"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?

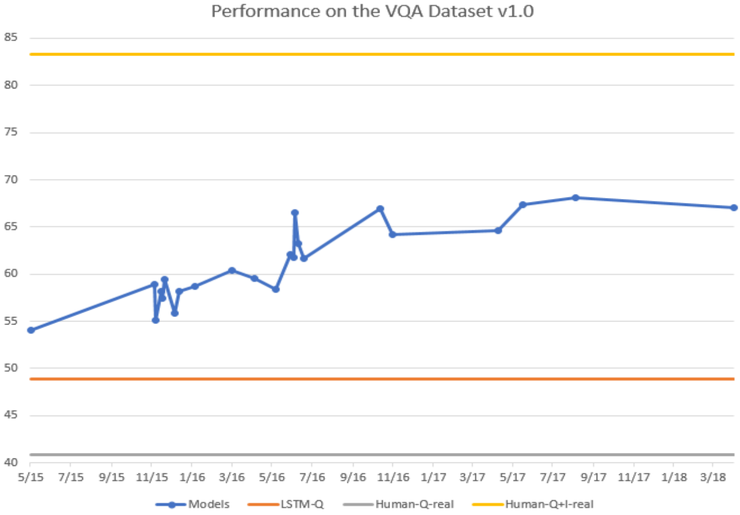
Is the dog begging for food?

$\Rightarrow$ **Visual Turing test?**

Examples from VQA Dataset (`http://visualqa.org/browser/`)

Visual question answering

Performance over time



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

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...? $\Rightarrow$ umbrella
- What season...? $\Rightarrow$ summer
- What season of...? $\Rightarrow$ summer
- ...
- What season of year was this photo taken in? $\Rightarrow$ summer

- What does the red sign say? $\Rightarrow$ 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

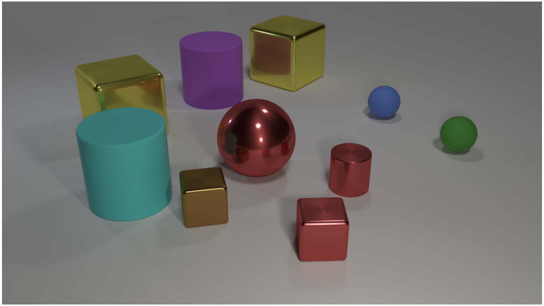
Sensitivity to question words

- How symmetrical are the white bricks on either side of the building? $\Rightarrow$ very
- How **spherical** are the white bricks on either side of the building? $\Rightarrow$ very
- How **soon** are the bricks **fading** on either side of the building? $\Rightarrow$ very
- How **fast** are the bricks **speaking** on either side of the building? $\Rightarrow$ very

Example from Mudrakarta et al. `https://arxiv.org/abs/1805.05492`).

Problems with the VQA Dataset

Low performance on CLEVR

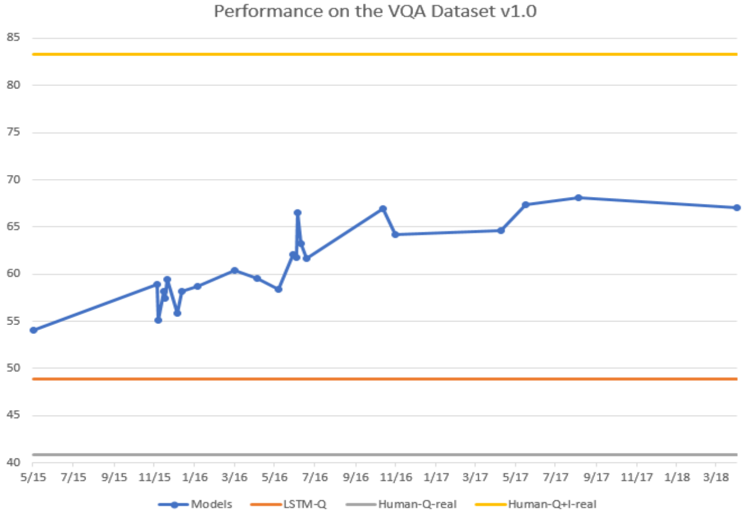


- How many small spheres are there? $\Rightarrow$ 2
- What number of cubes are small things or red metal objects? $\Rightarrow$ 2
- Does the metal sphere have the same color as the metal cylinder? $\Rightarrow$ Yes
- Are there more small cylinders than metal things? $\Rightarrow$ No

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

Evaluation methodology

Meaningful progress?



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.

Evaluate model capabilities?

?

Are these datasets appropriate to investigate this question?

- Natural?
- Difficult?
- Specific?

