Guiding Exploratory Behaviors for Multi-Modal Grounding of Linguistic Descriptions

Jesse Thomason
Raymond J. Mooney, Peter Stone
University of Texas at Austin

Jivko Sinapov
Tufts University
Grounding linguistic descriptions

Grab the heavy mug.
Grounding linguistic descriptions

Grab the heavy mug.
Grounding linguistic descriptions

heavy:

mug:
Grounding linguistic descriptions

heavy:
mug:
Grounding linguistic descriptions

Grab the heavy mug.
Multi-modal grounding

heavy:

mug:

tall:

+
Exploratory Behaviors

- grasp
- lift
- lower
- drop
- push
- press

+ hold
+ look
Exploratory Behaviors - Haptic and Audio Feedback

lift, hold, lower

drop
Exploratory Behaviors - Haptic and Audio Feedback

press

push
Exploratory Behaviors

- Grasp (22s)
- Lift (11.1s)
- Lower (10.6s)
- Drop (9.8s)
- Push (22s)
- Press (22s)

104 seconds (minimum) to explore one object once

520 seconds to explore an object five times

4.5 hours to fully explore 32 objects (best case!)
Guiding Exploratory Behaviors

rigid:

press:

look:

Haptics, Audio

RGB, HSV, VGG
Guiding Exploratory Behaviors

rigid:

press:

look:
Guiding Exploratory Behaviors

<table>
<thead>
<tr>
<th>squishy?</th>
</tr>
</thead>
<tbody>
<tr>
<td>press?</td>
</tr>
<tr>
<td>look?</td>
</tr>
</tbody>
</table>
Guiding Exploratory Behaviors

**Rigid:**
- press:
  - ball
  - string

**Squishy?**
- press?
  - mug

**Look:**
- mug
- ball

**Look?**
- ball
- string

*Image of objects: sponge, ball, string, mug, ball, mug, fork, plate, bowl.*
Guiding Exploratory Behaviors

<table>
<thead>
<tr>
<th>rigid</th>
<th>press?</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>rigid</th>
<th>look?</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

squishy?
Guiding Exploratory Behaviors

squishy:

press:

rigid

look?
Guiding Exploratory Behaviors

squishy:

press:

look:
Neural Word Embeddings - Word2Vec\textsuperscript{1}

“... had a rigid surface layer ...”

Neural Word Embeddings - Word2Vec
Neural Word Embeddings - Word2Vec

\[ \text{similarity}(\text{rigid}, \text{squishy}) = \cos(\theta) \]
Guiding Exploratory Behaviors using Embeddings

squishy?

press?

look?
Guiding Exploratory Behaviors using Embeddings

squishy?

press?
similarity(rigid, squishy) = 0.8
similarity(tall, squishy) = 0.4
similarity(mug, squishy) = 0.4

look?
Guiding Exploratory Behaviors using Embeddings

| $G_{p,c}(o)$ | SVM trained for predicate $p$ and behavior context $c$ result on object $o$ |

$$d(p, o) = sgn \left( \sum_{c \in C} w_{p,c} G_{p,c}(o) \right)$$
Guiding Exploratory Behaviors using Embeddings

\[ G_{p,c}(o) \] SVM trained for predicate \( p \) and behavior context \( c \) result on object \( o \)

\[
d(p, o) = sgn \left( \sum_{c \in C} w_{p,c} G_{p,c}(o) \right)
\]

Decision
Guiding Exploratory Behaviors using Embeddings

| $G_{p,c}(o)$ | SVM trained for predicate $p$ and behavior context $c$ result on object $o$ |

$$d(p, o) = sgn \left( \sum_{c \in C} w_{p,c} G_{p,c}(o) \right)$$

| Decision | Behavior Contexts |
Guiding Exploratory Behaviors using Embeddings

\[ G_{p,c}(o) \]

SVM trained for predicate \( p \) and behavior context \( c \) result on object \( o \)

\[
d(p, o) = sgn \left( \sum_{c \in C} w_{p,c} G_{p,c}(o) \right)
\]

Decision  Behavior  Context

Contexts  SVM result
Guiding Exploratory Behaviors using Embeddings

\[ G_{p,c}(o) \] SVM trained for predicate \( p \) and behavior context \( c \) result on object \( o \)

\[ d(p, o) = \text{sgn} \left( \sum_{c \in C} w_{p,c} G_{p,c}(o) \right) \]

Decision  Behavior Contexts  Reliability Weight  Context SVM result
Guiding Exploratory Behaviors using Embeddings

\[ G_{p,c}(o) \] SVM trained for predicate \( p \) and behavior context \( c \) result on object \( o \)

\[ d(p, o) = \text{sgn} \left( \sum_{c \in C} w_{p,c} G_{p,c}(o) \right) \]

Reliability weights for trained classifiers

<table>
<thead>
<tr>
<th>context</th>
<th>( w_{p,c} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>rigid press-haptics</td>
<td>0.5</td>
</tr>
<tr>
<td>grasp-haptics</td>
<td>0.3</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>look-rgb</td>
<td>0.01</td>
</tr>
</tbody>
</table>
Guiding Exploratory Behaviors using Embeddings

\[ d(p, o) = sgn \left( \sum_{c \in C} w_{p,c} G_{p,c}(o) \right) \]

\[ w_{q,c} \approx \frac{1}{|P_q|} \sum_{p \in P_q} \text{poscos}(p, q) w_{p,c} \]

Surrogate reliability weights for new word-embedding classifiers for \( q \)

Nearest word-embedding predicates to \( q \)

