Neurosymbolic AI for Situated Language Understanding

Nikhil Krishnaswamy and James Pustejovsky

Advances in Cognitive Systems
August 10, 2020

NLP Is Everywhere!

- In the last 15 years, pipelines for natural language processing (NLP) have matured;
  - e.g., speech recognition, tokenization, parsing, etc.
- NLP in 2010s: deep learning revolution:
  - Question answering, chatbots, text generation, etc.
  - NLP in 2020: well-handled pipelines.

Natural language processing (NLP) is a subfield of linguistics, computer science, information engineering, and artificial intelligence concerned with the interactions between computers and human (natural) languages, in particular how to program computers to process and analyze large amounts of natural language data. NLP encompasses methods to manipulate and categorize textual data and information in ways similar to biological organisms. As such, NLP is an excellent use case for learning intelligent bots, as you can teach a computer how to process natural language data in order to explore, build, test, validate, and apply advanced learning methods in response to the data. Furthermore, a bot can be trained on thousands of sentences of text, with the result appearing in a natural way on a search engine like Google, as well as in a social media

Figure: GPT-2-enabled text generation, courtesy of talktotransformer.com

What's Left?

Figure: "What am I pointing at?"

Big Data Enabled the NLP Revolution

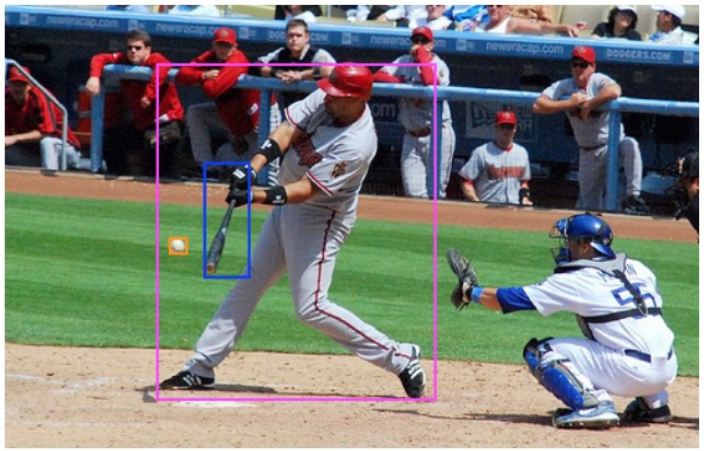
- Thanks to big data we have high-performing NLP.
- Compute power gives us usable interactive systems;
  - e.g., smartphones, personal digital assistants, smart homes, internet of things.
  - Many of these have some NLP capability in them!
- These systems exist in a situated context;
  - "A situation" (Lat. *situare*, to place) implies alternate modalities of description (e.g., configurations, deictic reference, object properties);
  - We lack: cross-modal validation, background knowledge.
- How do you interface in these scenarios if "What am I pointing at?" fails?
- This talk covers work over the last 5 years directed toward answering this question.

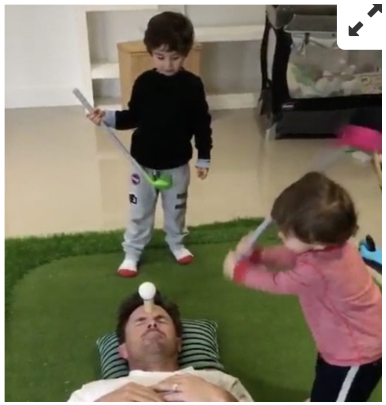
Big Data vs. Effective Modeling

- Robust communicative interaction between humans and computers requires the following three capabilities:
  1. Recognition and generation within multiple modalities, e.g., language, gesture, vision, action;
  2. Understanding of contextual grounding and co-situatedness in conversation;
  3. Appreciation of consequences of actions taken throughout the dialogue.
- Central to these is "semantically grounding" a concept to a situation;
  - Certain modalities are better at grounding certain types of information (e.g., deixis to locations, language to attributives or concept labels).

Big Data vs. Effective Modeling

- "Grounding" in NLP ≈ multimodal linking;





a group of young children
playing a game of soccer .

- *Situated grounding* entails knowledge of entities in context.

Big Data vs. Effective Modeling

- Current data doesn't solve this problem.
- What we need is context (conversational, situational, etc.).
  - Encoding context tends to be hit or miss.
- Language models might have billions of parameters (e.g., GPT-3, 175B parameters);
  - millions of those parameters could be incorrect for the task.

What Can't Data Do?

- Human reasoning is sensitive to *contextual* models
- DARPA's "3 waves of AI" as applied to contextual representation:
  - 1st wave AI: logical-symbolic models of context
    - (Lots of `if` statements)
  - 2nd wave AI: probabilistic learning of context
    - (Data-intensive vector similarity)
  - 3rd wave AI: *structured learned contextual representation*
    - "Neuro-symbolic AI" (LeCun, 2020; Cox, 2020; Kautz, 2020)
- What are the analytic units of context?
- What is the structure of context?

Common Ground

For example, take an interaction...

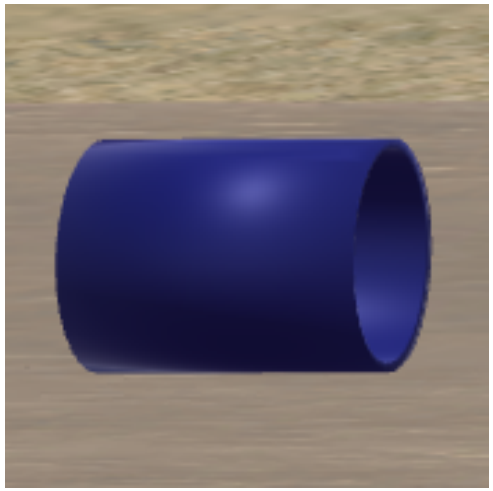
Entity Type	**Examples**
Agents	"you," "I," "us," etc.
Beliefs, desires, intentions	$\textbf{Know}_{mother}$smile(son), $\exists x \exists y \mathcal{R}(x,y)$, etc., goals under discussion
Objects	cups, plates, knives, "it," "them," etc.
Space	$\mathcal{E}$ (Embedding space)

Table: "Common ground" entities in peer-to-peer interaction

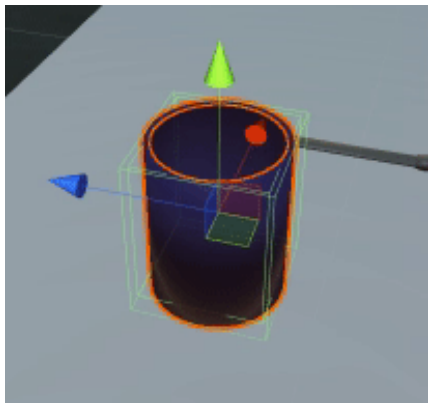
Focus on Objects

- Context of objects is described by their properties.
- Object properties cannot be decoupled from the events they facilitate.
  - *Affordances* (Gibson, 1977)

"He **slid** the cup across the table. Liquid spilled out."
"He **rolled** the cup across the table. Liquid spilled out."



VoxML



- Pustejovsky and Krishnaswamy, 2016;
- Captures common-sense object/event semantics;
- **Habitat**: Conditioning environment affecting object **affordances** (behaviors attached due to object structure or purpose);
- Ontological information difficult to learn from corpora.

