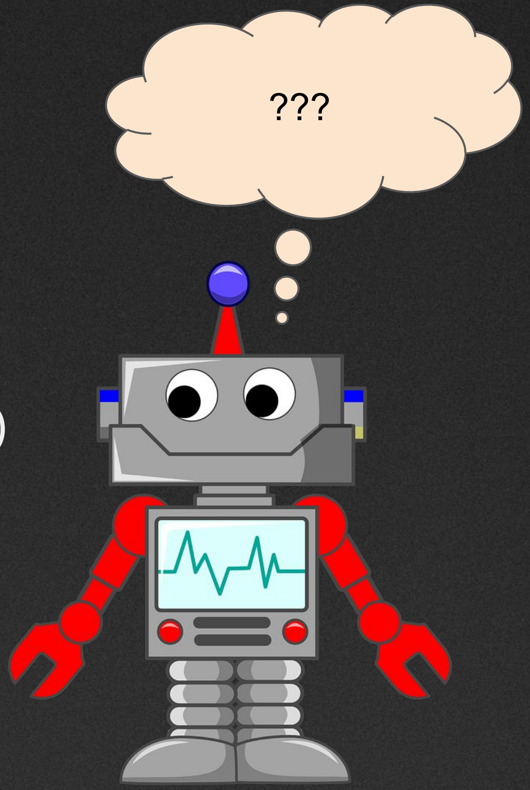
How do machines learn language?

Labeled data (eg. sentiment analysis, classification)

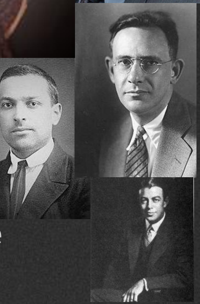
Language corpora (eg. MT, word embeddings)

Parsing / knowledge graphs (eg. QA, Penn Treebank)

Image and text pairs (eg. classification, caption generation)



How do we learn language?



Skinner: Behaviourist, **reinforcement** and imitation; operator conditioning (trial and error)

Chomsky: Skinner is wrong. Poverty of stimulus, **innate ability**

Snow / Bruner: **Social Interactionist**, language acquisition in the context of parent-child interaction

Bates / MacWhinney: Chomsky wrong. Emergentist view, acquisition through **cognitive process competition**

Piaget: **Cognitive theory**, four stages of development, symbolic reasoning → language acquisition

Tomasello: Functional theory of language acquisition, **shared understanding of intention**, structure from usage

BF Skinner, Noam Chomsky, Catherine Snow, Jerome Bruner, Brian MacWhinney, Elizabeth Bates, Jean Piaget, Anne Fernald, Michael Tomasello, Annick De Houwer, Kim Plunkett, Lev Vygotsky, Edward Sapir, Benjamin Lee Whorf, ...

"Humans learn language by inferring statistics from a billion tokens worth of text."

- Nobody, ever

No problem!

Machines are doing just fine.

- Most NLP systems are trained on tons of text data only.
- Large progress in machine translation, question answering, language modelling, information retrieval, you name it …
- So, maybe our approach is not too bad?

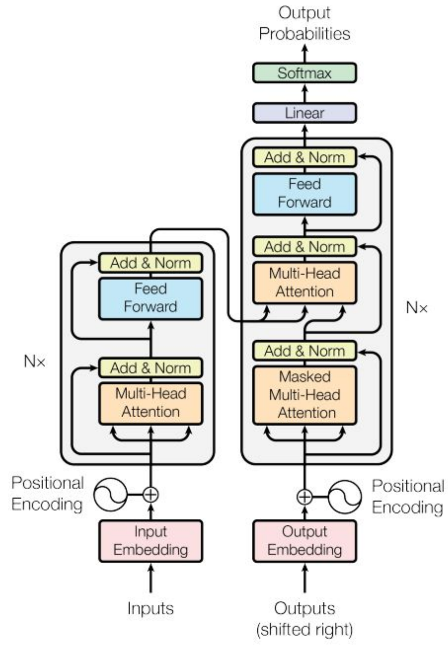
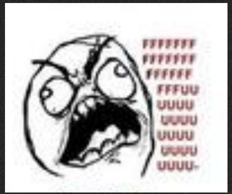


Figure 1: The Transformer - model architecture.

No, problem!



"A young woman in front of an old man"

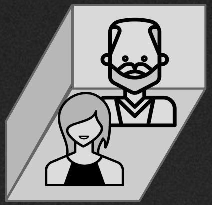
Which of these is the most similar in meaning?

1) A young man in front of an old woman
   MV- / RA / Pro / Sum cosine
2) An old woman in front of a young man
   RAE / Syntactic / Syntactic sum cosine
3) A young woman behind an old man
   MV-RNN / Syntacti / Prod regression
4) An old man behind a young woman
   All humans say this

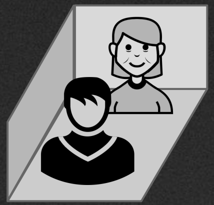
Gershman & Tenenbaum, 2015

Why do humans and machines disagree on this?

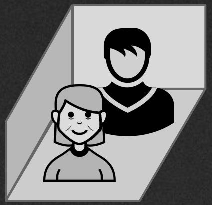
"A young woman in front of an old man"



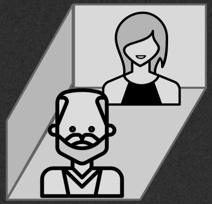
1) A young man in front of an old woman



2) An old woman in front of a young man



3) A young woman behind an old man



4) An old man behind a young woman

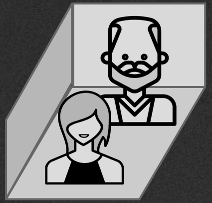


Image Credit - arif fajar yulianto, dDara, Creative Stall, Royyan Wijaya

How can we fix this?

Of course this is just one problem. There are many things that are difficult to learn for machines these days, such as:

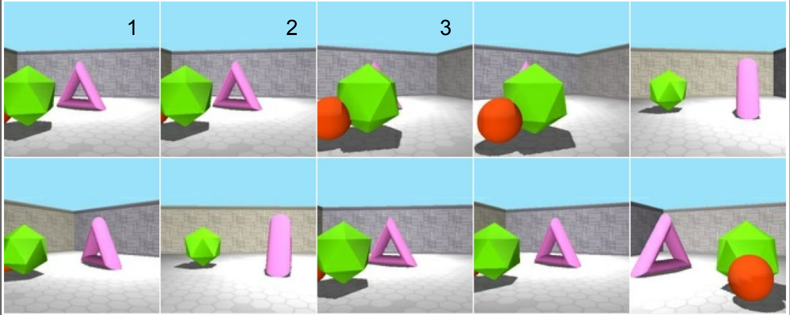
- Spatial reasoning
- Resolving ambiguity
- Coherence
- Forms of humour
- ...

Grounding language learning in other types of information should allow us to learn better semantics.

Grounding for the spatial reference problem

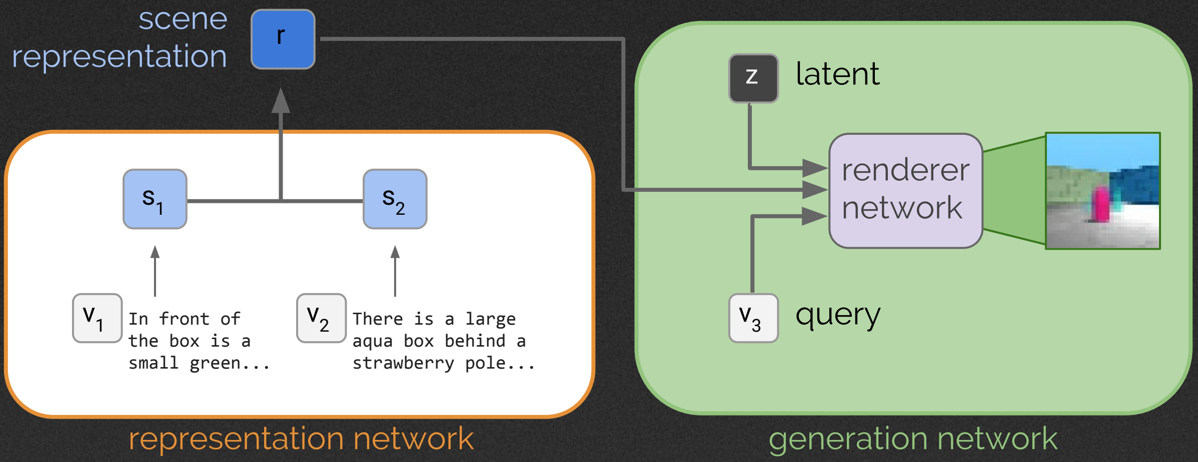
We built a large corpus of 3D scenes ...

- Multiple objects per scene
- Randomly placed
- Synthetic descriptions
- Natural language descriptions



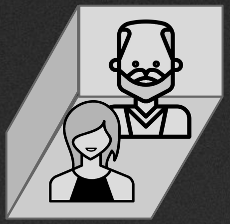
1: (NL) To the far left is a big green hexigon. The the right and back of that is a taller pink triangle.
(SYN) There is a pink torus behind a lime icosahedron.
2: (NL) On the left side forefront is a medium sized green three dimensional hexagon. Another object is to the left and is slightly visible. Moving straight back towards the right wall is a medium sized pink hollow triangle.
(SYN) There is a pink torus behind a lime icosahedron.
3: (NL) There is a red ball in front of a pink triangle. In the middle of them is a green diamond.
(SYN) There is a red sphere to the left of a pink torus. The sphere is in front of the torus. There is a red sphere to the left of a lime icosahedron. There is a pink torus behind a lime icosahedron

… trained with a bi-modal encoder-decoder setup …

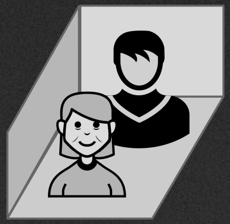


… and repeated the experiments on that model.

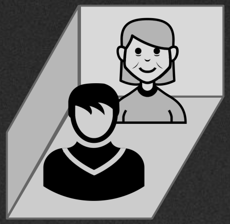
Base Scene



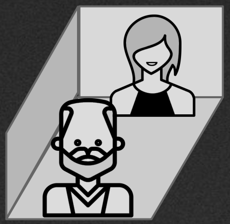
Adj Change



Noun Change



Prep Change



Invariant / Meaning Preserving

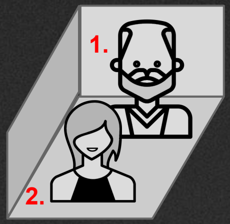
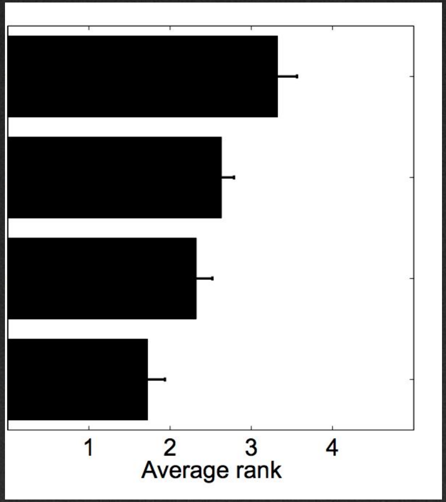


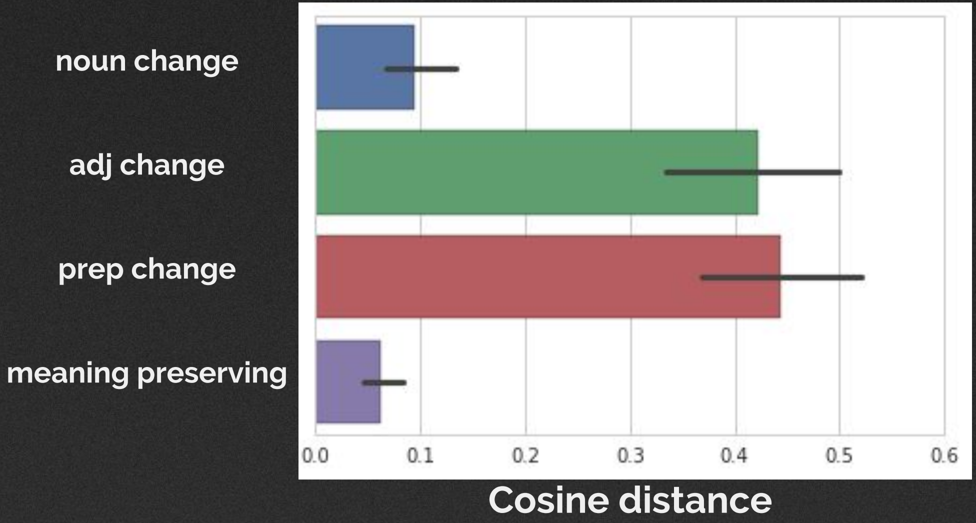
Image Credit - arif fajar yulianto, dDara, Creative Stall, Royyan Wijaya

Much better.