$\Rightarrow$ **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
- Unexpectedly simple data and strong baselines
- Adversarial examples with unintuitive model behavior
- Replication and task/dataset transfer failure

$\Rightarrow$ **Symptoms of insufficient/inappropriate evaluation**

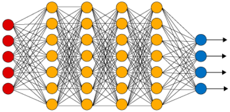
Evaluation methodology

Current approach

labeled dataset

(>100k data points)

~90% training split



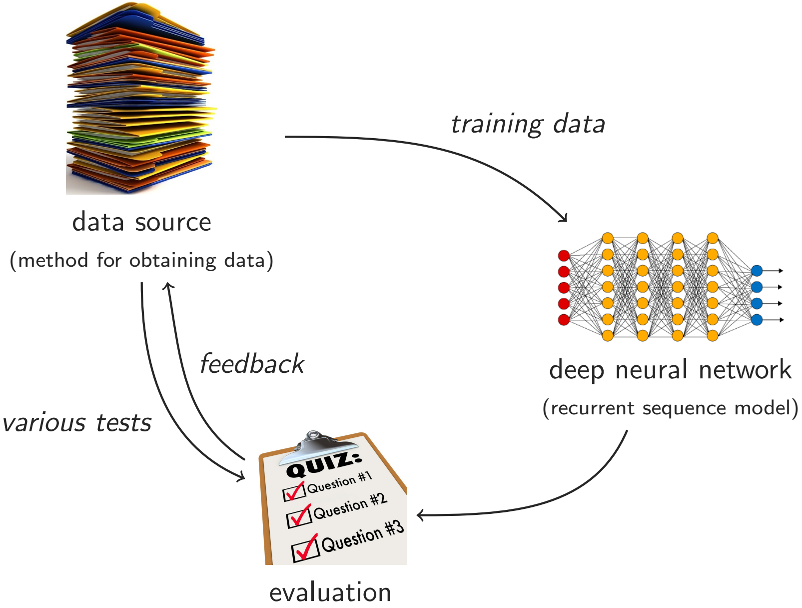
deep neural network

(recurrent sequence model)

~10% test split

evaluation

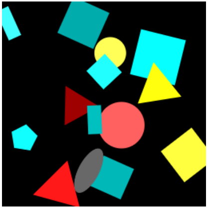
Evaluation methodology



ShapeWorld generation framework

Examples: relations and quantifiers

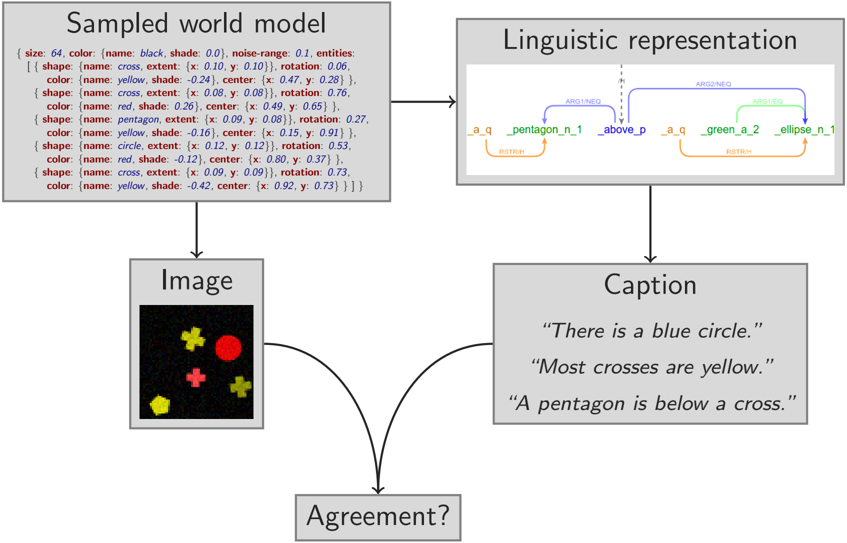
- A magenta square is to the right of a green shape.
- A yellow shape is not in front of a square.
- A circle is farther from an ellipse than a gray cross.
- A cross is not the same color as a green rectangle.
- The lowermost green shape is a cross.
- A red shape is the same shape as a green shape.



- 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.

