“Unit-testing” deep learning with synthetic data for more informative evaluation

Alexander Kuhnle

Department of Computer Science and Technology
University of Cambridge (UK)

MILA – 10th October 2018
Overview

- Visual question answering
- Problems with the VQA Dataset
- Evaluation methodology
- ShapeWorld generation framework
- Evaluation of FiLM on ShapeWorld
Visual question answering

Examples

Where is this cat laying?
Is the cat awake?
What color is the cat?

Is the cat facing the computer?
Is the cat typing?
Is the cat playing with the mouse?

What object is shining on the animal?
What objects is the cat sitting behind?
How many cats?

How many items are on the bookcase?
Are these two children related?
Is the dog begging for food?
Visual question answering

Examples

Where is this cat laying?  
Is the cat awake?  
What color is the cat?  

Is the cat facing the computer?  
Is the cat typing?  
Is the cat playing with the mouse?  

What object is shining on the animal?  
What objects is the cat sitting behind?  
How many cats?  

How many items are on the bookcase?  
Are these two children related?  
Is the dog begging for food?  

⇒ Visual Turing test?

Examples from VQA Dataset (http://visualqa.org/browser/)
Visual question answering

Performance over time

Performance on the VQA Dataset v1.0

Based on (incomplete) list of VQA papers with arXiv publication dates
Problems with the VQA Dataset

Question-answer biases

- What sport is...?

Problems with the VQA Dataset

Question-answer biases

- What sport is...? ⇒ tennis (41%)
Problems with the VQA Dataset
Question-answer biases

⚠️ What sport is...? ⇒ tennis (41%)

⚠️ How many...?
Problems with the VQA Dataset

Question-answer biases

- **What sport is...?** ⇒ tennis (41%)

- **How many...?** ⇒ two (39%)
Problems with the VQA Dataset

Question-answer biases

- What sport is...? ⇒ tennis (41%)
- How many...? ⇒ two (39%)
- Do you see a...?
Problems with the VQA Dataset

Question-answer biases

- What sport is...? ⇒ tennis (41%)
- How many...? ⇒ two (39%)
- Do you see a...? ⇒ yes (87%)

Examples from Goyal et al. (https://arxiv.org/abs/1612.00837)
Problems with the VQA Dataset
Complete question/image understanding

- What...?
Problems with the VQA Dataset

Complete question/image understanding

- What...? ⇒ umbrella
Problems with the VQA Dataset
Complete question/image understanding

► What...? ⇒ umbrella
► What season...?
Problems with the VQA Dataset

Complete question/image understanding

- What...? ⇒ umbrella
- What season...? ⇒ summer

Problems with the VQA Dataset

Complete question/image understanding

- What...? ⇒ umbrella
- What season...? ⇒ summer
- What season of...? ⇒ summer
- ...
- What season of year was this photo taken in? ⇒ summer
Problems with the VQA Dataset

Complete question/image understanding

- What...?  ⇒ umbrella
- What season...?  ⇒ summer
- What season of...?  ⇒ summer
- ...
- What season of year was this photo taken in?  ⇒ summer

- What does the red sign say?  ⇒ stop
Problems with the VQA Dataset

Complete question/image understanding

- What...? ⇒ umbrella
- What season...? ⇒ summer
- What season of...? ⇒ summer
- ...
- What season of year was this photo taken in? ⇒ summer

- What does the red sign say? ⇒ stop

Examples from Agrawal et al. (https://arxiv.org/abs/1606.07356) and Devi Parikh’s slides (https://newgeneralization.github.io/)
Problems with the VQA Dataset

Sensitivity to question words

- How symmetrical are the white bricks on either side of the building?
Problems with the VQA Dataset

Sensitivity to question words

- How symmetrical are the white bricks on either side of the building? ⇒ very

Problems with the VQA Dataset

Sensitivity to question words

- How symmetrical are the white bricks on either side of the building? ⇒ very
- How *spherical* are the white bricks on either side of the building? ⇒ very

Problems with the VQA Dataset

Sensitivity to question words

- How symmetrical are the white bricks on either side of the building? ⇒ very
- How *spherical* are the white bricks on either side of the building? ⇒ very
- How *soon* are the bricks *fading* on either side of the building? ⇒ very
Problems with the VQA Dataset

