Flexible intelligence

Andrei Barbu


Boris Katz


Shimon Ullman

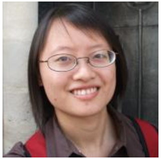

Josh Tenebaum


Gabriel Kreiman

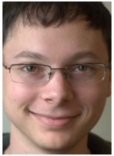

Andrei Barbu

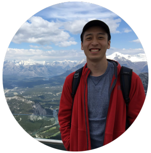

Ignacio Cases

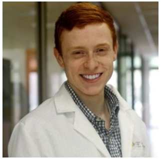

Candace Ross


Yen-Ling Kuo


Adam Yaari


David Mayo


Christopher Wang


Julian Alverio


Emily Cheng


Dylan Sleeper


Vighnesh Subramaniam


Aaditya Singh


Ravi Tejwani


Dana Rosenfarb

Flexible intelligence

Flexible intelligence

Knew who the adult is talking to

Flexible intelligence

Knew who the adult is talking to

What “That’s your train” is referring to

Flexible intelligence

Knew who the adult is talking to

What “That’s your train” is referring to

What trains do on tracks

Flexible intelligence

Knew who the adult is talking to

What "That's your train" is referring to

What trains do on tracks

Tried to put trains on tracks

Flexible intelligence

Knew who the adult is talking to

What “That’s your train” is referring to

What trains do on tracks

Tried to put trains on tracks

Knew what the adult was referring to when they waved their hand

Flexible intelligence

Knew who the adult is talking to

What “That’s your train” is referring to

What trains do on tracks

Tried to put trains on tracks

Knew what the adult was referring to when they waved their hand

Decided what object to get next

Flexible intelligence

Knew who the adult is talking to

What “That’s your train” is referring to

What trains do on tracks

Tried to put trains on tracks

Knew what the adult was referring to when they waved their hand

Decided what object to get next

Knew what Santa brought

Flexible intelligence

Knew who the adult is talking to

What “That’s your train” is referring to

What trains do on tracks

Tried to put trains on tracks

Knew what the adult was referring to when they waved their hand

Decided what object to get next

Knew what Santa brought

Answered questions about objects and abstract concepts

Flexible intelligence

Knew who the adult is talking to

What "That's your train" is referring to

What trains do on tracks

Tried to put trains on tracks

Knew what the adult was referring to when they waved their hand

Decided what object to get next

Knew what Santa brought

Answered questions about objects and abstract concepts

Demonstrated "it goes"

Flexible intelligence

Knew who the adult is talking to

What "That's your train" is referring to

What trains do on tracks

Tried to put trains on tracks

Knew what the adult was referring to when they waved their hand

Decided what object to get next

Knew what Santa brought

Answered questions about objects and abstract concepts

Demonstrated "it goes"

Understood that "doggie" and "dog" are similar

Flexible intelligence

Knew who the adult is talking to

What "That's your train" is referring to

What trains do on tracks

Tried to put trains on tracks

Knew what the adult was referring to when they waved their hand

Decided what object to get next

Knew what Santa brought

Answered questions about objects and abstract concepts

Demonstrated "it goes"

Understood that "doggie" and "dog" are similar

etc.

Machines on the other hand

Generated by GPT-2 large

Machines on the other hand

Generated by GPT-2 large

Q: Leader of Canada A: Trudeau
Q: Leader of Germany A: **Merkel**
Q: Leader of India A: Modi

Machines on the other hand

Generated by GPT-2 large

Q: Leader of Canada A: Trudeau
Q: Leader of Germany A: **Merkel**
Q: Leader of India A: Modi

Canada is a country that**'s become very strong over the last five or six years and has been at the heart of our growth over the last decade or so," he told host Andrew Coyne**

Machines on the other hand

Generated by GPT-2 large

Q: Leader of Canada A: Trudeau
Q: Leader of Germany A: **Merkel**
Q: Leader of India A: Modi

Canada is a country that**'s become very strong over the last five or six years and has been at the heart of our growth over the last decade or so," he told host Andrew Coyne**

I was on the glideslope and **looked down and there's a lot of wind and clouds and there's lightning coming from the west, so my flight path went off to the left in that direction.**

Get some things terribly wrong

GPT-2 large
The first 3 completions for

I was holding something heavy and my friend . . .

Get some things terribly wrong

GPT-2 large
The first 3 completions for

I was holding something heavy and my friend . . .
thought I was getting too close.

Get some things terribly wrong

GPT-2 large
The first 3 completions for

I was holding something heavy and my friend . . .

thought I was getting too close.
held my back so I wasn't able to grab anything.

Get some things terribly wrong

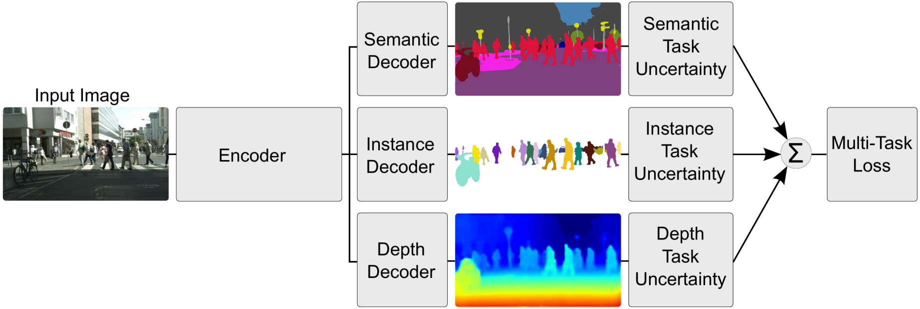
GPT-2 large
The first 3 completions for

I was holding something heavy and my friend . . .
thought I was getting too close.
held my back so I wasn't able to grab anything.
took it and hit me in the head,"

Multi-task Vision

Kendall *et al*. 2019

Multi-task Vision



Kendall *et al.* 2019

Recognition

Retrieval

Recognition

Retrieval

Generation

Recognition

Retrieval

Generation

Question answering

Recognition

Retrieval

Generation

Question answering

Disambiguation

Recognition

Retrieval

Generation

Question answering

Disambiguation

Language acquisition

Recognition

Retrieval

Generation

Question answering

Disambiguation

Language acquisition

Paraphrasing

Recognition

Retrieval

Generation

Question answering

Disambiguation

Language acquisition

Paraphrasing

Translation

Recognition

Retrieval

Generation

Question answering

Disambiguation

Language acquisition

Paraphrasing

Translation

Common sense reasoning

Recognition

Retrieval

Generation

Question answering

Disambiguation

Language acquisition

Paraphrasing

Translation

Common sense reasoning

Planning

Recognition

Retrieval

Generation

Question answering

Disambiguation

Language acquisition

Paraphrasing

Translation

Common sense reasoning

Planning

Command following

Recognition

Retrieval

Generation

Question answering

Disambiguation

Language acquisition

Paraphrasing

Translation

Common sense reasoning

Planning

Command following

...

Recognition

Retrieval

Generation

Question answering

Disambiguation

Language acquisition

Paraphrasing

Translation

Common sense reasoning

Planning

Command following

...

Computer vision

NLP

Robotics

AI

Representation

$P(\text{sentence}, \text{video})$

Yu, Siddharth, Barbu, Siskind JAIR 2015

Representation

$$P(\text{sentence}, \text{video})$$

Sentence $\rightarrow$ First order temporal logic $\rightarrow$ video detector

Yu, Siddharth, Barbu, Siskind JAIR 2015

Representation

$P(\text{sentence}, \text{video})$

Sentence → First order temporal logic → video detector

Danny and Andrei move a chair.

Yu, Siddharth, Barbu, Siskind JAIR 2015

Representation

$$P(\text{sentence}, \text{video})$$

Sentence → First order temporal logic → video detector

Danny and Andrei move a chair.

$$\exists xyz\ \textbf{chair}(x), \textbf{person}(y), \textbf{person}(z), y \neq z, \textbf{move}(y, x), \textbf{move}(z, x)$$

Yu, Siddharth, Barbu, Siskind JAIR 2015

Representation

$P(\text{sentence}, \text{video})$

Sentence $\rightarrow$ First order temporal logic $\rightarrow$ video detector

Danny and Andrei move a chair.

$\exists xyz\ \textbf{chair}(x), \textbf{person}(y), \textbf{person}(z), y \neq z, \textbf{move}(y, x), \textbf{move}(z, x)$

Don't build in attributes like *shape*, *color*, etc.

Yu, Siddharth, Barbu, Siskind JAIR 2015

Representation

$P(\text{sentence}, \text{video})$

Sentence $\rightarrow$ First order temporal logic $\rightarrow$ video detector

Danny and Andrei move a chair.

$\exists xyz\ \textbf{chair}(x), \textbf{person}(y), \textbf{person}(z), y \neq z, \textbf{move}(y, x), \textbf{move}(z, x)$

Don't build in attributes like *shape*, *color*, etc.

Build in mechanics! Physics, Interactions, ToM, out of domain goals, etc.

Yu, Siddharth, Barbu, Siskind JAIR 2015

Representation

$P(\text{sentence}, \text{video})$

Sentence $\rightarrow$ First order temporal logic $\rightarrow$ video detector

Danny and Andrei move a chair.

$\exists xyz\ \mathbf{chair}(x), \mathbf{person}(y), \mathbf{person}(z), y \neq z, \mathbf{move}(y, x), \mathbf{move}(z, x)$

Don't build in attributes like *shape*, *color*, etc.

Build in mechanics! Physics, Interactions, ToM, out of domain goals, etc.

Attributes exist because they matter, mechanisms are built in

Yu, Siddharth, Barbu, Siskind JAIR 2015

Representation

$P(\text{sentence}, \text{video})$

Sentence $\rightarrow$ First order temporal logic $\rightarrow$ video detector

Danny and Andrei move a chair.

$\exists xyz\ \mathbf{chair}(x), \mathbf{person}(y), \mathbf{person}(z), y \neq z, \mathbf{move}(y, x), \mathbf{move}(z, x)$

Don't build in attributes like *shape*, *color*, etc.
Build in mechanics! Physics, Interactions, ToM, out of domain goals, etc.
Attributes exist because they matter, mechanisms are built in

Compositions of networks or graphical models:

Yu, Siddharth, Barbu, Siskind JAIR 2015

Representation

$P(\text{sentence}, \text{video})$

Sentence $\rightarrow$ First order temporal logic $\rightarrow$ video detector

Danny and Andrei move a chair.

$\exists xyz\ \mathbf{chair}(x), \mathbf{person}(y), \mathbf{person}(z), y \neq z, \mathbf{move}(y, x), \mathbf{move}(z, x)$

Don't build in attributes like *shape*, *color*, etc.
Build in mechanics! Physics, Interactions, ToM, out of domain goals, etc.
Attributes exist because they matter, mechanisms are built in

Compositions of networks or graphical models:

...

Yu, Siddharth, Barbu, Siskind JAIR 2015

Representation

$P(\text{sentence}, \text{video})$

Sentence $\rightarrow$ First order temporal logic $\rightarrow$ video detector

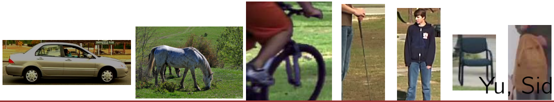
Danny and Andrei move a chair.

$\exists xyz\ \textbf{chair}(x), \textbf{person}(y), \textbf{person}(z), y \neq z, \textbf{move}(y, x), \textbf{move}(z, x)$

Don't build in attributes like *shape*, *color*, etc.
Build in mechanics! Physics, Interactions, ToM, out of domain goals, etc.
Attributes exist because they matter, mechanisms are built in

Compositions of networks or graphical models:

x-tracker y-tracker z-tracker



...

Yu, Siddharth, Barbu, Siskind JAIR 2015

Representation

$P(\text{sentence}, \text{video})$

Sentence $\rightarrow$ First order temporal logic $\rightarrow$ video detector

Danny and Andrei move a chair.

$\exists xyz\ \textbf{chair}(x), \textbf{person}(y), \textbf{person}(z), y \neq z, \textbf{move}(y, x), \textbf{move}(z, x)$

Don't build in attributes like *shape*, *color*, etc.
Build in mechanics! Physics, Interactions, ToM, out of domain goals, etc.
Attributes exist because they matter, mechanisms are built in

Compositions of networks or graphical models:

chair **person** **person** $\neq$ **move** **move**

x-tracker y-tracker z-tracker

...

Yu, Siddharth, Barbu, Siskind JAIR 2015

Representation

$P(\text{sentence}, \text{video})$

Sentence $\rightarrow$ First order temporal logic $\rightarrow$ video detector

Danny and Andrei move a chair.

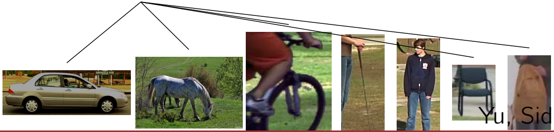
$\exists xyz\ \textbf{chair}(x), \textbf{person}(y), \textbf{person}(z), y \neq z, \textbf{move}(y, x), \textbf{move}(z, x)$

Don't build in attributes like *shape*, *color*, etc.
Build in mechanics! Physics, Interactions, ToM, out of domain goals, etc.
Attributes exist because they matter, mechanisms are built in

Compositions of networks or graphical models:

chair **person** **person** $\neq$ **move** **move**

x-tracker y-tracker z-tracker



. . .

Yu, Siddharth, Barbu, Siskind JAIR 2015

Representation

$P(\text{sentence}, \text{video})$

Sentence $\rightarrow$ First order temporal logic $\rightarrow$ video detector

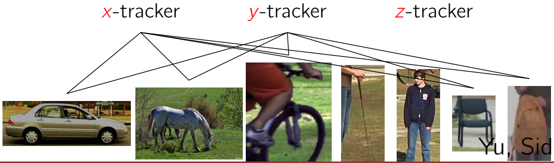
Danny and Andrei move a chair.

$\exists xyz\ \textbf{chair}(x), \textbf{person}(y), \textbf{person}(z), y \neq z, \textbf{move}(y, x), \textbf{move}(z, x)$

Don't build in attributes like *shape*, *color*, etc.
Build in mechanics! Physics, Interactions, ToM, out of domain goals, etc.
Attributes exist because they matter, mechanisms are built in

Compositions of networks or graphical models:

chair **person** **person** $\neq$ **move** **move**



...

Yu, Siddharth, Barbu, Siskind JAIR 2015

Representation

$P(\text{sentence}, \text{video})$

Sentence $\rightarrow$ First order temporal logic $\rightarrow$ video detector

Danny and Andrei move a chair.

$\exists xyz$ **chair**(x), **person**(y), **person**(z), $y \neq z$, **move**(y, x), **move**(z, x)

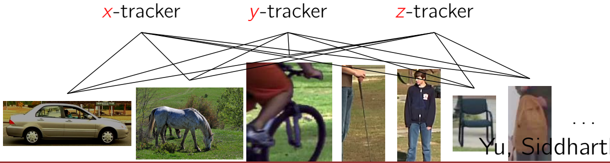
Don't build in attributes like *shape*, *color*, etc.

Build in mechanics! Physics, Interactions, ToM, out of domain goals, etc.

Attributes exist because they matter, mechanisms are built in

Compositions of networks or graphical models:

chair **person** **person** $\neq$ **move** **move**



Yu, Siddharth, Barbu, Siskind JAIR 2015

Representation

$P(\text{sentence}, \text{video})$

Sentence → First order temporal logic → video detector

Danny and Andrei move a chair.

$\exists xyz\ \textbf{chair}(x), \textbf{person}(y), \textbf{person}(z), y \neq z, \textbf{move}(y,x), \textbf{move}(z,x)$

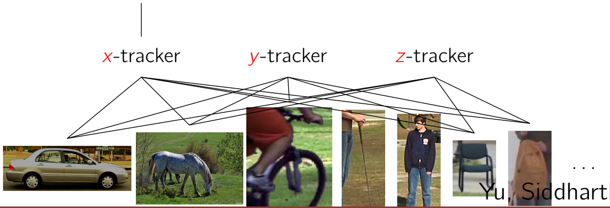
Don't build in attributes like *shape*, *color*, etc.

Build in mechanics! Physics, Interactions, ToM, out of domain goals, etc.

Attributes exist because they matter, mechanisms are built in

Compositions of networks or graphical models:

chair **person** **person** $\neq$ **move** **move**



Yu, Siddharth, Barbu, Siskind JAIR 2015

Representation

$P(\text{sentence}, \text{video})$

Sentence → First order temporal logic → video detector

Danny and Andrei move a chair.

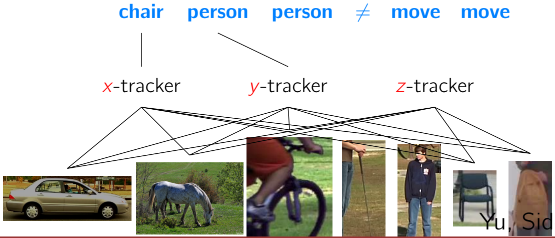
$\exists xyz\ \textbf{chair}(x), \textbf{person}(y), \textbf{person}(z), y \neq z, \textbf{move}(y, x), \textbf{move}(z, x)$

Don't build in attributes like *shape*, *color*, etc.

Build in mechanics! Physics, Interactions, ToM, out of domain goals, etc.

Attributes exist because they matter, mechanisms are built in

Compositions of networks or graphical models:



Yu, Siddharth, Barbu, Siskind JAIR 2015

Representation

$P(\text{sentence}, \text{video})$

Sentence → First order temporal logic → video detector

Danny and Andrei move a chair.

$\exists xyz\ \textbf{chair}(x), \textbf{person}(y), \textbf{person}(z), y \neq z, \textbf{move}(y, x), \textbf{move}(z, x)$

Don't build in attributes like *shape*, *color*, etc.

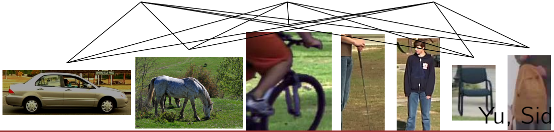
Build in mechanics! Physics, Interactions, ToM, out of domain goals, etc.

Attributes exist because they matter, mechanisms are built in

Compositions of networks or graphical models:

chair **person** **person** $\neq$ **move** **move**

x-tracker y-tracker z-tracker



. . .

Yu, Siddharth, Barbu, Siskind JAIR 2015

Representation

$P(\text{sentence}, \text{video})$

Sentence → First order temporal logic → video detector

Danny and Andrei move a chair.

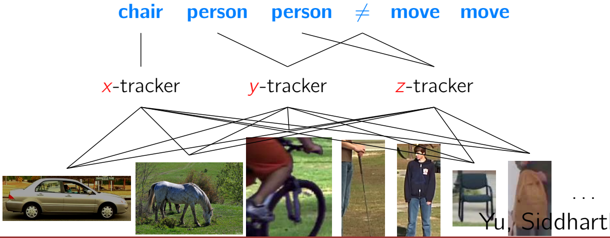
$$\exists xyz\ \mathbf{chair}(x), \mathbf{person}(y), \mathbf{person}(z), y \neq z, \mathbf{move}(y, x), \mathbf{move}(z, x)$$

Don't build in attributes like *shape*, *color*, etc.

Build in mechanics! Physics, Interactions, ToM, out of domain goals, etc.

Attributes exist because they matter, mechanisms are built in

Compositions of networks or graphical models:



Yu, Siddharth, Barbu, Siskind JAIR 2015

Representation

$P(\text{sentence}, \text{video})$

Sentence → First order temporal logic → video detector

Danny and Andrei move a chair.

