Source-Target Inference Models for Spatial Instruction Understanding

Hao Tan          Mohit Bansal

AAAI 2018

{haotan, mbansal}@cs.unc.edu
The University of North Carolina at Chapel Hill
Problem Description

Instruction
Move the block closest to the right table edge so it is to the left of the stack near the front left table corner.

Challenges

 Triumph: Long, Complex and Diverse instructions in natural language
 - Average length of instructions: 22

 Scene: Blank blocks

 Data: Small dataset
 - 397 different scenes
 - 3573 different instructions
Subtasks Trained Jointly

Move the block closest to the right table edge so it is to the left of the stack near the front left table corner.
Subtask: Source Block Selection

Input

Move the block closest to the right table edge so it is to the left of the stack near the front left table corner.

Source

Metric: Accuracy
The frequency that our model selects the correct source block.
Subtask: Target Position Prediction

Move the block closest to the right table edge so it is to the left of the stack near the front left table corner.

Metric: Mean Distance
The average distance between the prediction and the ground truth
Joint Model

Put the topmost block to the right of the stack.

---

LSTM-RNN

---

Attention Module

---

Offset

Reference

Source

Target

---

Sampling Loss

or

Expectation Loss

---

Cross Entropy Loss

---

Policy Gradient

---

Put the topmost block to the right of the stack.
Language Encoder

LSTM-RNN Encoder:

\[ \{h_1, h_2, \ldots, h_T\} = \text{LSTM}\{w_1, w_2, \ldots, w_T\} \]
Block Encoder

---

Put the topmost block to the right of the stack.

---

Fully-Connected Layer:

\[ c_i = \sigma(Wb_i + a) \]
Alignment (Attention)

- Put the topmost block to the right of the stack.

- **Attention Module**
  - LSTM-RNN
  - Source
  - Reference
  - Offset

- Policy Gradient
- Sampling Loss
- Expectation Loss
- Cross Entropy Loss
- Target

- Put the topmost block to the right of the stack.
Basic Attention Module

Instruction Embedding

\[ h_t \quad \ldots \ldots \quad h_T \]

Block Embedding

\[ c_i \]

\[ W \]

\[ W \]

Softmax

\[ \text{prob}(b_i) \]

\[ \text{score}(b_i) = c_i^T W_A h_T \]

\[ \text{prob}(b_i) = \text{soft max} (\{ \text{score}(b_i) \}) \]
Dual Attention

Instruction Embedding

Block Embedding

\[ a_{i,t} = c_i^T W_{\text{WORD}} h_t \]
\[ \alpha_{i,t} = \text{softmax}(\{a_{i,t}\}) \]
\[ z_i = \sum_t \alpha_{i,t} h_t \]

\[ \text{score}(b_i) = c_i^T W_{\text{BLOCK}} z_i \]
\[ \text{prob}(b_i) = \text{softmax}(\{\text{score}(b_i)\}) \]
CNN Attention

\[
\text{score}(b_i) = c_i^\top W_A h_{\text{CNN}}
\]

\[
\text{prob}(b_i) = \text{soft max}(\{\text{score}(b_i)\})
\]
Attention Module: Ablation Analysis

Dual / CNN vs Bilinear Attention

- 4% improvement in source accuracy
- 0.15 improvement in target mean distance
Source Loss:

$$\mathcal{L}_{\text{SOURCE}} = -\log(\text{prob}(b_{GT}))$$
Target Prediction

- Put the topmost block to the right of the stack.

- LSTM-RNN

- Attention Module

- FC

- Source

- Cross Entropy Loss

- Offset

- Reference

- Target

- Sampling Loss

- Expectation Loss

- Policy Gradient

- or

- Cross Entropy Loss

- Sampling Loss
Target = Reference + Offset

Move the block closest to the right table edge so it is to the left of the stack near the front left table corner.
Problem: No intermediate ground truth

Reference: \( P(R = b_i | I) \propto \exp(A_R(c_i, H)) \)

Offset: \( P(O = o | I) \propto N(\mu_o, \Sigma_o) \)

\[
\mu_o = W_2 \sigma(W_1 h_{CNN} + a_1) + a_2
\]
Target Prediction (Expectation Method)