Reliability weights for trained neighbor classifiers \( p \)
Shared Structure: Embeddings and Features

2D-projection of word embeddings

2D-projection of behavior context features
Problem Formulation

Given:
- A set of explored training objects
- A set of predicates $P$ and trained classifiers with reliability weights
- A new, unseen predicate $q$
- A set of unexplored test objects, each labeled with whether $q$ applies

$\{\text{rigid, mug, tall}\}$

squishy
Problem Formulation

Given:
- A set of explored training objects
- A set of predicates $P$ and trained classifiers with reliability weights
- A new, unseen predicate $q$
- A set of unexplored test objects, each labeled with whether $q$ applies

Produce:
- Surrogate reliability weights for behaviors on $q$
- Classifier for $q$ using only a subset of behaviors
Problem Formulation

rigid:

press:

look:

squishy:

press:

look?
Evaluation: colors, weights, contents\textsuperscript{[2]}

Given:
- 12 explored objects (identical except for color, weight, or content)
- $P$: \{red, blue, green, light, medium, heavy, rice, beans, glass, screws\}
- Unseen $q$ from $P$
- 24 unexplored objects labeled for $q$

Produce:
- Best classifier for $N$ behaviors

\textsuperscript{[2]} Sinapov \textit{et al.} Learning relational object categories using behavioral exploration and multimodal perception. 2014.
Evaluation: colors, weights, contents

Color predicates

Weight predicates

Contents predicates

(dotted lines show standard error)
Evaluation: everyday words$^{[3,4]}$

Given:
- 16 explored household objects
- $P$: 48 predicates from a language game
- Unseen $q$ from $P$
- 16 unexplored objects labeled for $q$

Evaluation: everyday words$^{[3,4]}$  

Produce:  
- A sequence of behaviors to maximize classifier performance **given a fixed time budget**; allowed to repeat each behavior 5 times for new observations

Exploratory Behaviors with Time Constraints

- **grasp (22s)**
- **lift (11.1s)**
- **lower (10.6s)**
- **drop (9.8s)**
- **push (22s)**
- **press (22s)**
- + **hold (5.7s)**
- + **look (0.8s)**
Exploratory Behaviors with Sequence Constraints

- on table
- grasped
- raised

- grasp
- lift
- lower

- look
- press
- push

- drop
- hold

Images showing robotic activities at different stages.
Exploratory Behaviors with Sequence Constraints

\[ v(b) = w_{q,b} + \epsilon \]

\[ p(\text{press}) = \frac{v(\text{press})}{v(\text{look}) + v(\text{press}) + v(\text{push}) + v(\text{grasp})} \]
Utilize Human Annotations of Behaviors

![Graph showing behavior annotation scores for cylindrical, heavy, and squishy objects across different behaviors: look, grasp, lift, hold, lower, drop, push, and press. The graph demonstrates varying scores for each behavior and object type.]
Guiding Behaviors using Embeddings + Annotations

(lex): \[ w_{q,c} \approx \frac{1}{|P_q|} \sum_{p \in P_q} poscos(p, q) w_{p,c} \]

Surrogate reliability weights for new classifiers for \( q \)

Nearest word-embedding predicates to \( q \)

Reliability weights for trained neighbor classifiers \( p \)
Guiding Behaviors using Embeddings + Annotations

(lex): $w_{q,c} \approx \frac{1}{|P_q|} \sum_{p \in P_q} \text{poscos}(p, q) w_{p,c}$

(ba): $w_{q,c} \approx \frac{1}{|C_b|} A(q, b_c)$

Nearest word-embedding predicates to $q$

Surrogate reliability weights for new classifiers for $q$

Behavior applicability annotation score

Reliability weights for trained neighbor classifiers $p$
Guiding Behaviors using Embeddings + Annotations

(lex): \[ w_{q,c} \approx \frac{1}{|P_q|} \sum_{p \in P_q} \text{poscos}(p, q) w_{p,c} \]

(ba): \[ w_{q,c} \approx \frac{1}{|C_b|} A(q, b_c) \]

(ba+lex): \[ w_{q,c} \approx \frac{1}{|C_b|} A(q, b_c) \times \frac{1}{|P_q|} \sum_{p \in P_q} \text{poscos}(p, q) w_{p,c} \]

Nearest word-embedding predicates to \( q \)

Surrogate reliability weights for new classifiers for \( q \)

Behavior applicability annotation score

Reliability weights for trained neighbor classifiers \( p \)
Fixed budget of 212 seconds to explore objects (bars give standard error):
Conclusions

squishy?

press?

look?
Conclusions

squishy?

press?

look?
Conclusions

squishy?

press?

look?

rigid:

press:

look:
Conclusions

How can I tell if something is squishy?

You could press down on it.
Conclusions

You could *press* down on it.
Conclusions

or
Guiding Exploratory Behaviors for Multi-Modal Grounding of Linguistic Descriptions

Jesse Thomason
Raymond J. Mooney, Peter Stone
University of Texas at Austin

Jivko Sinapov
Tufts University
Future Work: POMDP-driven Exploration

Given observation of object from most recent behavior, decide **online** which behavior to perform next.

Asking for a human annotation can be considered an extra, expensive action in the policy.
Future Work: Concept Relationships

Colors are largely mutually-exclusive

Light and heavy are antonyms

Soft and squishy are synonyms
Differing Structures from Different Corpora

2D-projection of word embeddings (lexsub)

2D-projection of word embeddings (Google)
Evaluation: everyday words

Time budget in seconds versus classifier ensemble agreement with gold labels