Encoding Spatial Relations from Natural Language

Tiago Ramalho*, Tomáš Kociský*, Frederic Besse, S. M. Ali Eslami, Gábor Melis, Fabio Viola, Phil Blunsom, Karl Moritz Hermann

July 2018
Topics

1. Some terms
2. The problem: spatial relations
3. Related works
4. Key contributions of this research
5. The model
6. Experiments & results
Some terms

- **Representation**
  - Localist
  - Distributed
- **Autoencoder**

<table>
<thead>
<tr>
<th>Concept</th>
<th>Representation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Small Red Car</td>
<td>[1 0 0 0 0 0 0 0]</td>
</tr>
<tr>
<td>Large Blue SUV</td>
<td>[0 1 0 0 0 0 0 0]</td>
</tr>
<tr>
<td>Large Red SUV</td>
<td>[0 0 1 0 0 0 0 0]</td>
</tr>
<tr>
<td>Green Apple</td>
<td>[0 0 0 1 0 0 0 0]</td>
</tr>
<tr>
<td>Bumble Bee</td>
<td>[0 0 0 0 1 0 0 0]</td>
</tr>
<tr>
<td>Tall Building</td>
<td>[0 0 0 0 0 1 0 0]</td>
</tr>
<tr>
<td>Small Fish</td>
<td>[0 0 0 0 0 0 1 0]</td>
</tr>
<tr>
<td>Banana</td>
<td>[0 0 0 0 0 0 0 1]</td>
</tr>
</tbody>
</table>

Localist: One-Hot Encoding [2]
Some terms

- **Representation**
  - Localist
  - Distributed
- **Autoencoder**

**Distributed: word embedding** [2]

\[
W(\text{"King"}) - W(\text{"Man"}) + W(\text{"Woman"}) = W(\text{"Queen"})
\]

<table>
<thead>
<tr>
<th>Concept</th>
<th>Representation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Small Red Car</td>
<td>[0.555 0.761 0.243 0.812]</td>
</tr>
<tr>
<td>Large Blue SUV</td>
<td>[0.773 0.309 0.289 0.835]</td>
</tr>
<tr>
<td>Large Red SUV</td>
<td>[0.766 0.780 0.294 0.834]</td>
</tr>
<tr>
<td>Green Apple</td>
<td>[0.153 0.022 0.654 0.513]</td>
</tr>
<tr>
<td>Bumble Bee</td>
<td>[0.045 0.219 0.488 0.647]</td>
</tr>
<tr>
<td>Tall Building</td>
<td>[0.955 0.085 0.900 0.773]</td>
</tr>
<tr>
<td>Small Fish</td>
<td>[0.118 0.192 0.432 0.618]</td>
</tr>
<tr>
<td>Banana</td>
<td>[0.184 0.232 0.671 0.589]</td>
</tr>
</tbody>
</table>
### Some terms

- **Representation**
  - Localist
  - Distributed

- **Autoencoder**
  - Img $\rightarrow$ repr
  - Repr $\rightarrow$ img

<table>
<thead>
<tr>
<th>Concept</th>
<th>Representation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Small Red Car</td>
<td>[0.555 0.761 0.243 0.812]</td>
</tr>
<tr>
<td>Large Blue SUV</td>
<td>[0.773 0.309 0.289 0.835]</td>
</tr>
<tr>
<td>Large Red SUV</td>
<td>[0.766 0.780 0.294 0.834]</td>
</tr>
<tr>
<td>Green Apple</td>
<td>[0.153 0.022 0.654 0.513]</td>
</tr>
<tr>
<td>Bumble Bee</td>
<td>[0.045 0.219 0.488 0.647]</td>
</tr>
<tr>
<td>Tall Building</td>
<td>[0.955 0.085 0.900 0.773]</td>
</tr>
<tr>
<td>Small Fish</td>
<td>[0.118 0.192 0.432 0.618]</td>
</tr>
<tr>
<td>Banana</td>
<td>[0.184 0.232 0.671 0.589]</td>
</tr>
</tbody>
</table>

Distributed: word embedding [2]
Spatial relations

● In front of, behind, ...
● Different viewpoints
● Hard to capture with models
  ○ Grading vs categorical
  ○ Lexicalization
  ○ Representation

3D objects [1]
Problem: description → image

“There is a green ball in front of a blue pyramid. A red cube is to the left of the pyramid, slightly farther from the green ball.”
Related work

- Semantic equivalences
- Spatial relation, perception and semantics
- Extracting and encoding relations to formal languages
- Heavily hand-engineered visualization methods
Key contributions

- Better representations
- Multi-modal architecture
- Large dataset of 3D scenes with descriptions
General process

1. Learn representations of text + camera angles
2. Generate images from **unseen angles**
3. Compare with descriptions
Dataset

- 10 million 3D scenes
  - 10 camera viewpoints: 3D view and description(s)
    - SYN for all scenes
    - NL for 6604 scenes
- 2-3 objects per scene

Table 1: Dataset statistics.

<table>
<thead>
<tr>
<th></th>
<th>Synthetic</th>
<th>Natural</th>
</tr>
</thead>
<tbody>
<tr>
<td># Training Scenes</td>
<td>10M</td>
<td>5,604</td>
</tr>
<tr>
<td># Validation Scenes</td>
<td>1M</td>
<td>432</td>
</tr>
<tr>
<td># Test Scenes</td>
<td>1M</td>
<td>568</td>
</tr>
<tr>
<td>Vocabulary Size</td>
<td>42</td>
<td>1,023</td>
</tr>
<tr>
<td>Tokens per Description</td>
<td>60</td>
<td>90</td>
</tr>
</tbody>
</table>
Dataset: example

**NL** There is a small purple cone far to the right slightly cut off at the edge of the screen. There is a lighter purple object that almost looks like a sideways outlined triangle slightly farther away near the center of the screen.

