Synthetic data for Visual Question Answering

Alexander Kuhnle

Department of Computer Science and Technology
University of Cambridge

NLIP Seminar
18th May 2018

Overview

1. Visual Question Answering
2. Problems with the VQA Dataset
3. ShapeWorld
4. Other synthetic datasets
5. Experiments

Visual Question Answering

The VQA Dataset

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?

Statistics: 205k MSCOCO images, 50k abstract scenes, 3 questions per image

Images: `http://visualqa.org/browser/`

Visual Question Answering

Motivation

- Multimodal grounded language task
- Generalisation of object recognition
- More specific and clearer evaluation than image captioning
- 'Visual formal semantics' in a natural setup
- Covers a wide variety of linguistic phenomena

Visual Turing Test?

Visual Question Answering

Basic CNN-LSTM model

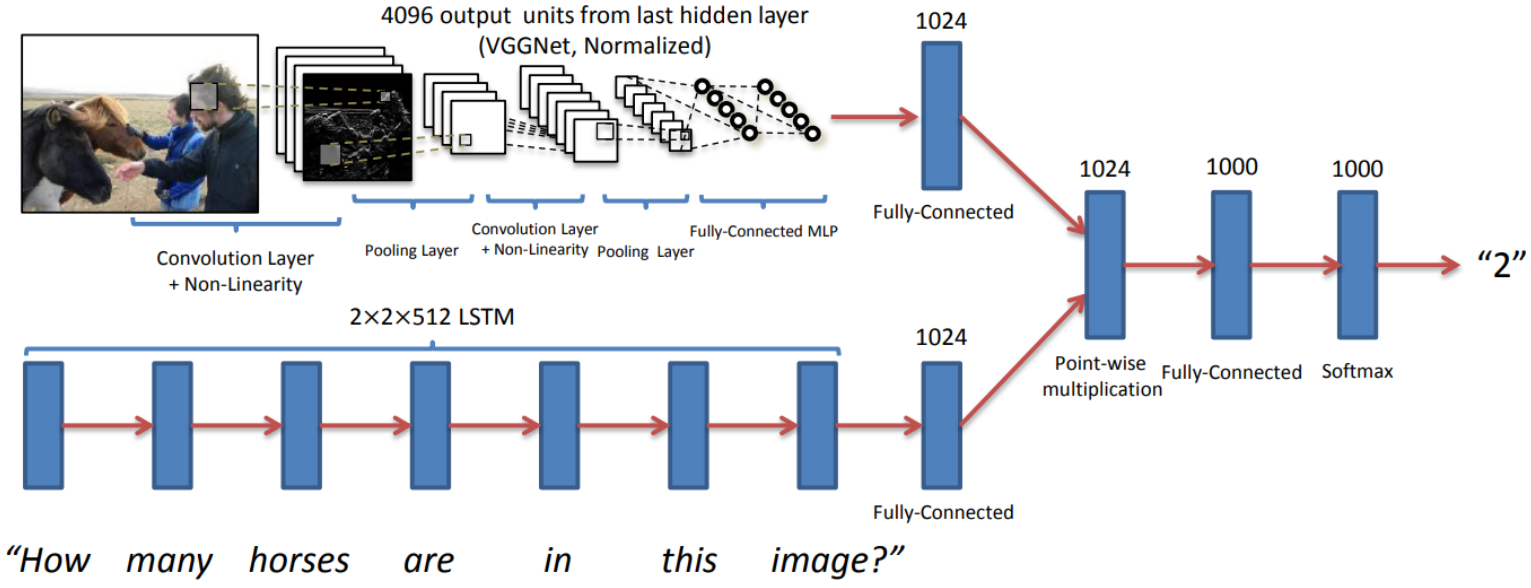
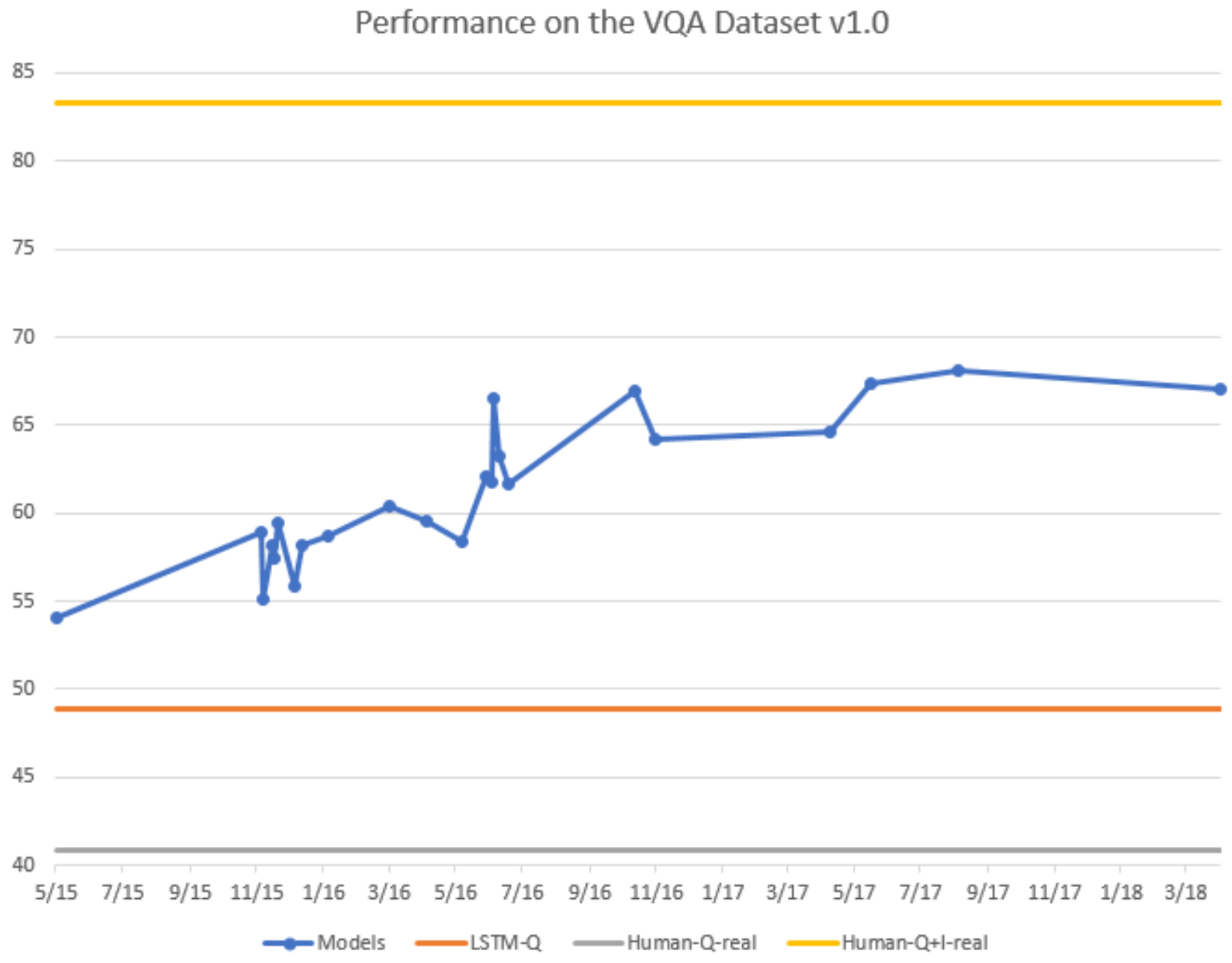


Image: `https://arxiv.org/abs/1505.00468`

Visual Question Answering

Performance over time



Problems with the VQA Dataset

Examples revisited

Where is this cat laying?

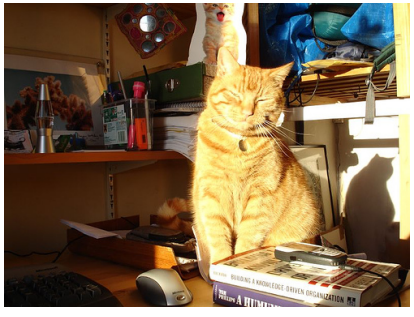
Is the cat awake?