Sensitivity to question words

- How symmetrical are the white bricks on either side of the building? ⇒ very
- How spherical are the white bricks on either side of the building? ⇒ very
- How soon are the bricks fading on either side of the building? ⇒ very
- How fast are the bricks speaking on either side of the building? ⇒ very

Problems with the VQA Dataset
Low performance on CLEVR

▶ How many small spheres are there? ⇒ 2
▶ What number of cubes are small things or red metal objects? ⇒ 2
▶ Does the metal sphere have the same color as the metal cylinder? ⇒ Yes
▶ Are there more small cylinders than metal things? ⇒ No

Images from https://github.com/facebookresearch/clevr-dataset-gen
Evaluation methodology

Meaningful progress?

Performance on the VQA Dataset v1.0

Based on (incomplete) list of VQA papers with arXiv publication dates
Evaluation methodology
Pros and cons of crowd-sourced real-world datasets

Solve the problem/dataset?

✓

Deep learning will find a way to make effective use of the data.
Evaluation methodology
Pros and cons of crowd-sourced real-world datasets

Solve the problem/dataset? ✓
Deep learning will find a way to make effective use of the data.

Evaluate model capabilities? ?
Are these datasets appropriate to investigate this question?
- Natural?
- Difficult?
- Specific?
Evaluation methodology
Pros and cons of crowd-sourced real-world datasets

Solve the problem/dataset?
✓
Deep learning will find a way to make effective use of the data.

Evaluate model capabilities?
?
Are these datasets appropriate to investigate this question?
► Natural?
► Difficult?
► Specific?
⇒ Synthetic data!
Evaluation methodology

Other popular datasets with similar issues

**SNLI – Stanford Natural Language Inference Corpus**

C: A soccer game with multiple males playing.
H: Some men are playing a sport.
→ **entailment**

C: A smiling costumed woman is holding an umbrella.
H: A happy woman in a fairy costume holds an umbrella.
→ **neutral**

C: A man inspects the uniform of a figure in some East Asian country.
H: The man is sleeping
→ **contradiction**

**SQuAD – Stanford Question Answering Dataset**

In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under **gravity**. The main forms of precipitation include drizzle, rain, sleet, snow, graupel and hail... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals **within a cloud**. Short, intense periods of rain in scattered locations are called “showers”.

(1) What causes precipitation to fall?
⇒ **gravity**

(2) What is another main form of precipitation besides drizzle, rain, snow, sleet and hail?
⇒ **graupel**

(3) Where do water droplets collide with ice crystals to form precipitation?
⇒ **within a cloud**

Examples from Bowman et al. (https://arxiv.org/abs/1508.05326) and Rajpurkar et al. (https://arxiv.org/abs/1606.05250)
Evaluation methodology

“Growing pains” for deep learning evaluation

▶ Dataset bias and “cheating” models
Evaluation methodology

“Growing pains” for deep learning evaluation

- Dataset bias and “cheating” models
- Unexpectedly simple data and strong baselines
Evaluation methodology

“Growing pains” for deep learning evaluation

- Dataset bias and “cheating” models
- Unexpectedly simple data and strong baselines
- Adversarial examples with unintuitive model behavior
Evaluation methodology

“Growing pains” for deep learning evaluation

- Dataset bias and “cheating” models
- Unexpectedly simple data and strong baselines
- Adversarial examples with unintuitive model behavior
- Replication and task/dataset transfer failure
Evaluation methodology

“Growing pains” for deep learning evaluation

- Dataset bias and “cheating” models
- Unexpectedly simple data and strong baselines
- Adversarial examples with unintuitive model behavior
- Replication and task/dataset transfer failure

⇒ Symptoms of insufficient/inappropriate evaluation
Evaluation methodology

Current approach

labeled dataset
 (>100k data points)

∼10% test split

∼90% training split

deep neural network
(recurrent sequence model)

evaluation
Evaluation methodology

data source
(method for obtaining data)

various tests

feedback

evaluation

deep neural network
(recurrent sequence model)

training data
ShapeWorld generation framework
Examples: relations and quantifiers

- A magenta square is to the right of a green shape.
- A yellow shape is not in front of a square.
- A circle is farther from an ellipse than a gray cross.
- A cross is not the same color as a green rectangle.
- The lowermost green shape is a cross.
- A red shape is the same shape as a green shape.