Human ranking



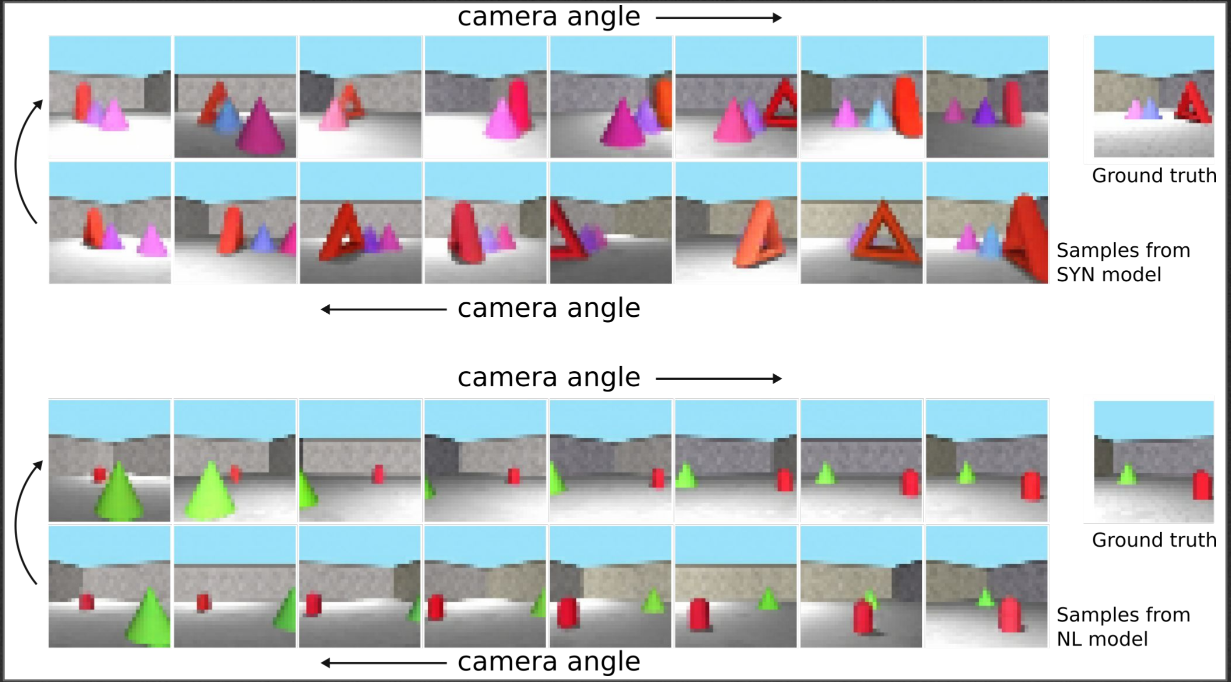
Model ranking



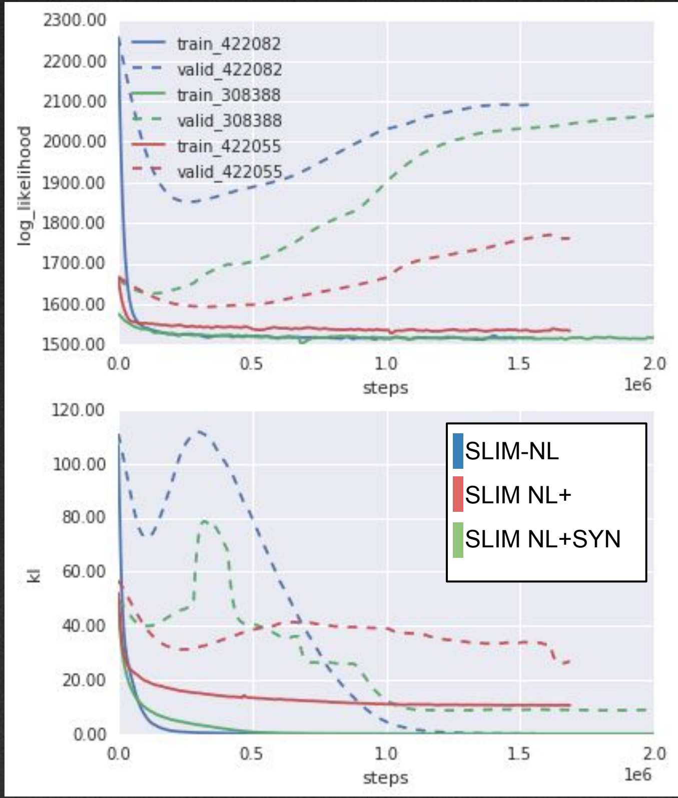
We do not match human rankings, but get the key bit - meaning preserving change - right!

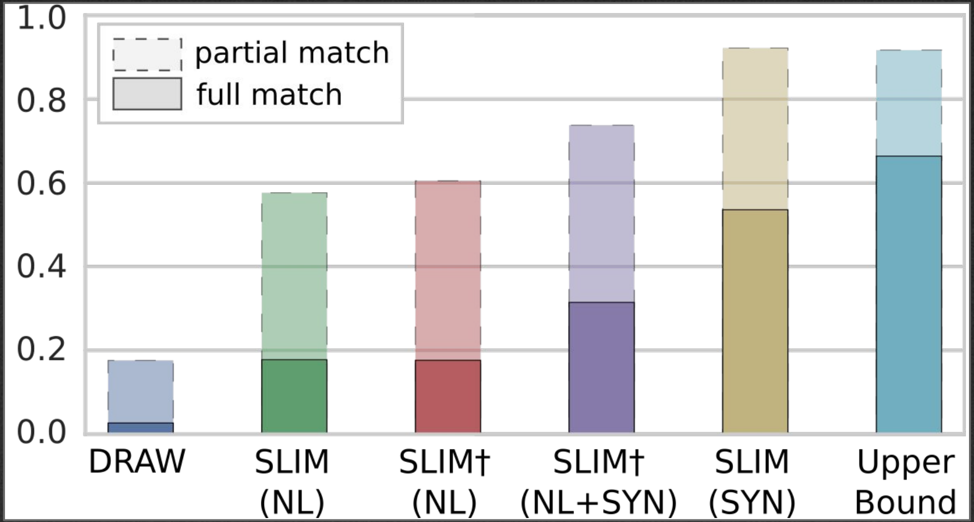
The model generates convincing scenes from text

Changing camera angle shows actual scene learning rather than just a flat image

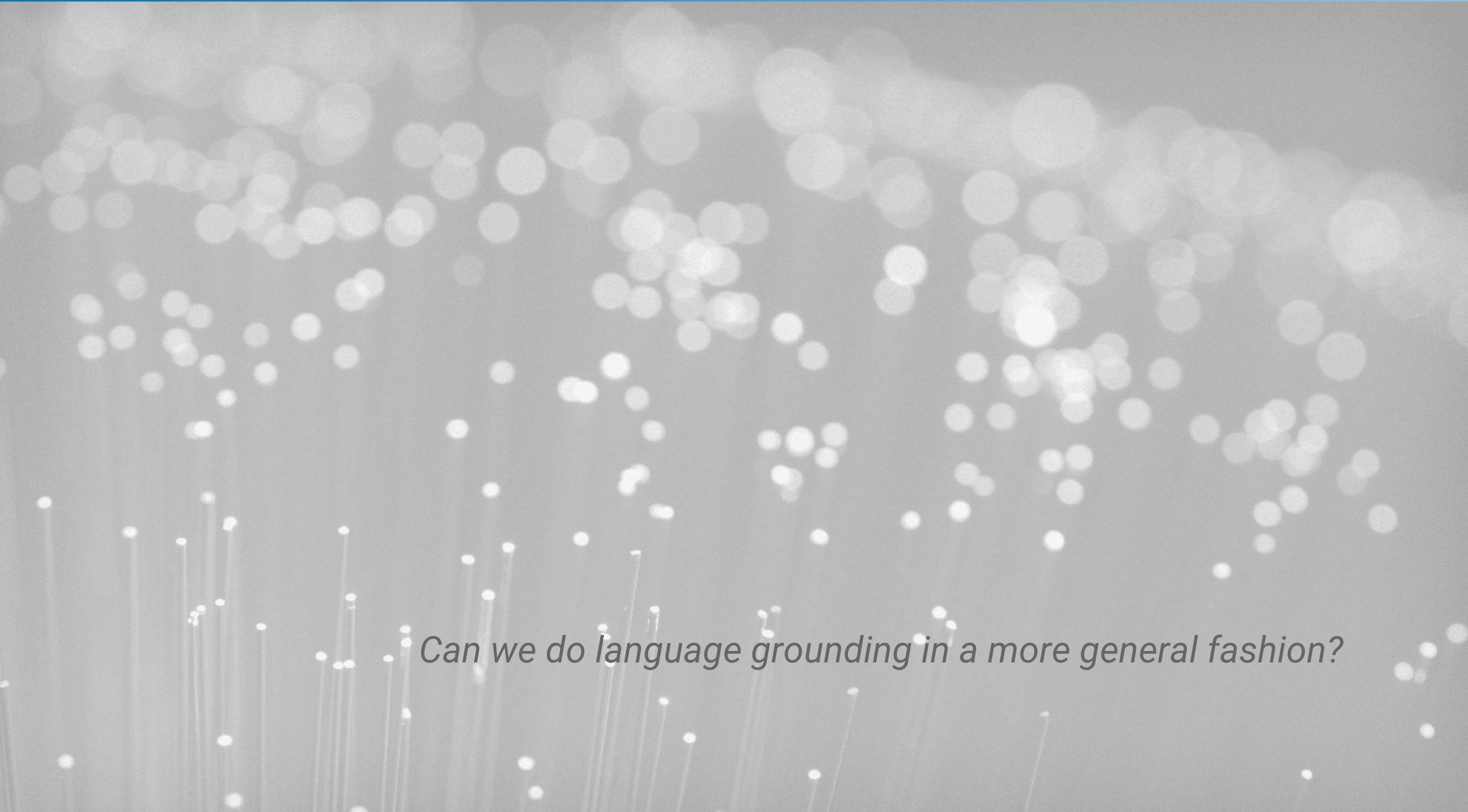


This works for synthetic and for natural language





Human accuracy evaluation

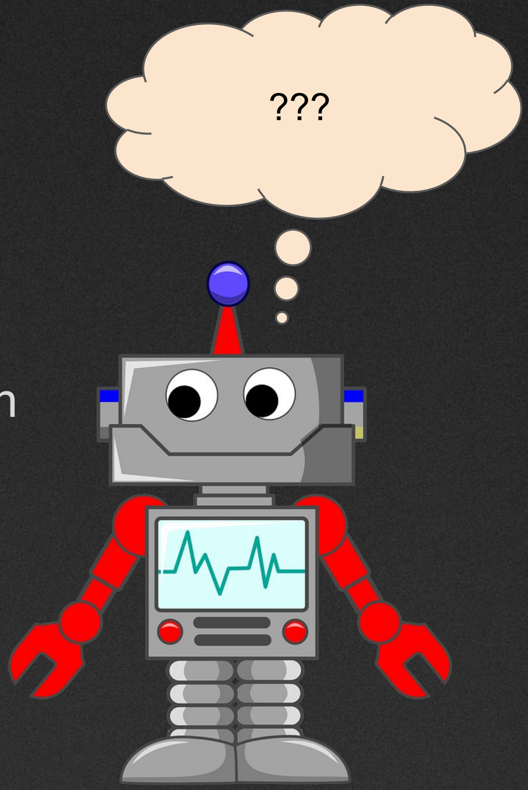

Can we do language grounding in a more general fashion?