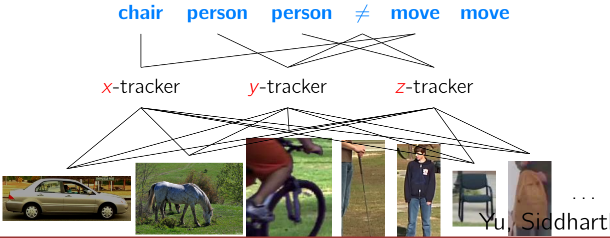
$\exists xyz$ **chair**(x), **person**(y), **person**(z), $y \neq z$, **move**(y, x), **move**(z, x)

Don't build in attributes like *shape*, *color*, etc.

Build in mechanics! Physics, Interactions, ToM, out of domain goals, etc.

Attributes exist because they matter, mechanisms are built in

Compositions of networks or graphical models:



Yu, Siddharth, Barbu, Siskind JAIR 2015

Representation

$P(\text{sentence}, \text{video})$

Sentence → First order temporal logic → video detector

Danny and Andrei move a chair.

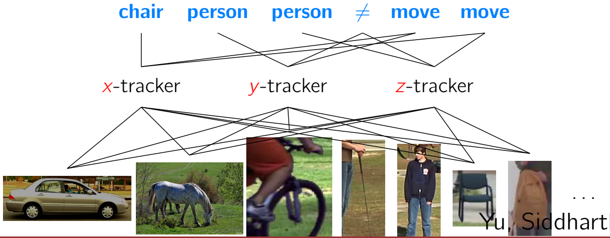
$\exists xyz\ \textbf{chair}(x), \textbf{person}(y), \textbf{person}(z), y \neq z, \textbf{move}(y, x), \textbf{move}(z, x)$

Don't build in attributes like *shape*, *color*, etc.

Build in mechanics! Physics, Interactions, ToM, out of domain goals, etc.

Attributes exist because they matter, mechanisms are built in

Compositions of networks or graphical models:



Yu, Siddharth, Barbu, Siskind JAIR 2015

Recognition $P(\text{sentence}, \text{video})$ Narayanaswamy *et al.* 2014

Retrieval

Generation

Question answering

Disambiguation

Language acquisition

Paraphrasing

Translation

Common sense reasoning

Planning

Command following

...

Recognition $P(\text{sentence}, \text{video})$ Narayanaswamy *et al.* 2014

Retrieval $\underset{v \in V}{\text{argmax}}\ P(s, v)$ Barret *et al.* 2016

Generation

Question answering

Disambiguation

Language acquisition

Paraphrasing

Translation

Common sense reasoning

Planning

Command following

...

Sentential retrieval

Barrett, Barbu, Siddharth, Siskind PAMI 2016

Sentential retrieval

Barrett, Barbu, Siddharth, Siskind PAMI 2016

Recognition $P(\text{sentence}, \text{video})$ Narayanaswamy *et al.* 2014

Retrieval $\underset{v \in V}{\text{argmax}}\ P(s, v)$ Barret *et al.* 2016

Generation

Question answering

Disambiguation

Language acquisition

Paraphrasing

Translation

Common sense reasoning

Planning

Command following

...

Recognition $\boxed{P(\text{sentence}, \text{video})}$ Narayanaswamy *et al.* 2014

Retrieval $\underset{v \in V}{\text{argmax}} \boxed{P(s, v)}$ Barret *et al.* 2016

Generation $\underset{s \in L}{\text{argmax}} \boxed{P(s, v)}$ Yu *et al.* 2015, N. *et al.* 2014

Question answering

Disambiguation

Language acquisition

Paraphrasing

Translation

Common sense reasoning

Planning

Command following

...

Generating sentences

Generating sentences

$$
\begin{aligned}
\mathrm{S} &\rightarrow \mathrm{NP\ VP} \\
\mathrm{NP} &\rightarrow \mathrm{D\ [A]\ N\ [PP]} \\
\mathrm{D} &\rightarrow \textit{an} \mid \textit{the} \\
\mathrm{A} &\rightarrow \textit{blue} \mid \textit{red} \\
\mathrm{N} &\rightarrow \textit{person} \mid \textit{backpack} \mid \textit{chair} \mid \textit{bin} \mid \textit{object} \\
\mathrm{PP} &\rightarrow \mathrm{P\ NP} \\
\mathrm{P} &\rightarrow \textit{to the left of} \mid \textit{to the right of} \\
\mathrm{VP} &\rightarrow \mathrm{V\ NP\ [Adv]\ [PP_M]} \\
\mathrm{V} &\rightarrow \textit{approached} \mid \textit{carried} \mid \textit{picked up} \mid \textit{put down} \\
\mathrm{Adv} &\rightarrow \textit{quickly} \mid \textit{slowly} \\
\mathrm{PP_M} &\rightarrow \mathrm{P_M\ NP} \\
\mathrm{P_M} &\rightarrow \textit{towards} \mid \textit{away from}
\end{aligned}
$$

Generating sentences

$$
\begin{aligned}
\mathrm{S} &\rightarrow \mathrm{NP\ VP} \\
\mathrm{NP} &\rightarrow \mathrm{D\ [A]\ N\ [PP]} \\
\mathrm{D} &\rightarrow \textit{an} \mid \textit{the} \\
\mathrm{A} &\rightarrow \textit{blue} \mid \textit{red} \\
\mathrm{N} &\rightarrow \textit{person} \mid \textit{backpack} \mid \textit{chair} \mid \textit{bin} \mid \textit{object} \\
\mathrm{PP} &\rightarrow \mathrm{P\ NP} \\
\mathrm{P} &\rightarrow \textit{to the left of} \mid \textit{to the right of} \\
\mathrm{VP} &\rightarrow \mathrm{V\ NP\ [Adv]\ [PP_M]} \\
\mathrm{V} &\rightarrow \textit{approached} \mid \textit{carried} \mid \textit{picked up} \mid \textit{put down} \\
\mathrm{Adv} &\rightarrow \textit{quickly} \mid \textit{slowly} \\
\mathrm{PP_M} &\rightarrow \mathrm{P_M\ NP} \\
\mathrm{P_M} &\rightarrow \textit{towards} \mid \textit{away from}
\end{aligned}
$$

147,123,874,800 sentences without recursion

Generating sentences

$$
\begin{aligned}
\mathrm{S} &\rightarrow \mathrm{NP}\ \mathrm{VP} \\
\mathrm{NP} &\rightarrow \mathrm{D}\ [\mathrm{A}]\ \mathrm{N}\ [\mathrm{PP}] \\
\mathrm{D} &\rightarrow \textit{an} \mid \textit{the} \\
\mathrm{A} &\rightarrow \textit{blue} \mid \textit{red} \\
\mathrm{N} &\rightarrow \textit{person} \mid \textit{backpack} \mid \textit{chair} \mid \textit{bin} \mid \textit{object} \\
\mathrm{PP} &\rightarrow \mathrm{P}\ \mathrm{NP} \\
\mathrm{P} &\rightarrow \textit{to the left of} \mid \textit{to the right of} \\
\mathrm{VP} &\rightarrow \mathrm{V}\ \mathrm{NP}\ [\mathrm{Adv}]\ [\mathrm{PP_M}] \\
\mathrm{V} &\rightarrow \textit{approached} \mid \textit{carried} \mid \textit{picked up} \mid \textit{put down} \\
\mathrm{Adv} &\rightarrow \textit{quickly} \mid \textit{slowly} \\
\mathrm{PP_M} &\rightarrow \mathrm{P_M}\ \mathrm{NP} \\
\mathrm{P_M} &\rightarrow \textit{towards} \mid \textit{away from}
\end{aligned}
$$

147,123,874,800 sentences without recursion

$\emptyset$

Generating sentences

$$\begin{aligned}
\text{S} &\rightarrow \text{NP VP} \\
\text{NP} &\rightarrow \text{D [A] N [PP]} \\
\text{D} &\rightarrow \textit{an} \mid \textit{the} \\
\text{A} &\rightarrow \textit{blue} \mid \textit{red} \\
\text{N} &\rightarrow \textit{person} \mid \textit{backpack} \mid \textit{chair} \mid \textit{bin} \mid \textit{object} \\
\text{PP} &\rightarrow \text{P NP} \\
\text{P} &\rightarrow \textit{to the left of} \mid \textit{to the right of} \\
\text{VP} &\rightarrow \text{V NP [Adv] [PP}_\text{M}\text{]} \\
\text{V} &\rightarrow \textit{approached} \mid \textit{carried} \mid \textit{picked up} \mid \textit{put down} \\
\text{Adv} &\rightarrow \textit{quickly} \mid \textit{slowly} \\
\text{PP}_\text{M} &\rightarrow \text{P}_\text{M}\text{ NP} \\
\text{P}_\text{M} &\rightarrow \textit{towards} \mid \textit{away from}
\end{aligned}$$

147,123,874,800 sentences without recursion

“carried”

Generating sentences

$$
\begin{aligned}
\mathrm{S} &\rightarrow \mathrm{NP\ VP} \\
\mathrm{NP} &\rightarrow \mathrm{D\ [A]\ N\ [PP]} \\
\mathrm{D} &\rightarrow \textit{an} \mid \textit{the} \\
\mathrm{A} &\rightarrow \textit{blue} \mid \textit{red} \\
\mathrm{N} &\rightarrow \textit{person} \mid \textit{backpack} \mid \textit{chair} \mid \textit{bin} \mid \textit{object} \\
\mathrm{PP} &\rightarrow \mathrm{P\ NP} \\
\mathrm{P} &\rightarrow \textit{to the left of} \mid \textit{to the right of} \\
\mathrm{VP} &\rightarrow \mathrm{V\ NP\ [Adv]\ [PP_M]} \\
\mathrm{V} &\rightarrow \textit{approached} \mid \textit{carried} \mid \textit{picked up} \mid \textit{put down} \\
\mathrm{Adv} &\rightarrow \textit{quickly} \mid \textit{slowly} \\
\mathrm{PP_M} &\rightarrow \mathrm{P_M\ NP} \\
\mathrm{P_M} &\rightarrow \textit{towards} \mid \textit{away from}
\end{aligned}
$$

147,123,874,800 sentences without recursion

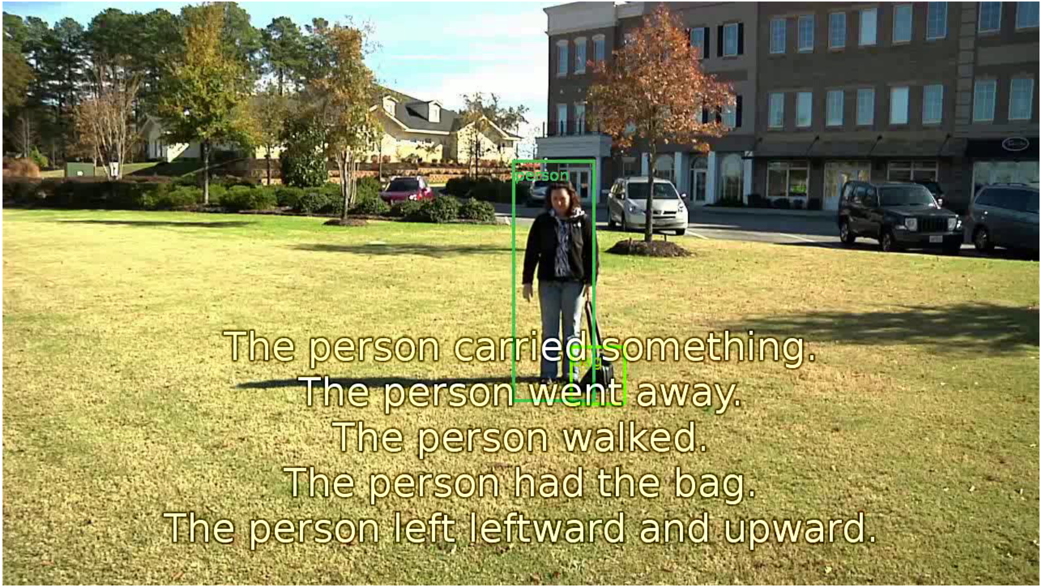
"the person carried"

Generating sentences

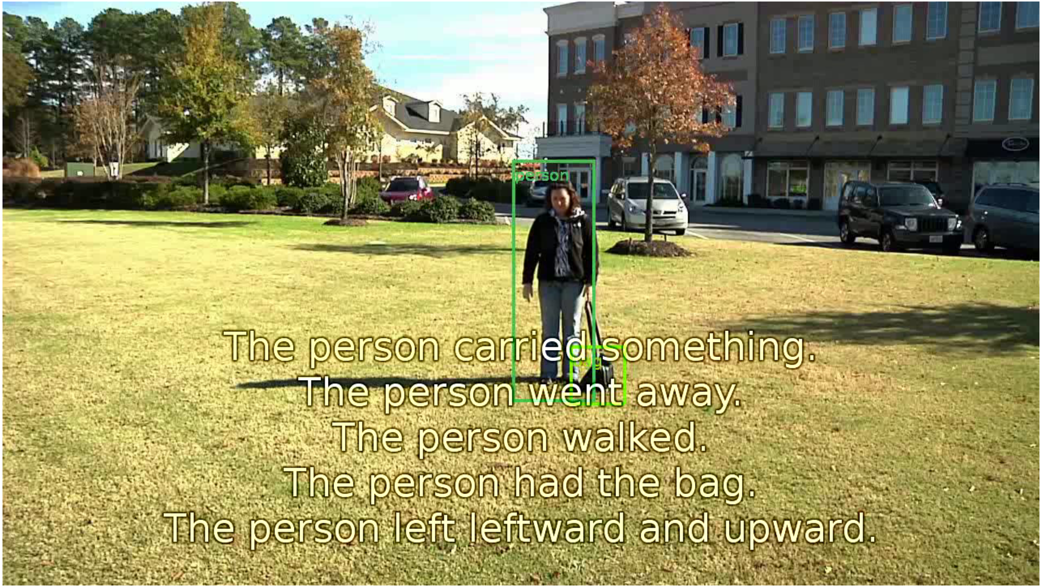
$$
\begin{aligned}
\mathrm{S} &\rightarrow \mathrm{NP}\ \mathrm{VP}\\
\mathrm{NP} &\rightarrow \mathrm{D}\ [\mathrm{A}]\ \mathrm{N}\ [\mathrm{PP}]\\
\mathrm{D} &\rightarrow \textit{an} \mid \textit{the}\\
\mathrm{A} &\rightarrow \textit{blue} \mid \textit{red}\\
\mathrm{N} &\rightarrow \textit{person} \mid \textit{backpack} \mid \textit{chair} \mid \textit{bin} \mid \textit{object}\\
\mathrm{PP} &\rightarrow \mathrm{P}\ \mathrm{NP}\\
\mathrm{P} &\rightarrow \textit{to the left of} \mid \textit{to the right of}\\
\mathrm{VP} &\rightarrow \mathrm{V}\ \mathrm{NP}\ [\mathrm{Adv}]\ [\mathrm{PP_M}]\\
\mathrm{V} &\rightarrow \textit{approached} \mid \textit{carried} \mid \textit{picked up} \mid \textit{put down}\\
\mathrm{Adv} &\rightarrow \textit{quickly} \mid \textit{slowly}\\
\mathrm{PP_M} &\rightarrow \mathrm{P_M}\ \mathrm{NP}\\
\mathrm{P_M} &\rightarrow \textit{towards} \mid \textit{away from}
\end{aligned}
$$

147,123,874,800 sentences without recursion

"the person carried the backpack"



Barbu *et al.* UAI 2012



Barbu *et al.* UAI 2012

Recognition	$\boxed{P(\text{sentence}, \text{video})}$	Narayanaswamy *et al.* 2014
Retrieval	$\operatorname{argmax}_{v \in V} \boxed{P(s, v)}$	Barret *et al.* 2016
Generation	$\operatorname{argmax}_{s \in L} \boxed{P(s, v)}$	Yu *et al.* 2015, N. *et al.* 2014
Question answering		
Disambiguation		
Language acquisition		
Paraphrasing		
Translation		
Common sense reasoning		
Planning		
Command following		

...

Recognition	$P(\text{sentence}, \text{video})$	Narayanaswamy *et al.* 2014
Retrieval	$\operatorname*{argmax}_{v \in V} P(s, v)$	Barret *et al.* 2016
Generation	$\operatorname*{argmax}_{s \in L} P(s, v)$	Yu *et al.* 2015, N. *et al.* 2014
Question answering	$\operatorname*{argmax}_{s \in L} P(Q(s, s_q), v)$	Barbu *et al.* in prep.
Disambiguation		
Language acquisition		
Paraphrasing		
Translation		
Common sense reasoning		
Planning		
Command following		
...		

Question answering

Question answering

Question answering

Question answering

What did the person put on top of the red car?

Question answering

What did the person put on top of the red car?

The person put NP on top of the red car.

Question answering

What did the person put on top of the red car?

The person put NP on top of the red car.

The person put the pear on top of the red car.

I saw the man with the telescope.

I saw the man with the telescope.

I saw the man with the telescope.

Danny looked at the man with a telescope.

Danny looked at the man with a telescope.

Danny has the telescope

Danny looked at the man with a telescope.

Danny has the telescope

The man has the telescope

Danny looked at the man with a telescope.

Danny has the telescope

The man has the telescope

Danny looked at the man with a telescope.

Danny has the telescope

The man has the telescope

Danny looked at the man with a telescope.

Danny has the telescope

The man has the telescope

Ambiguities

Berzak, Barbu, Katz, and Ullman EMNLP 2016

Ambiguities

PP Attachment

Danny looked at the man with a telescope.

Berzak, Barbu, Katz, and Ullman EMNLP 2016

Ambiguities

PP Attachment

VP Attachment

Andrei approached the person holding a green chair.

Berzak, Barbu, Katz, and Ullman EMNLP 2016

Ambiguities

PP Attachment

VP Attachment

Conjunction

Danny and Andrei picked up the yellow bag and chair.

Berzak, Barbu, Katz, and Ullman EMNLP 2016

Ambiguities

PP Attachment

VP Attachment

Conjunction

Logical Form

Someone put down the bags.

Berzak, Barbu, Katz, and Ullman EMNLP 2016

Ambiguities

PP Attachment

VP Attachment

Conjunction

Logical Form

Anaphora

Danny picked up the bag and the chair. It is yellow.

Berzak, Barbu, Katz, and Ullman EMNLP 2016

Ambiguities

PP Attachment

VP Attachment

Conjunction

Logical Form

Anaphora

Ellipsis

Danny left Andrei. Also Yevgeni.

Berzak, Barbu, Katz, and Ullman EMNLP 2016

Recognition	$\boxed{P(\text{sentence}, \text{video})}$	Narayanaswamy *et al.* 2014
Retrieval	$\operatorname*{argmax}_{v \in V} \boxed{P(s, v)}$	Barret *et al.* 2016
Generation	$\operatorname*{argmax}_{s \in L} \boxed{P(s, v)}$	Yu *et al.* 2015, N. *et al.* 2014
Question answering	$\operatorname*{argmax}_{s \in L} \boxed{P(Q(s, s_q), v)}$	Barbu *et al.* in prep.
Disambiguation	$\operatorname*{argmax}_{i \in \text{parser}(s)} \boxed{P(i, v)}$	Berzak *et al.* 2015
Language acquisition		
Paraphrasing		
Translation		
Common sense reasoning		
Planning		
Command following		

...