- Less than one triangle is cyan.
- At least half the triangles are red.
- More than a third of the shapes are cyan squares.
- Exactly all the five squares are red.
- More than one of the seven cyan shapes is a square.
- Twice as many red shapes as yellow shapes are circles.
ShapeWorld generation framework

System overview

Sampled world model

```
{ 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 } } ] }
```

Linguistic representation

```
| a_q  | _pentagon_n_1 | _above_p | a_q  | _green_a_2 | _ellipse_n_1 |
```

Caption

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

Agreement?
ShapeWorld generation framework

Language generation

[Diagram showing the generation process from World model to Surface string]

- World model
- JSON spec: `[attr]:_blue_a_sw e? =1=> [type]:node`
- Caption objects
- DMRS snippets
- DMRS graph
- MRS structure
- Grammar
- Captioner: `sample`
- RegularTypeCaptioner
- Surface string: "There is a blue shape."
ShapeWorld generation framework
Compositionality

“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.)shape=pg</th>
<th>(a.y&gt;b.y)</th>
<th>(\exists b)</th>
<th>(b.)color=gr</th>
<th>(b.)shape=el</th>
<th>&amp;</th>
<th>¬(\exists c)</th>
<th>(c.)color=bl</th>
<th>true</th>
<th>(c=d)</th>
<th>(\exists d)</th>
<th>(d.)shape=el</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\exists a)</td>
<td>(a.)shape=pg</td>
<td>(a.y&gt;b.y)</td>
<td>(\exists b)</td>
<td>(b.)color=gr &amp; (b.)shape=el</td>
<td>&amp;</td>
<td>¬(\exists c)</td>
<td>(c.)color=bl</td>
<td>true</td>
<td>(c=d)</td>
<td>(\exists d)</td>
<td>(d.)shape=el</td>
<td></td>
</tr>
<tr>
<td>(\exists a)</td>
<td>(a.)shape=pg &amp; [(\exists b)</td>
<td>(b.)color=gr &amp; (b.)shape=el &amp; (a.y&gt;b.y)]</td>
<td>&amp;</td>
<td>¬(\exists c)</td>
<td>(c.)color=bl &amp; (\exists d)</td>
<td>(d.)shape=el &amp; (c=d)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
ShapeWorld generation framework

Design choices

- Caption is extracted from image, i.e. world model
- Incorrect caption via minimal modification of correct one
- Three agreement values to avoid ambiguous cases
- Initialize generator/captioner values before sampling
- Various tautology/contradiction checks
- Modular and configurable
ShapeWorld generation framework

What type of generalization do we expect/desire?

- magenta square
- cyan circle
- magenta circle
- three crosses
- four triangles
- four crosses
Evaluation of FiLM on ShapeWorld

Results per instance type

<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>
Evaluation of FiLM on ShapeWorld

Results per instance type

<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 87.2</td>
</tr>
<tr>
<td>existential</td>
<td>100.0 81.1</td>
<td>100.0 99.7</td>
<td>100.0 99.9</td>
</tr>
<tr>
<td>logical</td>
<td>79.7 62.2</td>
<td>76.5 58.4</td>
<td>99.9 98.9</td>
</tr>
<tr>
<td>numbers</td>
<td>75.0 66.4</td>
<td>99.1 98.2</td>
<td>99.6 99.3</td>
</tr>
<tr>
<td>quantifiers</td>
<td>72.1 69.1</td>
<td>84.8 80.8</td>
<td>97.7 97.0</td>
</tr>
<tr>
<td>(simple-spatial)</td>
<td>81.4 64.8</td>
<td>81.9 57.7</td>
<td>85.1 61.3</td>
</tr>
<tr>
<td>relational</td>
<td>—</td>
<td>—</td>
<td>50.6 51.0</td>
</tr>
<tr>
<td>implicit-rel</td>
<td>—</td>
<td>—</td>
<td>52.9 53.2</td>
</tr>
<tr>
<td>superlatives</td>
<td>—</td>
<td>—</td>
<td>50.8 50.2</td>
</tr>
</tbody>
</table>