The paradigm problem

What is a minimally **adequate** training paradigm for an intelligent agent to learn to comprehend language?

The previous model integrated language, vision and location (camera angle), but it could not act in the world.

For an agent to interpret the meaning of an utterance, linguistic symbols must be **grounded** in an environment in which the agent can act and learn.



The paradigm problem

All approaches to natural language **understanding** so far have failed. **Except us (humanity).**

Question: What is the (data) environment within which humans learn language?

There is no one answer to this question. But we know this:

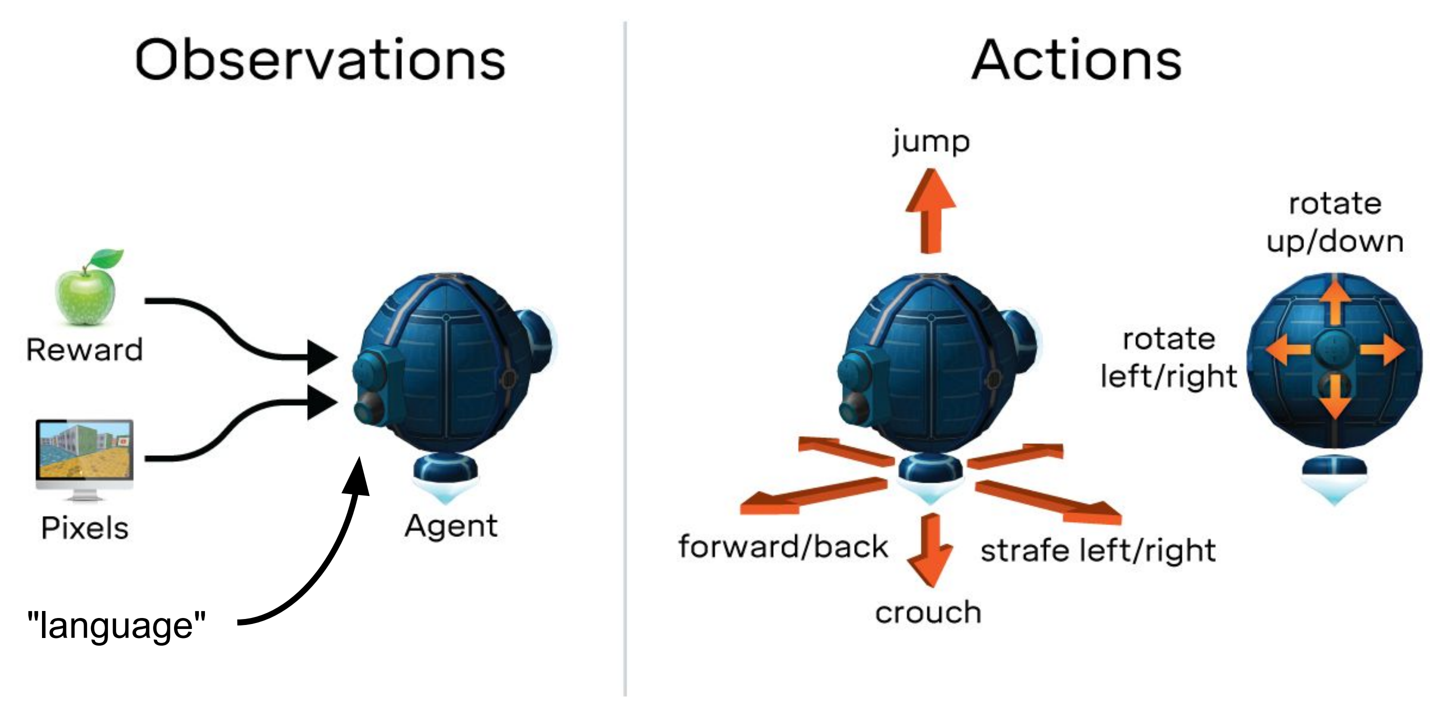
- Children learn language with **minimal direct teaching** and with incredible variations in data,
- they learn amazingly quickly from **sparse and ambiguous data**,
- children learn language in adverse circumstances, despite blindness, brain injury, or the inability to move or speak.

This would be great for artificial agents, too!

An agent that solves language tasks in simulation

- Solve billions of different (but closely related) tasks
- hundreds of fine-grained actions per task
- grounding: language -> solution space
- From scratch!

DeepMind Lab Environment



Beattie et al. DeepMind Lab. arXiv 2016. (https://github.com/deepmind/lab)

Language in DeepMind Lab: The Object Inventory

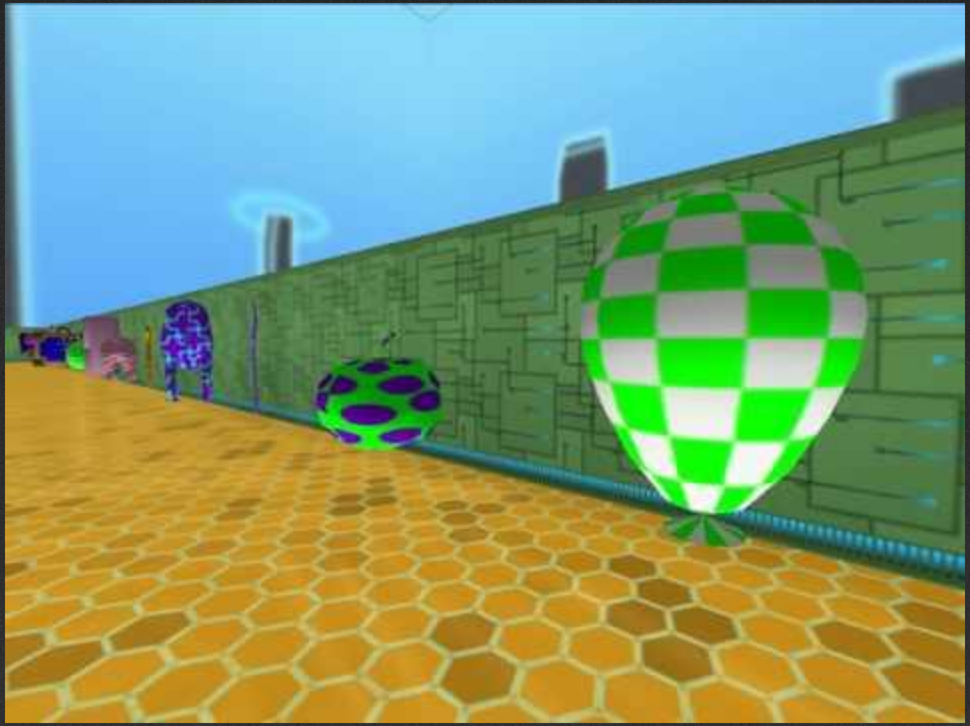
Shapes (40)
tv, ball, balloon, cake, can, cassette, chair, guitar, hairbrush, hat, ice lolly, ladder, mug, pencil, suitcase, toothbrush, key, bottle, car, cherries, fork, fridge, hammer, knife, spoon, apple, banana, cow, flower, jug, pig, pincer, plant, saxophone, shoe, tennis racket, tomato, tree, wine glass, zebra.

Colours (13)
red, blue, white, grey, cyan, pink, orange, black, green, magenta, brown, purple, yellow.

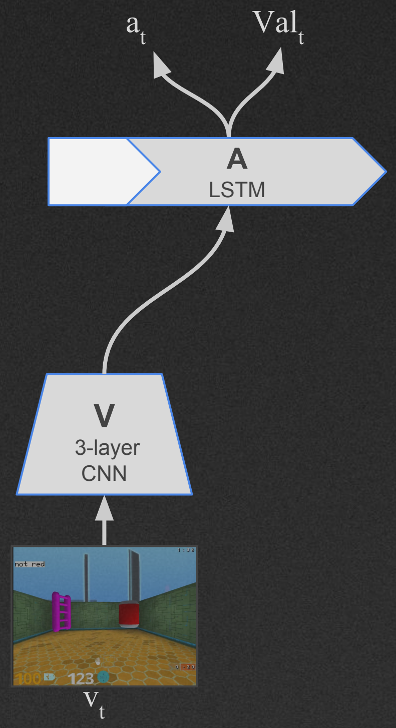
Patterns (9)
plain, chequered, crosses, stripes, discs, hex, pinstripe, spots, swirls.

Shades (3) light, dark, neutral.

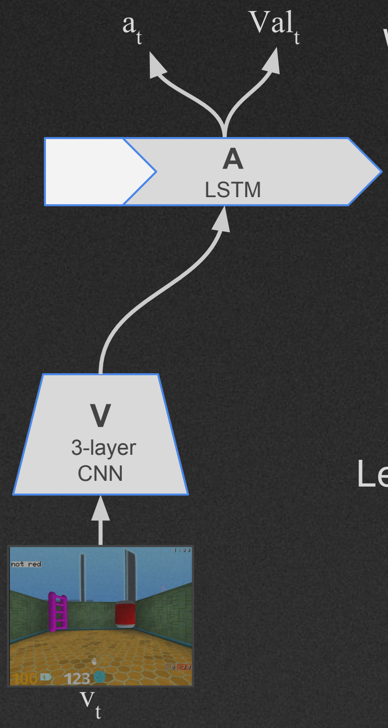
Sizes (3) small, large, medium.