**SYN** There is a violet cone to the far right of a purple torus. The cone is in front of the torus.

**NL** A blue cone is sitting near the right side of the screen. To the right of it and slightly behind is a 3d polygon sharp edged object, slightly cut off by the right edge of the screen. In the background and to the left of these objects is a green cube.

**SYN** A blue icosahedron is in front of a right green cylinder. There is a green cylinder far left of a blue cone. The icosahedron is close to the cone.
SLIM: Spatial Language Integrating Model

- $n$ viewpoint descriptions $\rightarrow$ single representation $\rightarrow$ new data
  - Generated viewpoint unknown
- Two parts:
  - representation network
  - generation network
Representation network

Each text observation $d_i$ ($i = 1 \ldots n$ views)

→ conv. language encoder
Representation network

Each text observation $d_i \ (i = 1...n \text{ views})$
→ conv. language encoder
→ concat. with embedded camera angle
Representation network

Each text observation $d_i$ ($i = 1...n$ views)
→ conv. language encoder
→ concat. with embedded camera angle
→ MLP
→ viewpoint embedding $h_i$
Representation network

Each text observation $d_i \ (i = 1 \ldots n \ \text{views})$
→ conv. language encoder
→ concat. with embedded camera angle
→ MLP
→ viewpoint embedding $h_i$
→ scene representation $r = \frac{1}{n} \sum_{i=1}^{n} h_i$
Generation network (training)

We have $r \rightarrow$ sample unseen pair $d_t$ and $c_t$
Generation network

We have $r$

$\rightarrow$ sample unseen pair $d_t$ and $c_t$

$\rightarrow$ encoder function $P(z) = e(d_t, c_t)$
Generation network

We have $r$

→ sample unseen pair $d_t$ and $c_t$

→ encoder function $P(z) = e(d_t, c_t)$

→ decoder function $\hat{d}_t = g(z, r, c_t)$

Minimizing ELBO loss $\mathcal{L} = -\log D(x|r) + \sum_{k=1}^{K} KL(Q(z_k|h_{k}^{enc})||P(z_k))$
SLIM

There is a violet torus behind a peach icosahedron. There is a pink cylinder to the left of a peach icosahedron. The cylinder is ...

There is a large pink cylinder to the right of a peach icosahedron. The cylinder is close to the icosahedron. There is a violet torus in front ...

representation network

encoder network

c_1

c_2

encoder network

...
Experiments

1. SYN experiments
2. NL experiments
Evaluation

Human annotators given model samples & corresponding descriptions

1. Matched? Partially match? Did not match?
2. Binary choice
   a. all object shapes in the image?
   b. all object colours in the image?
   c. all combinations correct?
   d. all positions correct?
Synthetic language experiments

For each scene 9 views as inputs, 1 as target

"There is a pink cone to the left of a red torus. There is a pink cone close to a purple cone. The cone is to the left of the cone. There is a red torus to the right of a purple cone."
"There are two objects in the image. In the back left corner is a light green cone, about half the height of the wall. On the right side of the image is a bright red capsule. It is about the same height as the cone, but it is more forward in the plane of the image."
Natural language experiments

- SLIM (NL) -- only NL
- SLIM\(^+\) (NL) -- generation SYN, representation NL
- SLIM\(^+\) (NL + SYN) -- generation SYN, representation 50/50 NL+SYN
ELBO loss and human ranking

![Graph showing ELBO loss and human ranking for various methods and partitions. The graph includes bars for train, valid, and test partitions, with different methods (DRAW, SLIM, SLIM\dagger) and configurations (NL, NL+SYN, SYN).]
Representation analysis

Semantic coherence ("Car" vs "Banana" and "Car" vs "SUV")
## Representation analysis

Semantic coherence ("Car" vs "Banana" and "Car" vs "SUV")

<table>
<thead>
<tr>
<th></th>
<th>Base sentence</th>
<th>&quot;a young woman in front of an old man.&quot;</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>Noun change</td>
<td>&quot;a young man in front of an old woman.&quot;</td>
</tr>
<tr>
<td>N</td>
<td>Adjective change</td>
<td>&quot;an old woman in front of a young man.&quot;</td>
</tr>
<tr>
<td>A</td>
<td>Preposition change</td>
<td>&quot;a young woman behind an old man.&quot;</td>
</tr>
<tr>
<td>P</td>
<td>Meaning preservation</td>
<td>&quot;an old man behind a young woman.&quot;</td>
</tr>
</tbody>
</table>

Table 2: Gershman and Tenenbaum (2015) transformations.
Representation analysis

Semantic coherence (“Car” vs “Banana” and “Car” vs “SUV”)

Table 2: Gershman and Tenenbaum (2015) transformations.

<table>
<thead>
<tr>
<th></th>
<th>Base sentence</th>
<th>&quot;a young woman in front of an old man.&quot;</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>Noun change</td>
<td>&quot;a young man in front of an old woman.&quot;</td>
</tr>
<tr>
<td>N</td>
<td>Adjective change</td>
<td>&quot;an old woman in front of a young man.&quot;</td>
</tr>
<tr>
<td>A</td>
<td>Preposition change</td>
<td>&quot;a young woman behind an old man.&quot;</td>
</tr>
<tr>
<td>P</td>
<td>Meaning preservation</td>
<td>&quot;an old man behind a young woman.&quot;</td>
</tr>
</tbody>
</table>

Most to least similar: B = M > P > A > N
Semantic coherence analysis results

$$B = M > P > A > N$$
Conclusion

The model is able to

- represent scenes from language descriptions
- independently from viewpoint and paraphrasing
- can reconstruct reasonable images from the representations
- improve NL performance by using SYN data
References