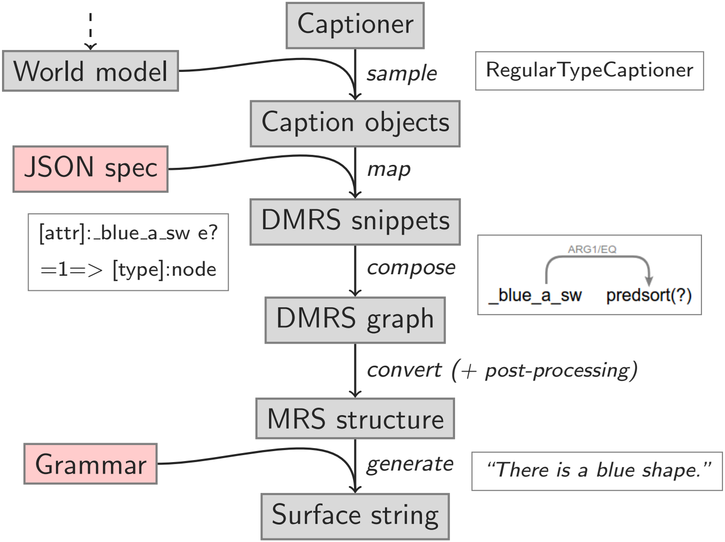
ShapeWorld generation framework

System overview



ShapeWorld generation framework

Language generation

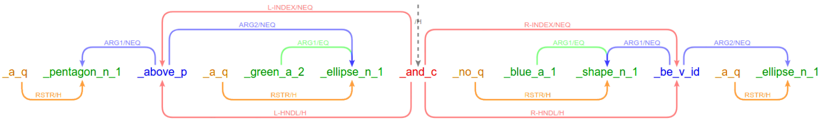


ShapeWorld generation framework

Compositionality

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

⇑ ERG + ACE realization ⇑



⇑ Internal DMRS mapping ⇑

<table>
<tr><td>$\exists a$</td><td>a.shape=pg</td><td>a.y>b.y</td><td>$\exists b$</td><td>b.color=gr</td><td>b.shape=el</td><td>$\wedge$</td><td>$\neg\exists c$</td><td>c.color=bl</td><td>true</td><td>$c=d$</td><td>$\exists d$</td><td>d.shape=el</td></tr>
<tr><td colspan="2">$\exists a:$ a.shape=pg</td><td>a.y>b.y</td><td colspan="3">$\exists b:$ b.color=gr $\wedge$ b.shape=el</td><td>$\wedge$</td><td colspan="3">$\neg\exists c:$ c.color=bl</td><td>$c=d$</td><td colspan="2">$\exists d:$ d.shape=el</td></tr>
<tr><td colspan="6">$\exists a:$ a.shape=pg $\wedge$ [$\exists b:$ b.color=gr $\wedge$ b.shape=el $\wedge$ a.y>b.y]</td><td>$\wedge$</td><td colspan="6">$\neg\exists c:$ c.color=bl $\wedge$ [$\exists d:$ d.shape=el $\wedge$ $c=d$]</td></tr>
<tr><td colspan="13">($\exists a:$ a.shape=pg $\wedge$ [$\exists b:$ b.color=gr $\wedge$ b.shape=el $\wedge$ a.y>b.y]) $\wedge$ ($\neg\exists c:$ c.color=bl $\wedge$ [$\exists d:$ d.shape=el $\wedge$ $c=d$])</td></tr>
</table>

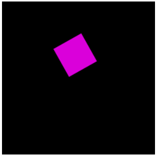
ShapeWorld generation framework

Design choices

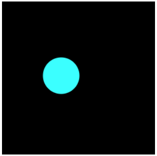
- Caption is extracted from image, i.e. world model
- Incorrect caption via minimal modification of correct one
- Three agreement values to avoid ambiguous cases
- Initialize generator/captioner values before sampling
- Various tautology/contradiction checks
- Modular and configurable

ShapeWorld generation framework

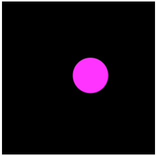
What type of generalization do we expect/desire?