▶ Can relational-like instances implicitly be learned when training on a broader set of instances?
▶ Can relational-like instances be learned when (pre)training on simpler pedagogical instances?
Evaluation of FiLM on ShapeWorld
Learning from a broader set of instances

relational

implicit-relational

superlatives
Evaluation of FiLM on ShapeWorld
Learning from a broader set of instances

relational

implicit-relational

superlatives

relation + implicit-rel

relational + superlat

implicit-rel + superlat
Evaluation of FiLM on ShapeWorld
Learning bootstrapped by simpler instances

![Graph showing augmentation vs pretraining]

Legend:
- Augmented rel
- Augmented rel-neg
- Augmented exist+num
- Pretrained rel
- Pretrained rel-neg
- Pretrained exist+num
Evaluation of FiLM on ShapeWorld
Learning bootstrapped by simpler instances

augmentation vs pretraining

augmentation distributions

augmented rel
augm. rel-neg
augm. exist+num
pretrained rel
pretr. rel-neg
pretr. exist+num

45%
47.5%
50%
52.5%
55%
57.5%
60%
Evaluation of FiLM on ShapeWorld

Additional findings

- pretrained ResNet doesn’t work
- overlapping objects impede learning

Graphs showing performance metrics for different configurations:
- existential fixed
- existential trainable
- numbers fixed
- numbers trainable
- relational fixed
- relational trainable
- overlap-free
- 5% numbers
- 10% numbers
- 17.5% numbers
- 25% numbers
- 25% existential
Conclusion

real-world data vs synthetic data
Conclusion

real-world data vs synthetic data

limited and expensive $\leftrightarrow$ unlimited amount

Complementary evaluation paradigms
Conclusion

real-world data vs synthetic data

limited and expensive $\leftrightarrow$ unlimited amount

uncontrolled content $\leftrightarrow$ clean content
Conclusion

real-world data \hspace{1cm} vs \hspace{1cm} synthetic data

\begin{align*}
\text{limited and expensive} & \leftrightarrow \text{unlimited amount} \\
\text{uncontrolled content} & \leftrightarrow \text{clean content} \\
\text{sparse instance coverage} & \leftrightarrow \text{targeted instance coverage}
\end{align*}
Conclusion

<table>
<thead>
<tr>
<th>real-world data</th>
<th>vs</th>
<th>synthetic data</th>
</tr>
</thead>
<tbody>
<tr>
<td>limited and expensive</td>
<td>$\leftrightarrow$</td>
<td>unlimited amount</td>
</tr>
<tr>
<td>uncontrolled content</td>
<td>$\leftrightarrow$</td>
<td>clean content</td>
</tr>
<tr>
<td>sparse instance coverage</td>
<td>$\leftrightarrow$</td>
<td>targeted instance coverage</td>
</tr>
<tr>
<td>monolithic benchmark</td>
<td>$\leftrightarrow$</td>
<td>tailored unit tests</td>
</tr>
</tbody>
</table>
## Conclusion

<table>
<thead>
<tr>
<th>real-world data</th>
<th>vs</th>
<th>synthetic data</th>
</tr>
</thead>
<tbody>
<tr>
<td>limited and expensive</td>
<td>↔</td>
<td>unlimited amount</td>
</tr>
<tr>
<td>uncontrolled content</td>
<td>↔</td>
<td>clean content</td>
</tr>
<tr>
<td>sparse instance coverage</td>
<td>↔</td>
<td>targeted instance coverage</td>
</tr>
<tr>
<td>monolithic benchmark</td>
<td>↔</td>
<td>tailored unit tests</td>
</tr>
<tr>
<td>test interpolation ability</td>
<td>↔</td>
<td>test extrapolation ability</td>
</tr>
</tbody>
</table>
Conclusion

real-world data vs synthetic data

limited and expensive $\leftrightarrow$ unlimited amount
uncontrolled content $\leftrightarrow$ clean content
sparse instance coverage $\leftrightarrow$ targeted instance coverage
monolithic benchmark $\leftrightarrow$ tailored unit tests
test interpolation ability $\leftrightarrow$ test extrapolation ability

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

Questions?