Building on the thing that can learn to play 2D games



Building on the thing that can learn to play 2D games

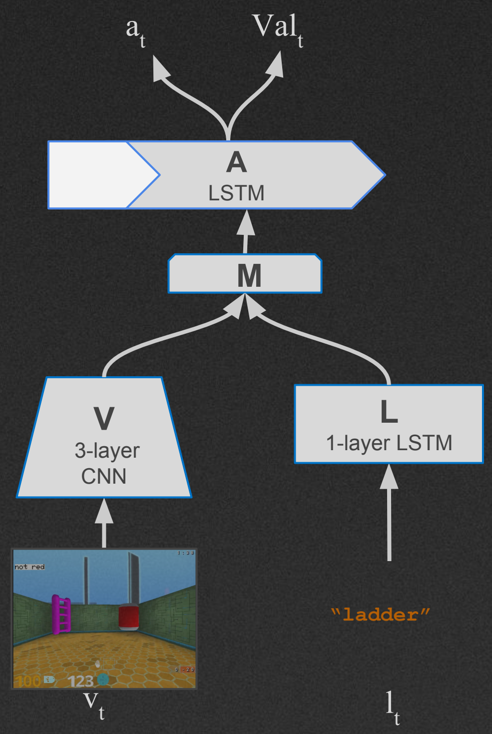


Williams, 1992

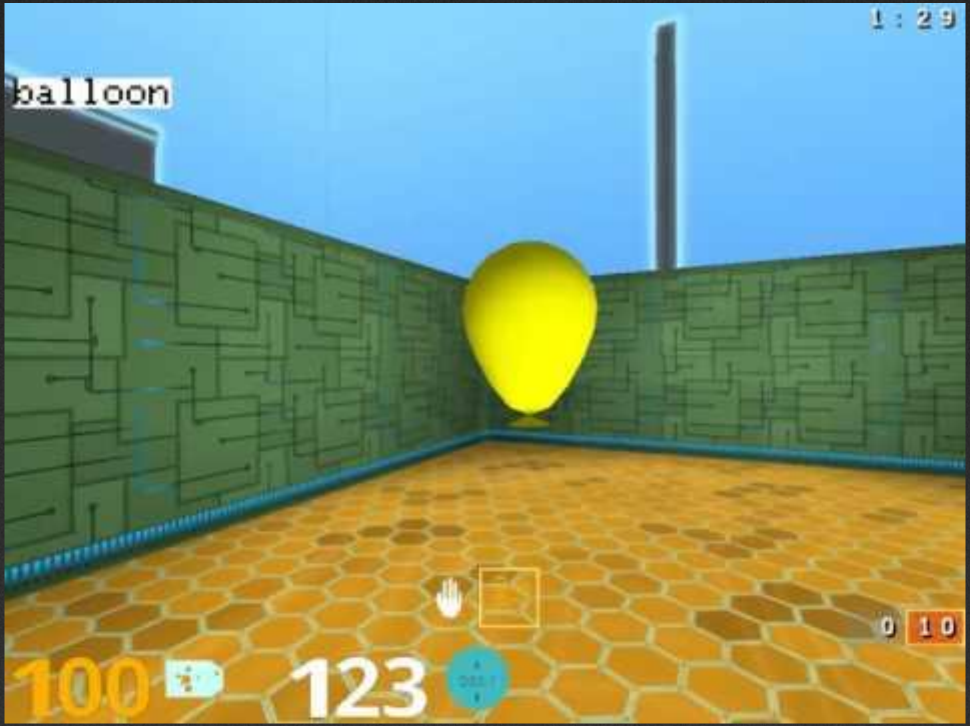
Elman 1991, Hochreiter 1997

Le Cun, 1988

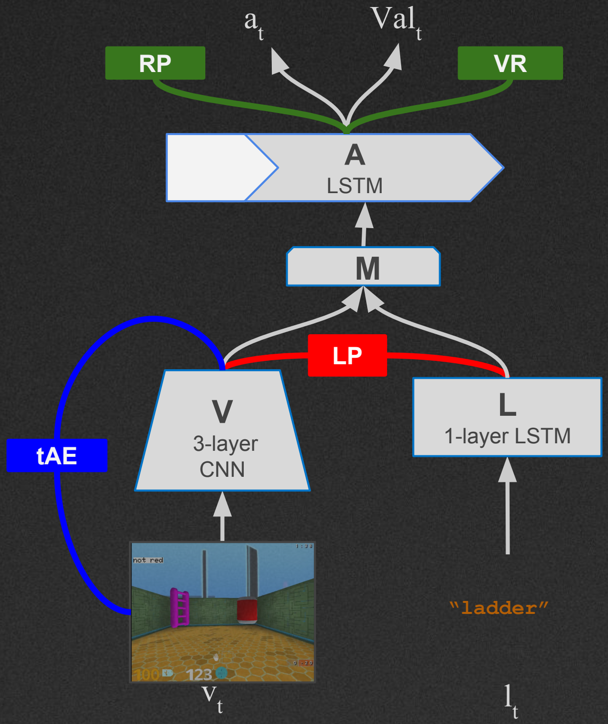
Building on the thing that can learn to play 2D games

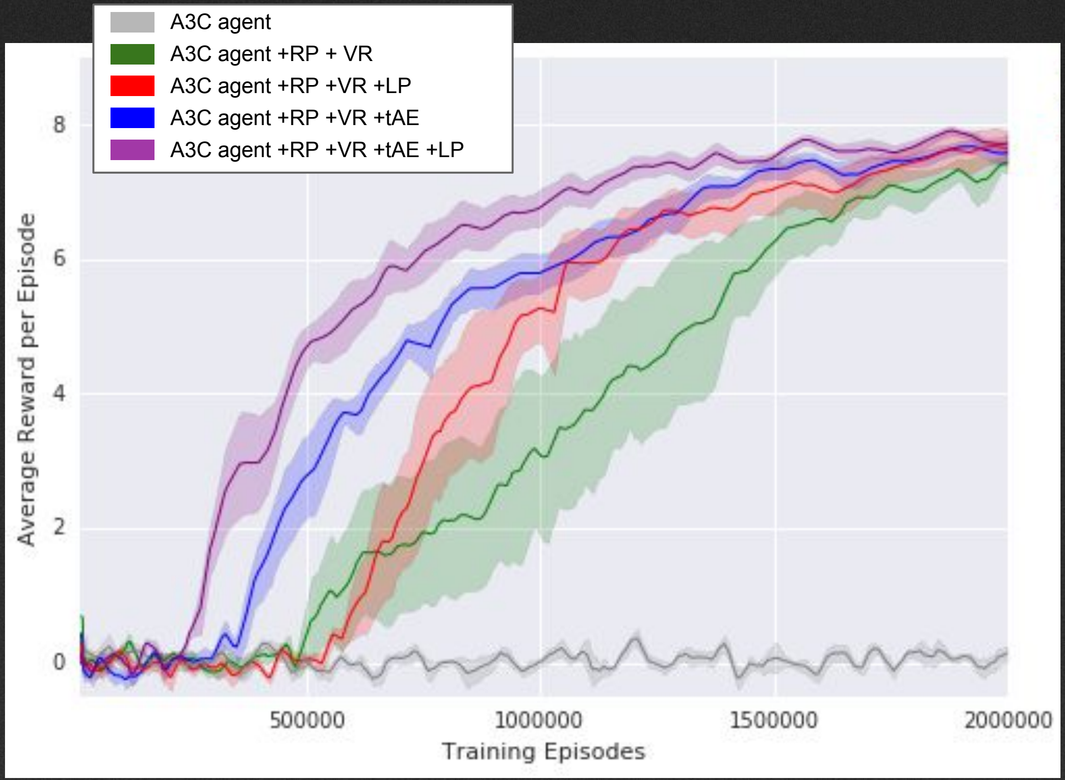


A simple test case: single words



Auxiliary objectives help agent learning





Language prediction provides 'interpretability'

Maybe you noticed....

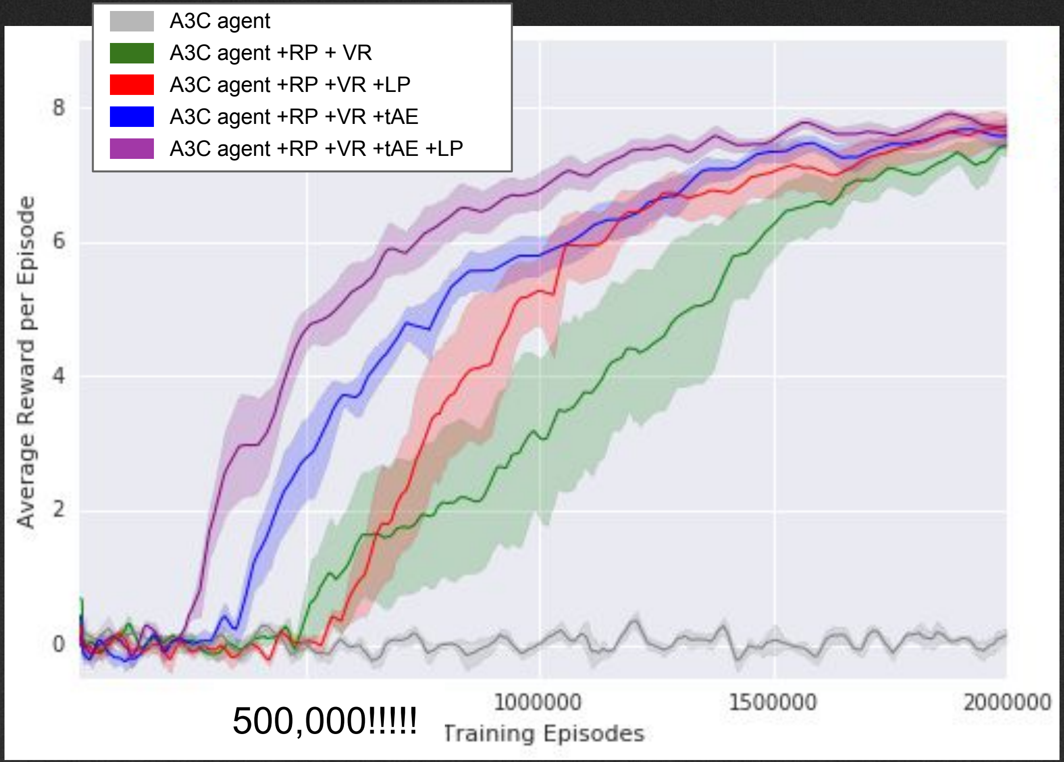
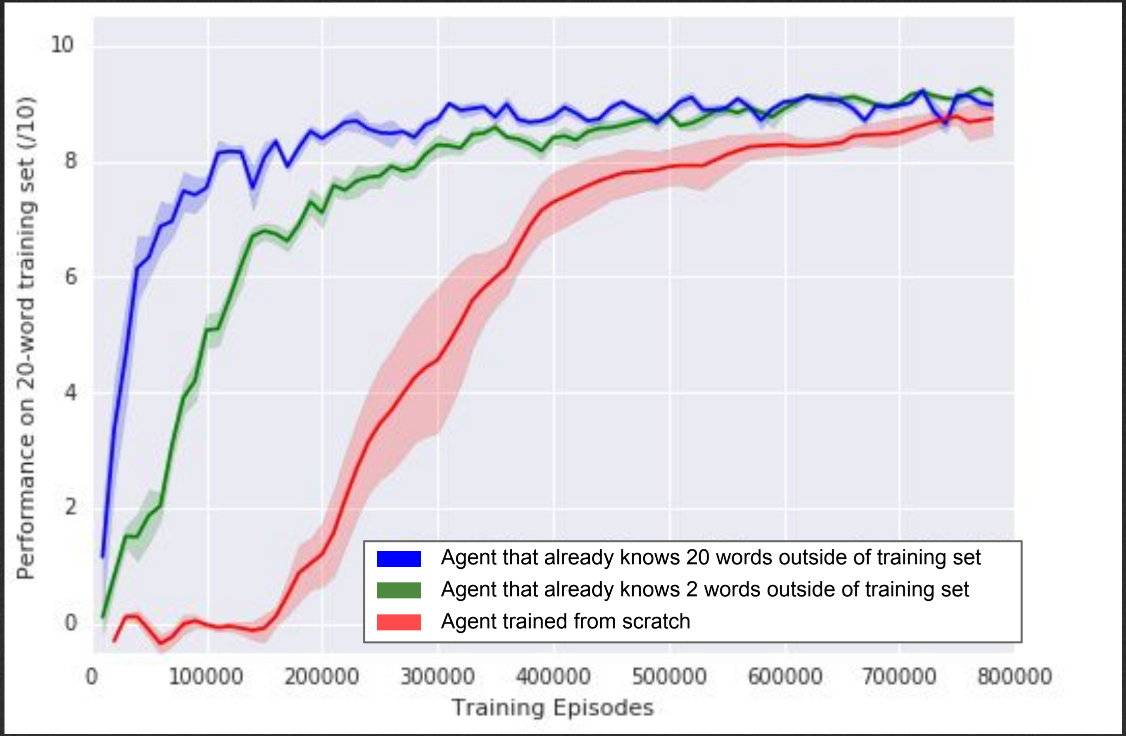
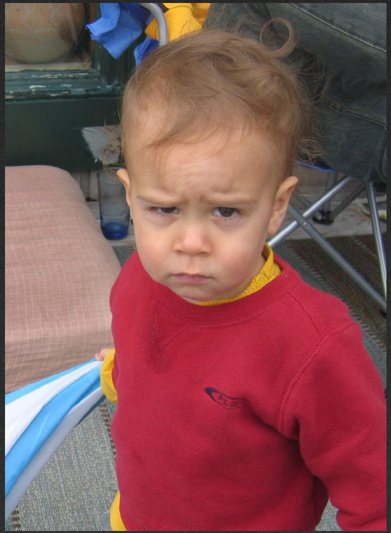


