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

Alexander Kuhnle

Supervisor: Prof Ann Copestake
Department of Computer Science and Technology
University of Cambridge

Language Sciences Symposium
19th June 2018
Overview

- Visual question answering
- Problems with the VQA Dataset
- Potential of synthetic data
- Evaluation and generalization
Deep learning in Natural Language Processing

labeled dataset
(>100k data points)

~90% training split

deep neural network
(recurrent sequence model)

evaluation

~10% test split
Visual question answering

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/)
### Other popular datasets

<table>
<thead>
<tr>
<th>SNLI – Stanford Natural Language Inference Corpus</th>
<th>SQuAD – Stanford Question Answering Dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td>C: A soccer game with multiple males playing.</td>
<td>In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under <strong>gravity</strong>. The main forms of precipitation include drizzle, rain, sleet, snow, <strong>graupel</strong> and hail... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals <strong>within a cloud</strong>. Short, intense periods of rain in scattered locations are called “showers”.</td>
</tr>
<tr>
<td>H: Some men are playing a sport. → <strong>entailment</strong></td>
<td>(1) What causes precipitation to fall? ⇒ <strong>gravity</strong></td>
</tr>
<tr>
<td>C: A smiling costumed woman is holding an umbrella.</td>
<td>(2) What is another main form of precipitation besides drizzle, rain, snow, sleet and hail? ⇒ <strong>graupel</strong></td>
</tr>
<tr>
<td>H: A happy woman in a fairy costume holds an umbrella. → <strong>neutral</strong></td>
<td>(3) Where do water droplets collide with ice crystals to form precipitation? ⇒ <strong>within a cloud</strong></td>
</tr>
<tr>
<td>C: A man inspects the uniform of a figure in some East Asian country.</td>
<td></td>
</tr>
<tr>
<td>H: The man is sleeping → <strong>contradiction</strong></td>
<td></td>
</tr>
</tbody>
</table>

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)
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/)
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

Crowd-sourced real-world datasets

Solve the problem/dataset? ✓

Evaluate model capabilities? ❓

Deep learning will find a way to make effective use of the data.

Are these datasets appropriate to investigate this question?

- Natural?
- Difficult?
- Specific?

⇒ Synthetic data!
ShapeWorld 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.
Properties and comparison

real-world data vs synthetic data

uncontrolled content ←→ clean content
sparse instance coverage ←→ targeted instance coverage
monolithic benchmark ←→ tailored unit tests
test interpolation ability ←→ test extrapolation ability

⇒ Complementary evaluation paradigms
What type of generalization do we expect/desire?

- magenta square
- cyan circle
- magenta circle

- three crosses
- four triangles
- four crosses
Example use case: replication of psychology experiment
(灵感 by The meaning of “most”, Pietroski et al., 2009)

Random

Paired

Partitioned

“More/less than half the shapes are X?”

ratio

accuracy

1.0

0.75

0.5

1.0

1.25

1.5

random

paired

part.
Thank you for your attention!

Questions?