Deep learning evaluation using ShapeWorld

Alexander Kuhnle

Department of Computer Science and Technology
University of Cambridge
Evaluation methodology

labeled dataset
(>100k data points)

∼90% training split

deep neural network
(recurrent sequence model)

∼10% test split

evaluation
Evaluation methodology

data source
(method for obtaining data)

feedback & adjustment

series of probing tests

evaluation

training data

deep neural network
(recurrent sequence model)
ShapeWorld generation framework

Sampled world model

```json
{
  size: 64, color: {name: black, shade: 0.0}, noise-range: 0.1, entities:
  [{
    shape: {name: cross, extent: {x: 0.10, y: 0.10}}, rotation: 0.06,
    color: {name: yellow, shade: -0.24}, center: {x: 0.47, y: 0.28}
  },
    {shape: {name: cross, extent: {x: 0.08, y: 0.08}}, rotation: 0.76,
    color: {name: red, shade: 0.26}, center: {x: 0.49, y: 0.65}
  },
    {shape: {name: pentagon, extent: {x: 0.09, y: 0.08}}, rotation: 0.27,
    color: {name: yellow, shade: -0.16}, center: {x: 0.15, y: 0.91}
  },
    {shape: {name: circle, extent: {x: 0.12, y: 0.12}}, rotation: 0.53,
    color: {name: red, shade: -0.12}, center: {x: 0.80, y: 0.37}
  },
    {shape: {name: cross, extent: {x: 0.09, y: 0.09}}, rotation: 0.73,
    color: {name: yellow, shade: -0.42}, center: {x: 0.92, y: 0.73}}
}
```

Image

Linguistic representation

Caption

“There is a blue circle.”
“Most crosses are yellow.”
“A pentagon is below a cross.”

Agreement?
“A pentagon is above a green ellipse, and no blue shape is an ellipse.”

↑ ERG + ACE realization ↑

↑ Internal DMRS mapping ↑

<table>
<thead>
<tr>
<th>( \exists a )</th>
<th>( a.\text{shape}=\text{pg} )</th>
<th>( a.\text{y}&gt;b.\text{y} )</th>
<th>( \exists b )</th>
<th>( b.\text{color}=\text{gr} )</th>
<th>( b.\text{shape}=\text{el} )</th>
<th>( \land )</th>
<th>( \neg \exists c )</th>
<th>( c.\text{color}=\text{bl} )</th>
<th>( \text{true} )</th>
<th>( c=d )</th>
<th>( \exists d )</th>
<th>( d.\text{shape}=\text{el} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \exists a ): ( a.\text{shape}=\text{pg} )</td>
<td>( a.\text{y}&gt;b.\text{y} )</td>
<td>( \exists b ): ( b.\text{color}=\text{gr} \land b.\text{shape}=\text{el} )</td>
<td>( \land )</td>
<td>( \neg \exists c ): ( c.\text{color}=\text{bl} )</td>
<td>( \text{true} )</td>
<td>( c=d )</td>
<td>( \exists d ): ( d.\text{shape}=\text{el} )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \exists a ): ( a.\text{shape}=\text{pg} \land [\exists b : b.\text{color}=\text{gr} \land b.\text{shape}=\text{el} \land a.\text{y}&gt;b.\text{y}] )</td>
<td>( \land )</td>
<td>( \neg \exists c ): ( c.\text{color}=\text{bl} \land [\exists d : d.\text{shape}=\text{el} \land c=d] )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
ShapeWorld: language generation

- **World model**
  - **JSON spec**
    - [attr]: _blue_a_sw e?
    - =1=> [type]: node

- **Captioner**
  - **DMRS snippets**
    - **DMRS graph**
      - convert (+ post-processing)
        - _blue_a_sw predsort(?)

- **RegularTypeCaptioner**

- **Grammar**
  - **Surf ace string**
    - "There is a blue shape."
Performance breakdown and generalisation

<table>
<thead>
<tr>
<th>Dataset</th>
<th>CNN-LSTM</th>
<th>CNN-LSTM-SA</th>
<th>FiLM</th>
</tr>
</thead>
<tbody>
<tr>
<td>(single-shape)</td>
<td>—</td>
<td>—</td>
<td>100.0</td>
</tr>
<tr>
<td>existential</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
</tr>
<tr>
<td>logical</td>
<td>79.7</td>
<td>76.5</td>
<td>99.9</td>
</tr>
<tr>
<td>numbers</td>
<td>75.0</td>
<td>99.1</td>
<td>99.6</td>
</tr>
<tr>
<td>quantifiers</td>
<td>72.1</td>
<td>84.8</td>
<td>97.7</td>
</tr>
<tr>
<td>(simple-spatial)</td>
<td>81.4</td>
<td>81.9</td>
<td>85.1</td>
</tr>
<tr>
<td>relational</td>
<td>—</td>
<td>—</td>
<td>50.6</td>
</tr>
<tr>
<td>implicit-rel</td>
<td>—</td>
<td>—</td>
<td>52.9</td>
</tr>
<tr>
<td>superlatives</td>
<td>—</td>
<td>—</td>
<td>50.8</td>
</tr>
</tbody>
</table>

- Three crosses
- Four triangles
- Four crosses
Replication of psycholinguistic experiments

Random
Paired
Partitioned

“More/less than half the shapes are X?”
Intermediate representations and multilingual data

Existential [ObjectType1 Attribute-shape-pentagon] [Relation-y-rel--1 [ObjectType Attribute-color-green] Attribute-shape-ellipse] ]

“A pentagon is above a green ellipse.”

有某一个红色正方形
有一个圆形
有某一个绿色半圆形
有某一个紫色十字形
有某一个红色半圆形
Real-world vs artificial data

- **real-world data** vs **artificial data**
  - limited and expensive $\leftrightarrow$ unlimited amount
  - uncontrolled content $\leftrightarrow$ configurable content
  - sparse instance coverage $\leftrightarrow$ targeted instance coverage
  - monolithic benchmark $\leftrightarrow$ set of tailored probing tests
  - test interpolation ability $\leftrightarrow$ test extrapolation ability

$\Rightarrow$ **Complementary evaluation paradigms**
Thank you for your attention!

Questions?