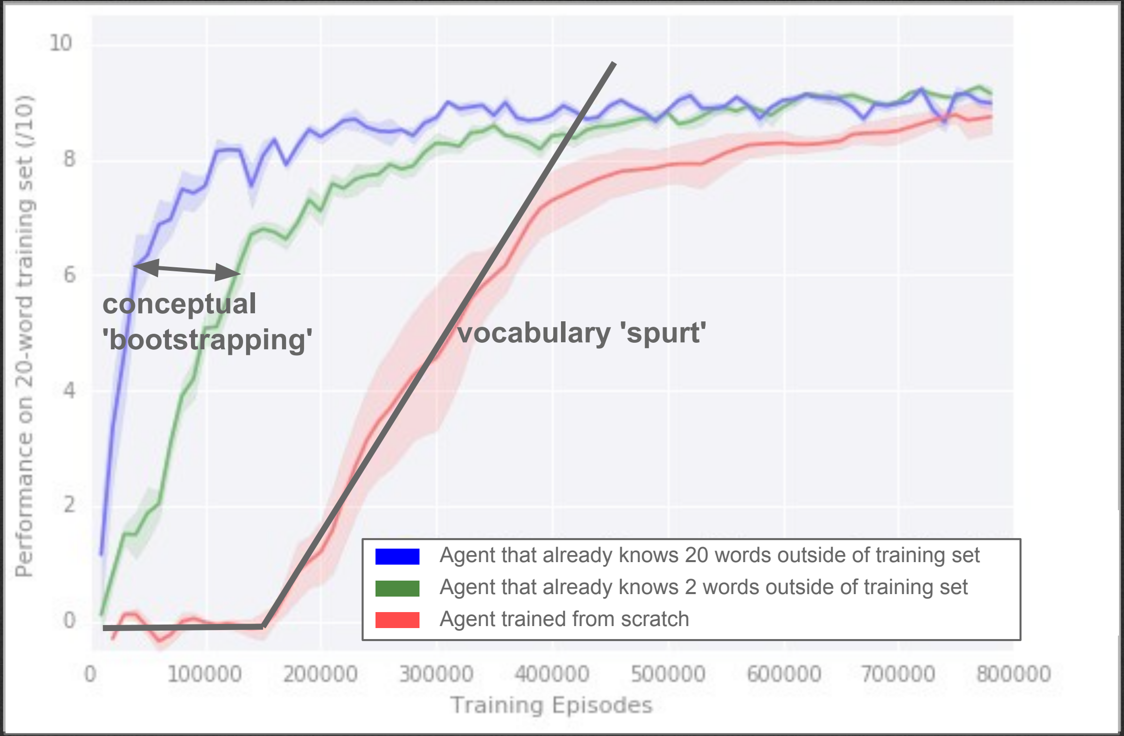
Image Credit - James Fenton

Knowing some words makes learning faster

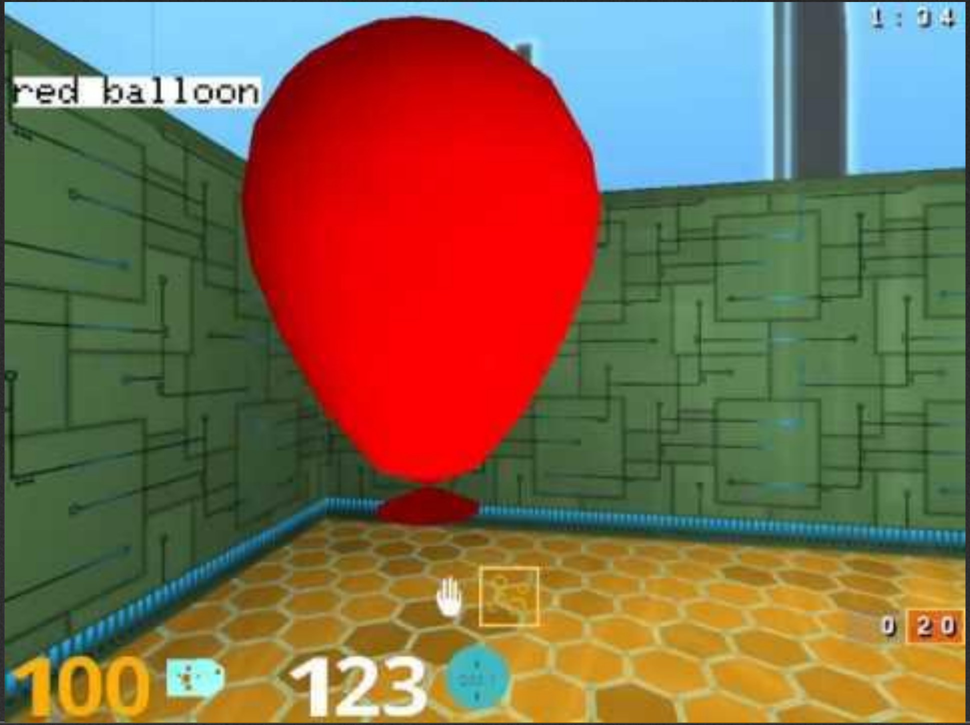


Much like little people

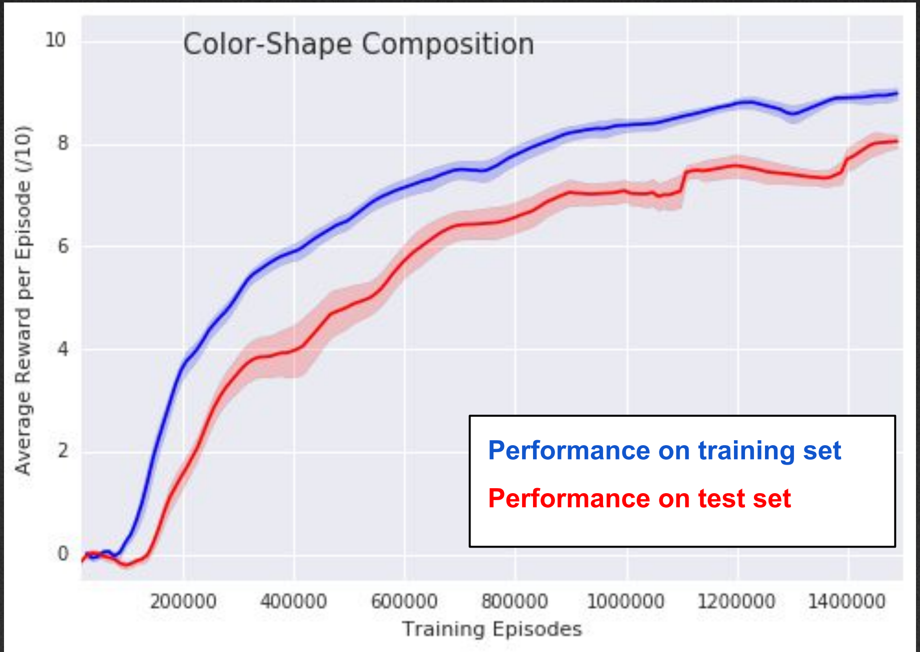




Moving towards longer sequences...



Agents naturally generalise word composition...



Training

"ladder"
"mug"
"pencil"
"suitcase"
"toothbrush"

x 40

"red"
"green"
"blue"
"pink"

x 13

"red ladder"
"green mug"
"blue pencil"

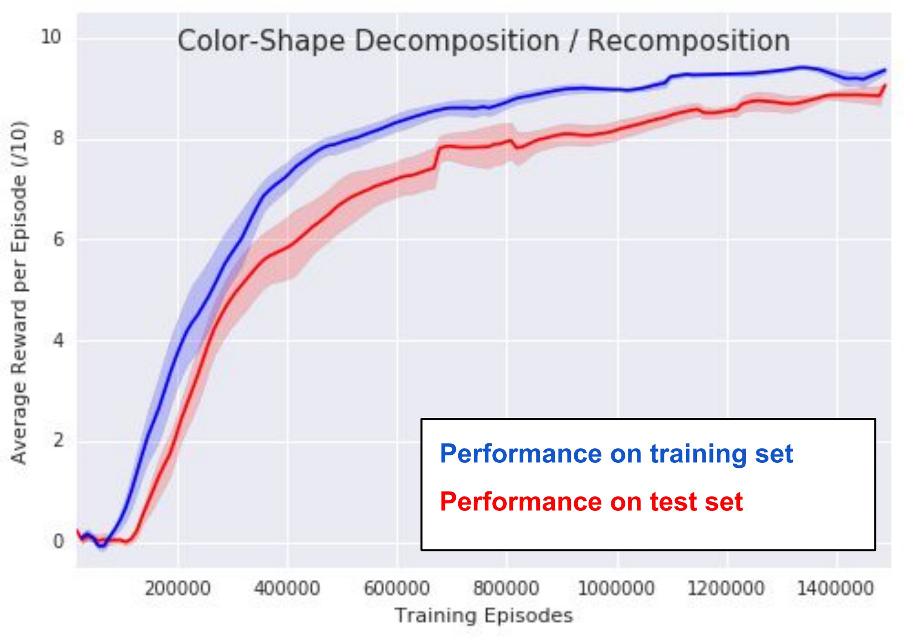
x 400

Test

"pink ladder"
"yellow mug"
"green pencil"

x 120

Decompose before re-compose



Training

"red ladder"
"**green** mug"
"blue pencil"

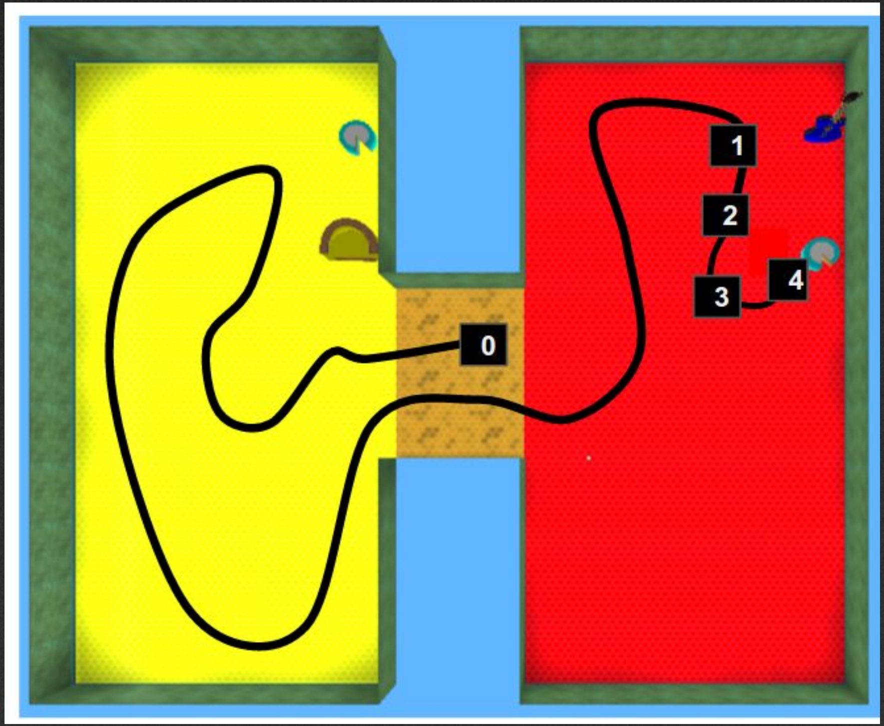
x 400

Test

"pink ladder"
"yellow mug"
"**green** pencil"