Recognition	$P(\text{sentence}, \text{video})$	Narayanaswamy *et al.* 2014
Retrieval	$\underset{v \in V}{\text{argmax}}\ P(s, v)$	Barret *et al.* 2016
Generation	$\underset{s \in L}{\text{argmax}}\ P(s, v)$	Yu *et al.* 2015, N. *et al.* 2014
Question answering	$\underset{s \in L}{\text{argmax}}\ P(\mathrm{Q}(s, s_q), v)$	Barbu *et al.* in prep.
Disambiguation	$\underset{i \in \text{parser}(s)}{\text{argmax}}\ P(i, v)$	Berzak *et al.* 2015
Language acquisition	$\underset{\theta}{\text{argmax}} \prod_{s,v} P(s(\theta), v)$	Yu *et al.* 2015, Ross *et al.* 2018
Paraphrasing		
Translation		
Common sense reasoning		
Planning		
Command following		
...		

Language acquisition from 10,000 feet

Language acquisition from 10,000 feet

Take this apple.

Language acquisition from 10,000 feet

Take this apple.

Language acquisition from 10,000 feet

(*Take this apple.* , )

Language acquisition from 10,000 feet

$\{($ *Take this apple.* ,  $)\}$

Language acquisition from 10,000 feet

$\underbrace{\{\,(\,\textit{Take this apple.}\,,\ \text{[image]}\,)\,\}}$

$\Downarrow$

Representation

Language acquisition from 10,000 feet

{ (*Take this apple.* , ) }

Recognition

Representation

Language acquisition from 10,000 feet

{ (*Take this apple.* , ) }

Representation

Recognition

Generation

Language acquisition from 10,000 feet

{ (*Take this apple.* , ) }

Representation

Recognition

Generation

Disambiguation

Language acquisition from 10,000 feet

{ (*Take this apple.* , ) }

Representation

Recognition

Generation

Disambiguation

Paraphrasing

Language acquisition from 10,000 feet

{ (*Take this apple.* , ) }

Representation

Recognition
Generation
Disambiguation
Paraphrasing
Planning

Language acquisition from 10,000 feet

{ (*Take this apple.* , ) }

Representation

Recognition
Generation
Disambiguation
Paraphrasing
Planning
Command following

Language acquisition from 10,000 feet

$\underbrace{\{(\textit{Take this apple.}\ ,}$  $)\}$

$\Downarrow$

Representation → Recognition
Representation → Generation
Representation → Disambiguation
Representation → Paraphrasing
Representation → Planning
Representation → Command following
⋮

Language acquisition from 10,000 feet

{ (*Take this apple.* , ) }

Representation

Recognition
Generation
Disambiguation
Paraphrasing
Planning
Command following
⋮

Language acquisition from 10,000 feet

By age 4

{ (*Take this apple.* , ) }

Representation

Recognition
Generation
Disambiguation
Paraphrasing
Planning
Command following
⋮

Language acquisition from 10,000 feet

By age 4

heard ≈12000 words per day

{ (*Take this apple.* , ) }

Representation

Recognition

Generation

Disambiguation

Paraphrasing

Planning

Command following

⋮

Language acquisition from 10,000 feet

By age 4

heard ≈12000 words per day

≈1-2 million utterances

{ (*Take this apple.* , ) }

Representation

Recognition

Generation

Disambiguation

Paraphrasing

Planning

Command following

⋮

Language acquisition from 10,000 feet

By age 4

- heard ≈12000 words per day
- ≈1-2 million utterances
- ≈550 daily conversational turns

{ (*Take this apple.* ,

) }

Representation

- Recognition
- Generation
- Disambiguation
- Paraphrasing
- Planning
- Command following
- ⋮

Grounded language acquisition

Ross, Barbu, Berzak, Myanganbayar, Katz EMNLP 2018

Grounded language acquisition

Danny approached the chair with a bag.

Ross, Barbu, Berzak, Myanganbayar, Katz EMNLP 2018

Grounded language acquisition

Danny approached the chair with a bag.

↓

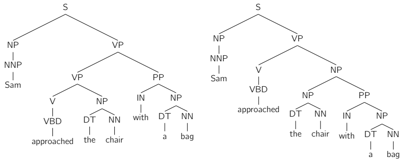
parser

Ross, Barbu, Berzak, Myanganbayar, Katz EMNLP 2018

Grounded language acquisition

Danny approached the chair with a bag.

parser

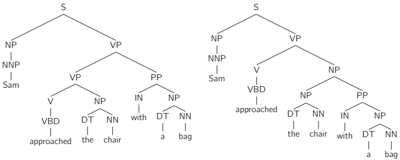


Ross, Barbu, Berzak, Myanganbayar, Katz EMNLP 2018

Grounded language acquisition

Danny approached the chair with a bag.

parser

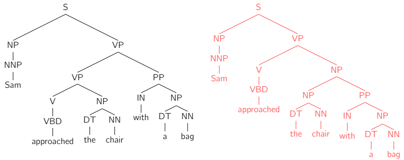


Ross, Barbu, Berzak, Myanganbayar, Katz EMNLP 2018

Grounded language acquisition

Danny approached the chair with a bag.

parser

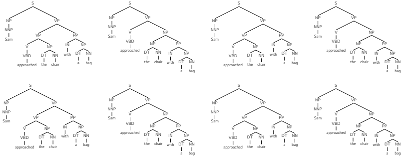


Ross, Barbu, Berzak, Myanganbayar, Katz EMNLP 2018

Grounded language acquisition

Danny approached the chair with a bag.

parser



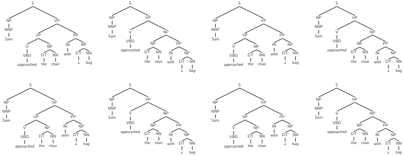
. . .

Ross, Barbu, Berzak, Myanganbayar, Katz EMNLP 2018

Grounded language acquisition

Danny approached the chair with a bag.

≈parser



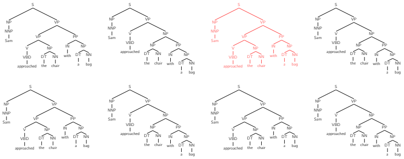
. . .

Ross, Barbu, Berzak, Myanganbayar, Katz EMNLP 2018

Grounded language acquisition

Danny approached the chair with a bag.

≈parser



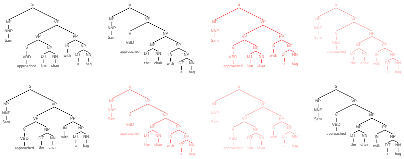
. . .

Ross, Barbu, Berzak, Myanganbayar, Katz EMNLP 2018

Grounded language acquisition

Danny approached the chair with a bag.

≈parser



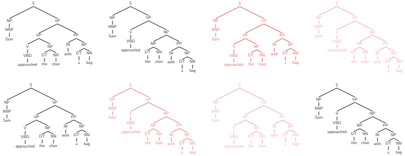
. . .

Ross, Barbu, Berzak, Myanganbayar, Katz EMNLP 2018

Grounded language acquisition

Danny approached the chair with a bag.

≈parser



• • •

Ross, Barbu, Berzak, Myanganbayar, Katz EMNLP 2018

The acquired syntactic & semantic parser

A CCG-based parser with an acquired lexicon and
a small network that ranks derivations

The acquired syntactic & semantic parser

A CCG-based parser with an acquired lexicon and
a small network that ranks derivations

She places the toy car down on the table.

The acquired syntactic & semantic parser

A CCG-based parser with an acquired lexicon and
a small network that ranks derivations

She places the toy car down on the table.

$\lambda xyz.\texttt{person}\ x, \texttt{put-down}\ x\ y, \texttt{toy}\ y, \texttt{car}\ y, \texttt{table}\ z, \texttt{on}\ y\ z$

The acquired syntactic & semantic parser

A CCG-based parser with an acquired lexicon and
a small network that ranks derivations

She places the toy car down on the table.

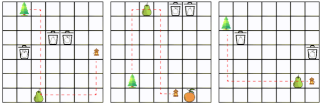
$\lambda xyz.\texttt{person}\ x, \texttt{put-down}\ x\ y, \texttt{toy}\ y, \texttt{car}\ y, \texttt{table}\ z, \texttt{on}\ y\ z$

Fully-supervised: 93% accuracy
Unsupervised: ≈1% accuracy
Ours, videos without annotations: 60% accuracy

Learning a parser-generator pair using a robotic simulator

Wang, Ross, Katz, Barbu CoRL 2020

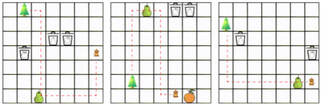
Learning a parser-generator pair using a robotic simulator

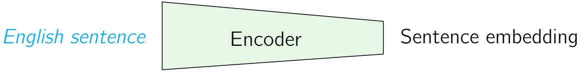


English sentence

Wang, Ross, Katz, Barbu CoRL 2020

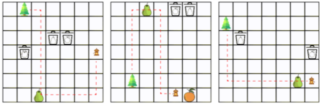
Learning a parser-generator pair using a robotic simulator

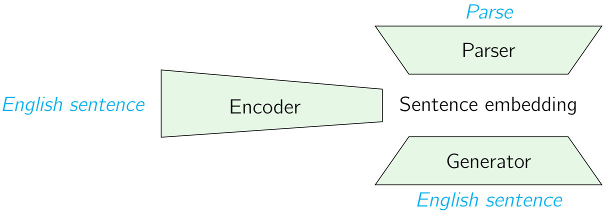




Wang, Ross, Katz, Barbu CoRL 2020

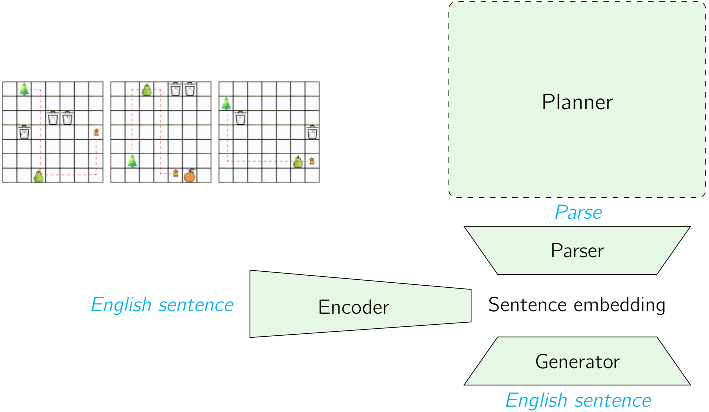
Learning a parser-generator pair using a robotic simulator





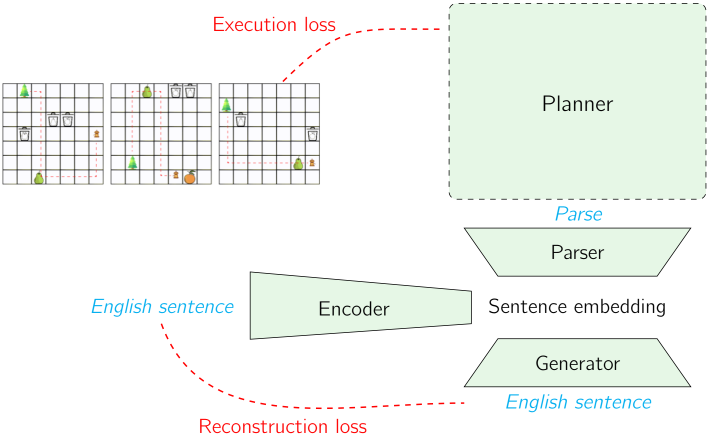
Wang, Ross, Katz, Barbu CoRL 2020

Learning a parser-generator pair using a robotic simulator



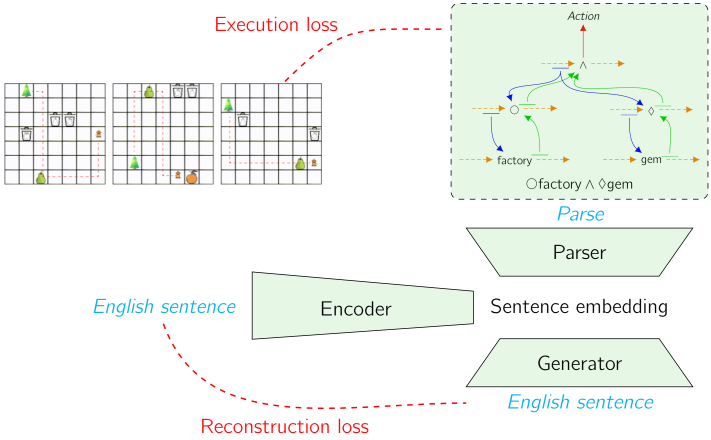
Wang, Ross, Katz, Barbu CoRL 2020

Learning a parser-generator pair using a robotic simulator



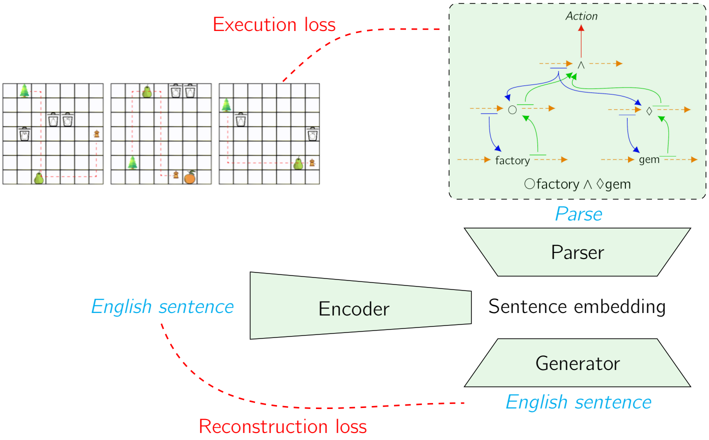
Wang, Ross, Katz, Barbu CoRL 2020

Learning a parser-generator pair using a robotic simulator



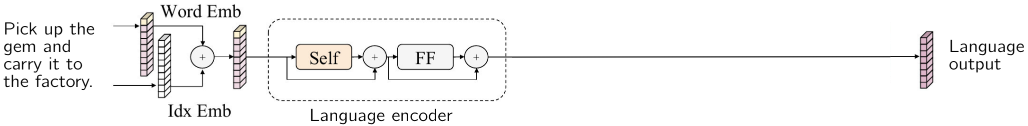
Wang, Ross, Katz, Barbu CoRL 2020

Use a robot to discover the structure of language



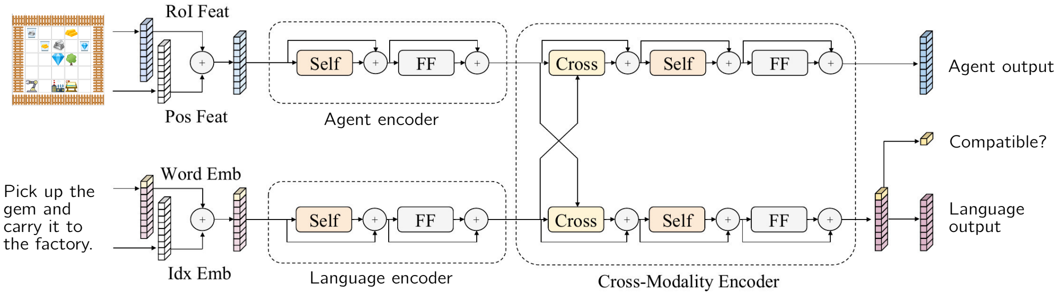
Wang, Ross, Katz, Barbu CoRL 2020

Language models that interact physically



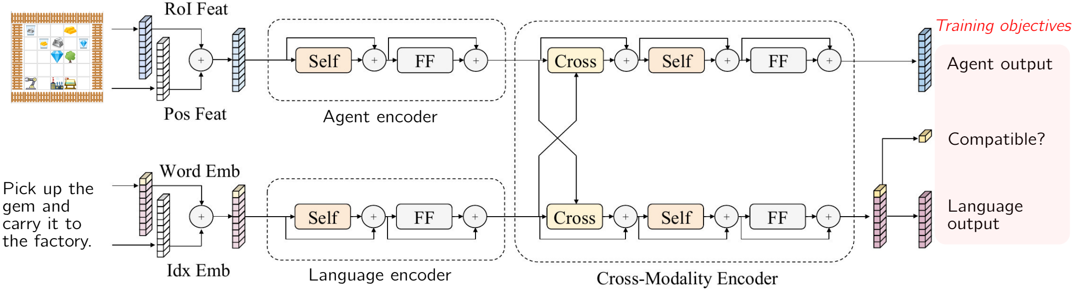
Wang, Ross, Sleeper, Katz, Barbu in prep

Language models that interact physically



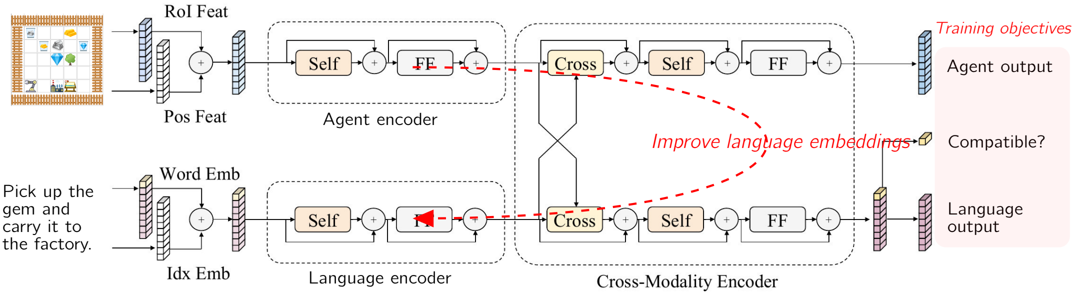
Wang, Ross, Sleeper, Katz, Barbu in prep

Language models that interact physically



Wang, Ross, Sleeper, Katz, Barbu in prep

Language models that interact physically



Wang, Ross, Sleeper, Katz, Barbu in prep

What about social interactions?

We have no social simulators

Netanyahu, Shu, Barbu, Katz, Tenenbaum AAAI 2021

What about social interactions?

We now have social simulators

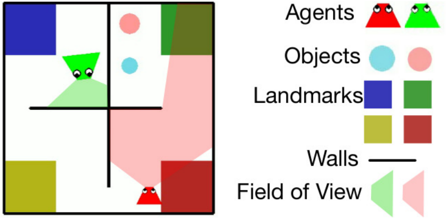
PHASE: PHysically-grounded Abstract Social Events for Machine Social Perception

Netanyahu, Shu, Barbu, Katz, Tenenbaum AAAI 2021

What about social interactions?

We now have social simulators

PHASE: PHysically-grounded Abstract Social Events for Machine Social Perception

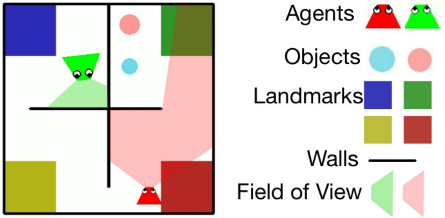


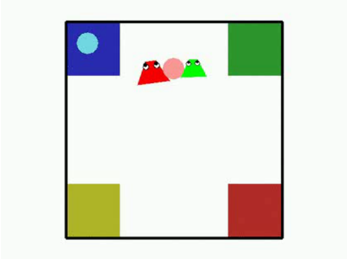
Netanyahu, Shu, Barbu, Katz, Tenenbaum AAAI 2021

What about social interactions?