Expectation Loss:

\[ \mathcal{L}_{\text{EXP}} = \| t_{GT} - \mathbb{E}[R + O] \|^2 \]
Target Prediction (Sampling Method)

- $r \sim P(R = r)$
- $o \sim P(O = o)$
- $t_s = r + o$

**Sampling**

$$\mathcal{L}_{SMP} = \mathbb{E} \left\| t_{GT} - R - O \right\|$$

Problem: Not fully-differentiable
Policy Gradient (REINFORCE)

- Sampling R (the reference block)
- Sampling O (the offset)
- Calculate the reward $\|t_{GT} - R - O\|$ 

Policy Gradient

$$\frac{\partial}{\partial \theta} L_{SMP} \approx \frac{1}{K} \sum_i \left[ \left( \frac{\partial \log P(R = r_i)}{\partial \theta} \right) - 1 \frac{\partial (o_i - \mu_O)\Sigma_O^{-1}(o_i - \mu_O)}{\partial \theta} \right] \cdot \|t_{GT} - r_i - o_i\|$$
Techniques in Policy Gradient

- Using linear-regressor baseline (based on the instruction)
  - Actual reward: R
  - Baseline: b (expected reward)
    \[ b = W h_{\text{CNN}} + a \]
  - Reward used in policy gradient: R - b

- Annealing method
  - Gradually anneal the expectation loss to the sampling loss
Jointly Training: the losses of source and target are summed.

Parameter Sharing: Share the weights in language encoder and block encoder.
Result

- Source Accuracy: 56.8% vs 10%
- Target Mean Distance: 2.90 (in block length) vs 3.70

<table>
<thead>
<tr>
<th>Model</th>
<th>Source Accuracy</th>
<th>Source Median</th>
<th>Source Mean</th>
<th>Target Median</th>
<th>Target Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>End-to-End FFN (Bisk, Yuret, and Marcu 2016)</td>
<td>9.0%</td>
<td>3.45</td>
<td>3.52</td>
<td>3.60</td>
<td>3.94</td>
</tr>
<tr>
<td>End-to-End RNN (Bisk, Yuret, and Marcu 2016)</td>
<td>10.0%</td>
<td>3.29</td>
<td>3.47</td>
<td>3.60</td>
<td>3.70</td>
</tr>
<tr>
<td>Our Expectation Model</td>
<td>56.1%</td>
<td>0.00</td>
<td>2.21</td>
<td>2.78</td>
<td>3.07</td>
</tr>
<tr>
<td>Our Sampling Model</td>
<td>56.3%</td>
<td>0.00</td>
<td>2.18</td>
<td>3.12</td>
<td>3.18</td>
</tr>
<tr>
<td>Our Expectation Model w/ Ensemble</td>
<td>56.6%</td>
<td>0.00</td>
<td>2.12</td>
<td>2.65</td>
<td>2.91</td>
</tr>
<tr>
<td>Our Sampling Model w/ Ensemble</td>
<td>56.8%</td>
<td>0.00</td>
<td>2.11</td>
<td>2.71</td>
<td>2.90</td>
</tr>
</tbody>
</table>
Positive Examples

The box in the bottom right, slightly right of center, moves one space north of the tower.

Move the highest block down to below and in front of the right stack of blocks.

Take the leftmost front block and place it on top of the stack of two blocks furthest to the back.

Take the block from the last row and hide it behind the tower.

❌ is the predicted target position
Negative Examples

- is the predicted target position
- is the wrong source block

Move the block closest to the bottom left corner so that it is on top of the block at the top of the backwards L.

The box next to the Tetris structure moves two spaces left and one half up.

Slide the block left of the two in the top right over and on top of the block in front of the tower.

In the 3-piece-long line, the middle box takes a second story from the middle box in the top row.
Negative Example 1

Move the block closest to the bottom left corner so that it is on top of the block at the top of the backwards L.

Source: Correct
Target: 0.437

Source: Correct
Target: 5.914
Negative Example 2

The box next to the Tetris structure moves two spaces left and one half up.
Thank you!

Hao Tan  Mohit Bansal

{haotan, mbansal}@cs.unc.edu
The University of North Carolina at Chapel Hill