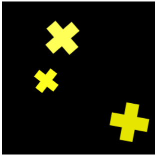
magenta square



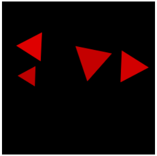
cyan circle



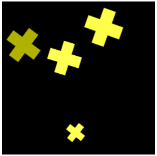
magenta circle



three crosses



four triangles

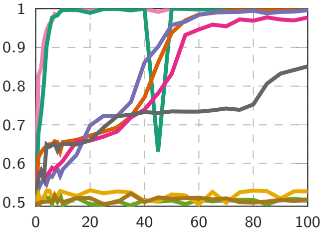


four crosses

Evaluation of FiLM on ShapeWorld

Results per instance type

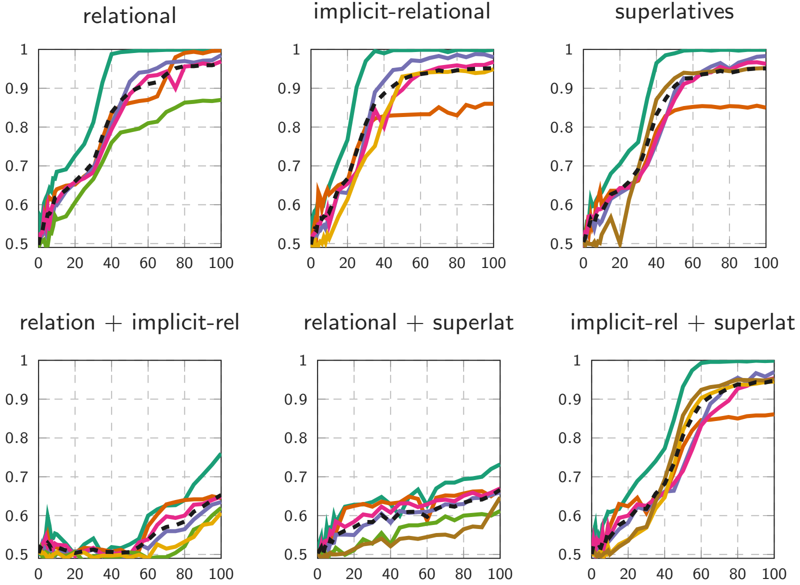
Dataset	CNN-LSTM		CNN-LSTM-SA		FiLM	
(single-shape)	—		—		100.0	87.2
existential	100.0	81.1	100.0	99.7	100.0	99.9
logical	79.7	62.2	76.5	58.4	99.9	98.9
numbers	75.0	66.4	99.1	98.2	99.6	99.3
quantifiers	72.1	69.1	84.8	80.8	97.7	97.0
(simple-spatial)	81.4	64.8	81.9	57.7	85.1	61.3
relational	—		—		50.6	51.0
implicit-rel	—		—		52.9	53.2
superlatives	—		—		50.8	50.2



- 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?

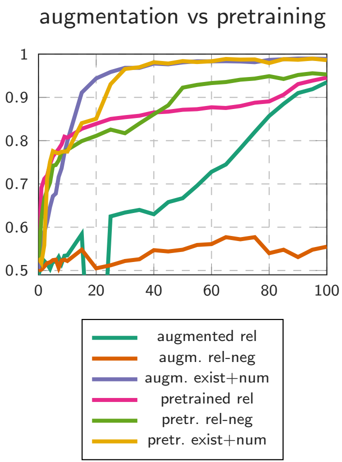
Evaluation of FiLM on ShapeWorld

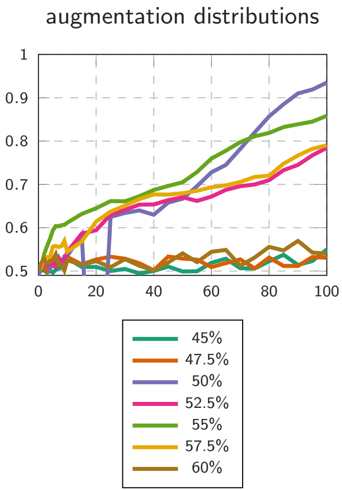
Learning from a broader set of instances



Evaluation of FiLM on ShapeWorld

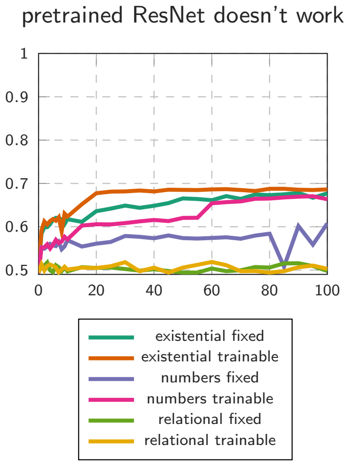
Learning bootstrapped by simpler instances

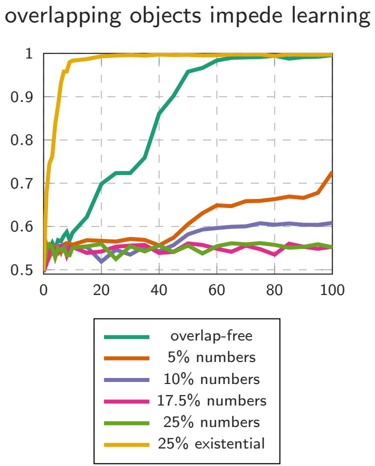




Evaluation of FiLM on ShapeWorld

Additional findings





Conclusion

real-world data	**vs**	**synthetic data**
limited and expensive	$\longleftrightarrow$	unlimited amount
uncontrolled content	$\longleftrightarrow$	clean content
sparse instance coverage	$\longleftrightarrow$	targeted instance coverage
monolithic benchmark	$\longleftrightarrow$	tailored unit tests
test interpolation ability	$\longleftrightarrow$	test extrapolation ability

$\Rightarrow$ **Complementary evaluation paradigms**

Thank you for your attention!

Questions?