We now have social simulators

PHASE: PHysically-grounded Abstract Social Events for Machine Social Perception



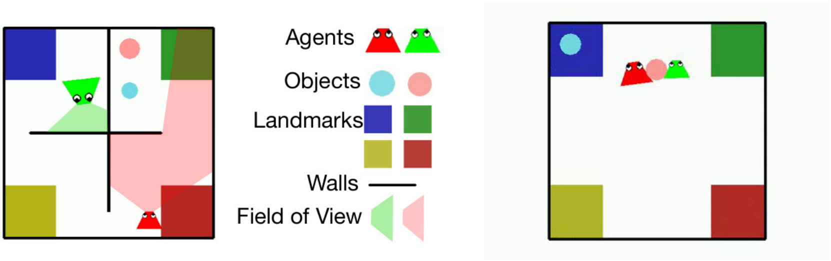


Netanyahu, Shu, Barbu, Katz, Tenenbaum AAAI 2021

What about social interactions?

We now have social simulators

PHASE: PHysically-grounded Abstract Social Events for Machine Social Perception



We have no mathematical theories theories of social interactions

Netanyahu, Shu, Barbu, Katz, Tenenbaum AAAI 2021

Pilley and Reid 2011

Pilley and Reid 2011

Pilley and Reid 2011

Recognition	$P(\text{sentence}, \text{video})$	Narayanaswamy *et al.* 2014
Retrieval	$\underset{v \in V}{\text{argmax}}\ P(s, v)$	Barret *et al.* 2016
Generation	$\underset{s \in L}{\text{argmax}}\ P(s, v)$	Yu *et al.* 2015, N. *et al.* 2014
Question answering	$\underset{s \in L}{\text{argmax}}\ P(Q(s, s_q), v)$	Barbu *et al.* in prep.
Disambiguation	$\underset{i \in \text{parser}(s)}{\text{argmax}}\ P(i, v)$	Berzak *et al.* 2015
Language acquisition	$\underset{\theta}{\text{argmax}} \prod_{s,v} P(s(\theta), v)$	Yu *et al.* 2015, Ross *et al.* 2018
Paraphrasing		
Translation		
Common sense reasoning		
Planning		
Command following		

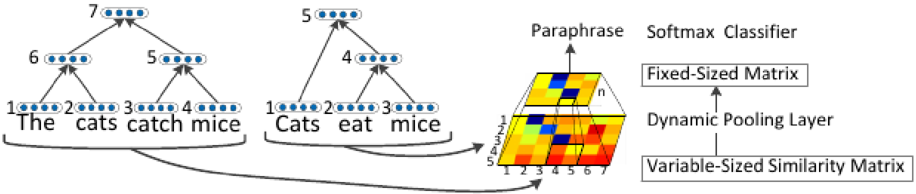
...

Recognition	$\boxed{P(\text{sentence}, \text{video})}$	Narayanaswamy *et al.* 2014
Retrieval	$\underset{v \in V}{\text{argmax}} \boxed{P(s, v)}$	Barret *et al.* 2016
Generation	$\underset{s \in L}{\text{argmax}} \boxed{P(s, v)}$	Yu *et al.* 2015, N. *et al.* 2014
Question answering	$\underset{s \in L}{\text{argmax}} \boxed{P(Q(s, s_q), v)}$	Barbu *et al.* in prep.
Disambiguation	$\underset{i \in \text{parser}(s)}{\text{argmax}} \boxed{P(i, v)}$	Berzak *et al.* 2015
Language acquisition	$\underset{\theta}{\text{argmax}} \prod_{s,v} \boxed{P(s(\theta), v)}$	Yu *et al.* 2015, Ross *et al.* 2018
Paraphrasing	$\int_v \left\vert \boxed{P(s, v)} - \boxed{P(s', v)} \right\vert$	Mao *et al.* in review
Translation		
Common sense reasoning		
Planning		
Command following		
...		

Paraphrasing today

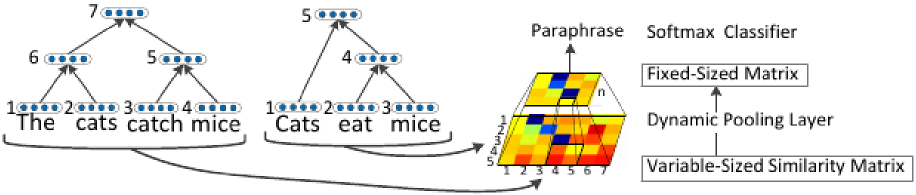
Socher *et al*. 2011, Cheng and Kartsaklis 2015

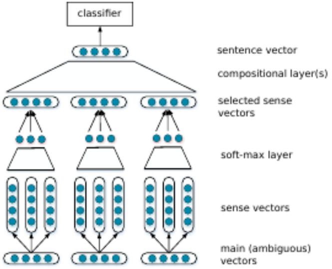
Paraphrasing today

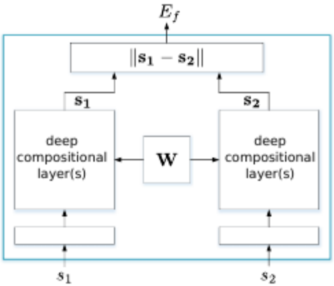


Socher *et al.* 2011, Cheng and Kartsaklis 2015

Paraphrasing today







Socher *et al.* 2011, Cheng and Kartsaklis 2015

Paraphrasing with vision

Mao, Katz, Barbu; IJCAI in review

Paraphrasing with vision

$$s \overset{?}{\Rightarrow} s'$$

Mao, Katz, Barbu; IJCAI in review

Paraphrasing with vision

$$s \overset{?}{\Rightarrow} s'$$

s

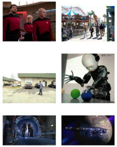
Mao, Katz, Barbu; IJCAI in review

Paraphrasing with vision

$$s \overset{?}{\Longrightarrow} s'$$

YouTube

$s \longrightarrow$



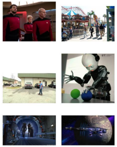
⋮

Mao, Katz, Barbu; IJCAI in review

Paraphrasing with vision

$$s \overset{?}{\Longrightarrow} s'$$

YouTube

$s \longrightarrow$  $\longrightarrow$ Videos v

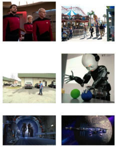
$\vdots$

Mao, Katz, Barbu; IJCAI in review

Paraphrasing with vision

$$s \overset{?}{\Longrightarrow} s'$$

YouTube

$s \longrightarrow$  $\longrightarrow$ Videos v $\quad s'$

$\vdots$

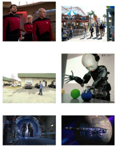
Mao, Katz, Barbu; IJCAI in review

Paraphrasing with vision

$$s \overset{?}{\Rightarrow} s'$$

YouTube

$s \longrightarrow$



$\vdots$

$\longrightarrow$ Videos v $\qquad s'$

$$\underbrace{\hphantom{s \longrightarrow \text{Videos } v \quad s'}}_{\sum_{v} |S(s,v) - S(s',v)|^2}$$

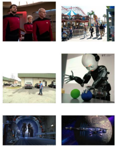
Mao, Katz, Barbu; IJCAI in review

Paraphrasing with vision

$$s \overset{?}{\Longrightarrow} s'$$

YouTube

$s \longrightarrow$



$\longrightarrow$ Videos v $\qquad s'$

$$\underbrace{\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad}_{\sum_{v} |S(s,v) - S(s',v)|^2}$$

Mao, Katz, Barbu; IJCAI in review

Paraphrasing with vision

$$s \overset{?}{\Rightarrow} s'$$

YouTube

$s \longrightarrow$ YouTube $\longrightarrow$ Videos v $\qquad s'$

$$\sum_{v} |S(s,v) - S(s',v)|^2$$

Mao, Katz, Barbu; IJCAI in review

Paraphrasing with vision

$$s \overset{?}{\Rightarrow} s'$$

$$\underbrace{s \xrightarrow{\text{sample}} \text{Videos } v \qquad s'}_{\int_v |S(s,v) - S(s',v)|^2}$$

Mao, Katz, Barbu; IJCAI in review

Paraphrasing with imagination

$$s \overset{?}{\Rightarrow} s'$$

$$\underbrace{s \xrightarrow{\text{sample}} \text{Videos } v \qquad s'}_{\int_v |S(s,v) - S(s',v)|^2}$$

Mao, Katz, Barbu; IJCAI in review

Generated “videos”

Alice carried the chair away from the backpack.

Mao, Katz, Barbu; IJCAI in review

Generated “videos”

Alice carried the chair away from the backpack.

Mao, Katz, Barbu; IJCAI in review

Comparing sentences vs Parikh *et al*. 2016 + ELMo

Does the sentence above imply the one below, and vice versa?

Mao *et al*. in review

Comparing sentences vs Parikh *et al*. 2016 + ELMo

Does the sentence above imply the one below, and vice versa?

Ground Ours Theirs

Mao *et al*. in review

Comparing sentences vs Parikh *et al*. 2016 + ELMo

Does the sentence above imply the one below, and vice versa?

Ground Ours Theirs

Alice carried the chair
Alice held the chair

Mao *et al*. in review

Comparing sentences vs Parikh *et al*. 2016 + ELMo

Does the sentence above imply the one below, and vice versa?

		Ground	Ours	Theirs
Alice carried the chair / *Alice held the chair*	$\Downarrow$	Y		

Mao *et al*. in review

Comparing sentences vs Parikh *et al.* 2016 + ELMo

Does the sentence above imply the one below, and vice versa?

		Ground	Ours	Theirs
Alice carried the chair *Alice held the chair*	$\Downarrow$	Y	Y	

Mao *et al.* in review

Comparing sentences vs Parikh *et al.* 2016 + ELMo

Does the sentence above imply the one below, and vice versa?

		Ground	Ours	Theirs
Alice carried the chair *Alice held the chair*	$\Downarrow$	Y	Y	N

Mao *et al.* in review

Comparing sentences vs Parikh *et al*. 2016 + ELMo

Does the sentence above imply the one below, and vice versa?

		Ground	Ours	Theirs
Alice carried the chair	$\Downarrow$	Y	Y	N
Alice held the chair	$\Uparrow$	N		

Mao *et al*. in review

Comparing sentences vs Parikh *et al.* 2016 + ELMo

Does the sentence above imply the one below, and vice versa?

		Ground	Ours	Theirs
Alice carried the chair	$\Downarrow$	Y	Y	N
Alice held the chair	$\Uparrow$	N	N	

Mao *et al.* in review

Comparing sentences vs Parikh *et al.* 2016 + ELMo

Does the sentence above imply the one below, and vice versa?

		Ground	Ours	Theirs
Alice carried the chair	$\Downarrow$	Y	Y	N
Alice held the chair	$\Uparrow$	N	N	Y

Mao *et al.* in review

Comparing sentences vs Parikh *et al.* 2016 + ELMo

Does the sentence above imply the one below, and vice versa?

		Ground	Ours	Theirs
Alice carried the chair	$\Downarrow$	Y	Y	N
Alice held the chair	$\Uparrow$	N	N	Y

Alice carried the chair towards Ben
Alice approached Ben

Mao *et al.* in review

Comparing sentences vs Parikh *et al*. 2016 + ELMo

Does the sentence above imply the one below, and vice versa?

		Ground	Ours	Theirs
Alice carried the chair	$\Downarrow$	Y	Y	N
Alice held the chair	$\Uparrow$	N	N	Y
Alice carried the chair towards Ben	$\Downarrow$	Y	Y	N
Alice approached Ben	$\Uparrow$	N	N	Y?

Mao *et al*. in review

Comparing sentences vs Parikh *et al.* 2016 + ELMo

Does the sentence above imply the one below, and vice versa?

		Ground	Ours	Theirs
Alice carried the chair	$\Downarrow$	Y	Y	N
Alice held the chair	$\Uparrow$	N	N	Y
Alice carried the chair towards Ben	$\Downarrow$	Y	Y	N
Alice approached Ben	$\Uparrow$	N	N	Y?
Alice carried the chair towards Ben	$\Downarrow$	N	N	Y?
Alice left Ben	$\Uparrow$	N	N	Y?

Mao *et al.* in review

Comparing sentences vs Parikh *et al.* 2016 + ELMo

Does the sentence above imply the one below, and vice versa?

		Ground	Ours	Theirs
Alice carried the chair	$\Downarrow$	Y	Y	N
Alice held the chair	$\Uparrow$	N	N	Y
Alice carried the chair towards Ben	$\Downarrow$	Y	Y	N
Alice approached Ben	$\Uparrow$	N	N	Y?
Alice carried the chair towards Ben	$\Downarrow$	N	N	Y?
Alice left Ben	$\Uparrow$	N	N	Y?
Alice picked up the chair, and Ben put down the bag	$\Downarrow$	N	N	Y
Ben picked up the chair, and Alice put down the bag	$\Uparrow$	N	N	Y

Mao *et al.* in review

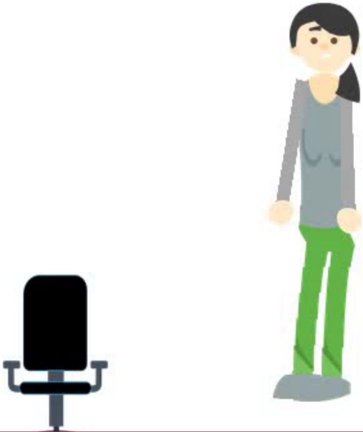
What about other languages?

What about other languages?

爱丽丝靠近了一把椅子。

What about other languages?

爱丽丝靠近了一把椅子。



Word Embedding Association Test (WEAT)

Word Embedding Association Test (WEAT)

$$\mathit{Women} = \{\text{Anna, Mary}\} \qquad \mathit{Work} = \{\text{office, desk}\}$$

$$\mathit{Men} = \{\text{Dave, John}\} \qquad \mathit{Home} = \{\text{children, home}\}$$

Word Embedding Association Test (WEAT)

$$\underbrace{\begin{array}{c} \textit{Women} = \{\text{Anna, Mary}\} \\ \textit{Men} = \{\text{Dave, John}\} \end{array}}_{\text{Targets}} \qquad \underbrace{\begin{array}{c} \textit{Work} = \{\text{office, desk}\} \\ \textit{Home} = \{\text{children, home}\} \end{array}}_{\text{Attributes}}$$

Word Embedding Association Test (WEAT)

$$\underbrace{\begin{array}{c} \textit{Women} = \{\text{Anna, Mary}\} \\ \textit{Men} = \{\text{Dave, John}\} \end{array}}_{\text{Targets}} \qquad \underbrace{\begin{array}{c} \textit{Work} = \{\text{office, desk}\} \\ \textit{Home} = \{\text{children, home}\} \end{array}}_{\text{Attributes}}$$

Compare the distances between *Women* *Work* and *Home*
and between *Men* *Work* and *Home*

Word Embedding Association Test (WEAT)

$$\underbrace{\begin{array}{c} \textit{Women} = \{\text{Anna, Mary}\} \\ \textit{Men} = \{\text{Dave, John}\} \end{array}}_{\text{Targets}} \qquad \underbrace{\begin{array}{c} \textit{Work} = \{\text{office, desk}\} \\ \textit{Home} = \{\text{children, home}\} \end{array}}_{\text{Attributes}}$$

Compare the distances between *Women* *Work* and *Home*
and between *Men* *Work* and *Home*

If they are very different then there is some bias
because this has practical consequences for the inferences that networks make.
There are variants like SEAT which test whole sentences.

Grounded WEAT

Dave 

John 

Steve 

Anna 

Mary 

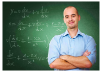
Beth 

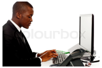
doctor 

lawyer 

police 

librarian 

teacher 

secretary 

doctor 

lawyer 

police 

librarian 

teacher 

secretary 

Ross, Katz, Barbu NAACL 2021

Different groupings answer different questions

Ross, Katz, Barbu NAACL 2021

Different groupings answer different questions

3 tests: word, sentence, and context
6 gender bias tests and 7 racial bias tests
4 popular models (VisualBERT, VL-BERT, ViLBERT, LXMERT

Ross, Katz, Barbu NAACL 2021

Different groupings answer different questions

3 tests: word, sentence, and context

6 gender bias tests and 7 racial bias tests

4 popular models (VisualBERT, VL-BERT, ViLBERT, LXMERT

Do multimodal models have social biases?

Ross, Katz, Barbu NAACL 2021

Different groupings answer different questions

3 tests: word, sentence, and context
6 gender bias tests and 7 racial bias tests
4 popular models (VisualBERT, VL-BERT, ViLBERT, LXMERT

Do multimodal models have social biases? They all do

Ross, Katz, Barbu NAACL 2021

Different groupings answer different questions

3 tests: word, sentence, and context
6 gender bias tests and 7 racial bias tests
4 popular models (VisualBERT, VL-BERT, ViLBERT, LXMERT

Do multimodal models have social biases? They all do
Can counterstereotypical visual evidence offset a bias?

Ross, Katz, Barbu NAACL 2021

Different groupings answer different questions

3 tests: word, sentence, and context
6 gender bias tests and 7 racial bias tests
4 popular models (VisualBERT, VL-BERT, ViLBERT, LXMERT

Do multimodal models have social biases? They all do
Can counterstereotypical visual evidence offset a bias? No
Do biases come from vision or language?

Ross, Katz, Barbu NAACL 2021

Different groupings answer different questions

3 tests: word, sentence, and context
6 gender bias tests and 7 racial bias tests
4 popular models (VisualBERT, VL-BERT, ViLBERT, LXMERT

Do multimodal models have social biases? They all do
Can counterstereotypical visual evidence offset a bias? No
Do biases come from vision or language? ViLBERT clearly language, mix for the others

Ross, Katz, Barbu NAACL 2021

A missing algorithm in our language learning story

Ross, Katz, and Barbu in prep

A missing algorithm in our language learning story

Ross, Katz, and Barbu in prep

A missing algorithm in our language learning story

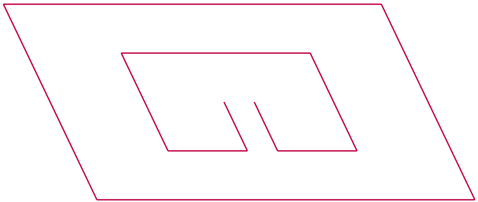
Overwhelm with realistic but synthetic visual evidence against biases.