VoxML

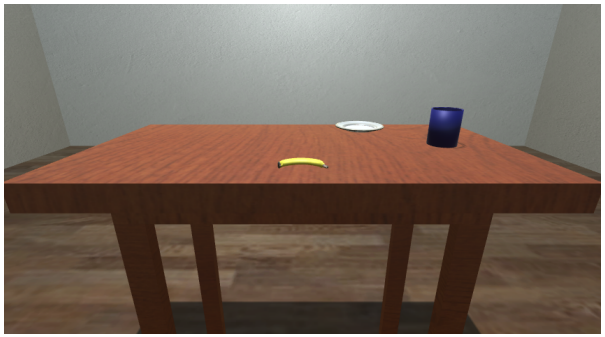
$$
\begin{bmatrix}
\textbf{cup} \\
\text{LEX} = \begin{bmatrix} \text{PRED} = \textbf{cup} \\ \text{TYPE} = \textbf{physobj} \end{bmatrix} \\
\text{TYPE} = \begin{bmatrix} \text{HEAD} = \textbf{cylindroid[1]} \\ \text{COMPONENTS} = \textbf{surface,interior} \\ \text{CONCAVITY} = \textbf{concave} \\ \text{ROTATSYM} = \{Y\} \\ \text{REFLECTSYM} = \{XY, YZ\} \end{bmatrix} \\
\text{HABITAT} = \begin{bmatrix} \text{INTR} = {}_{[2]}\begin{bmatrix} \text{UP} = align(Y, \mathcal{E}_Y) \\ \text{TOP} = top(+Y) \end{bmatrix} \\ \text{EXTR} = \ldots \end{bmatrix} \\
\text{AFFORD\_STR} = \begin{bmatrix} \text{A}_1 = H[2] \rightarrow [put(x, on([1]))] support([1], x) \\ \text{A}_2 = H[2] \rightarrow [put(x, in([1]))] contain([1], x) \\ \text{A}_3 = H[2] \rightarrow [grasp(x, [1])] \end{bmatrix} \\
\text{EMBODIMENT} = \begin{bmatrix} \text{SCALE} = \textbf{<agent} \\ \text{MOVABLE} = \textbf{true} \end{bmatrix}
\end{bmatrix}
$$

VoxML encoding for [[CUP]] object, showing habituated knowledge of object properties

Multimodal Simulations

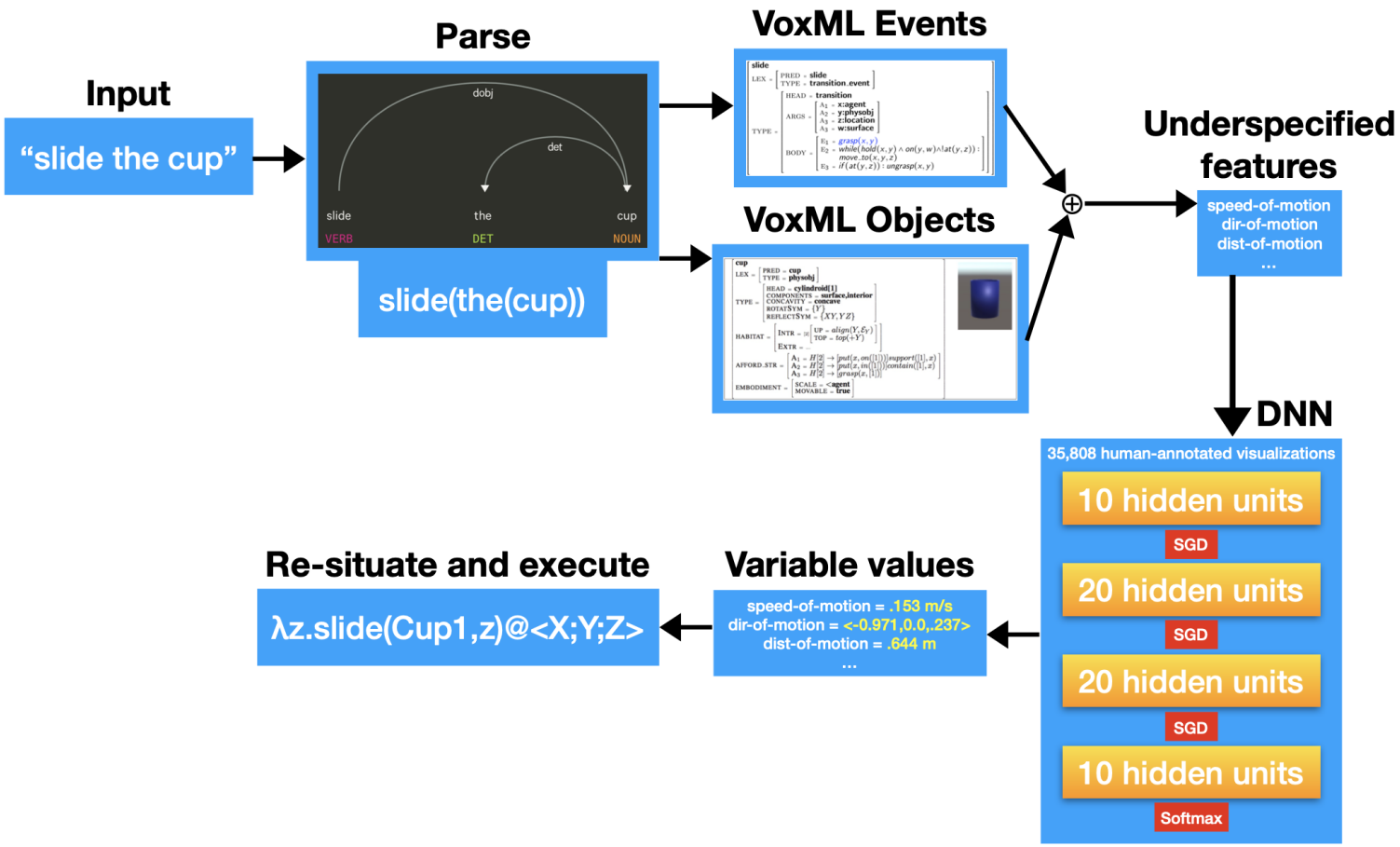
- Human understanding depends on a wealth of common-sense knowledge; humans perform much reasoning qualitatively.
- To simulate events, every parameter must have a value
  - "Roll the ball." How fast? In which direction?
  - "Roll the block." Can this be done?
  - "Roll the cup." Only possible in a certain orientation.
- VoxML: Formal semantic encoding of properties of objects, events, attributes, relations, functions.
- VoxSim: What can situated grounding do? (Krishnaswamy, 2017)
  - Exploit numerical information demanded by 3D visualization;
  - Perform qualitative reasoning about objects and events;
  - Capture semantic context often overlooked by unimodal language processing.