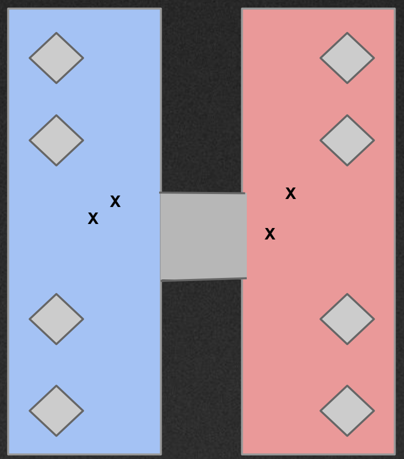
x 120

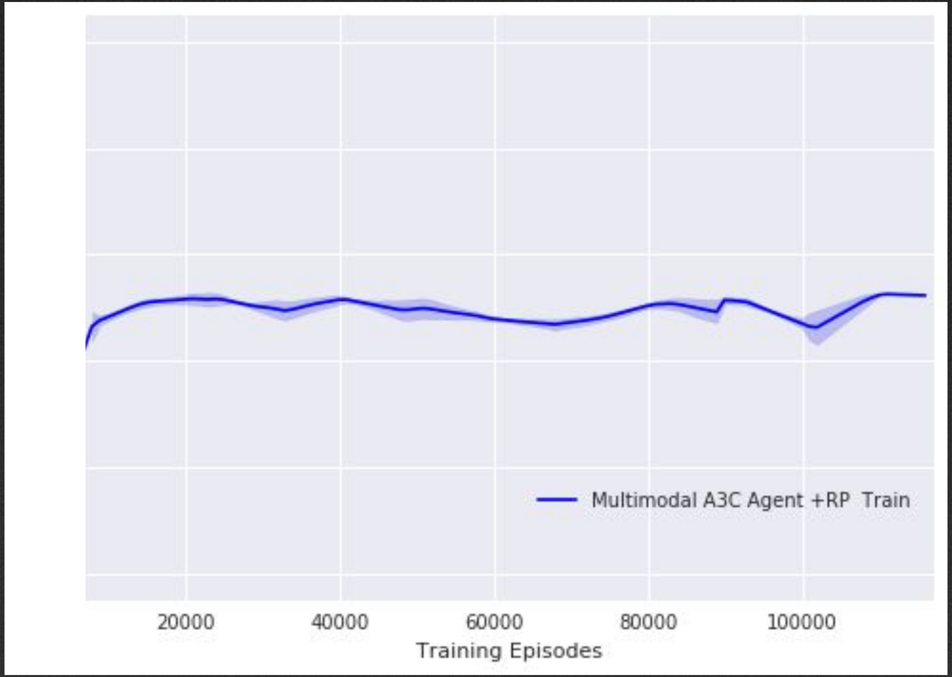
How far can we go...



Curriculum Learning for Complex Tasks

Top-down view of the level

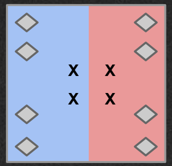




Curriculum Learning for Complex Tasks

single-room layout

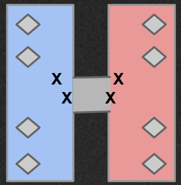
1



two object words and room descriptors

two room layout

2

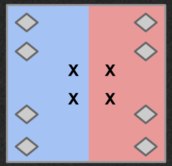


two object words and room descriptors

Curriculum Learning for Complex Tasks

single-room layout

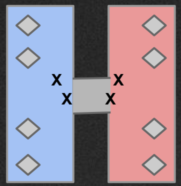
1



two object words and room descriptors

two room layout

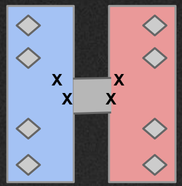
2



two object words and room descriptors

two room layout

3

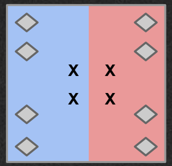


medium object word / room descriptor vocabulary

Curriculum Learning for Complex Tasks

single-room layout

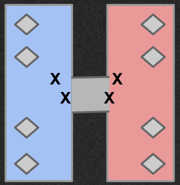
1



two object words and room descriptors

two room layout

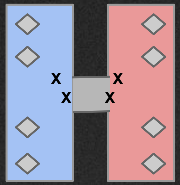
2



two object words and room descriptors

two room layout

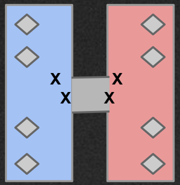
3



medium object word / room descriptor vocabulary

two room layout

4



full object word / room descriptor vocabulary

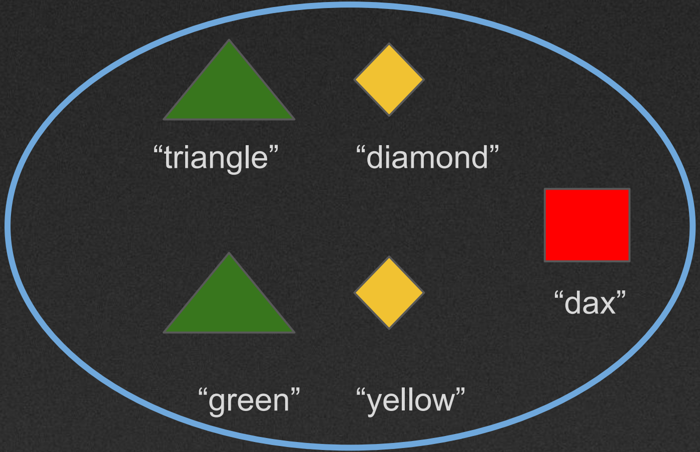

pick the chequered hair_brush
0
4 0

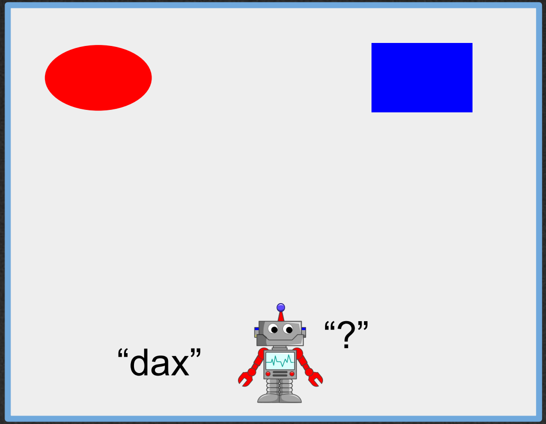
How can we understand (and learn from) this model…

Does this have anything to do with human language learning?

'Shape bias' helps humans to resolve reference

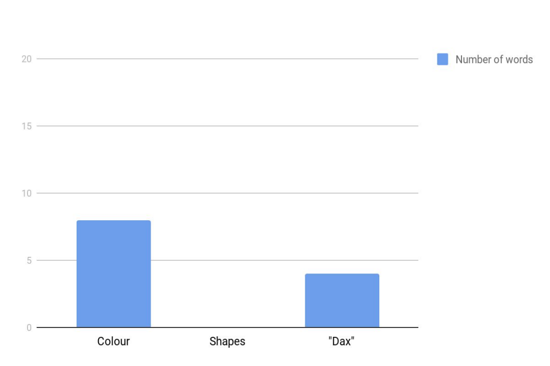
example training **example test**

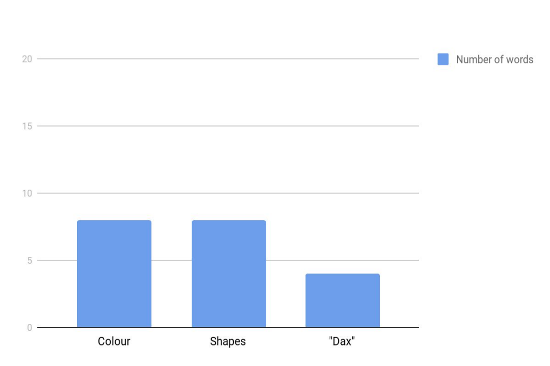


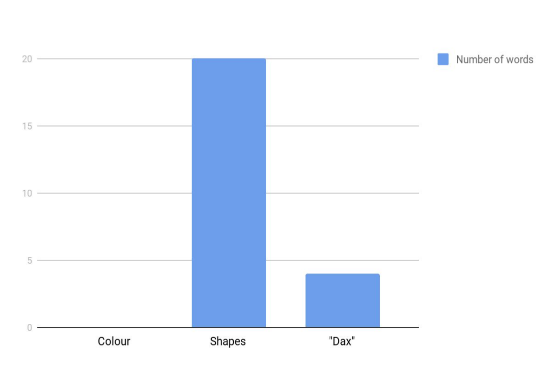


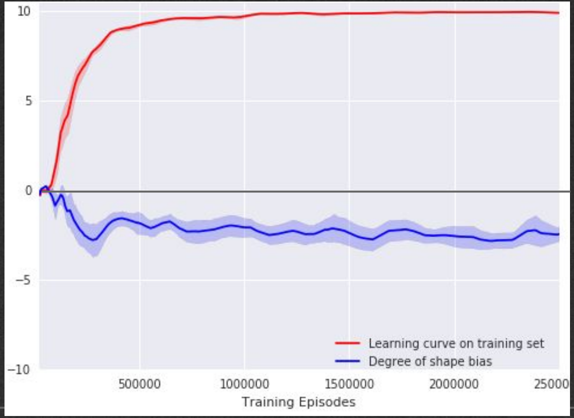
B Landau et al. 1988

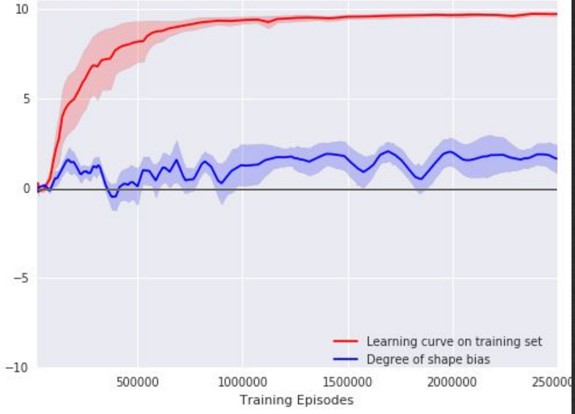
Agent learns bias from the training distribution