Ross, Katz, and Barbu in prep

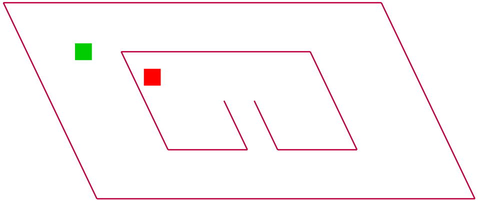
Communication isn't all verbal

Communication isn't all verbal

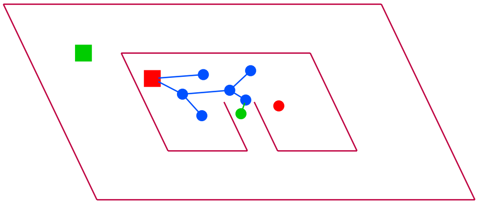
Sampling-based planning with language



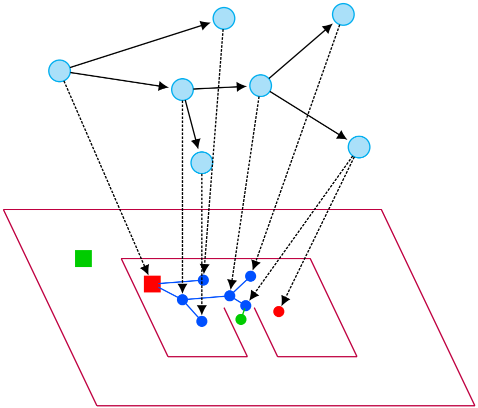
Sampling-based planning with language



Sampling-based planning with language

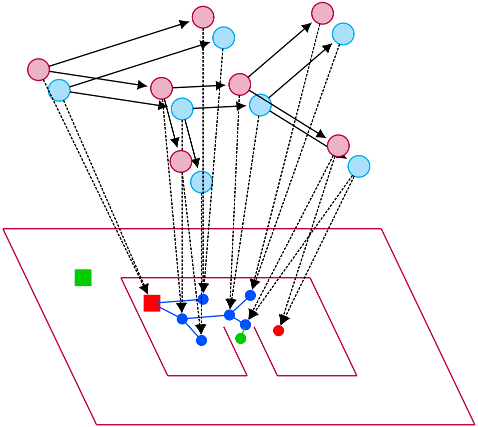


Sampling-based planning with language



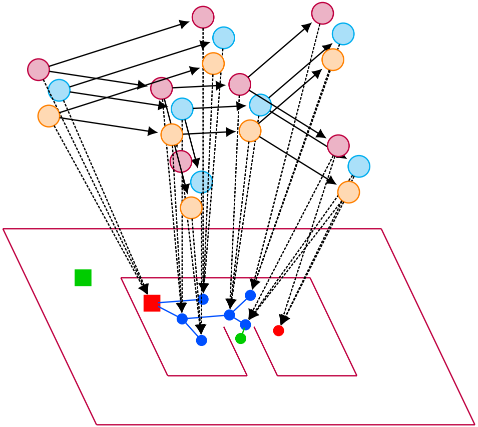
Kuo, Katz, Barbu 2018; Kuo, Katz, Barbu 2020

Sampling-based planning with language



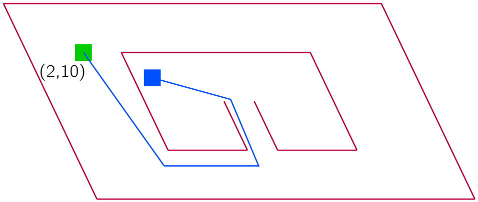
Kuo, Katz, Barbu 2018; Kuo, Katz, Barbu 2020

Sampling-based planning with language



Kuo, Katz, Barbu 2018; Kuo, Katz, Barbu 2020

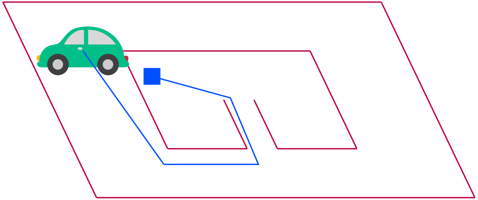
Planning and language



Find a short plan to get to (2, 10).

Kuo, Katz, Barbu 2020

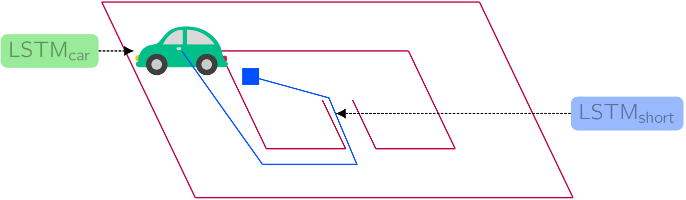
Planning and language



Go to the car quickly.

Kuo, Katz, Barbu 2020

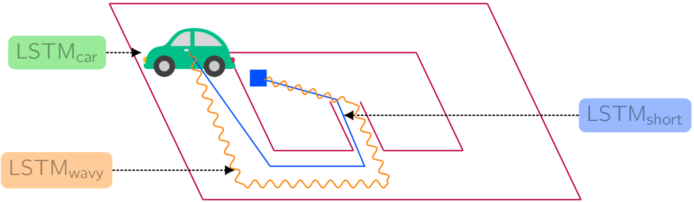
Planning and language



Go to the car quickly.

Kuo, Katz, Barbu 2020

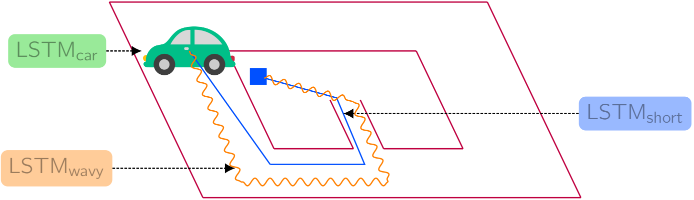
Planning and language



Weave to the green car.

Kuo, Katz, Barbu 2020

Planning and language



Weave to the green car.

Kuo, Katz, Barbu 2020

Turn parses into networks that encode sentences

Kuo, Katz, Barbu 2020

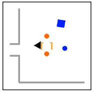
Turn parses into networks that encode sentences

Pick up the black triangle below the orange ball.

Kuo, Katz, Barbu 2020

Turn parses into networks that encode sentences

Pick up the black triangle below the orange ball.

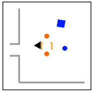


Kuo, Katz, Barbu 2020

Turn parses into networks that encode sentences

Pick up the black triangle below the orange ball.

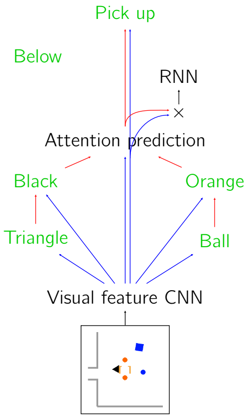
Visual feature CNN



Kuo, Katz, Barbu 2020

Turn parses into networks that encode sentences

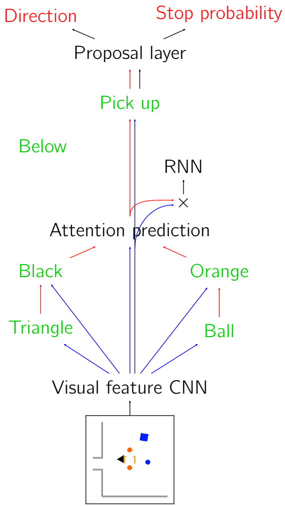
Pick up the black triangle below the orange ball.



Kuo, Katz, Barbu 2020

Turn parses into networks that encode sentences

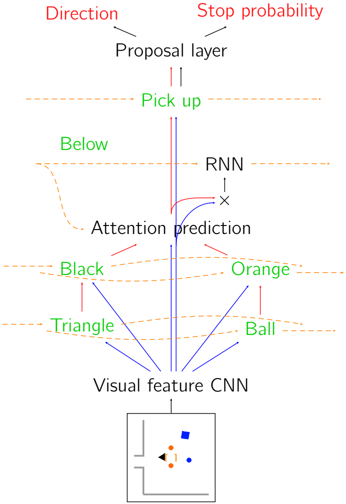
Pick up the black triangle below the orange ball.



Kuo, Katz, Barbu 2020

Turn parses into networks that encode sentences

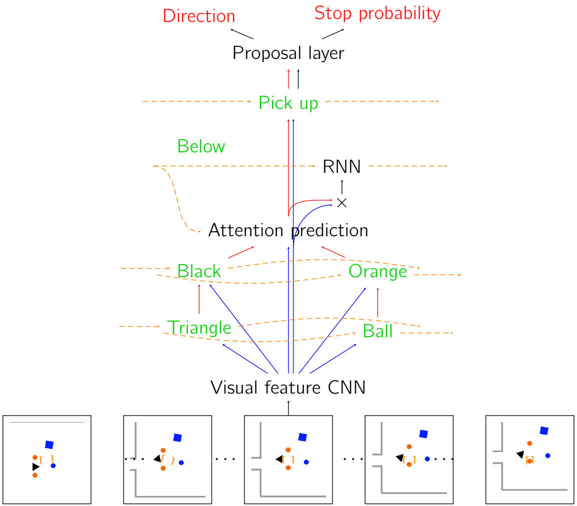
Pick up the black triangle below the orange ball.



Kuo, Katz, Barbu 2020

Turn parses into networks that encode sentences

Pick up the black triangle below the orange ball.



Kuo, Katz, Barbu 2020

Compositionality matters, but how is unclear

Planner accuracy for different models

Same number of parameters, same implementation, same optimizer, same hyperparameters

Kuo, Katz, Barbu in review

Compositionality matters, but how is unclear

Planner accuracy for different models

Same number of parameters, same implementation, same optimizer, same hyperparameters

Dataset	RNN	Compositional RNN	Compositional RNN (bad tree)
gSCAN in domain	97%	96%	96%
gSCAN out of domain	54%	96%	58%
gSCAN target length 15	95%	93%	
gSCAN target length 16	19%	91%	
gSCAN target length 17	1%	88%	
gSCAN target length 18	$\approx$0%	57%	

Kuo, Katz, Barbu in review

Compositionality matters, but how is unclear

Planner accuracy for different models
Same number of parameters, same implementation, same optimizer, same hyperparameters

Dataset	RNN	Compositional RNN	Compositional RNN (bad tree)
gSCAN in domain	97%	96%	96%
gSCAN out of domain	54%	96%	58%
gSCAN target length 15	95%	93%	
gSCAN target length 16	19%	91%	
gSCAN target length 17	1%	88%	
gSCAN target length 18	≈0%	57%	

Our compositional network is robust to new combinations
But why are we robust to longer sequences?

Kuo, Katz, Barbu in review

What theories can we generalize to?

Pick up the box.

Propositional Logic

What theories can we generalize to?

Pick up the box.	Propositional Logic
Pick up all the boxes.	FOL

What theories can we generalize to?

Pick up the box.	Propositional Logic
Pick up all the boxes.	FOL
Every time a box falls on the ground, pick it up.	LTL, STL, etc.

What theories can we generalize to?

Pick up the box.	Propositional Logic
Pick up all the boxes.	FOL
Every time a box falls on the ground, pick it up.	LTL, STL, etc.
If the box is stacked precariously, fix it.	Physics

What theories can we generalize to?

Pick up the box.	Propositional Logic
Pick up all the boxes.	FOL
Every time a box falls on the ground, pick it up.	LTL, STL, etc.
If the box is stacked precariously, fix it.	Physics
If someone wants to drop the box, stop them.	Inverse planning, ToM

What theories can we generalize to?

Pick up the box.	Propositional Logic
Pick up all the boxes.	FOL
Every time a box falls on the ground, pick it up.	LTL, STL, etc.
If the box is stacked precariously, fix it.	Physics
If someone wants to drop the box, stop them.	Inverse planning, ToM
If the box is about to fall, catch it.	Possibility, Modal logic

What theories can we generalize to?

Pick up the box.	Propositional Logic
Pick up all the boxes.	FOL
Every time a box falls on the ground, pick it up.	LTL, STL, etc.
If the box is stacked precariously, fix it.	Physics
If someone wants to drop the box, stop them.	Inverse planning, ToM
If the box is about to fall, catch it.	Possibility, Modal logic
Pick up the biggest box.	Scalar implicatures

What theories can we generalize to?

Pick up the box.	Propositional Logic
Pick up all the boxes.	FOL
Every time a box falls on the ground, pick it up.	LTL, STL, etc.
If the box is stacked precariously, fix it.	Physics
If someone wants to drop the box, stop them.	Inverse planning, ToM
If the box is about to fall, catch it.	Possibility, Modal logic
Pick up the biggest box.	Scalar implicatures
Get the box that won't leak.	Modification

What theories can we generalize to?

Pick up the box.	Propositional Logic
Pick up all the boxes.	FOL
Every time a box falls on the ground, pick it up.	LTL, STL, etc.
If the box is stacked precariously, fix it.	Physics
If someone wants to drop the box, stop them.	Inverse planning, ToM
If the box is about to fall, catch it.	Possibility, Modal logic
Pick up the biggest box.	Scalar implicatures
Get the box that won't leak.	Modification
Show me the shiny side.	Grounding

What theories can we generalize to?

Pick up the box.	Propositional Logic
Pick up all the boxes.	FOL
Every time a box falls on the ground, pick it up.	LTL, STL, etc.
If the box is stacked precariously, fix it.	Physics
If someone wants to drop the box, stop them.	Inverse planning, ToM
If the box is about to fall, catch it.	Possibility, Modal logic
Pick up the biggest box.	Scalar implicatures
Get the box that won't leak.	Modification
Show me the shiny side.	Grounding
Be friendly	Social interactions

What theories can we generalize to?

Pick up the box.	Propositional Logic
Pick up all the boxes.	FOL
Every time a box falls on the ground, pick it up.	LTL, STL, etc.
If the box is stacked precariously, fix it.	Physics
If someone wants to drop the box, stop them.	Inverse planning, ToM
If the box is about to fall, catch it.	Possibility, Modal logic
Pick up the biggest box.	Scalar implicatures
Get the box that won't leak.	Modification
Show me the shiny side.	Grounding
Be friendly	Social interactions

...

What theories can we generalize to?

Pick up the box.	Propositional Logic
Pick up all the boxes.	FOL
Every time a box falls on the ground, pick it up.	LTL, STL, etc.
If the box is stacked precariously, fix it.	Physics
If someone wants to drop the box, stop them.	Inverse planning, ToM
If the box is about to fall, catch it.	Possibility, Modal logic
Pick up the biggest box.	Scalar implicatures
Get the box that won't leak.	Modification
Show me the shiny side.	Grounding
Be friendly	Social interactions

...

You can't just hope to generalize between these domains!

But in the real world . . .

But in the real world . . .

Vision is integral to grounded language acquisition

Torralba & Efros, 2011; Berzak *et al.*, 2017

Vision is integral to grounded language acquisition

Machine performance on ImageNet is around 97%

Torralba & Efros, 2011; Berzak *et al.*, 2017

Vision is integral to grounded language acquisition

Machine performance on ImageNet is around 97%
Human-level performance on ImageNet is around 94%

Torralba & Efros, 2011; Berzak *et al.*, 2017

Vision is integral to grounded language acquisition

Machine performance on ImageNet is around 97%
Human-level performance on ImageNet is around 94%
Machines outperform humans according standard metrics!

Torralba & Efros, 2011; Berzak *et al.*, 2017

Vision is integral to grounded language acquisition

Machine performance on ImageNet is around 97%
Human-level performance on ImageNet is around 94%
Machines outperform humans according standard metrics!

Performance on datasets is not predictive of real-world performance.

Torralba & Efros, 2011; Berzak *et al.*, 2017

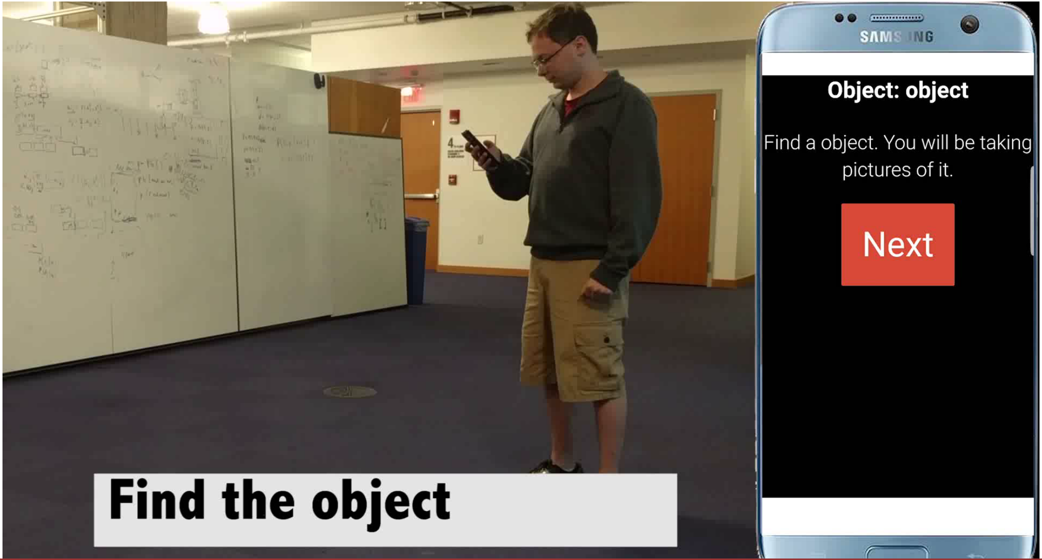
Vision is integral to grounded language acquisition

Machine performance on ImageNet is around 97%
Human-level performance on ImageNet is around 94%
Machines outperform humans according standard metrics!

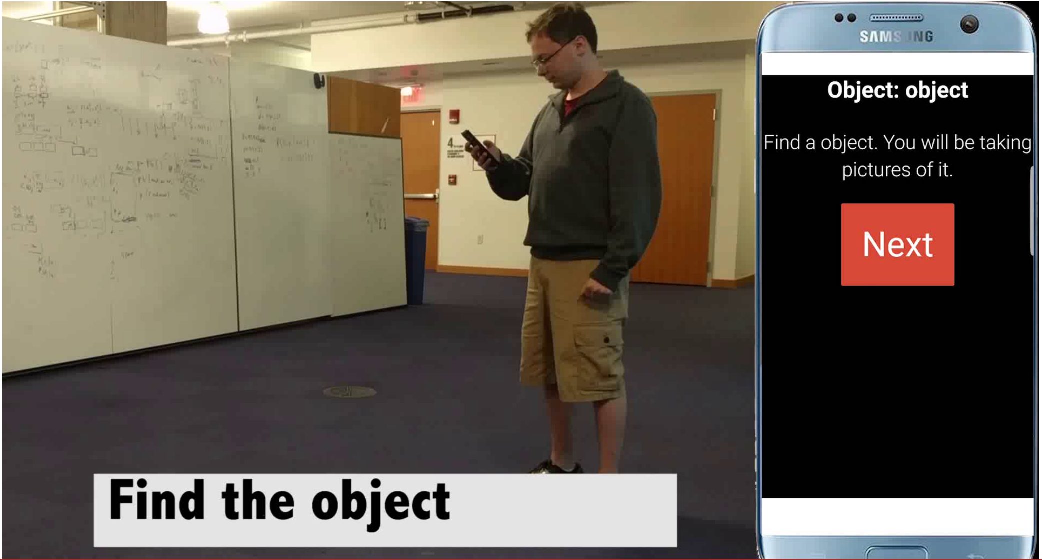
Performance on datasets is not predictive of real-world performance.
Or even performance on other datasets!