Multimodal Simulations



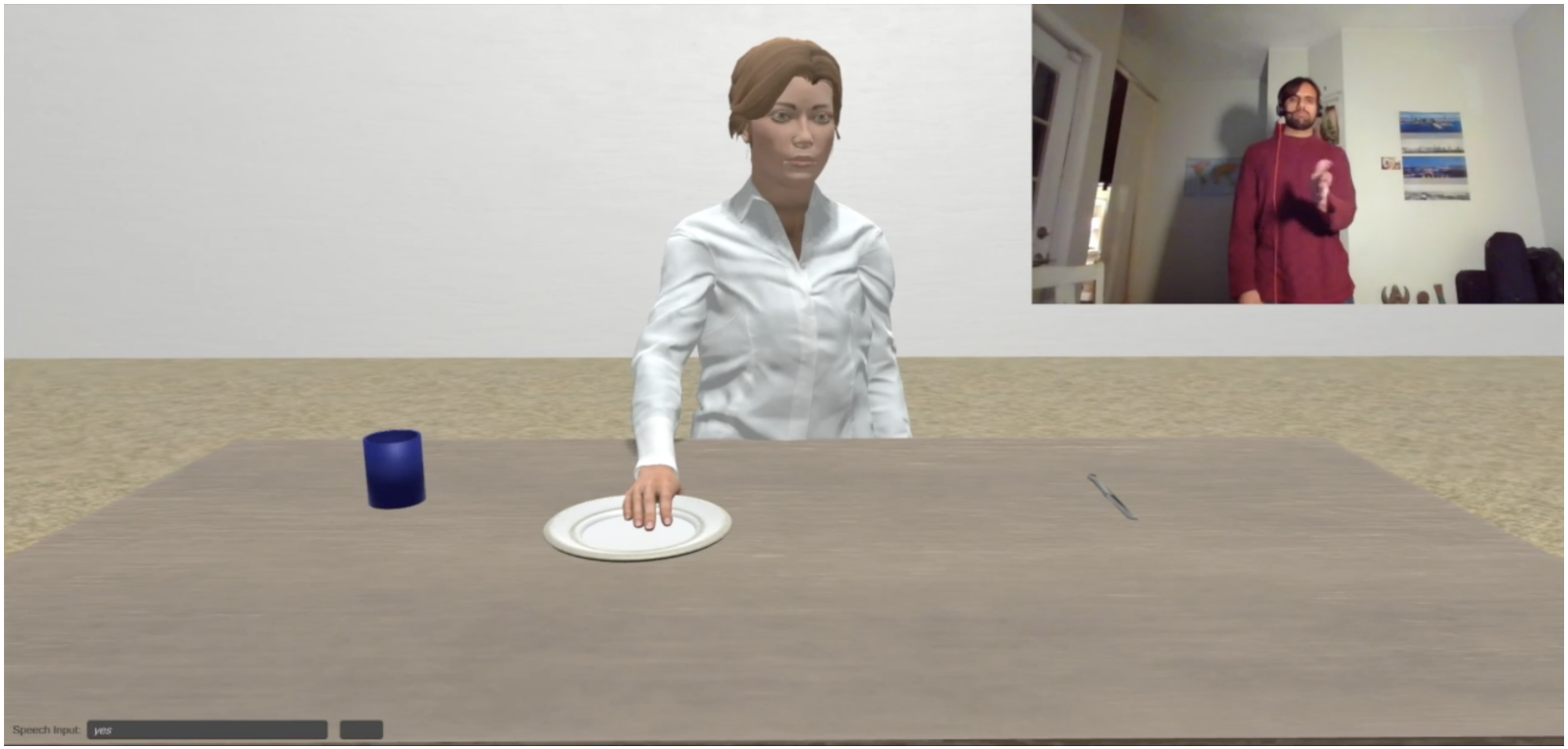
- Multiple meanings of "simulation":
  1. Parameter setting and testing in a computational model (e.g., models of physical systems);
  2. Virtual embodied interaction (e.g., video games);
  3. Language comprehension via "mental simulations" ("embodied theory of mind")
- Krishnaswamy, 2017: Brings all 3 together into Monte-Carlo visual simulation of underspecified motion predicates.

Multimodal Simulations: Neuro-Symbolic Intelligence



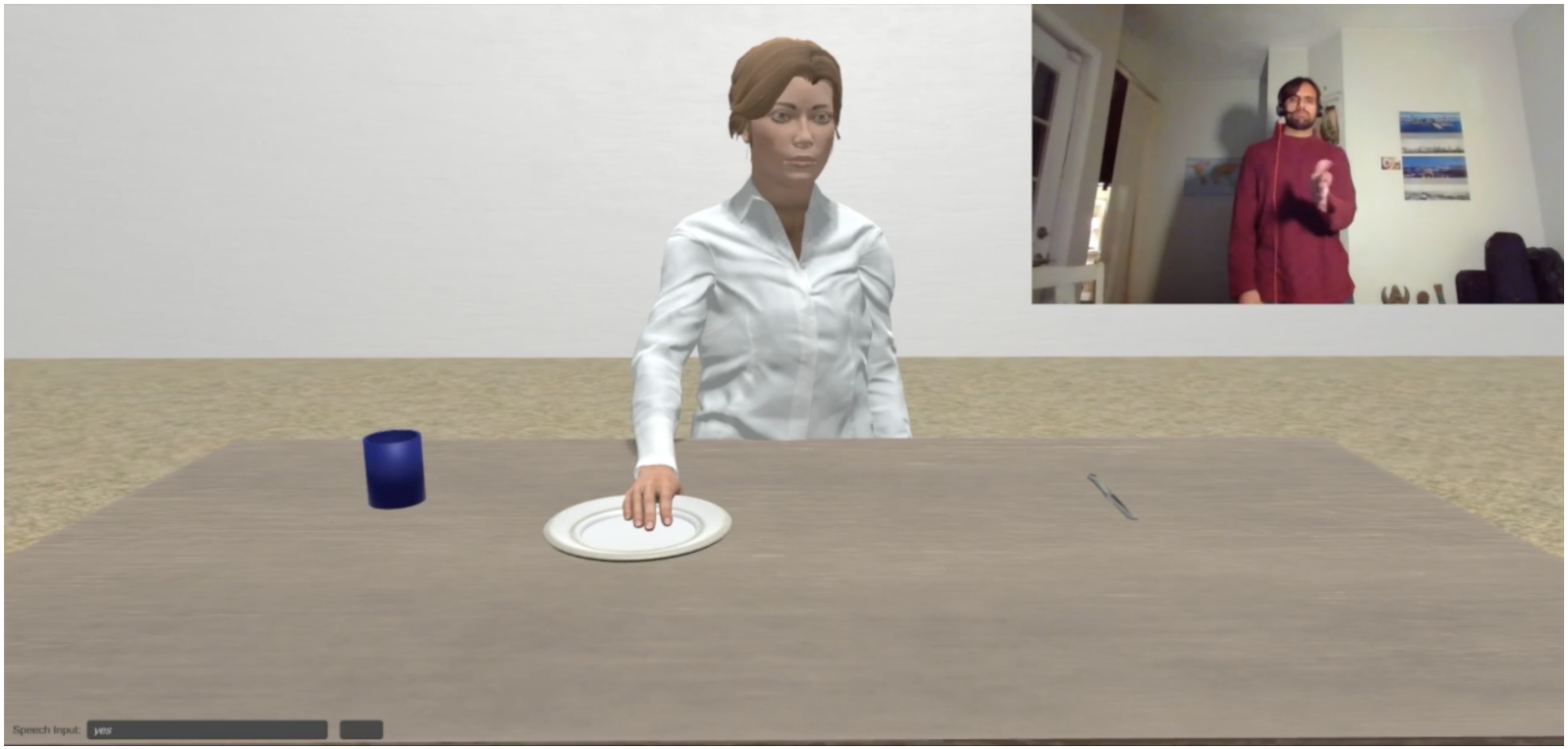
VoxWorld: AI Partners in VoxSim

Examining situatedness and situated context through an encounter between two "people" where we model multimodal dialogue.