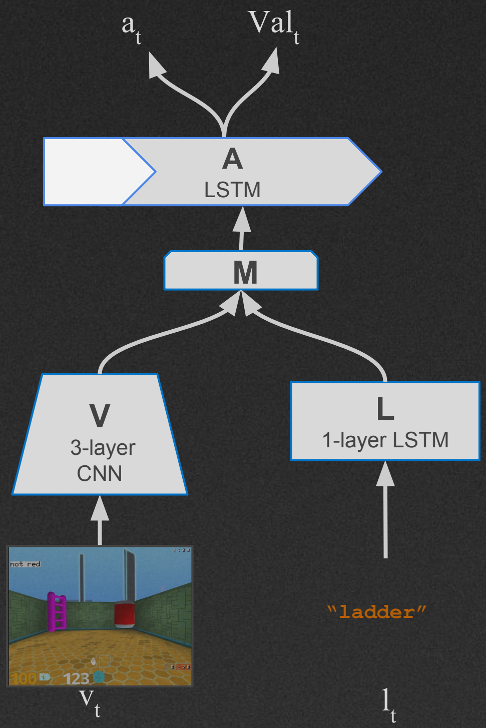




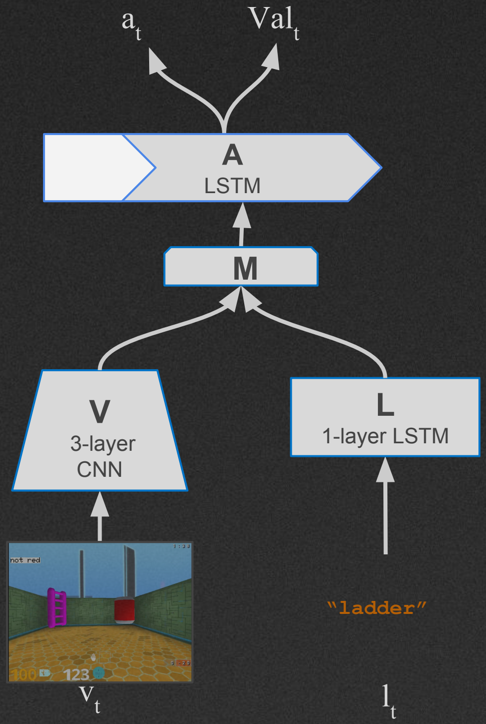


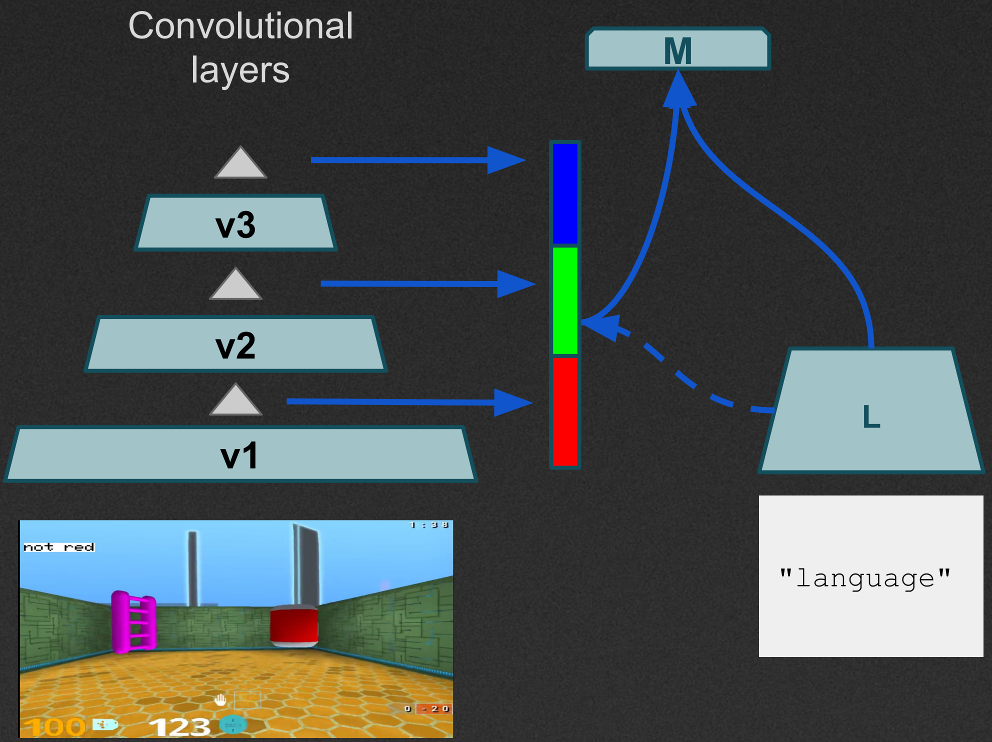


Layerwise attention

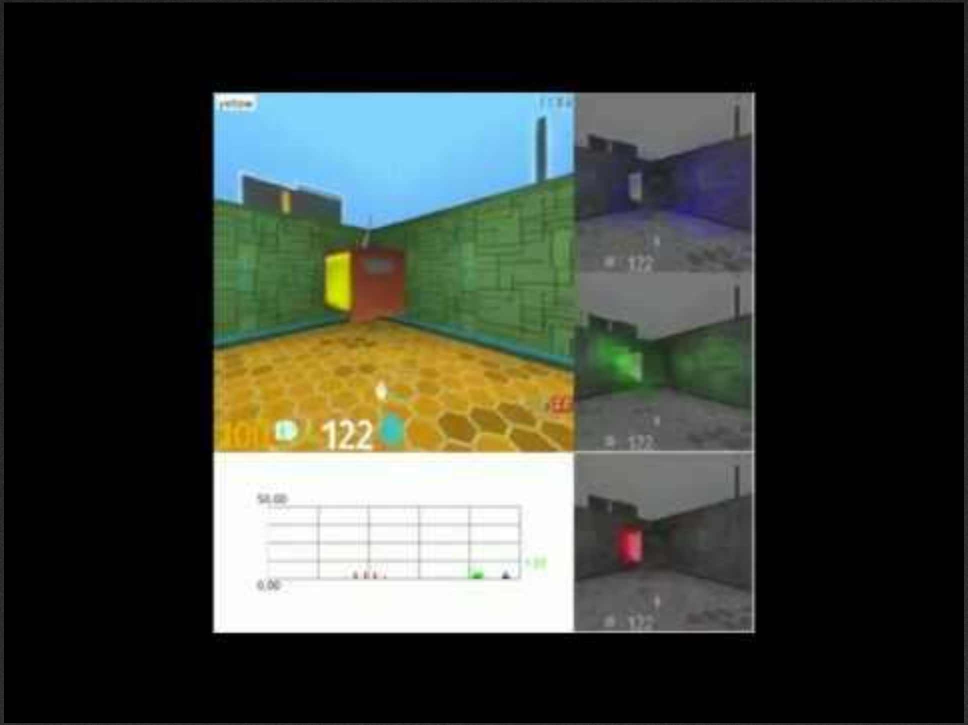


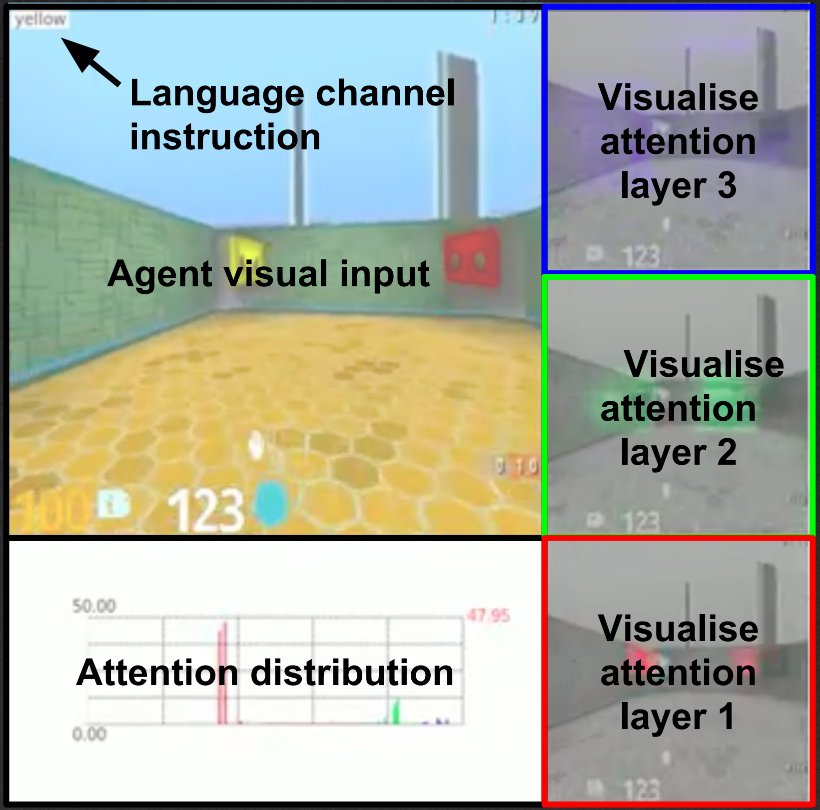
Layerwise attention



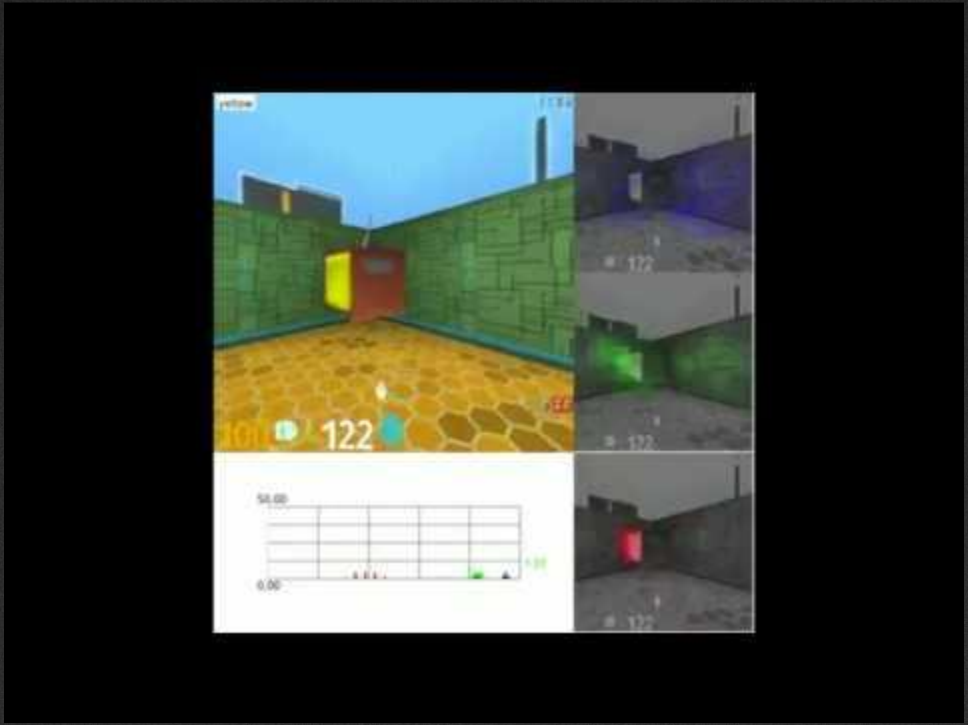


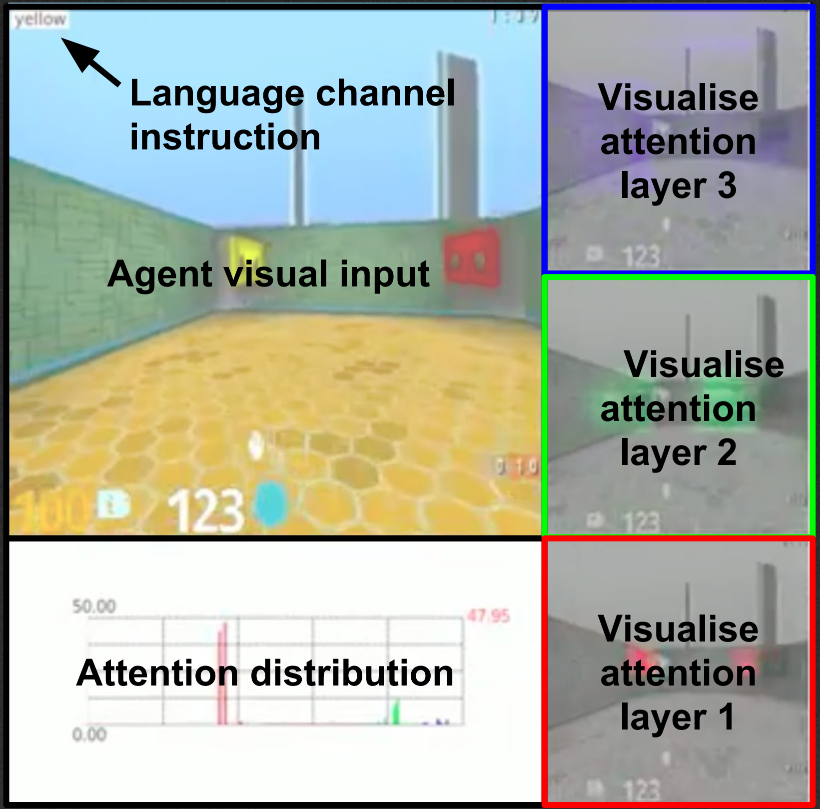
Processing colour words





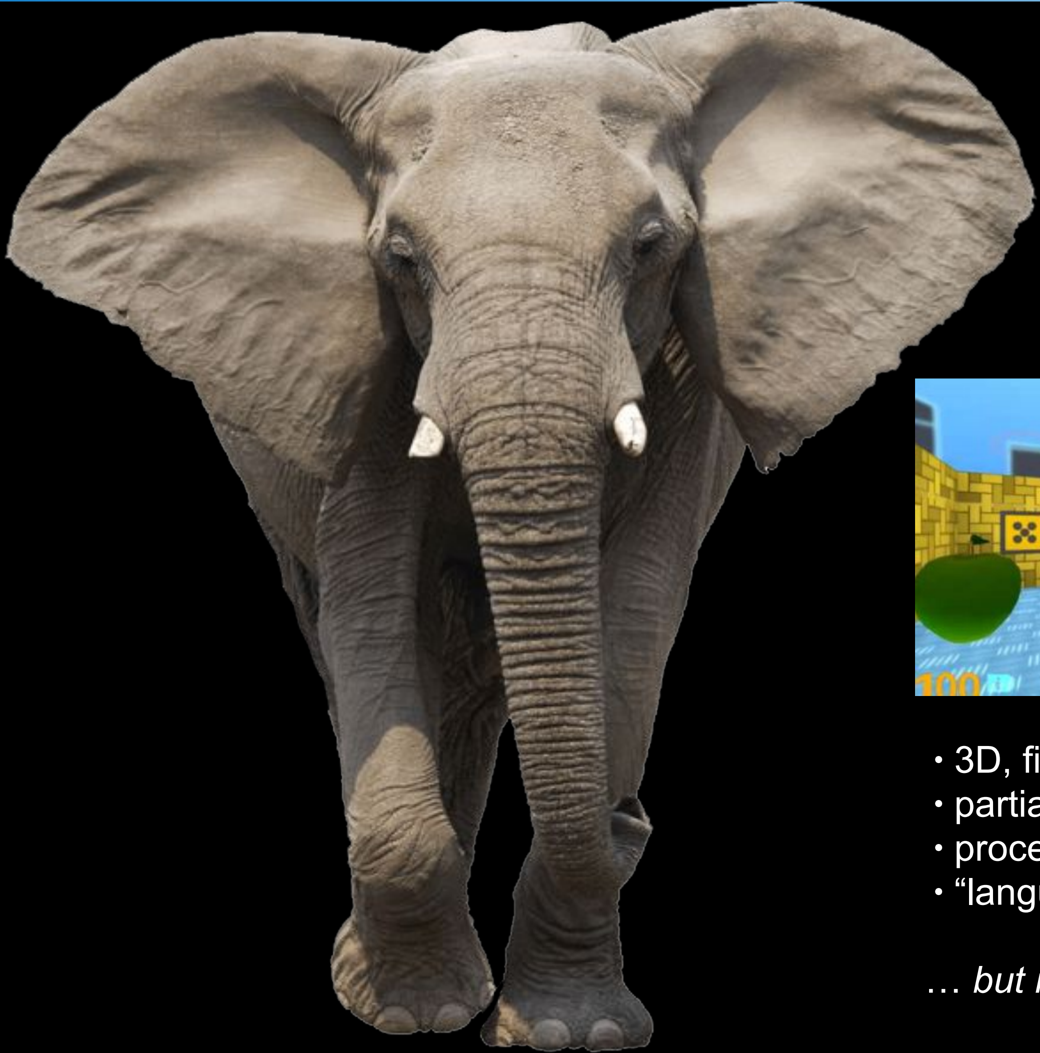
Processing shape words





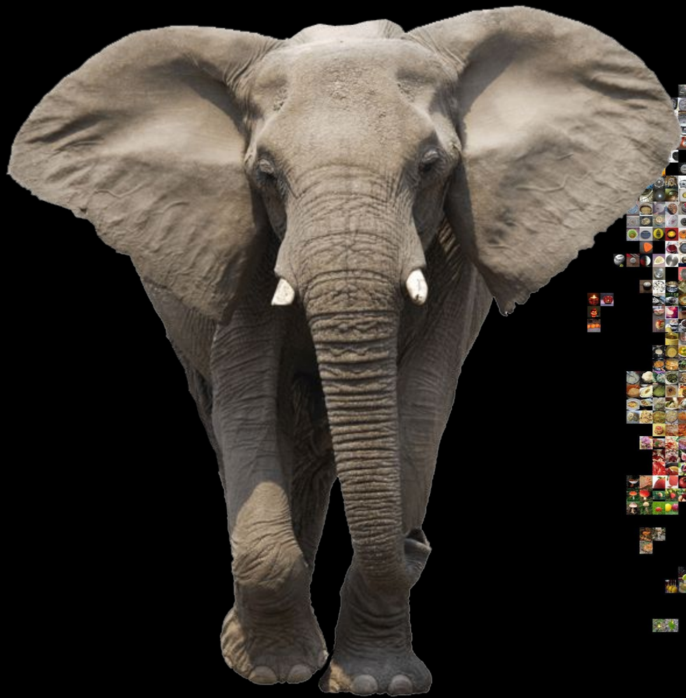
Summary

- We can learn to ground language in perception and actions
- Advantages of learning everything from scratch:
    - Avoid fallibility of human intuition
    - Generalisable representations and transferable knowledge
    - Accelerating learning
- Disadvantages:
    - Slow
    - Capacity / forgetting
- Let's understand these models by testing them like we test humans



- 3D, first person environment
- partially observed
- procedural variations
- “language” input

… *but it’s not real*

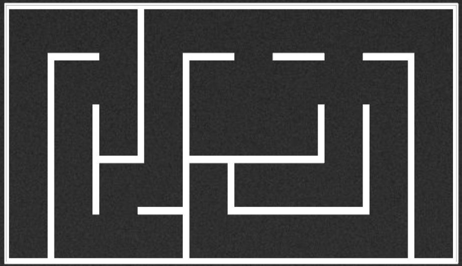



100
117

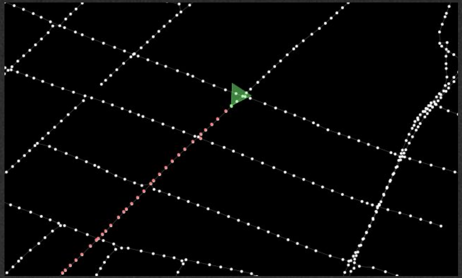
Can we ground language in the real world?

observation

observation



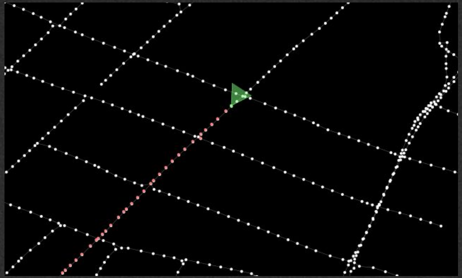
structure



structure

We can use StreetView as an RL environment ...

streetview images



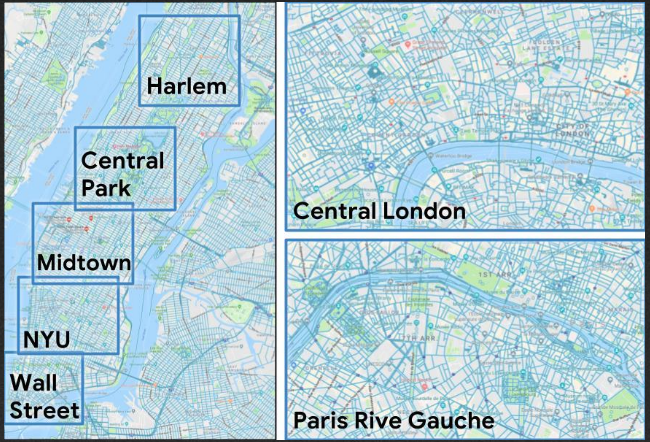
google maps

- RGB image cropped from panorama (84x84)

Actions: move to next node, rotate view 20° or 60°

Learning to navigate in cities without a map
Piotr Mirowski et al., arXiv 2018

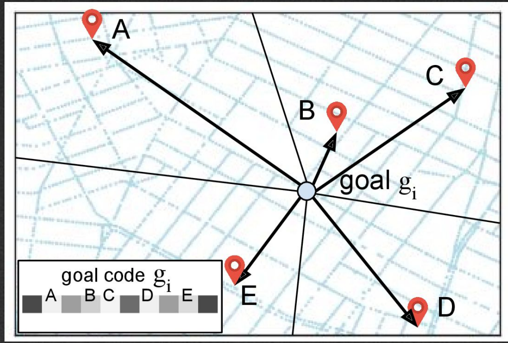
… grounding learning in the real world.