Torralba & Efros, 2011; Berzak *et al.*, 2017

Data collection in action



Data collection in action



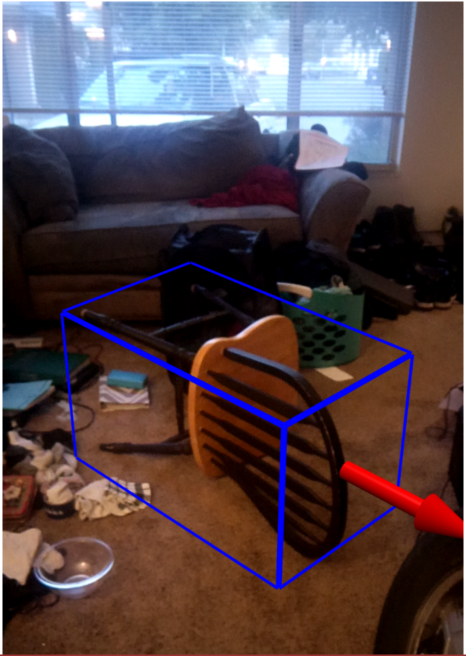
313 object classes

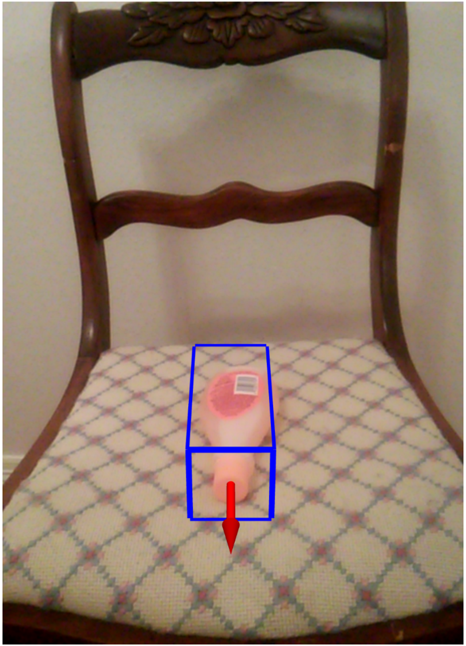
313 object classes 50k images

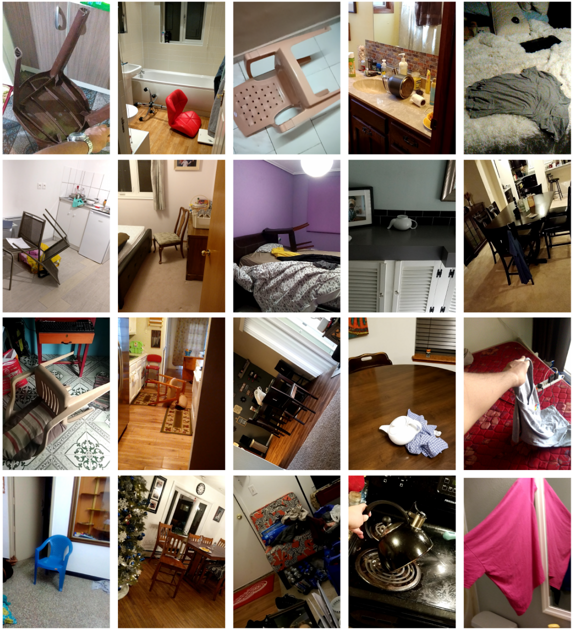
313 object classes 50k images No training set!

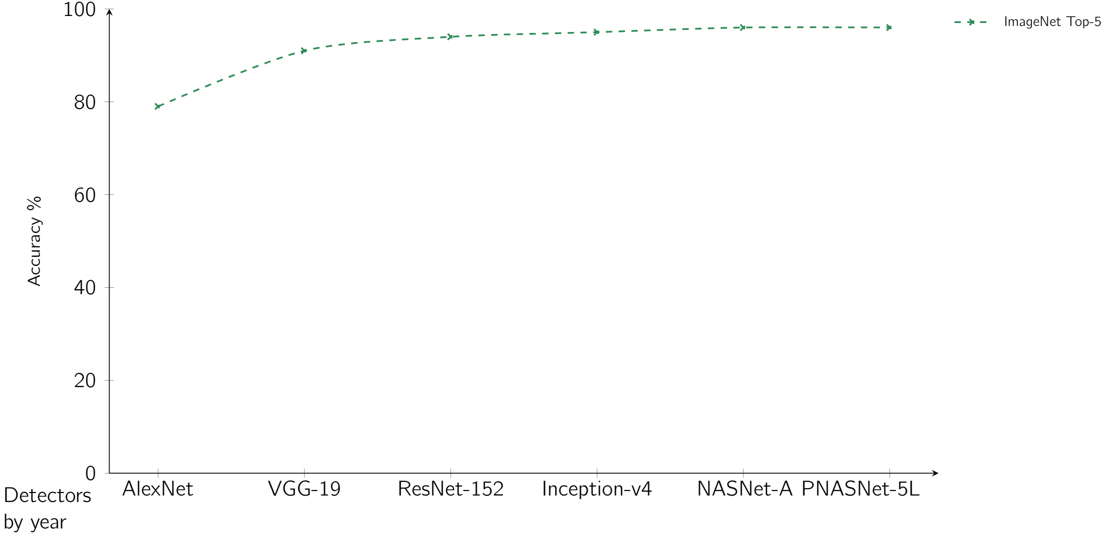
313 object classes 50k images No training set!

313 object classes 50k images No training set!

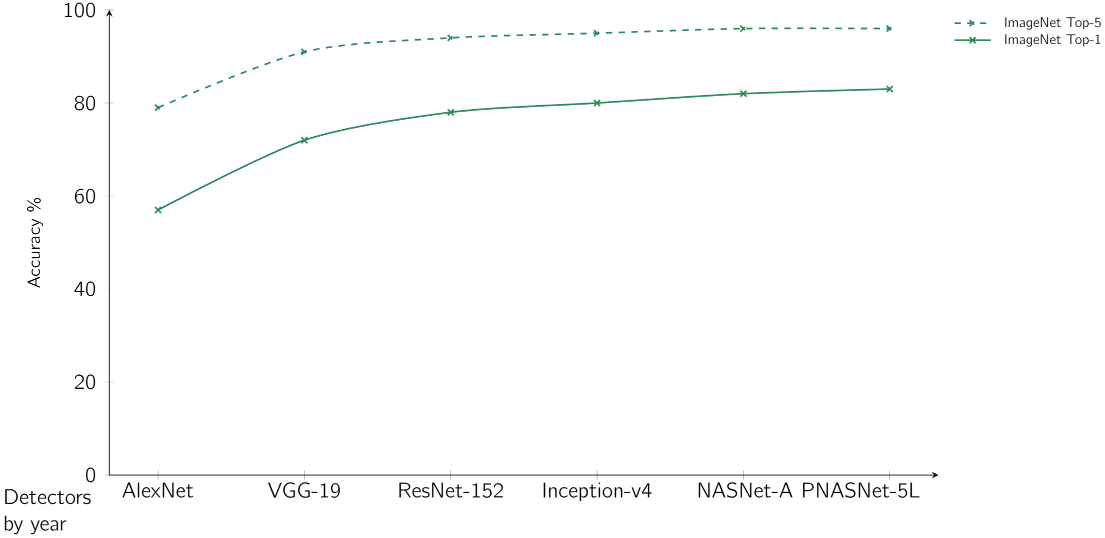




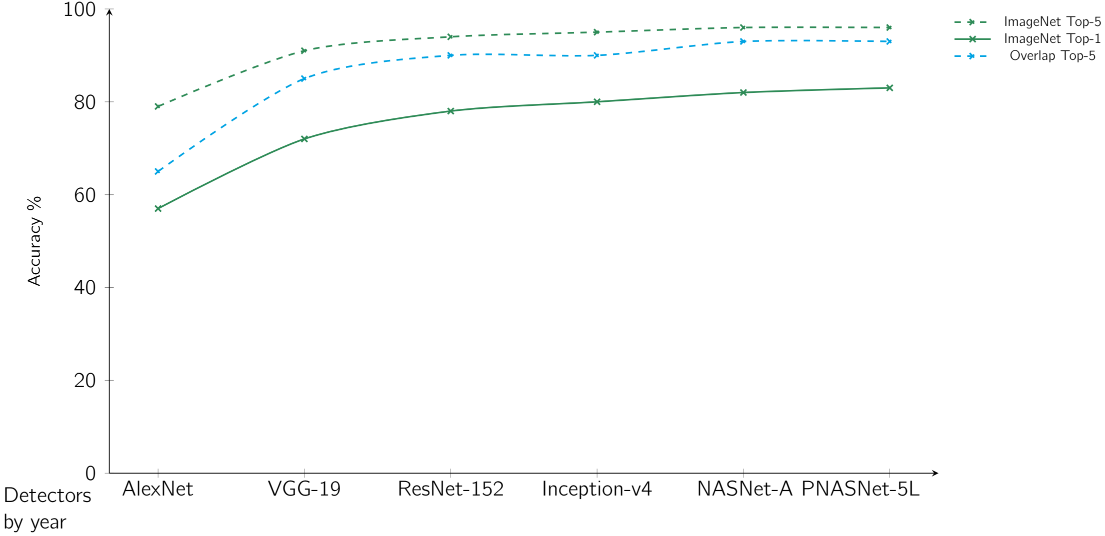




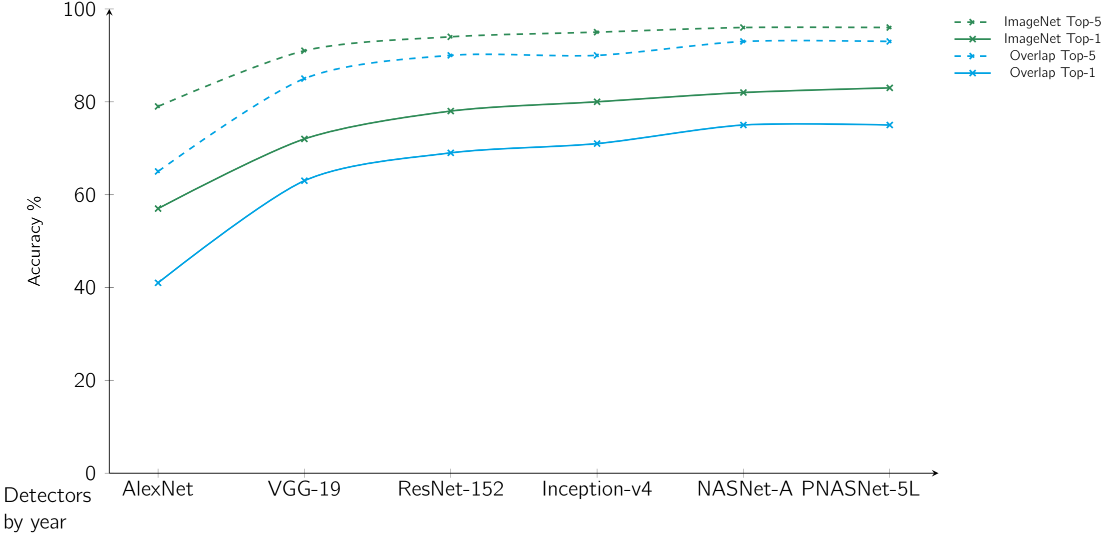
Barbu, Mayo, Alverio, Luo, Wang, Gutfreund, Tenenbaum, Katz. NeurIPS 2019



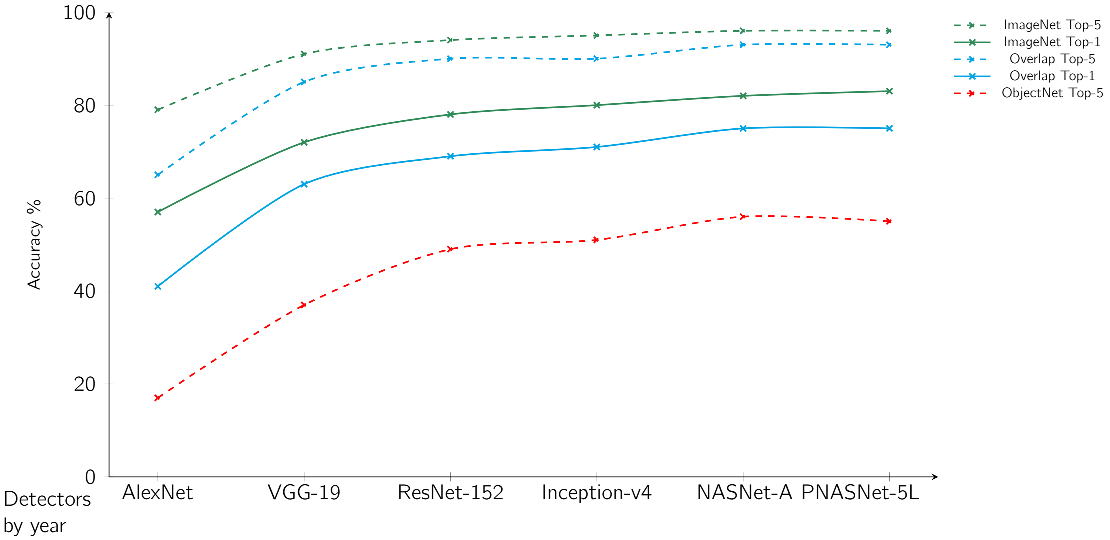
Barbu, Mayo, Alverio, Luo, Wang, Gutfreund, Tenenbaum, Katz. NeurIPS 2019



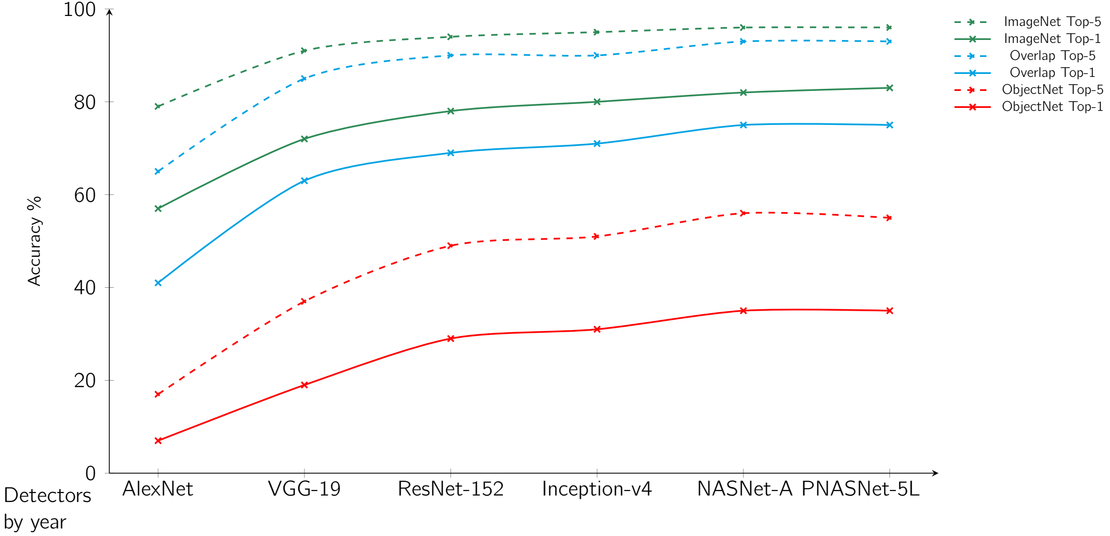
Barbu, Mayo, Alverio, Luo, Wang, Gutfreund, Tenenbaum, Katz. NeurIPS 2019



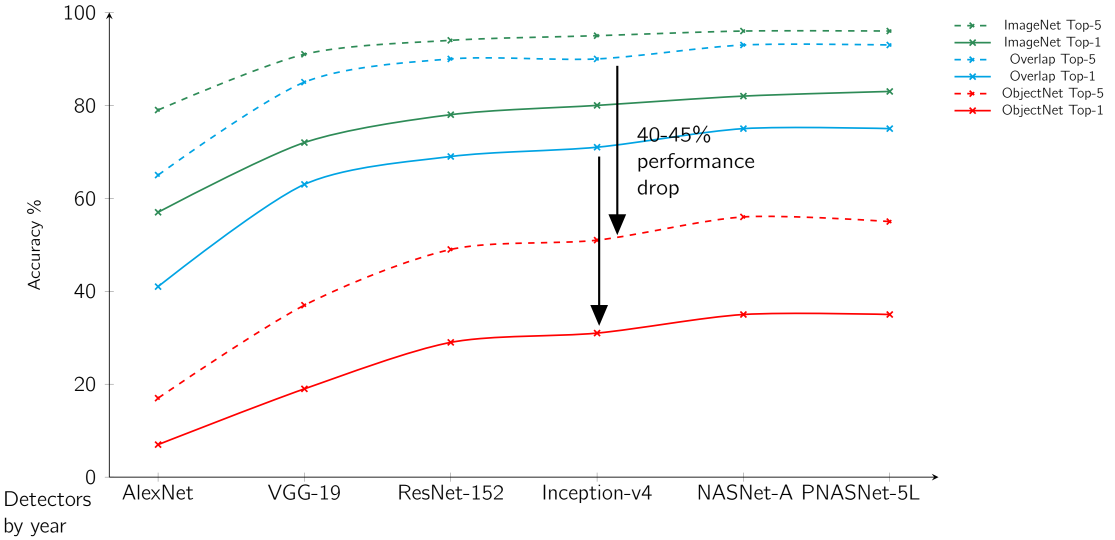
Barbu, Mayo, Alverio, Luo, Wang, Gutfreund, Tenenbaum, Katz. NeurIPS 2019



Barbu, Mayo, Alverio, Luo, Wang, Gutfreund, Tenenbaum, Katz. NeurIPS 2019



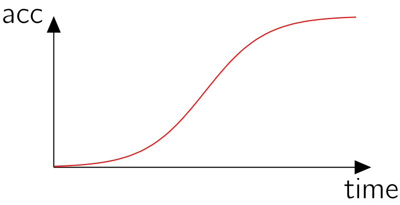
Barbu, Mayo, Alverio, Luo, Wang, Gutfreund, Tenenbaum, Katz. NeurIPS 2019



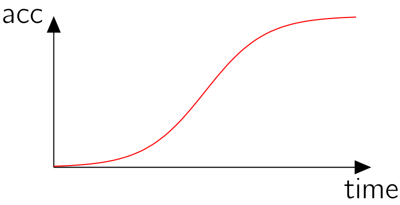
Barbu, Mayo, Alverio, Luo, Wang, Gutfreund, Tenenbaum, Katz. NeurIPS 2019

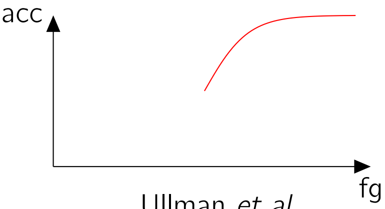
Next steps in understanding human vision

Next steps in understanding human vision

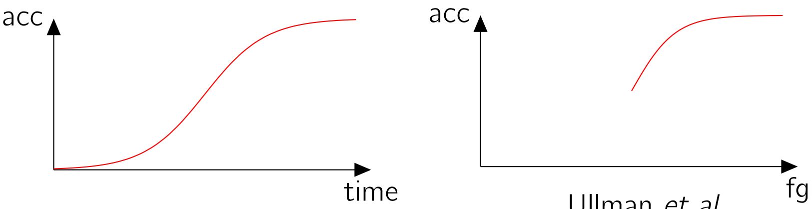


Next steps in understanding human vision





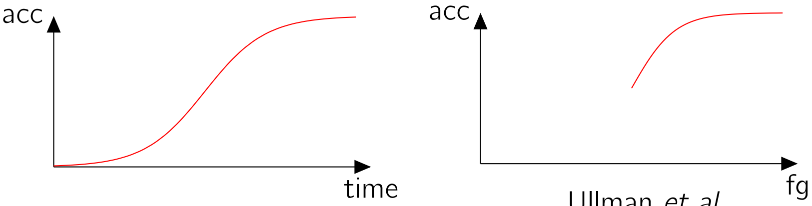
Next steps in understanding human vision



Ullman *et al.*

Accuracy(Foreground, Background, Time)

Next steps in understanding human vision

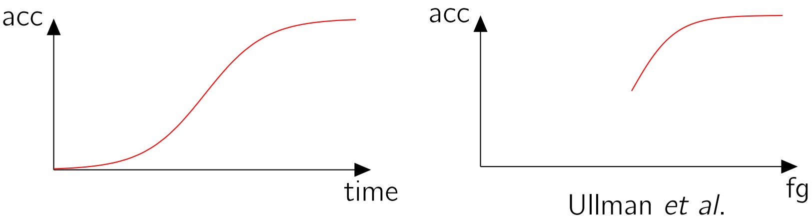


Ullman *et al.*

Accuracy(Foreground, Background, Time)

What object features (shape, texture, etc.) lead to similar response curves?