"Diana": *D*ynamic *I*nteractive *A*sy*n*chronous *A*gent

VoxWorld: Video Demo



▸ Play!

Learning Affordances

Figure: Using iconic "plate" gesture to signal "grasp the plate"

Having learned the gesture's correlated instruction, the human can instruct the avatar to grasp an object with 1 visual cue.

Learning Affordances

- Continuation-passing style allows learned reusable associations to fill in other action sequences.

$$\left[\begin{array}{l}\textbf{slide} \\ \text{LEX} = \left[\begin{array}{l}\text{PRED} = \textbf{slide} \\ \text{TYPE} = \textbf{transition\_event}\end{array}\right] \\ \text{TYPE} = \left[\begin{array}{l}\text{HEAD} = \textbf{transition} \\ \text{ARGS} = \left[\begin{array}{l}\text{A}_1 = \textbf{x:agent} \\ \text{A}_2 = \textbf{y:physobj} \\ \text{A}_3 = \textbf{z:location} \\ \text{A}_3 = \textbf{w:surface}\end{array}\right] \\ \text{BODY} = \left[\begin{array}{l}\text{E}_1 = \mathit{grasp}(x,y) \\ \text{E}_2 = \mathit{while}(\mathit{hold}(x,y) \wedge \mathit{on}(y,w) \wedge !\mathit{at}(y,z)) : \\ \quad\quad \mathit{move\_to}(x,y,z) \\ \text{E}_3 = \mathit{if}(\mathit{at}(y,z)) : \mathit{ungrasp}(x,y)\end{array}\right]\end{array}\right]\end{array}\right]$$

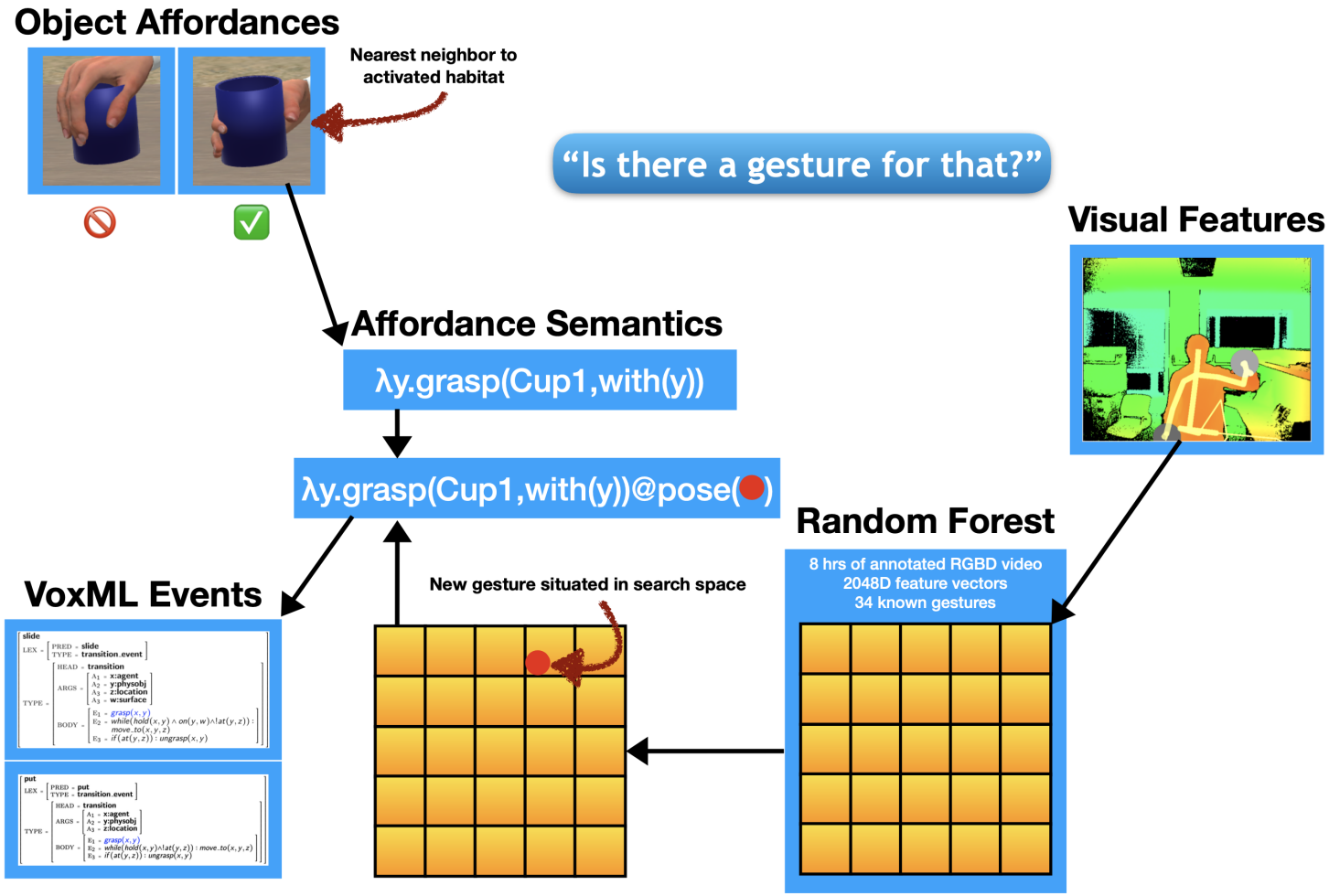
Figure: VoxML encoding for [[SLIDE]]

Learning Affordances

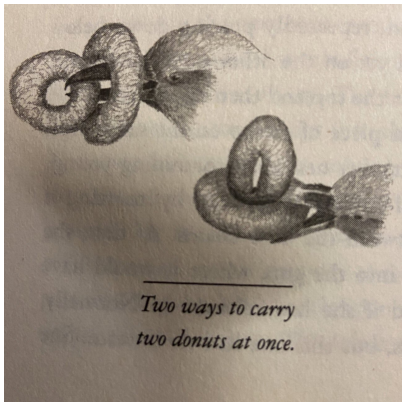
- Event encoding for [[SLIDE]] contains a [[GRASP]] subevent as precondition.
- If agent encounters context containing an action with an outstanding variable:
  - λb.slide(b,z)
  - Human supplies learned gesture for *grasp*(*plate*);
  - Directly lift subevent *grasp*(*plate*) to λb.slide(b,z);
  - Apply the argument plate to b: λb.slide(b,z)@plate $\Rightarrow$ slide(plate,z).

