Neural-Symbolic VQA: Disentangling Reasoning from Vision and Language Understanding

NeurIPS 2018: Yi et. al.

Presented by Victoria Proetsch
Visual Question Answering

CLEVR, CLEVR-Humans (Johnson et al. 2017), and Minecraft.
1. **Scene Parsing**: Convert the image into a symbolic representation

2. **Question Parsing**: Convert the natural language question into a set of instructions

3. **Program Execution**: Apply the program to the scene representation and obtain an answer
Advantages of Separation:

- Reduces amount of needed training data
- Scene representation and reasoning module are lightweight, easy to compute and store
- Reasoning process “can be analyzed and diagnosed step-by-step”
Architecture Overview

Figure from Yi et al., 2018
1. Segment proposals generated with Mask R-CNN
2. Categorical labels (size, shape, material, color) also predicted by Mask R-CNN
3. Segment proposals with score > 0.9 pass to next step
4. Each segment resized to 224x224
5. Paired with original image for location information and passed to next CNN (ResNet-34)
6. Predict continuous values: x, y, z coordinates
Architecture Overview

Figure from Yi et al., 2018.
Question Parsing

Detail of Figure from Yi et al., 2018.

- Natural language question $\Rightarrow$ Sequence of program functions
- Seq2seq translation problem
- Use Encoder-Decoder LSTM
Question Parsing

Encoder: Bi-directional LSTM

\[ e_i = [e_i^F, e_i^B], \quad \text{where} \quad e_i^F, h_i^F = \text{LSTM}(\Phi_E(x_i), h_{i-1}^F), \]
\[ e_i^B, h_i^B = \text{LSTM}(\Phi_E(x_i), h_{i+1}^B) \]

Decoder: LSTM

\[ q_t = \text{LSTM}(\Phi_D(y_{t-1})) \]

Output:

\[ \alpha_{ti} \propto \exp(q_t^T W_A e_i), \quad c_t = \sum_i \alpha_{ti} e_i, \quad y_t \sim \text{softmax}(W_O [q_t, c_t]) \]
Architecture Overview

Figure from Yi et al., 2018.
Program Execution

Applies the output of the Question Parser to the output of the Scene Parser.

Detail of figures from Yi et al., 2018
Architecture Review

Figure from Yi et al., 2018.
Training

Scene Parser:

- Mask R-CNN trained on CLEVR dataset with ground truth boxes
- Proposed segments from Mask R-CNN paired with ground truth annotations to train attribute extractor
- 30,000 iterations, 4,000 images in 8-image minibatches
Training

Question Parser:

1. Supervised: question-program ground truth pairs - 20,000 iterations
2. REINFORCE: Question Parser paired with program executor. Reinforcement training performed based on correctness of question answer. - Max. 2M iterations
Quantitative results on CLEVR

<table>
<thead>
<tr>
<th>Methods</th>
<th>Count</th>
<th>Exist</th>
<th>Compare Number</th>
<th>Compare Attribute</th>
<th>Query Attribute</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Humans [Johnson et al., 2017b]</td>
<td>86.7</td>
<td>96.6</td>
<td>86.4</td>
<td>96.0</td>
<td>95.0</td>
<td>92.6</td>
</tr>
<tr>
<td>CNN+LSTM+SAN [Johnson et al., 2017b]</td>
<td>59.7</td>
<td>77.9</td>
<td>75.1</td>
<td>70.8</td>
<td>80.9</td>
<td>73.2</td>
</tr>
<tr>
<td>N2NMN* [Hu et al., 2017]</td>
<td>68.5</td>
<td>85.7</td>
<td>84.9</td>
<td>88.7</td>
<td>90.0</td>
<td>83.7</td>
</tr>
<tr>
<td>Dependency Tree [Cao et al., 2018]</td>
<td>81.4</td>
<td>94.2</td>
<td>81.6</td>
<td>97.1</td>
<td>90.5</td>
<td>89.3</td>
</tr>
<tr>
<td>CNN+LSTM+RN [Santoro et al., 2017]</td>
<td>90.1</td>
<td>97.8</td>
<td>93.6</td>
<td>97.1</td>
<td>97.9</td>
<td>95.5</td>
</tr>
<tr>
<td>IEP* [Johnson et al., 2017b]</td>
<td>92.7</td>
<td>97.1</td>
<td>98.7</td>
<td>98.9</td>
<td>98.1</td>
<td>96.9</td>
</tr>
<tr>
<td>CNN+GRU+FiLM [Perez et al., 2018]</td>
<td>94.5</td>
<td>99.2</td>
<td>93.8</td>
<td>99.0</td>
<td>99.2</td>
<td>97.6</td>
</tr>
<tr>
<td>DDRprog* [Suarez et al., 2018]</td>
<td>96.5</td>
<td>98.8</td>
<td>98.4</td>
<td>99.0</td>
<td>99.1</td>
<td>98.3</td>
</tr>
<tr>
<td>MAC [Hudson and Manning, 2018]</td>
<td>97.1</td>
<td>99.5</td>
<td>99.1</td>
<td>99.5</td>
<td>99.5</td>
<td>98.9</td>
</tr>
<tr>
<td>TbD+reg+hres* [Mascharoka et al., 2018]</td>
<td>97.6</td>
<td>99.2</td>
<td>99.4</td>
<td>99.6</td>
<td>99.5</td>
<td>99.1</td>
</tr>
<tr>
<td>NS-VQA (ours, 90 programs)</td>
<td>64.5</td>
<td>87.4</td>
<td>53.7</td>
<td>77.4</td>
<td>79.7</td>
<td>74.4</td>
</tr>
<tr>
<td>NS-VQA (ours, 180 programs)</td>
<td>85.0</td>
<td>92.9</td>
<td>83.4</td>
<td>90.6</td>
<td>92.2</td>
<td>89.5</td>
</tr>
<tr>
<td>NS-VQA (ours, 270 programs)</td>
<td><strong>99.7</strong></td>
<td><strong>99.9</strong></td>
<td><strong>99.9</strong></td>
<td><strong>99.8</strong></td>
<td><strong>99.8</strong></td>
<td><strong>99.8</strong></td>
</tr>
</tbody>
</table>

