Vision Beyond Pixels: Visual Reasoning via Blocksworld Abstractions

Tejas Gokhale
tgokhale@asu.edu
https://www.public.asu.edu/~tgokhale/

IJCAI Doctoral Consortium
Macao
August 2019
The Physical World

has objects and actions

The world has *objects*. *Actions* move or change objects.
The world has *blocks*. *Actions* move or change blocks.
The world has *objects*. *Actions* move or change objects.
The world has *objects*. *Actions* move or change objects.
Task:

Given: a source image ($I^S$) and a target image ($I^T$)

Predict: an event-sequence $M = [m_1, m_2, \ldots, m_L]$ such that performing $M$ on $I^S$ leads to $I^T$

Predict intermediate actions between two scenes
## Blocksworld Image Reasoning Dataset (BIRD)

<table>
<thead>
<tr>
<th>Source</th>
<th>Target</th>
<th>GT Sequence</th>
</tr>
</thead>
</table>
| ![Source Image 1](image1.png) | ![Target Image 1](image2.png) | Move(B, out)  
Move(O, R)  
Move(G, O) |
| ![Source Image 2](image3.png) | ![Target Image 2](image4.png) | Move(B, out)  
Move(G, out)  
Move(P, out) |
| ![Source Image 3](image5.png) | ![Target Image 3](image6.png) | Move(R, table)  
Move(Y, R)  
Move(B, table)  
Move(O, G)  
Move(B, O) |
| ![Source Image 4](image7.png) | ![Target Image 4](image8.png) | Move(Y, table)  
Move(O, Y)  
Move(P, O)  
Move(G, P)  
Move(0, out) |

[https://asu-active-perception-group.github.io/bird_dataset_web/](https://asu-active-perception-group.github.io/bird_dataset_web/)
Modular Approach: Visual Perception + Inductive Logic Programming

Two-Stage Modular Approach

Combines learning (visual perception) and reasoning (event-sequencing)

- **Visual Perception Module** encodes each image into an interpretable representation
- **Event Sequencing Module** reasons about the intermediate actions between these configurations
Modular Approach: Results on BIRD

Metrics

FSA (full sequence accuracy): % of exact matches
\[ \text{FSA} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}\{y^i = \hat{y}^i\}. \]

SLA (step-level accuracy): % of common moves between predicted and ground-truth sequences
\[ \text{SLA} = \frac{1}{N} \sum_{i=1}^{N} \frac{\sum_{l=1}^{L} \mathbb{1}\{y_{l}^i = \hat{y}_{l}^i\}}{L}. \]

<table>
<thead>
<tr>
<th>Approach</th>
<th>Method</th>
<th>FSA</th>
<th>SLA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td></td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>End-to-End Learning</td>
<td>ResNet50</td>
<td>30.52</td>
<td>36.26</td>
</tr>
<tr>
<td></td>
<td>PSPNet</td>
<td>35.04</td>
<td>56.69</td>
</tr>
<tr>
<td></td>
<td>RN</td>
<td>34.37</td>
<td>52.09</td>
</tr>
<tr>
<td>Modular (GT Perception)</td>
<td>FC</td>
<td>68.87</td>
<td>72.58</td>
</tr>
<tr>
<td></td>
<td>QL</td>
<td>84.10</td>
<td>87.83</td>
</tr>
<tr>
<td></td>
<td><strong>ILP</strong></td>
<td><strong>100</strong></td>
<td><strong>100</strong></td>
</tr>
<tr>
<td>Modular (Noisy Perception)</td>
<td>FC</td>
<td>56.25</td>
<td>60.24</td>
</tr>
<tr>
<td></td>
<td>QL</td>
<td>68.98</td>
<td>71.17</td>
</tr>
<tr>
<td></td>
<td><strong>ILP</strong></td>
<td><strong>83.60</strong></td>
<td><strong>88.53</strong></td>
</tr>
</tbody>
</table>
Modular Approach: Evaluation of Inductive Generalization

Implications

Inductive Generalization for IES is:

The ability of predicting sequences of any length

(Even larger than the sequences encountered while training)

Training Set: max. seq. length = L
Test Set: min. seq. length = L +1
X-axis: max. seq. length in Training Set
Y-axis: accuracy on test set
Modular Approach: Evaluation of Inductive Generalization

**Implications**

Inductive Generalization for IES is: **The ability of predicting sequences of any length**

(Even larger than the sequences encountered while training)

**Training Set:** max. seq. length = L

**Test Set:** min. seq. length = L +1

**X-axis:** max. seq. length in Training Set

**Y-axis:** accuracy on **test set**
Replace Perception Module, Reuse Sequencing Module

- Domain specific perception module
- Configurations converted to blocksworld abstraction
- Re-use Blocksworld Sequencing for predicting intermediate steps !!!
Ongoing Work (Sneak Peak)
Ongoing Work (Sneak Peak)

Continuing work from Telluride 2019: http://tellurideneuromorphic.org

Agents

Agents have *desires, intentions and goals*

**Teleological Reasoning:** Agents perform actions in order to achieve a goal.

**Input:** videos of “agent” block moving over obstacles to reach the goal

**Task:** learn which goals are viewed as more “profitable” by the agent, i.e. model the agent’s preferences

**Dataset:** videos with varying obstacles, distances from goals and goal attributes
Ongoing Work (Sneak Peak)

Ongoing work in collaboration with Jacob Fang (ASU)

Commonsense Extraction from Videos

Going beyond caption generation

Caption: A soldier fights with his enemy.

Commonsense:
- **Attribute**: He is seen as brave.
- **Intention**: Because he wants to protect his friends.
- **Effect**: As a result, he goes to the drug store.

Learn to output the causes and effects of an agent’s actions directly from video
References:
