Embodied Language Grounding

Katerina Fragkiadaki
Carnegie Mellon University
What does it mean to comprehend natural language?
What does it mean to comprehend natural language?

Localize the referents it mentions
What does it mean to comprehend natural language?

Generate the image it describes
What does it mean to comprehend natural language?

"man in black shirt is playing guitar."
"construction worker in orange safety vest is working on road."
"two young girls are playing with lego toy."
"boy is doing backflip on wakeboard."

Caption an image
What does it mean to comprehend natural language?

- After wading barefoot in the lake, Erik used his shirt to dry his feet.
- After wading barefoot in the lake, Erik used his glasses to dry his feet.

To act upon it, and infer its affordability

Embodiment, simulation and meaning, Bergen, How reading comprehension is embodied and why that matters, Glenberg, Grounding language in action, Glenberg and Kaschak, Grounding Meaning in Affordances, Glenberg
Embodied Cognition

- Words and phrases are indexes to objects in the world or to prototypical symbols of those objects
- We derive affordance from those objects
- The derived affordances constrain the way ideas can be coherently combined

To act upon it, and infer its affordability
What does it mean to comprehend natural language?

- The bowl inside the cube
- The cube inside the bowl

To act upon it, and infer its affordability

Embodiment, simulation and meaning, Bergen, How reading comprehension is embodied and why that matters, Glenberg, Grounding language in action, Glenberg and Kaschak, Grounding Meaning in Affordances, Glenberg
Simulation Semantics

- Words and phrases are indexes to objects in the world or to prototypical symbols of those objects
- We derive affordance from those objects
- The derived affordances constrain the way ideas can be coherently combined

To act upon it, and infer its affordability

Embodiment, simulation and meaning, Bergen, How reading comprehension is embodied and why that matters, Glenberg, Grounding language in action, Glenberg and Kaschak, Grounding Meaning in Affordances, Glenberg
Many philosophers agree and support such simulation semantics, no computational model for language grounding.

*Visually grounded interaction and language, NIPS workshop 2018*
Language grounding to visual cues

Remains disconnected from affordances

2D boxes or 2D CNN features do not have any affordability attached
Are themselves ungrounded

- The bowl inside the cube
- The cube inside the bowl
Reward learning using natural language

Given a NL utterance, learn a visual detector

“Can is to the right of the mug”
Reward learning using natural language

Given a NL utterance, learn a visual detector

“Can is to the right of the mug”

Use the learned visual detector to guide policy learning for achieving the NL described goal

It did not really worked, the reward detector could not effectively generalize across camera placements

*Modeling Relationships in Referential Expressions with Compositional Modular Networks, Ronghang et al.*
“Can is to the right of the mug”

reward detector

Learned reward detector

score

time

Learned policy

“Can is to the right of the mug”
Prior work used manually coded rewards.

We ground rewards to the sensory input of the agent.
Affordable visual representations

We seek visual cues that obey basic common sense constraints and basic affordability reasoning:

- Objects have 3D extent
- Objects do not interpenetrate in 3D
- Objects come in regular sizes
- Objects persist over time

Grounding language on such representations would be able to support affordability inference
Affordandable visual representations

We seek visual cues that obey basic common sense constraints and basic affordability reasoning:

- Objects have 3D extent
- Objects do not interpenetrate in 3D
- Objects come in regular sizes
- Objects persist over time

Grounding language on such representations would be able to support affordability inference
Geometry-Aware Recurrent Networks
Geometry-Aware Recurrent Networks
Geometry-Aware Recurrent Networks
Geometry-Aware Recurrent Networks

2D-to-3D mapping

egomotion-stabilized update

unproject

t
View prediction

View Prediction

R, t

t
View prediction

3D-to-2D mapping

View Prediction

R, t
View prediction

geometry-aware RNN

2D RNN [1]
View prediction

3 input views

GRNN

2D RNN [1]

[1] Neural scene representation and rendering DeepMind, Science, 2018
View prediction

3 input views

GRNN

2D RNN [1]

Truly novel scenes

[1] Neural scene representation and rendering DeepMind, Science, 2018
Geometry-Aware Recurrent Networks (GRNNs)
3D object detection
3D object detection

# of input views
3D object detection

Results - 3D object detection

# of input views
3D object detection

# of input views
3D representations

- ask too much: high level of 3D details may be impossible to obtain
- ask too little: information about semantics of the objects is not present
1. We consider an embodied agent that can see a scene from multiple viewpoints

“The green rubber cylinder is on the right of the blue bowl”
1. We consider an embodied agent that can see a scene from multiple viewpoints.

