CSAIL
CBMM

This is a caption for the example and it may be this long or longer

Flexible Intelligence By Integrating Language and Perception

Candace Ross, Yen-Ling Kuo,
Andrei Barbu, Yevgeni Berzak, Rohan Paul
Boris Katz , Josh Tenenbaum, Nick Roy

Humans use language in extremely diverse ways

Caption, answer questions, understand a description, explain it to someone, engage in a conversation, give agents commands, imagine something different, recognize its description in a story, rewrite that description in another language, understand if someone is missing the point, reproduce it, intervene, etc.

Recognition	$P(\text{sentence}, \text{video})$
Retrieval	$\underset{v \in V}{\text{argmax}}\ P(s, v)$
Generation	$\underset{s \in L}{\text{argmax}}\ P(s, v)$
Question answering	$\underset{s \in L}{\text{argmax}}\ P(\text{Q}(s, s_q), v)$
Disambiguation	$\underset{i \in \text{parser}(s)}{\text{argmax}}\ P(i, v)$
Language acquisition	$\underset{\theta}{\text{argmax}} \prod_{s,v} P(s(\theta), v)$
Follow commands	$\underset{p}{\text{argmax}} \int_{v^+} P(C(s), v^+ v)\ E(v^+, p, v)$
Paraphrasing	$\int_v \lvert P(s, v) - P(s', v) \rvert$
Common sense reasoning	$\underset{s \in L}{\text{argmax}} \int_v P(s_q, v)\ \ P(\text{Q}(s, s_q), v)$
Planning	$\underset{s \in L}{\text{argmax}} \int_v P(s, v_0 : v : v_n)$

New tasks like visual disambiguation and zero-shot SNLI

Narayanaswamy et al. 2014 — Generated 250 sentences with 2 to 6 interpretations and filmed each. Over 2000 clips.

Barret et al. 2016

Narayanaswamy et al. 2014

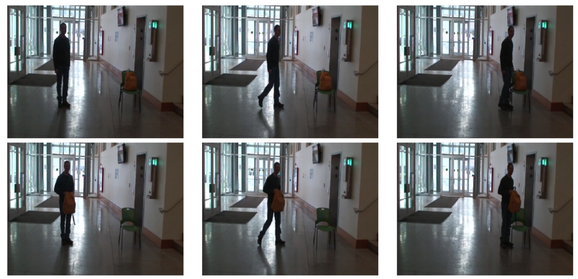
Barbu et al. in prep.

Berzak et al. 2015

Yu et al. 2015

Paul et al. 2017

Danny approached the chair with a yellow bag.

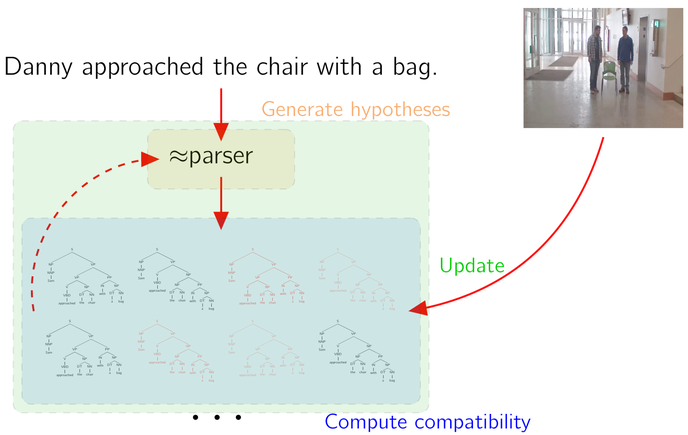


Comparing sentences vs Parikh *et al.* 2016 + ELMo

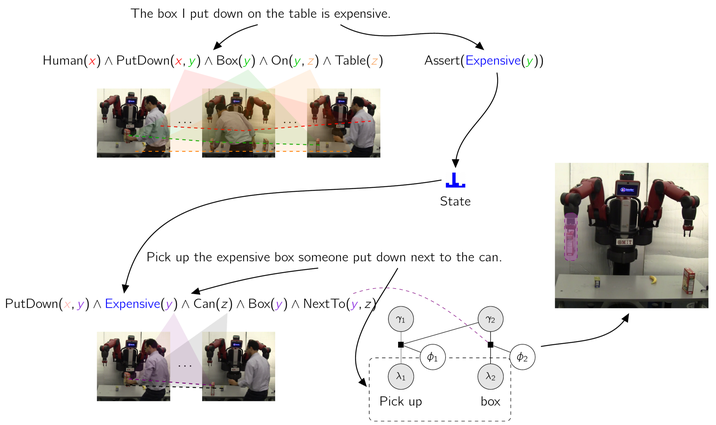
Does the sentence above imply the one below, and vice versa?

		Ground	Ours	Theirs
Alice carried the chair *Alice held the chair*	⇓ ⇑	Y N	Y N	N Y
Alice carried the chair towards Ben *Alice approached Ben*	⇓ ⇑	Y N	Y N	N Y?
Alice carried the chair towards Ben *Alice left Ben*	⇓ ⇑	N N	N N	Y? Y?
Alice picked up the chair, and Ben put down the bag *Ben picked up the chair, and Alice put down the bag*	⇓ ⇑	N N	N N	Y Y

Models of language learning that are grounded in perception



The future is in interactive models that understand intent & physics



2D and 3D velocities & distances are not invariant as the viewpoint changes.
The contact, forces, and support are invariant for many actions.