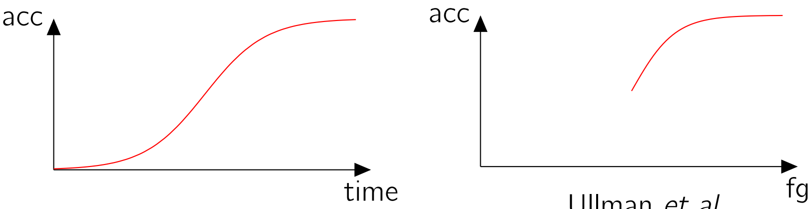
Next steps in understanding human vision



Accuracy(Foreground, Background, Time)

What object features (shape, texture, etc.) lead to similar response curves?
What are the parameters of these curves?

Next steps in understanding human vision

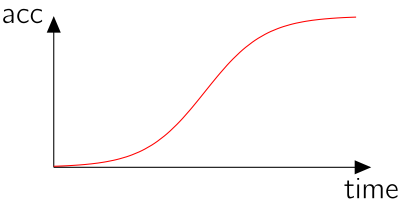


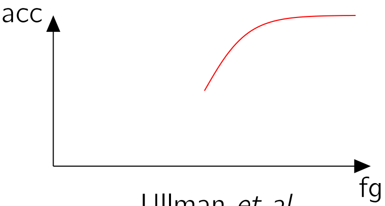
Ullman *et al.*

Accuracy(Foreground, Background, Time)

What object features (shape, texture, etc.) lead to similar response curves?
What are the parameters of these curves?
Are there any other discontinuities?

Next steps in understanding human vision

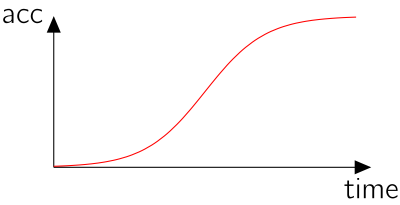


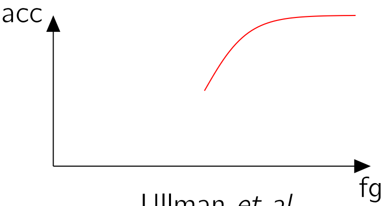


Accuracy(Foreground, Background, Time)

What object features (shape, texture, etc.) lead to similar response curves?
What are the parameters of these curves?
Are there any other discontinuities?
Can we find mode switches? (feedforward vs. feedback)

Next steps in understanding human vision





Ullman *et al.*

Accuracy(Foreground, Background, Time)

What object features (shape, texture, etc.) lead to similar response curves?
What are the parameters of these curves?
Are there any other discontinuities?
Can we find mode switches? (feedforward vs. feedback)
Can human data constrain networks?

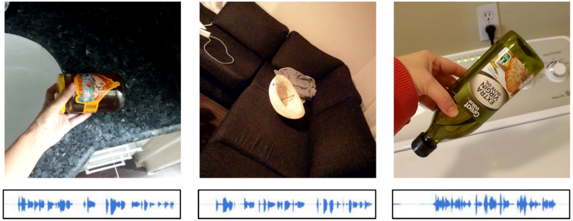
Captioning datasets are really biased . . .

Spoken ObjectNet to address this and to test out of domain generalization.

Palmer, Rouditchenko, Barbu, Katz, Glass in review

Captioning datasets are really biased . . .

Spoken ObjectNet to address this and to test out of domain generalization.



a jar of honey being held on its side of the yellow lid in front of old flowers in a vase on a bathroom sink countertop

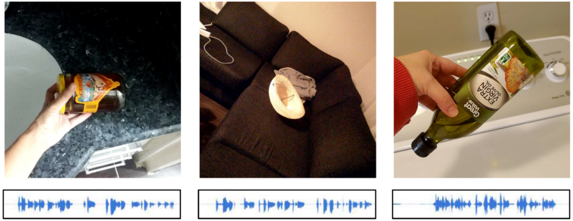
a straw like sun hat upside down on a brown sectional with a gray piece of clothing in the corner

someone holding a bottle of extra virgin olive oil upside down over a white washer with a plug-in behind it

Palmer, Rouditchenko, Barbu, Katz, Glass in review

Captioning datasets are really biased . . .

Spoken ObjectNet to address this and to test out of domain generalization.



a jar of honey being held on its side of the yellow lid in front of old flowers in a vase on a bathroom sink countertop

a straw like sun hat upside down on a brown sectional with a gray piece of clothing in the corner

someone holding a bottle of extra virgin olive oil upside down over a white washer with a plug-in behind it

On a retrieval task trained with Places-400k R@10 goes from 0.735 to 0.118!
A larger drop than for object recognition on ObjectNet

Palmer, Rouditchenko, Barbu, Katz, Glass in review

Big data neuroscience for language

Yaari, Singh, Cases, Subramaniam, Katz, Kreiman, Barbu in prep.

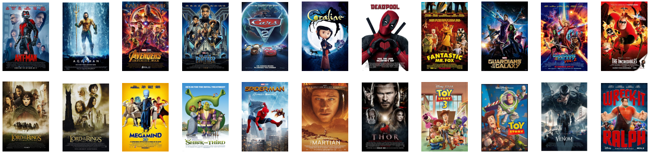
Big data neuroscience for language

Yaari, Singh, Cases, Subramaniam, Katz, Kreiman, Barbu in prep.

Big data neuroscience for language

Yaari, Singh, Cases, Subramaniam, Katz, Kreiman, Barbu in prep.

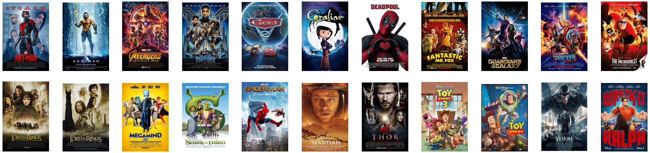
Big data neuroscience for language



Collected ≈40 hours of data while 7 subjects watched movies.

Yaari, Singh, Cases, Subramaniam, Katz, Kreiman, Barbu in prep.

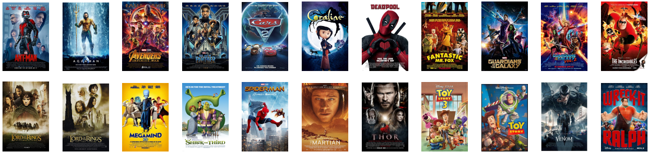
Big data neuroscience for language



Collected ≈40 hours of data while 7 subjects watched movies.
153 $\pm$ 26.6 electrodes per subject, 5 male / 2 female, 12.5 years old $\pm$ 5.23

Yaari, Singh, Cases, Subramaniam, Katz, Kreiman, Barbu in prep.

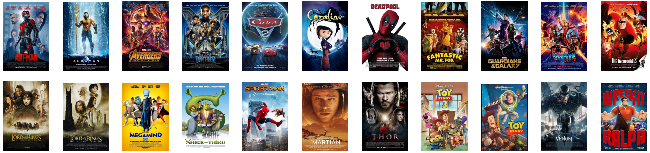
Big data neuroscience for language



Collected $\approx$40 hours of data while 7 subjects watched movies.
153 $\pm$ 26.6 electrodes per subject, 5 male / 2 female, 12.5 years old $\pm$ 5.23
1,079 electrodes, 27,981 sentences, 169,314 words

Yaari, Singh, Cases, Subramaniam, Katz, Kreiman, Barbu in prep.

Big data neuroscience for language



Collected ≈40 hours of data while 7 subjects watched movies.

153 ± 26.6 electrodes per subject, 5 male / 2 female, 12.5 years old ± 5.23

1,079 electrodes, 27,981 sentences, 169,314 words

10-100x more data per subject than previous language datasets.

Yaari, Singh, Cases, Subramaniam, Katz, Kreiman, Barbu in prep.

Some levels of linguistic analysis

I don't think Gandalf meant for us to come this way.

Some levels of linguistic analysis

Word segmentation | *I* | *don't* | *think* | *Gandalf* | *meant* | *for* | *us* | *to* | *come* | *this* | *way.* |

Some levels of linguistic analysis

Word segmentation | *I* | *don't* | *think* | *Gandalf* | *meant* | *for* | *us* | *to* | *come* | *this* | *way.* |

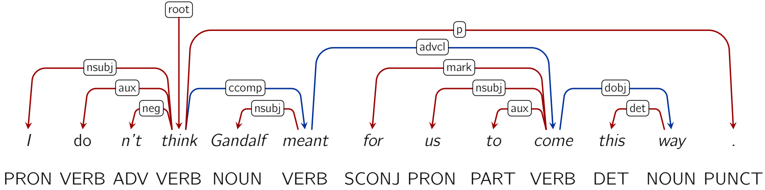
Phonemes aɪ doʊnt θɪŋk ˈgændɔlf mɛnt fər əs tə kʌm ðɪs weɪ

Some levels of linguistic analysis

Word segmentation |I|don't|think|Gandalf|meant|for|us|to|come|this|way.|

Phonemes aɪ doʊnt θɪŋk ˈgændɔlf mɛnt fər əs tə kʌm ðɪs weɪ

Parse structure

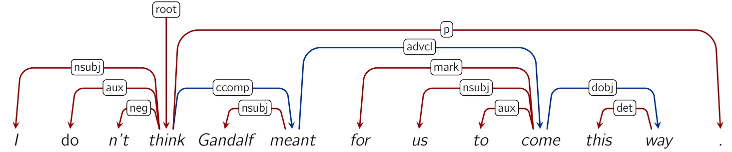


Some levels of linguistic analysis

Word segmentation | *I | don't | think | Gandalf | meant | for | us | to | come | this | way.*

Phonemes aɪ doʊnt θɪŋk ˈgændɔlf mɛnt fər əs tə kʌm ðɪs weɪ

Parse structure



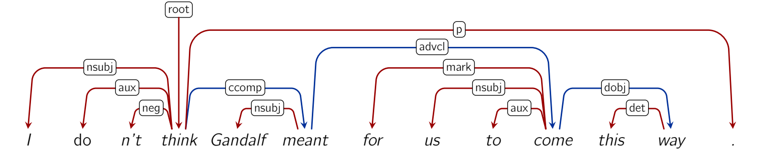
Part of speech PRON VERB ADV VERB NOUN VERB SCONJ PRON PART VERB DET NOUN PUNCT

Some levels of linguistic analysis

Word segmentation: *|I|don't|think|Gandalf|meant|for|us|to|come|this|way.|*

Phonemes: aɪ doʊnt θɪŋk ˈgændɔlf mɛnt fər əs tə kʌm ðɪs weɪ

Parse structure

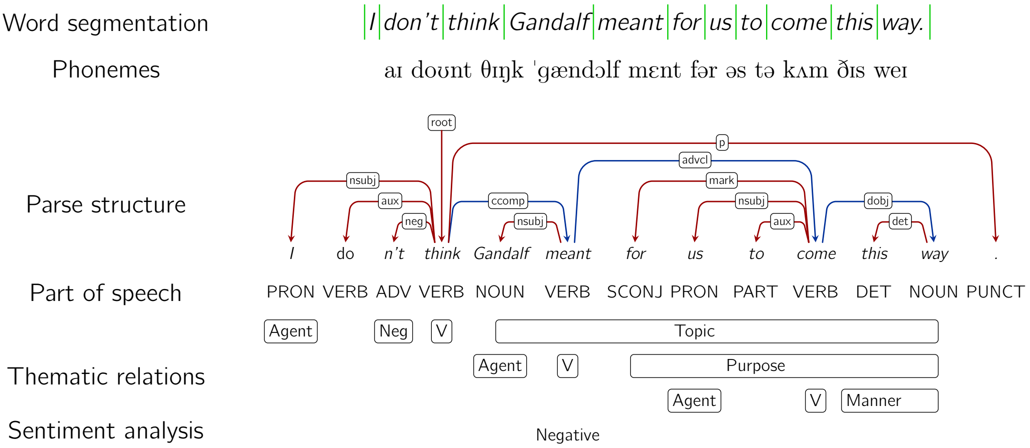


Part of speech: PRON VERB ADV VERB NOUN VERB SCONJ PRON PART VERB DET NOUN PUNCT

Thematic relations

I	do	n't	think	Gandalf	meant	for	us	to	come	this	way
Agent		Neg	V	Topic							
				Agent	V	Purpose					
							Agent		V	Manner	

Some levels of linguistic analysis

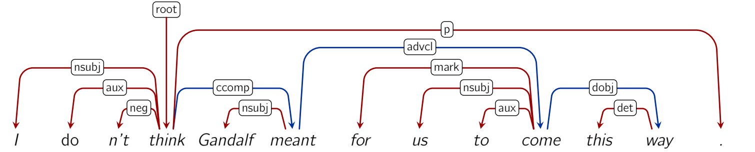


Some levels of linguistic analysis

Word segmentation: |I|don't|think|Gandalf|meant|for|us|to|come|this|way.|

Phonemes: aɪ doʊnt θɪŋk ˈgændɔlf mɛnt fər əs tə kʌm ðɪs weɪ

Parse structure



Part of speech: PRON VERB ADV VERB NOUN VERB SCONJ PRON PART VERB DET NOUN PUNCT

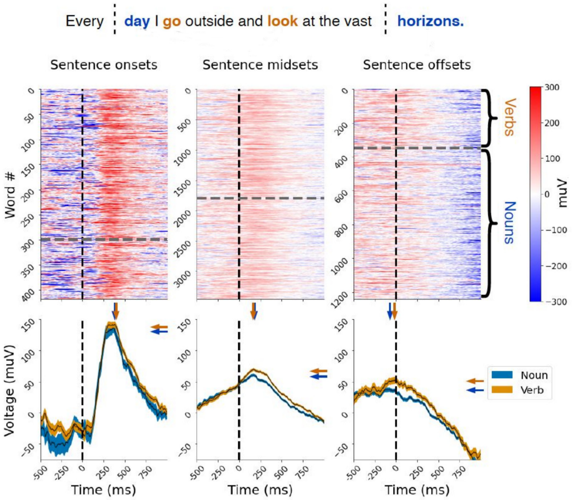
Thematic relations:

I	do	n't	think	Gandalf	meant	for	us	to	come	this	way
Agent		Neg	V	Topic							
				Agent	V	Purpose					
							Agent		V	Manner	

Sentiment analysis: Negative

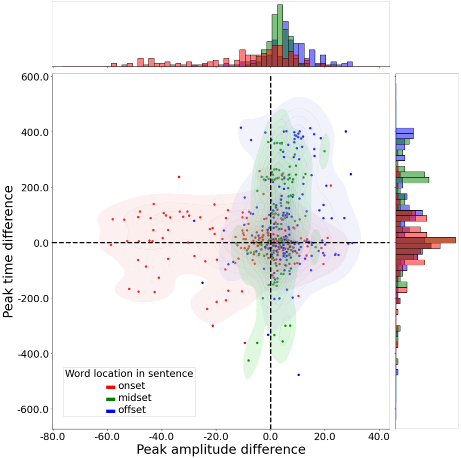
Semantics: . . .

Nouns and verbs, one electrode



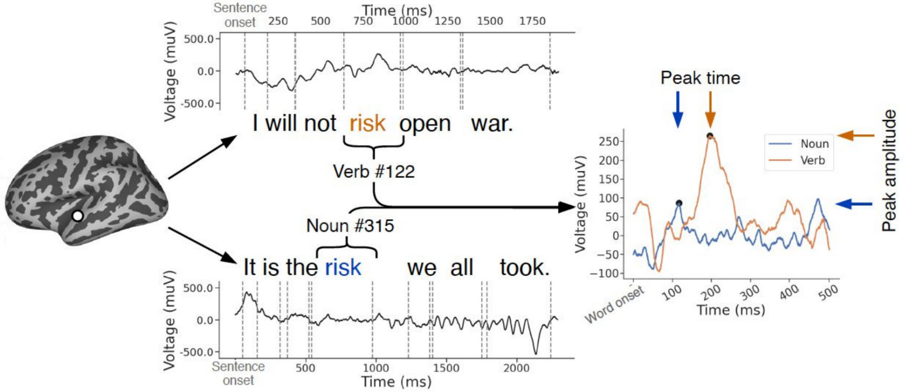
Yaari, Singh, Cases, Subramaniam, Katz, Kreiman, Barbu in prep.

Nouns and verbs, all electrodes



Yaari, Singh, Cases, Subramaniam, Katz, Kreiman, Barbu in prep.

Homonyms allow for controlled experiments

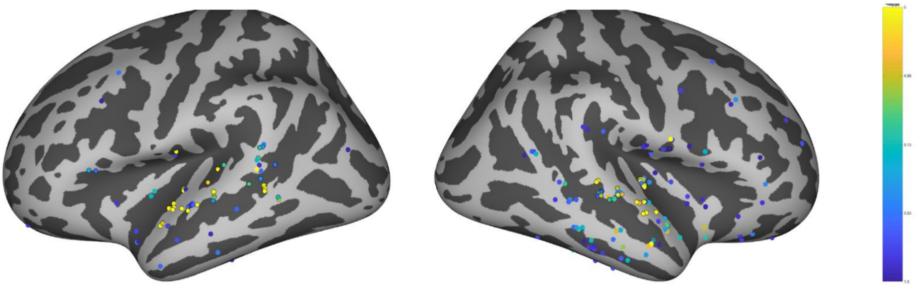


Yaari, Singh, Cases, Subramaniam, Katz, Kreiman, Barbu in prep.

Decoding nouns and verbs — homonyms

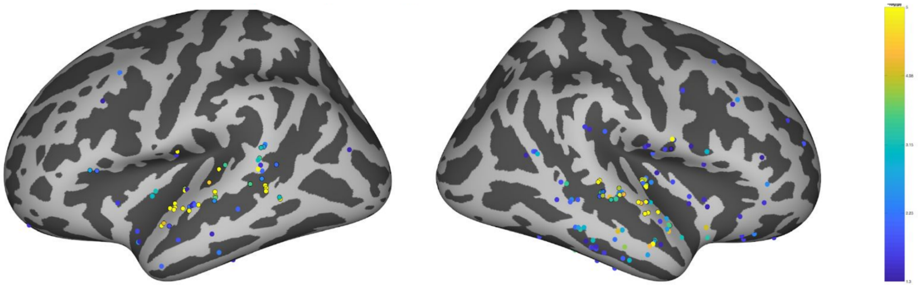
Yaari, Singh, Cases, Subramaniam, Katz, Kreiman, Barbu in prep.

Decoding nouns and verbs — homonyms



Yaari, Singh, Cases, Subramaniam, Katz, Kreiman, Barbu in prep.