- # of pretraining question/program pairs.
- Reported mean of 3 runs
Quantitative results on CLEVR

Figure from Yi et al., 2018
Qualitative Results - NS-VQA vs IEP

Figure from Yi et al., 2018
CLEVER-CoGenT

Ability to generalize to unseen combinations of attributes

Training
- Cubes are gray, blue, brown, or yellow
- Cylinders are red, green, purple, or cyan
- Spheres can have any color

Evaluation
- Cubes are red, green, purple, or cyan
- Cylinders are gray, blue, brown, or yellow
- Spheres can have any color

Figure from https://cs.stanford.edu/people/jcjohns/clevr/
CLEVER-CoGenT Results

- Scene Parser trained both sets, question parser only on set A.
- +Gray adds gray channel to image parser
- +Ori uses Scene parser trained on original set

<table>
<thead>
<tr>
<th>Methods</th>
<th>Not Fine-tuned</th>
<th>Fine-tune on</th>
<th>Fine-tuned</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A</td>
<td>B</td>
<td>B</td>
</tr>
<tr>
<td>CNN+LSTM+SA</td>
<td>80.3</td>
<td>68.7</td>
<td>B</td>
</tr>
<tr>
<td>IEP (18K programs)</td>
<td>96.6</td>
<td>73.7</td>
<td>B</td>
</tr>
<tr>
<td>CNN+GRU+FiLM</td>
<td>98.3</td>
<td>78.8</td>
<td>B</td>
</tr>
<tr>
<td>TbD+reg</td>
<td>98.8</td>
<td>75.4</td>
<td>B</td>
</tr>
<tr>
<td>NS-VQA (ours)</td>
<td>99.8</td>
<td>63.9</td>
<td>B</td>
</tr>
<tr>
<td>NS-VQA (ours)</td>
<td>99.8</td>
<td>63.9</td>
<td>A+B</td>
</tr>
<tr>
<td>NS-VQA+Gray (ours)</td>
<td>99.6</td>
<td>98.4</td>
<td>-</td>
</tr>
<tr>
<td>NS-VQA+Ori (ours)</td>
<td><strong>99.8</strong></td>
<td><strong>99.7</strong></td>
<td>-</td>
</tr>
</tbody>
</table>

Figure from Yi et al., 2018
CLEVR-Humans & Results

CLEVR images paired with human-generated questions.

Figure from https://cs.stanford.edu/people/jcjohns/iep/

<table>
<thead>
<tr>
<th># Programs</th>
<th>NS-VQA</th>
<th>IEP</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>60.2</td>
<td>38.7</td>
</tr>
<tr>
<td>200</td>
<td>65.2</td>
<td>40.1</td>
</tr>
<tr>
<td>500</td>
<td>67.8</td>
<td>49.2</td>
</tr>
<tr>
<td>1K</td>
<td>67.8</td>
<td>63.4</td>
</tr>
<tr>
<td>18K</td>
<td>67.0</td>
<td>66.6</td>
</tr>
</tbody>
</table>

Figure from Yi et al., 2018
New Scenes: Minecraft

- 10,000 scenes containing 3-6 objects, 12 entities.
- category, 2D coordinates, and direction facing
- Hierarchical attributes:
  - “wolf” and “pig” are also both “animals”

Figures from Yi et al., 2018
Summary

- Successfully disentangles visual perception from logical reasoning
- Two-step training of question parser reduces need for annotated training examples
- Somewhat generalizable to new contexts
- Drawbacks: requires building structured representations of scenes and sentence meanings