▸ Example

Learning Affordances: Neuro-Symbolic Intelligence

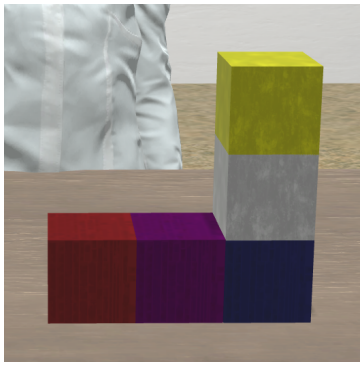
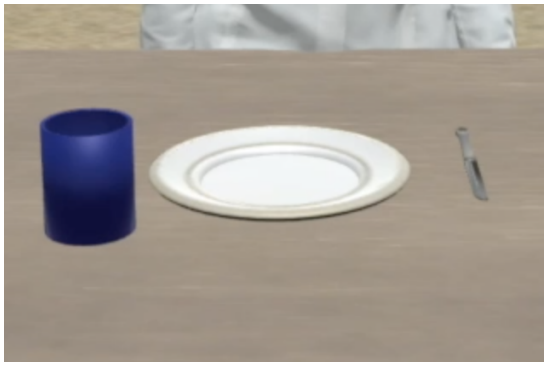


From Affordances to Configurations



- Intuition: Composing the resultant states of exploiting affordances creates configurations.
- A far more complex problem than learning affordances over single objects.
- Affordance composition increases with the complexity of the objects in the domain (naively $O(m^k)$).
- Test and development: back off to a simpler domain.

Learning Configurations

- Problem 1: Different constraints matter in different configurations.
  - Place setting: particular configuration matters;
  - Loading moving truck/holding donuts: overall configuration matters.
- Problem 2: Need representative data to train on.
  - One-shot gesture learning uses random forests over 2048D feature vectors from >8 hours of annotated RGBD video.
  - We don't have equivalent data for structure learning.

Learning Configurations: Data Gathering

Figure: "This is a staircase."

- Study data from Krishnaswamy and Pustejovsky (2018).
- 20 naïve users collaborated with Diana to build a 3-step staircase;
  - Told only that system uses natural language and gesture.
- Definition of success left up to user.
- Configuration and relative placement of the blocks varies.
- Can an algorithm infer and reproduce commonalities across a small, noisy sample?

Learning Configurations: Data Gathering

Figure: 17 samples: sparse and noisy data

- Qualitative relations *directly extractable* from situated environment (e.g., $left(x, y)$, $touching(x, y)$, $under \wedge touching \wedge support(x, y)$, etc.)
  - Subset of Region Connection Calculus (RCC) (Randell et al., 1992) and Ternary Point Configuration Calculus (TPCC) (Moratz, Nebel, and Freksa, 2002) from QSRLib (Gatsoulis et al., 2016)
  - 3D relations using RCC-3D (Albath et al., 2010) or by computing axial overlap with Separating Hyperplane Theorem

Learning Configurations: Framework (Krishnaswamy, Friedman, and Pustejovsky, 2019)

Given some sample configurations and a set of objects to place:

1. Place an object.
2. Look at current configuration, predict which known example you may be approaching (relational classification using CNN).
3. Predict remaining set of moves required to get there (sequence generation via LSTM).
4. Choose the best move (heuristic loss).
5. Repeat until out of objects.
6. Repeat 1-5.
7. Infer constraints satisfied by generated examples.

Learning Configurations: Validation

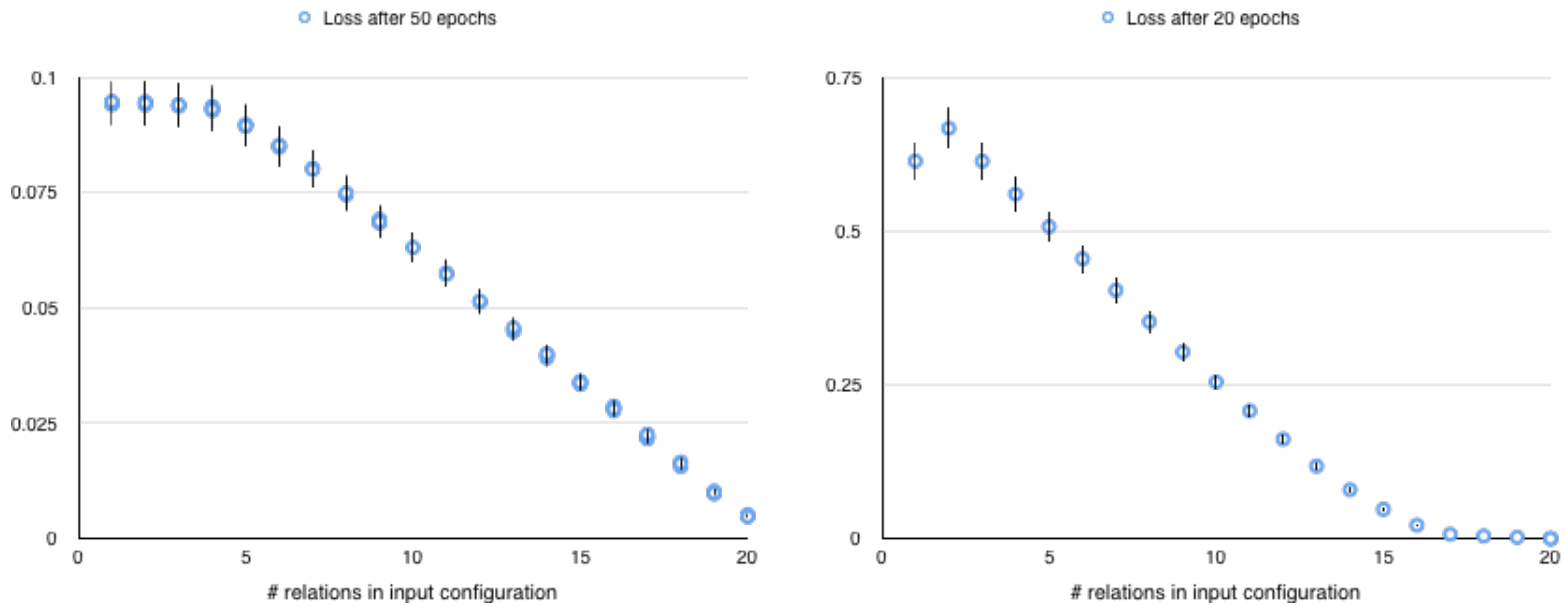


Figure: CNN (L)/LSTM (R) loss vs. input length

Classifiers are highly inaccurate at start, more accurate as more relations are created for input.

Learning Configurations: Margin-based Loss and Pruning

- Heuristics assessed for which selects the best moves toward the CNN-chosen goal state from LSTM-presented move options:
  - Random chance;
  - Jaccard distance (JD) (presence or absence of shared relation[s]);
  - Levenshtein distance (LD) (count of shared relation[s]);
  - Graph matching (SPIRE) (McLure, Friedman, and Forbus, 2015);
  - LD-pruned graph matching (Combined).

Learning Configurations: Evaluation

- 8 annotators—adult English speakers with college degree.
- "On a scale of 0-10 (10 being best), how much does the structure shown resemble a staircase?"
- No extra information provided;
  - Annotator to answer based on their particular notion of canonical staircase.
- Images viewed in random order.
- σ: standard deviation of average scores per structure generated using heuristic;
  - Lower corresponds to greater overall evaluator agreement.