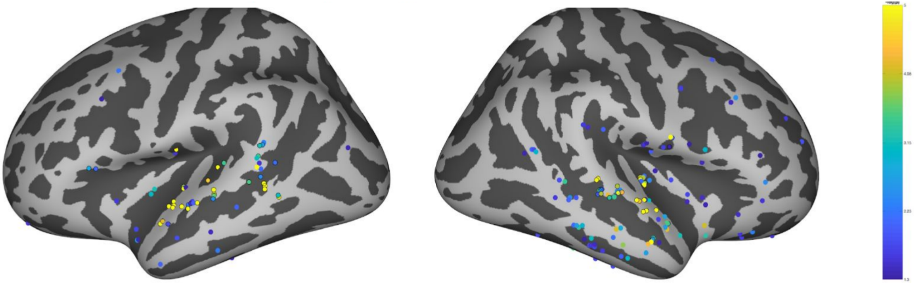
Decoding nouns and verbs — homonyms



Train on nouns and verbs, hold out all homonyms

Yaari, Singh, Cases, Subramaniam, Katz, Kreiman, Barbu in prep.

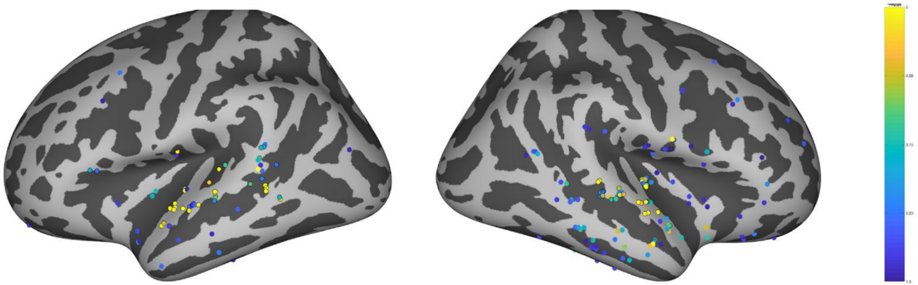
Decoding nouns and verbs — homonyms



Train on nouns and verbs, hold out all homonyms
At training time no word appears as both as noun and a verb

Yaari, Singh, Cases, Subramaniam, Katz, Kreiman, Barbu in prep.

Decoding nouns and verbs — homonyms



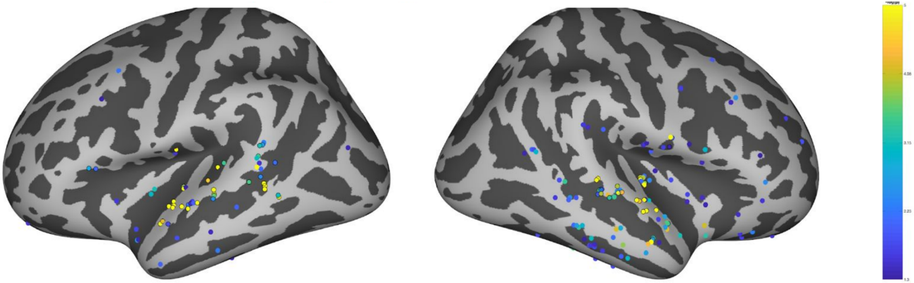
Train on nouns and verbs, hold out all homonyms

At training time no word appears as both as noun and a verb

At test time, test on only homonyms:

Yaari, Singh, Cases, Subramaniam, Katz, Kreiman, Barbu in prep.

Decoding nouns and verbs — homonyms



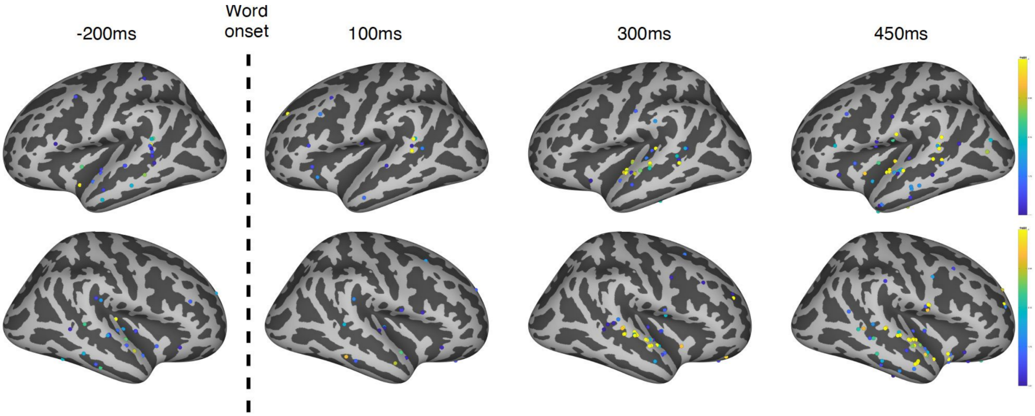
Train on nouns and verbs, hold out all homonyms
At training time no word appears as both as noun and a verb
At test time, test on only homonyms:
Generalize to unseen words and unseen word-POS pairs

Yaari, Singh, Cases, Subramaniam, Katz, Kreiman, Barbu in prep.

Decoding nouns and verbs in time

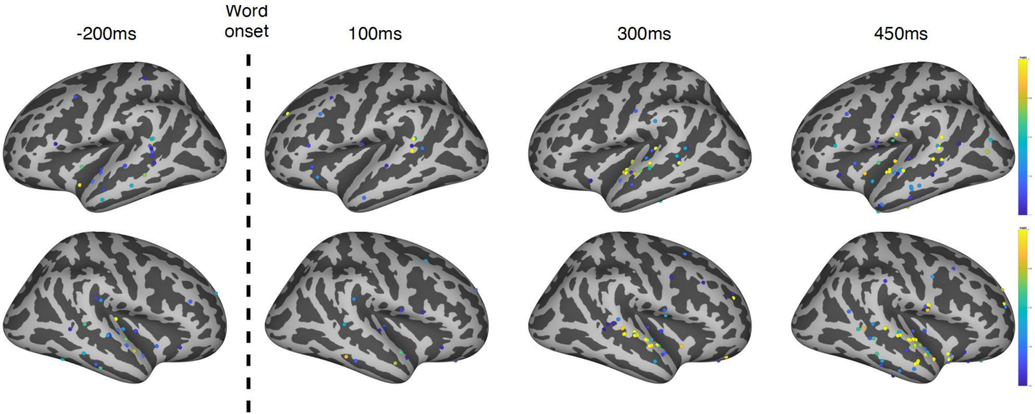
Yaari, Singh, Cases, Subramaniam, Katz, Kreiman, Barbu in prep.

Decoding nouns and verbs in time



Yaari, Singh, Cases, Subramaniam, Katz, Kreiman, Barbu in prep.

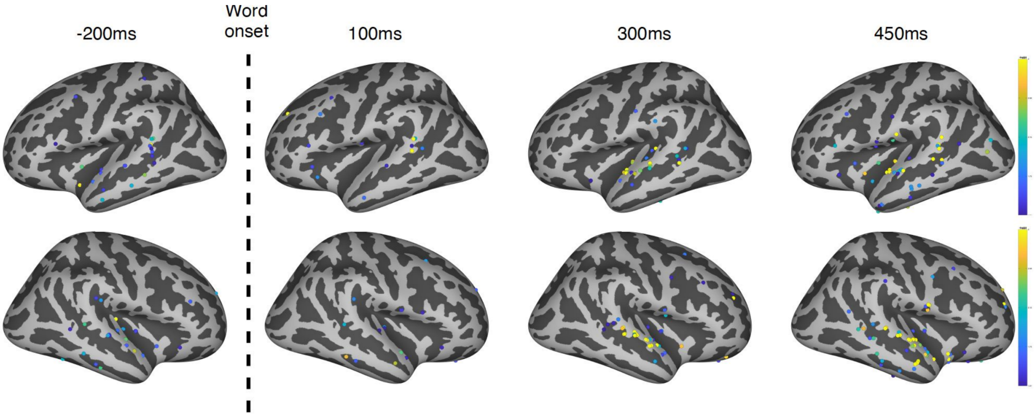
Decoding nouns and verbs in time



Well before a word is uttered you predict the POS of that word

Yaari, Singh, Cases, Subramaniam, Katz, Kreiman, Barbu in prep.

Decoding nouns and verbs in time



Well before a word is uttered you predict the POS of that word
When the word is uttered predictions are updated in STG

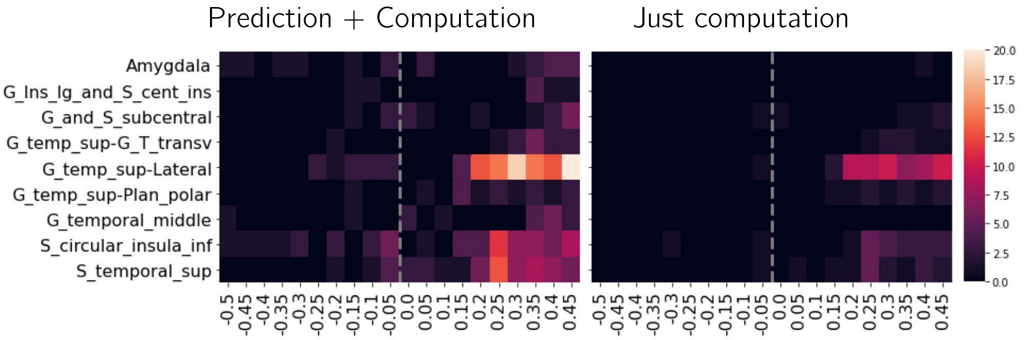
Yaari, Singh, Cases, Subramaniam, Katz, Kreiman, Barbu in prep.

Prediction vs computation

Prediction + Computation | Just computation

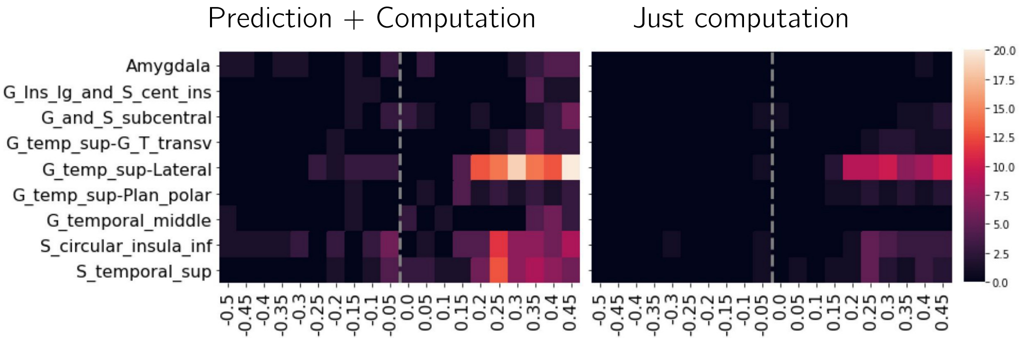
Yaari, Singh, Cases, Subramaniam, Katz, Kreiman, Barbu in prep.

Prediction vs computation



Yaari, Singh, Cases, Subramaniam, Katz, Kreiman, Barbu in prep.

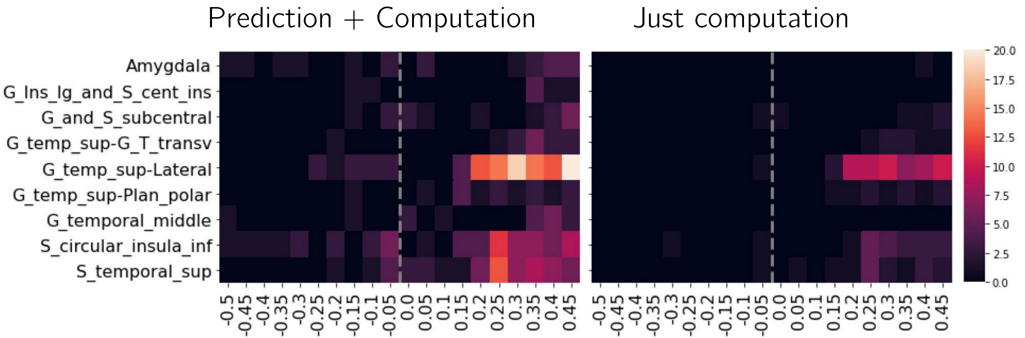
Prediction vs computation



Balance data to make previous POS not predictive about next POS

Yaari, Singh, Cases, Subramaniam, Katz, Kreiman, Barbu in prep.

Prediction vs computation



Balance data to make previous POS not predictive about next POS
Areas that predict POS are now gone, isolating POS computation

Yaari, Singh, Cases, Subramaniam, Katz, Kreiman, Barbu in prep.

The long road ahead . . .

The long road ahead . . .

Coherent stories

The long road ahead . . .

Coherent stories

3D

The long road ahead . . .

Coherent stories

3D

Physics: Forces & contact relations

Physics

Physics

The long road ahead . . .

Coherent stories

3D

Physics: Forces & contact relations

The long road ahead . . .

Coherent stories

3D

Physics: Forces & contact relations

Segmentation

The long road ahead . . .

Coherent stories

3D

Physics: Forces & contact relations

Segmentation

Parts and low-level features


Pooh

The long road ahead . . .

Coherent stories

3D

Physics: Forces & contact relations

Segmentation

Parts and low-level features

The long road ahead . . .

Coherent stories

3D

Physics: Forces & contact relations

Segmentation

Parts and low-level features

Theory of mind

The long road ahead . . .

Coherent stories

3D

Physics: Forces & contact relations

Segmentation

Parts and low-level features

Theory of mind

The long road ahead . . .

- Coherent stories
- 3D
- Physics: Forces & contact relations
- Segmentation
- Parts and low-level features
- Theory of mind
- Social understanding

Theory of mind

Theory of mind

The long road ahead . . .

Coherent stories

3D

Physics: Forces & contact relations

Segmentation

Parts and low-level features

Theory of mind

Social understanding

The long road ahead . . .

Coherent stories

3D

Physics: Forces & contact relations

Segmentation

Parts and low-level features

Theory of mind

Social understanding

Modification


Boris Katz


Shimon Ullman

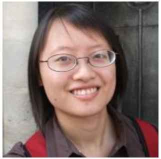

Josh Tenebaum


Gabriel Kreiman

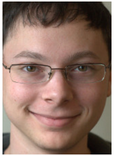

Andrei Barbu

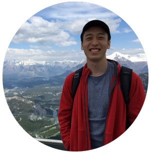

Ignacio Cases

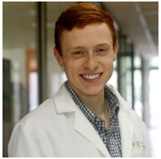

Candace Ross


Yen-Ling Kuo


Adam Yaari


David Mayo


Christopher Wang


Julian Alverio


Emily Cheng


Dylan Sleeper


Vighnesh Subramaniam


Aaditya Singh


Ravi Tejwani


Dana Rosenfarb

Recognition	$\boxed{P(\text{sentence}, \text{video})}$	Narayanaswamy *et al.* 2014
Retrieval	$\underset{v \in V}{\operatorname{argmax}} \boxed{P(s, v)}$	Barret *et al.* 2016
Generation	$\underset{s \in L}{\operatorname{argmax}} \boxed{P(s, v)}$	Yu *et al.* 2015, N. *et al.* 2014
Question answering	$\underset{s \in L}{\operatorname{argmax}} \boxed{P(\mathrm{Q}(s, s_q), v)}$	Barbu *et al.* in prep.
Disambiguation	$\underset{i \in \text{parser}(s)}{\operatorname{argmax}} \boxed{P(i, v)}$	Berzak *et al.* 2015
Language acquisition	$\underset{\theta}{\operatorname{argmax}} \prod_{s,v} \boxed{P(s(\theta), v)}$	Yu *et al.* 2015, Ross *et al.* 2018
Paraphrasing	$\int_v \lvert \boxed{P(s, v)} - \boxed{P(s', v)} \rvert$	Mao *et al.* in review
Translation	$\underset{s' \in L'}{\operatorname{argmin}} \int_v \lvert \boxed{P(s, v)} - \boxed{P(s', v)} \rvert$	Fu *et al.* in prep.
Common sense reasoning	$\underset{s \in L}{\operatorname{argmax}} \int_v \boxed{P(s_q, v) \;\; P(\mathrm{Q}(s, s_q), v)}$	
Planning	$\underset{s \in L}{\operatorname{argmax}} \int_v \boxed{P(s, v_0 : v : v_n)}$	Kuo *et al.* 2018, Kuo *et al.* 2020
Command following	$\underset{p}{\operatorname{argmax}} \int_{v^+} \boxed{P(C(s), v^+ v)}\, E(v^+, p, v)$	Paul *et al.* 2017, Kuo *et al.* in prep.

...

There are deep connections between language, vision, perception, and embodiment

There are deep connections between language, vision, perception, and embodiment
Most research is on supervised uni-modal uni-task in-domain models.

There are deep connections between language, vision, perception, and embodiment
Most research is on supervised uni-modal uni-task in-domain models.
But our brains are unlikely to be anything like those models!

There are deep connections between language, vision, perception, and embodiment

Most research is on supervised uni-modal uni-task in-domain models.

But our brains are unlikely to be anything like those models!

We need to take seriously zero-shot multi-modal multi-task out-of-domain learning.

There are deep connections between language, vision, perception, and embodiment

Most research is on supervised uni-modal uni-task in-domain models.

But our brains are unlikely to be anything like those models!

We need to take seriously zero-shot multi-modal multi-task out-of-domain learning.

The dataset crisis in vision & NLP is just a manifestation of this.

There are deep connections between language, vision, perception, and embodiment
Most research is on supervised uni-modal uni-task in-domain models.
But our brains are unlikely to be anything like those models!
We need to take seriously zero-shot multi-modal multi-task out-of-domain learning.
The dataset crisis in vision & NLP is just a manifestation of this.

These domains are all compositional and that looks to be our biggest hammer.

There are deep connections between language, vision, perception, and embodiment
Most research is on supervised uni-modal uni-task in-domain models.
But our brains are unlikely to be anything like those models!
We need to take seriously zero-shot multi-modal multi-task out-of-domain learning.
The dataset crisis in vision & NLP is just a manifestation of this.

These domains are all compositional and that looks to be our biggest hammer.
But we have no idea about why and how compositionality works.

There are deep connections between language, vision, perception, and embodiment
Most research is on supervised uni-modal uni-task in-domain models.
But our brains are unlikely to be anything like those models!
We need to take seriously zero-shot multi-modal multi-task out-of-domain learning.
The dataset crisis in vision & NLP is just a manifestation of this.

These domains are all compositional and that looks to be our biggest hammer.
But we have no idea about why and how compositionality works.
We use many crutches (like data augmentation) to avoid dealing with the theory of compositionality.

There are deep connections between language, vision, perception, and embodiment
Most research is on supervised uni-modal uni-task in-domain models.
But our brains are unlikely to be anything like those models!
We need to take seriously zero-shot multi-modal multi-task out-of-domain learning.
The dataset crisis in vision & NLP is just a manifestation of this.

These domains are all compositional and that looks to be our biggest hammer.
But we have no idea about why and how compositionality works.
We use many crutches (like data augmentation) to avoid dealing with the theory of compositionality.

Maybe flexible intelligence is possible because it's multi-modal and multi-task?

There are deep connections between language, vision, perception, and embodiment
Most research is on supervised uni-modal uni-task in-domain models.
But our brains are unlikely to be anything like those models!
We need to take seriously zero-shot multi-modal multi-task out-of-domain learning.
The dataset crisis in vision & NLP is just a manifestation of this.

These domains are all compositional and that looks to be our biggest hammer.
But we have no idea about why and how compositionality works.
We use many crutches (like data augmentation) to avoid dealing with the theory of compositionality.

Maybe flexible intelligence is possible because it's multi-modal and multi-task? Maybe our poor understanding of the brain is a consequence of building narrow models?