Vision-and-Language Navigation: Interpreting visually-grounded navigation instructions in real environments

Peter Anderson1, Qi Wu2, Damien Teney2, Jake Bruce3, Mark Johnson4, Niko Sünderhauf5, Ian Reid2, Stephen Gould1, Anton van den Hengel2
1Australian National University, 2University of Adelaide, 3Queensland University of Technology, 4Macquarie University, 5Transitioning to Georgia Tech

1. Motivation
- Connect language and vision to actions.
- Recent availability of 3D reconstructions at large scale is an enabler for research on embodied agents.
- Timely to refocus on the intersection of computer vision, NLP and robotics.

- Given a natural language navigation instruction, navigate through a real environment to find the goal location.

3. Matterport3D Simulator
- Simulator for embodied visual agents, based on the Matterport3D dataset containing:
  - 10,800 panoramas
  - 90 diverse buildings
- Discrete motion but with continuous camera control and real images.

4. Room-to-Room (R2R) Navigation Dataset
Data Collection:
- Sampled 7,169 shortest paths between locations (mostly) in different rooms.
- Collected 21,567 navigation instructions (3 per path) using crowd workers and a WebGL interface (1,600 hours).
- Environment splits:
  - 61 training / val-seen, 11 val-unseen, 18 test (unseen).
- Distribution of navigation instructions based on their first words:

5. Baseline Seq2Seq Agent
- Instruction encoder (with attention)
- Decoder observes the image and outputs action
- LSTM-based Seq2Seq baseline model outputting a distribution over 6 actions: left, right, up, down, forward & stop.
- Image features from ResNet-152.
- Training with ‘student-forcing’ (sampling the next action) outperforms ‘teacher-forcing’ (selecting the ground-truth action).

6. Evaluation
Clear Evaluation Protocol:
- Report navigation error (distance from goal) for each instruction in the unseen test environments.
- ‘Success’ when navigation error < 3m.
- Agent must choose to stop (also report success rate with oracle stopping).

Test (unseen) performance:

- Unseen environments prove very challenging for Seq2Seq.
  - Test server available
  - More data coming soon

<table>
<thead>
<tr>
<th>Trajectory Length (m)</th>
<th>Navigation Error (m)</th>
<th>Success (%)</th>
<th>Oracle Success (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>9.93</td>
<td>9.77</td>
<td>13.2</td>
</tr>
<tr>
<td>Seq2Seq</td>
<td>8.13</td>
<td>7.85</td>
<td>20.4</td>
</tr>
<tr>
<td>Human</td>
<td>11.90</td>
<td>1.61</td>
<td>86.4</td>
</tr>
<tr>
<td>Shortest Path</td>
<td>9.93</td>
<td>0.00</td>
<td>100</td>
</tr>
</tbody>
</table>

Examples of new vocabulary encountered in unseen environments:
- hieroglyphs
- Squiggle
- mannequins
- tawort

*Data collection was generously supported by a Facebook ParAI Research Award.

Simulator, dataset, models & test server available via: https://bringmeaspoon.org