- 14,000 to 60,000 nodes (panoramas) per city, covering range of 3-5km per city
- Discrete action space allows rotating in place and stepping to next node
- Multi-city dataset and RL environment will be released later this year

Without language this works: The Courier Task

- **Random start/end** navigation without a map
- Reward when close to goal
- Actions: rotate left, right, or step forward
- Inputs for the agent at every time point *t*:
  - 84x84 RGB **image observations**
  - landmark-based **goal description**




Dior

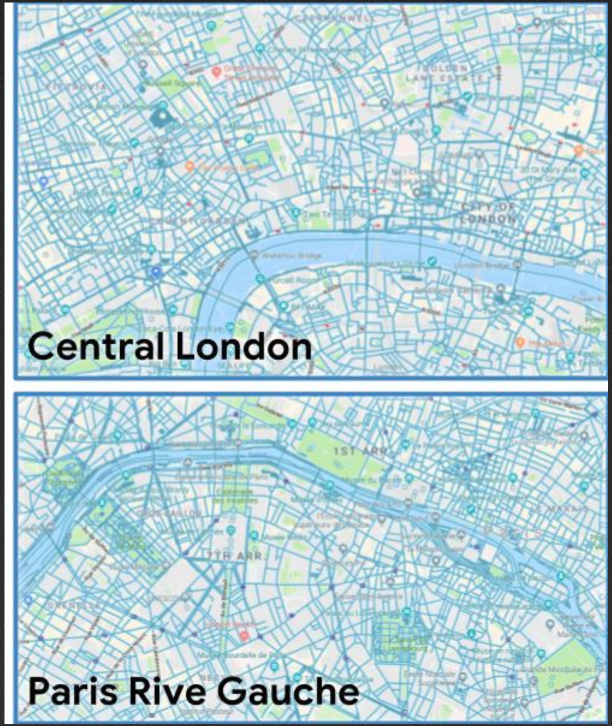
Can we also use this to learn language?

Short answer: **Yes**

Long answer:

We could construct language-based tasks in this environment, using information from e.g.

- Driving directions
- Tourist guidebooks and information
- Wikipedia
- ...?



This should be the next frontier for

grounded language learning

(if you want to get involved, e-mail me at kmh@google.com)

Thanks to my many collaborators!

Grounded language learning in a simulated 3D world & Understanding grounded language learning agents

Felix Hill, Simon Green, Fumin Wang, Ryan Faulkner, Hubert Soyer, David Szepesvari, Wojtek Czarnecki, Max Jaderberg, Denis Teplyashin, Stephen Clark, Marcus Wainwright, Chris Apps, Demis Hassabis, Phil Blunsom

Encoding Spatial Relations from Natural Language

Tomas Kocisky, Tiago Ramalho, Frederic Besse, Ali Eslami, Gabor Melis, Fabio Viola, Phil Blunsom

Learning to Navigate in Cities without a Map

Piotr Mirowski, Matthew Koichi Grimes, Keith Anderson, Denis Teplyashin, Mateusz Malinowski, Karen Simonyan, Koray Kavukcuoglu, Andrew Zisserman, Raia Hadsell