Learning Configurations: Results

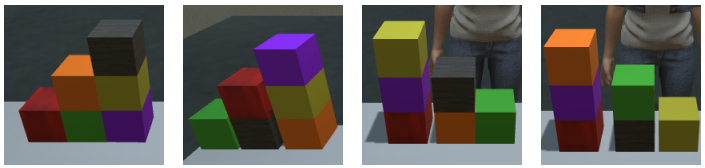


Figure: Generated staircases display desired inferences

Figure: Agent builds structure in VoxSim from generated move sequence

Learning Configurations: Results

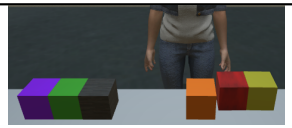
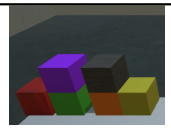
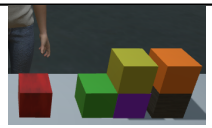
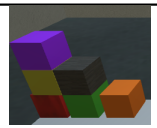
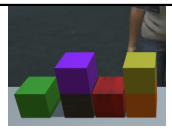
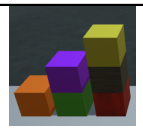
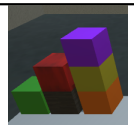
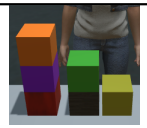
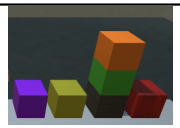
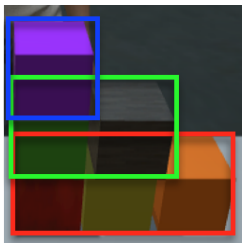
Heuristic	Median Ex.	Best Ex.	μ Score	Score σ
Chance			2.0375	1.0122
JD			4.3375	2.0387
LD			3.7688	2.1028
SPIRE:			**5.8313**	2.7173
Comb.			4.7188	2.4309

Figure: 50 structures, 5 per heuristic. Shown: median- (L) and highest-scored (R) structure generated using each heuristic with evaluator score (0-10).

Learning Configurations: Grounding Novel Semantics

- Identify components, encode constraints.
- What affordances of components satisfy the constraints that appear in the learned and generated examples?



build

LEX = [PRED = **build**
TYPE = **transition_event**]

TYPE = [HEAD = **transition**
ARGS = [A1 = **x:agent**
A2 = **y[]:physobj**
A3 = **z:model**]
BODY = [E1 = *for*(*o* ∈ *y*[1..*n*])
[*reify*(*predict*(*target*(*z*), *as*(*w*))),
reify(*predict*(*completion*(*w*), *as*(*v*))),
reify(*prune*(*v*), *as*(*u*)),
put(*x*, *o*, *at*(*u*)), *reify*((*z*, *o*), *as*(*z*))]]]

staircase

LEX = [PRED = **staircase**
TYPE = **physobj**]

TYPE = [HEAD = **assembly[1]**
COMPONENTS = **base[2]**,**step[3]***,**top[4]**
CONCAVITY = **nil**
ROTATSYM = **nil**
REFLECTSYM = {*XY*}
CONSTR = *Y*([3]) > *Y*([2]), *Y*([4]) ⩾ *Y*([3])]

HABITAT = [INTR = [5][BASE = *align*([2], $\mathcal{E}_X$)
UP = *align*(*vec*(*loc*([4]) − *loc*([2])), $\mathcal{E}_Y$)]]

AFFORD_STR = [A1 = $H_{[5]} \rightarrow$ [*put*(*x*, *on*([1]))]*component*(*x*, [1])
A2 = $H_{[5]} \rightarrow$ [*put*(*x*, *on*([2]))]*component*(*x*, [3])
A3 = $H_{[5]} \rightarrow$ [*put*(*x*, *left* ∨ *right*∨
touching([2]) ∧ ¬*on*([2])]*extend*(*x*, [2])
A4 = $H_{[5]} \rightarrow$ [*put*(*x*, *left* ∨ *right*∨
touching([3]) ∧ ¬*on*([3])]*extend*(*x*, [3])]

EMBODIMENT = [SCALE = <**agent**
MOVABLE = **true**]

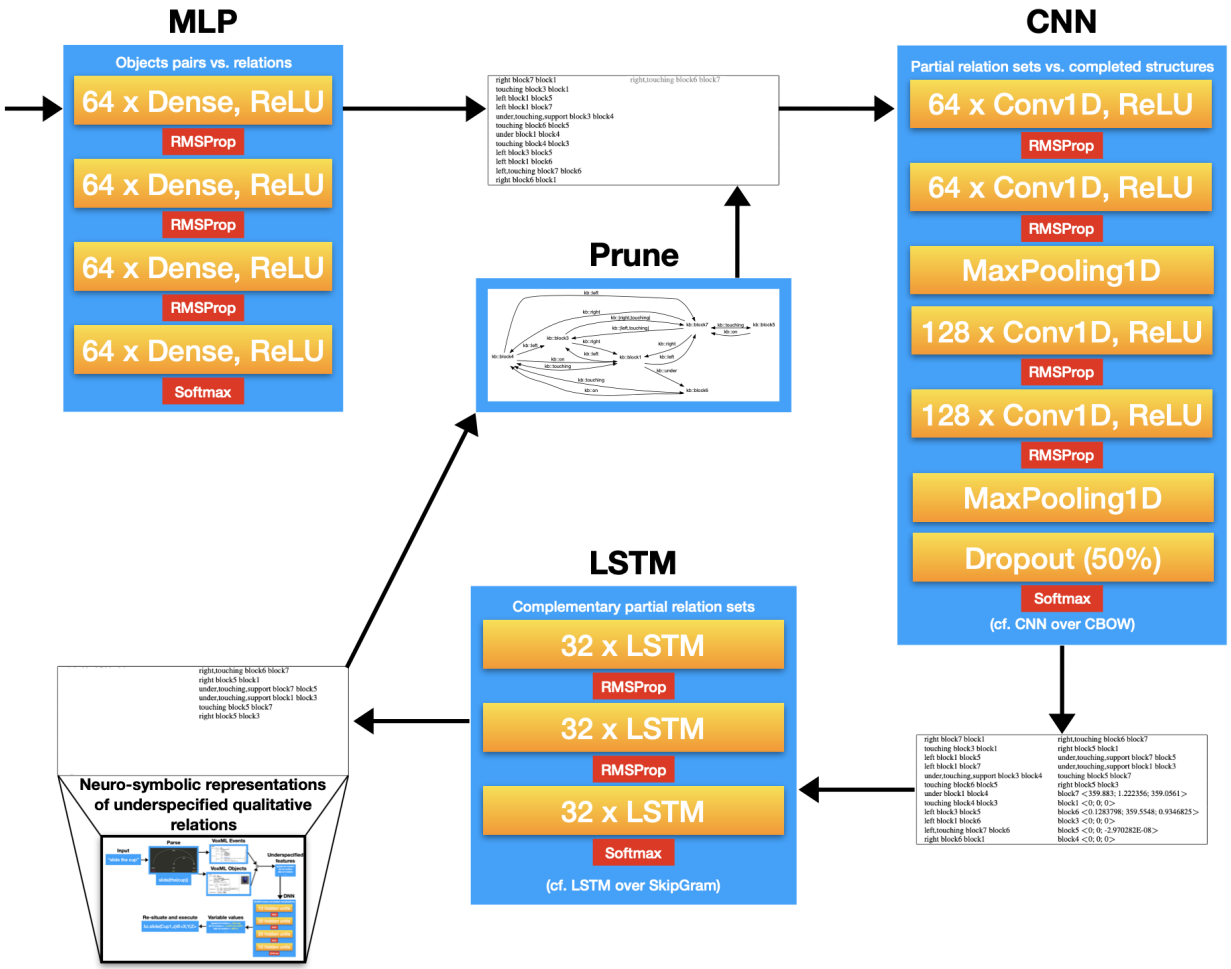
Table: VoxML for [[BUILD]] and [[STAIRCASE]]

