Multimodal Text Style Transfer for Outdoor Vision-and-Language Navigation

Wanrong Zhu¹, Xin Wang², Tsu-Jui Fu¹, An Yan³, Pradyumna Narayana⁴, Kazoo Sone⁴, Sugato Basu⁴, William Yang Wang¹
¹UC Santa Barbara, ²UC Santa Cruz, ³UC San Diego, ⁴Google
Outdoor Vision-and-Language Navigation

- Challenges:
  - Complicated visual input
  - Lack of annotated instructions

Orient yourself so that the umbrellas are to the right. Go straight and take a right at the first intersection. At the next intersection there should be an old-fashioned store to the left. There is also a dinosaur mural to the right. Touchdown is on the back of the dinosaur.
External Resource

- Google Street View
  - Street view images
  - Machine-generated instructions
Multimodal Text Style Transfer Framework Overlook

Touchdown Dataset
- Orient yourself so that the red deli awning is on your right. Turn left at the intersection. Go straight until you see a parking garage on your right.

StreetLearn Dataset
- Head southwest on 5th Ave toward E 49th St. Turn right onto W 47th St.

Multimodal Text Style Transfer Model
- generated by machine focus on street names and directions

Cross-Modal Reasoning Navigator
- annotated by human frequently refer to objects

Pre-train

Train

Inference Input

Inference Sample

StreetLearn Trajectory & Style-Modified Instructions
- Head down the street with traffic on your right. Turn right onto the street.
Multimodal Text Style Transfer: Inference

Head southwest on 5th Ave toward E 49th St. Turn right onto W 47th St.

Masking

[MASK] on [MASK] toward [MASK].

[MASK] right onto [MASK].

Transferring Text Style

Multimodal Text Style Transfer Model

Head down the street with traffic on your right. Turn right onto the street.
Cross Modal Reasoning Navigator

Action Predictor

concat

Cross-Modal Transformer

Instruction Encoder

Orient yourself so that the red dell awning is on your right.
Go forward.
Turn left at the intersection.
Go straight until you see a parking garage on your right.

View Encoder

t = 1  t = 2  t = 3
Tasks & Datasets

- Task: Touchdown dataset [1]
- External resource: StreetLearn dataset [2]

Datasets Comparison

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Trajectory source</th>
<th>Instruction source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Touchdown</td>
<td>Google Street Views</td>
<td>Human-written</td>
</tr>
<tr>
<td>StreetLearn</td>
<td>Google Street Views</td>
<td>Google Map API</td>
</tr>
</tbody>
</table>

Experiment Settings

- Models:
  - RCONCAT
  - GA
  - VLN-Transformer (ours)

- Metrics
  - TC: task completion rate
  - SED: success weighted by edit distance
  - CLS: coverage weighted by length score
Experiment Results

- \( +M50 \): pre-train on a StreetLearn subset with machine-generated instructions
- \( +M50 +\text{style} \): pre-train on a StreetLearn subset with style-modified instructions

Graphs showing comparisons in TC, SED, and CLS.
## Case Study

- **Red tokens**: contradictions with ground truth.
- **Blue tokens**: alignment with ground truth.

<table>
<thead>
<tr>
<th>StreetLearn</th>
<th><strong>Head northwest</strong> on W 35th St toward Hudson Blvd E. <strong>Turn right</strong> at the 1st cross street onto Hudson Blvd E.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original Speaker</td>
<td>Turn so the red construction is on your left and the red brick building is on your right. Go forward to the intersection and turn right. You'll have a red brick building with a red awning on your right.</td>
</tr>
<tr>
<td>Multimodal Text Style Transfer</td>
<td>Move forward with traffic on the right turn right at the light. Continue straight.</td>
</tr>
</tbody>
</table>
Thanks!