"The green rubber cylinder is on the right of the blue bowl"
“The green rubber cylinder is on the right of the blue bowl”

1. Our agent learns to map an RGB image to a set of 3D feature maps by training GRNNs to predict views
“The green rubber cylinder is on the right of the blue bowl”

1. Our agent maps noun phrases to object-centric 3D feature maps
1. Our agent maps noun phrases to object-centric 3D feature maps

“The green rubber cylinder is on the right of the blue bowl”
1. Our agent maps spatial expressions to relative 3D offsets

"The green rubber cylinder is on the right of the blue bowl"
“The green rubber cylinder is on the right of the blue bowl”

1. Our agent populates a 3D canvas with the predicted object tensors and their relative offsets.
1. Our agent populates a 3D canvas with the predicted object tensors and their relative offsets

“The green rubber cylinder is on the right of the blue bowl”
1. Our agent populates a 3D canvas with the predicted object feature maps and their relative spatial offsets.

“The green rubber cylinder is on the right of the blue bowl”
1. The generated canvas when projected should match the RGB image views

"The green rubber cylinder is on the right of the blue bowl"
Scene imagination

Red Rubber Cylinder to the left front of Blue Rubber Cube to the left front of Green Rubber Cylinder to right front of Blue Rubber Cube

Red Rubber Cube to the left front of the Blue Rubber Sphere to the right front of Cyan Metal Cylinder

Neural render

Blender render
# Scene imagination

<table>
<thead>
<tr>
<th>Natural language utterance</th>
<th>Neural render</th>
<th>Blender render</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Purple Cylinder to the left behind of Brown Cube to the left front of Purple Sphere</strong></td>
<td><img src="image1.png" alt="Neural render" /></td>
<td><img src="image2.png" alt="Blender render" /></td>
</tr>
<tr>
<td><strong>Purple Cylinder to the left behind of Cyan Cube to the left front of Cyan Cube</strong></td>
<td><img src="image3.png" alt="Neural render" /></td>
<td><img src="image4.png" alt="Blender render" /></td>
</tr>
</tbody>
</table>
Scene imagination

**Natural language utterance**

Cyan Cube to the left behind of Gray Sphere to the left front of Blue Cube

**Neural render**

![Images of Neural render](image1)

**Blender render**

![Images of Blender render](image2)

**Red Sphere to the left behind of Cyan Cylinder to the left front of Red Sphere**

![Images of Red sphere with Cyan Cylinder](image3)
Scene imagination

Natural language utterance

“red cylinder to the right behind of green cube”

Neural render

“pink cylinder to the left front of red cylinder”

Blender render
Scene alteration

<table>
<thead>
<tr>
<th>Natural language utterance</th>
<th>Neural render</th>
<th>Blender render</th>
</tr>
</thead>
<tbody>
<tr>
<td>“blue sphere to the right behind of green cube”</td>
<td><img src="image1.png" alt="Image" /></td>
<td><img src="image2.png" alt="Image" /></td>
</tr>
<tr>
<td>“green cube to the left front of cyan cylinder”</td>
<td><img src="image3.png" alt="Image" /></td>
<td><img src="image4.png" alt="Image" /></td>
</tr>
</tbody>
</table>
Affordability Inference

Natural language utterance: “blue sphere to the right behind of green cube”

“green cube to the left front of cyan cylinder”

Neural render

Blender render
Grounding Language on 3D visual feature representations

• Objects have regular sizes: object size is disentangled from the camera viewpoint
• Objects have 3D extent
• Objects do not interpenetrate in 3D: during iterative scene generation we can detect 3D intersection and continue sampling valid configurations
• Objects persist over time
Next steps

- Grounding action descriptions
- Use intuitive physics and dynamics beyond static spatial constraints
Thank you

- Embodied language grounding, arxiv
- Reward Learning from Narrated Demonstrations, Tung et al., CVPR 2018