Learning Configurations: Grounding Novel Semantics

Sample approaches:

- Weighted constraint satisfaction; partially-observable Markov Decision Process (POMDP) (Krishnaswamy and Pustejovsky, 2019);
- **Qualitative Constraint Network (QCN)**;
  - Tries to solve for combined qualitative spatial relations (e.g., "left," "right," "under," "support") with interval algebra distinctions along an axis (e.g., meets, overlaps).
- Features are not just neural network parameters, but also spatial features than can be talked about and composed independently of the network.

Learning Configurations: Neuro-Symbolic Intelligence



Transfer Learning

- Situatedness is particularly useful for transfer learning, because similar concepts often exist in similar situations (cf. analogical generalization, a la Forbus et al., 2017).
  - e.g., "Build an X out of *these*," "Put *all those* in that X."
- Associate affordances with abstract properties—spheres roll, sphere-like entities probably do too.
- This informs the way you can talk about items (in real or virtual situations).
- Q: "What am I pointing at?" A: "I don't know, but it looks like [a container, something that rolls, etc.]"
- Similar objects have similar habitats/affordances.
- What happens when Diana encounters a new object?

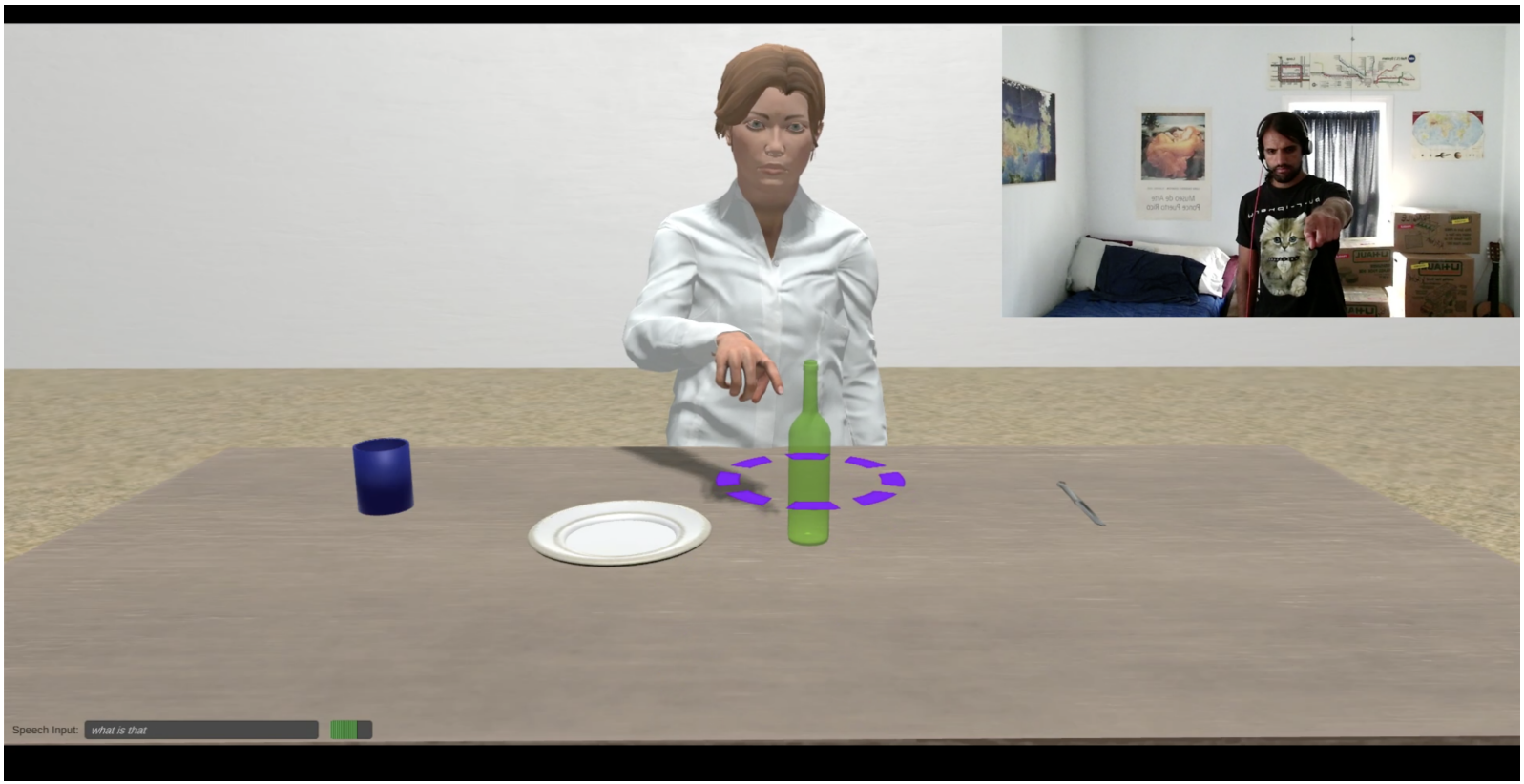
Transfer Learning

- Exploit the correlations between habitats and affordances over known objects, and map those correspondences to novel objects;
- 17 distinct VoxML objects (~22 distinct affordance encodings):
  - e.g., $H_{[3]} = [\text{UP} = \mathit{align}(\bar{Y}, \mathcal{E}_Y), \text{TOP} = \mathit{top}(+Y)]$, $H_{[3]} \rightarrow [\mathit{put}(x, \mathit{in}(\mathit{this}))]\mathit{contain}(\mathit{this}, x)$;
- Train 200-dimensional habitat or affordance embeddings using a Skip-Gram model;
- Represent objects as averaged habitat or affordance vectors.

Transfer Learning

- 2 architectures: 7-layer MLP and 4-layer CNN w/ 1D convolutions;
- Evaluate against a ground truth of k-means clustered objects derived from human annotators;
- Achieve ~80% accuracy with the predicted object clustering with the ground-truth object;
  - ~40% of the time the predicted object *always* clusters with the ground truth in 5 randomized trials.
- Habitat-based model typically better at capturing common behaviors (e.g., grasping), affordance-based model better at object-specific behaviors (e.g., rolling).

Transfer Learning



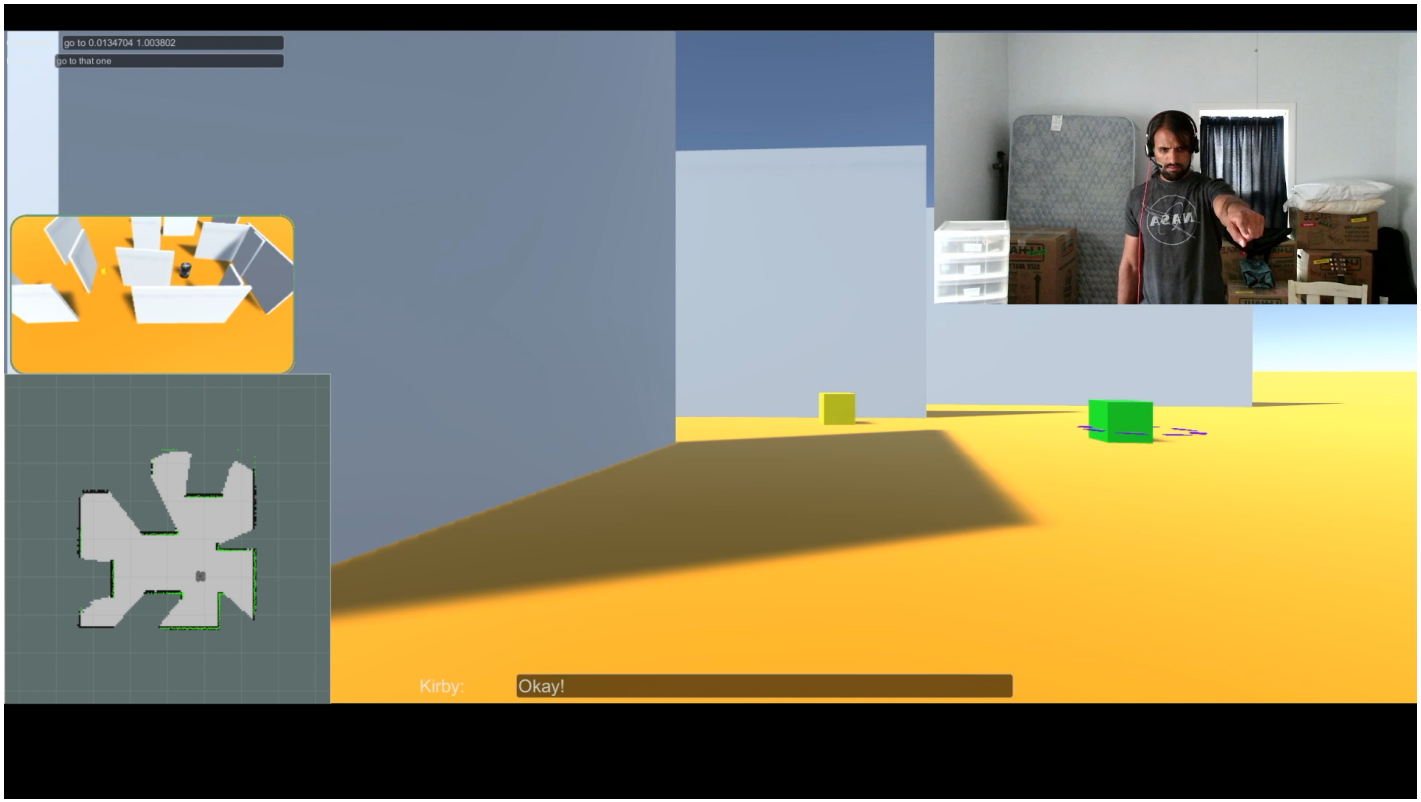
▸ Play!

Why Situated Grounding?

- Situatedness goes beyond visual grounding; it demonstrates meaning multimodally;
- Environmental awareness: each additional modality provides an orthogonal angle through which to validate models of other modalities;
- Many methods of encoding context, quantitatively and qualitatively;
  - Multimodal Semantic Grammar: rich model of multimodal contextual parsing.
- Provides a model to accommodate both neural and symbolic representations;
- Bridges data-modeling gap, adaptable to different tasks;
- A sustainable way toward more powerful AI.

Future Work

Domain transfer to robotics:



▸ Play!

Conclusion

July 11, 2020:

Figure: Who needs situated grounding anyway?

Don't Panic!

"Thought in action becomes word, word in action becomes speech. Speech in action becomes character Jesus."

Thank You!

This work was supported by the US Defense Advanced Research Projects Agency (DARPA) and the Army Research Office (ARO) under contract W911NF-15-C-0238 at Brandeis University

Thank You!

Brandeis Lab for Linguistics and Computation

- Marc Verhagen, Tuan Do, Kyeongmin Rim, Kelley Lynch, Ken Lai
  - Research assistants: Mark Hutchens, Jin Zhao, Daeja Showers, Katie Krajovic, Eli Goldner

Collaborators at Colorado State University

- **Bruce Draper**, **Ross Beveridge**, Pradyumna Narayana, David White, Rahul Bangar, Dhruva Patil, Joe Strout, Jason Yu, Heting Wang, Matt Dragan

Collaborators at University of Florida

- **Jaime Ruiz**, Isaac Wang, Daniel Delgado

Collaborators at SIFT

- Scott Friedman, **David McDonald**, **Mark Burstein**

Thank You!

Questions?

References I


Albath, Julia et al. (2010). "RCC-3D: Qualitative Spatial Reasoning in 3D.". In: *CAINE*, pp. 74–79.


Forbus, Kenneth D et al. (2017). "Extending SME to handle large-scale cognitive modeling". In: *Cognitive Science* 41.5, pp. 1152–1201.


Gatsoulis, Yiannis et al. (2016). "QSRlib: a software library for online acquisition of Qualitative Spatial Relations from Video". In:


Gibson, James J. (1977). "The Theory of Affordances". In: *Perceiving, Acting, and Knowing: Toward an ecological psychology*, pp. 67–82.

References II


Krishnaswamy, Nikhil (2017). "Monte-Carlo Simulation Generation Through Operationalization of Spatial Primitives". PhD thesis. Brandeis University.


Krishnaswamy, Nikhil, Scott Friedman, and James Pustejovsky (2019). "Combining Deep Learning and Qualitative Spatial Reasoning to Learn Complex Structures from Sparse Examples with Noise". In: *AAAI Conference on Artificial Intelligence (AAAI)*. AAAI.


Krishnaswamy, Nikhil and James Pustejovsky (2018). "An Evaluation Framework for Multimodal Interaction". In: *Proceedings of LREC*.

References III


Krishnaswamy, Nikhil and James Pustejovsky (2019). "Situated Grounding Facilitates Multimodal Concept Learning for AI". In: *Workshop on Visually Grounded Interaction and Language*.


McLure, Matthew D., Scott E. Friedman, and Kenneth D. Forbus (2015). "Extending Analogical Generalization with Near-Misses.". In: *AAAI*, pp. 565–571.


Moratz, Reinhard, Bernhard Nebel, and Christian Freksa (2002). "Qualitative spatial reasoning about relative position". In: *International Conference on Spatial Cognition*. Springer, pp. 385–400.

References IV

Pustejovsky, James and Nikhil Krishnaswamy (2016). "VoxML: A Visualization Modeling Language". In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. Ed. by Nicoletta Calzolari (Conference Chair) et al. Portoroz, Slovenia: European Language Resources Association (ELRA). ISBN: 978-2-9517408-9-1.

Randell, D.A. et al. (1992). "A spatial logic based on regions and connection". In: *KR'92. Principles of Knowledge Representation and Reasoning: Proceedings of the Third International Conference*. Morgan Kaufmann. San Mateo, pp. 165–176.