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?

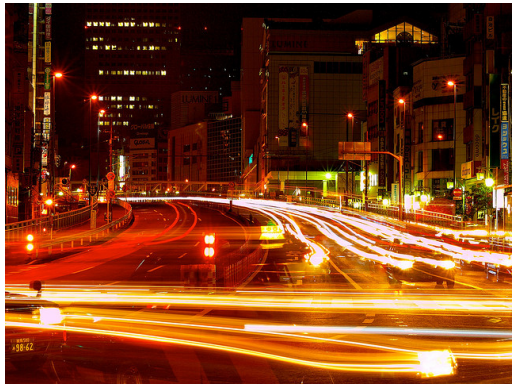
Is the dog begging for food?

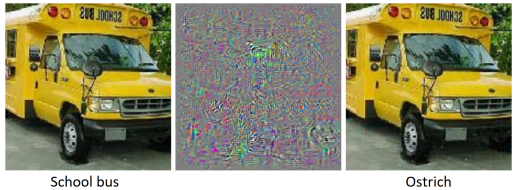
Images: `http://visualqa.org/browser/`

Problems with the VQA Dataset

Crowd-sourced real-world datasets

- Natural or repurposed?
- 'Zipfian' tendency to simplicity
- Unintended biases/correlations
- Adversarial examples





School bus Ostrich

Images: `http://visualqa.org/browser/`, `https://arxiv.org/abs/1312.6199`

Problems with the VQA Dataset

Remedies

Improvements of the VQA Dataset

- Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering (Goyal et al.)
- C-VQA: A Compositional Split of the Visual Question Answering (VQA) v1.0 Dataset (Agrawal et al.)

Modification of existing datasets

- Focused Evaluation for Image Description with Binary Forced-Choice Tasks (Hodosh & Hockenmaier)
- FOIL it! Find One mismatch between Image and Language caption (Shekhar et al.)

Problems with the VQA Dataset

Our proposal

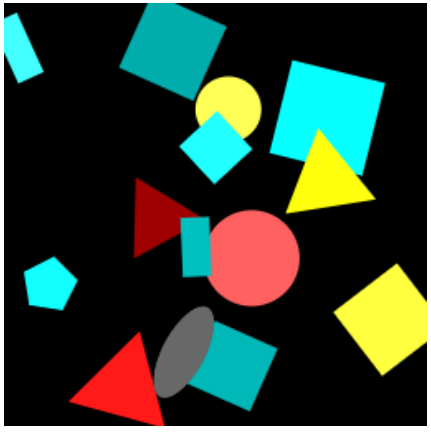
- No single general evaluation benchmark,
  but investigation tailored to the model
- Dissimilar train/test distributions,
  requiring compositional generalisation
- Clean data with clear image/text relationship,
  instead of uncontrolled content

$\Rightarrow$ Synthetic data as targeted 'unit-testing' evaluation step,
complementing general real-world benchmark datasets

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.

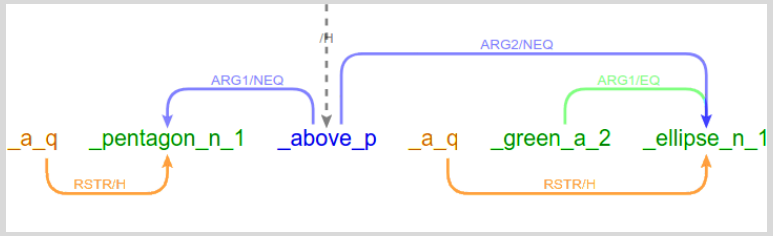
ShapeWorld

System overview

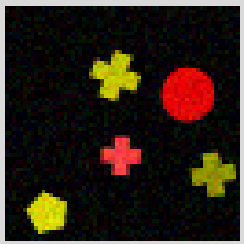
Sampled world model

{ **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



Image



Caption

"There is a blue circle."

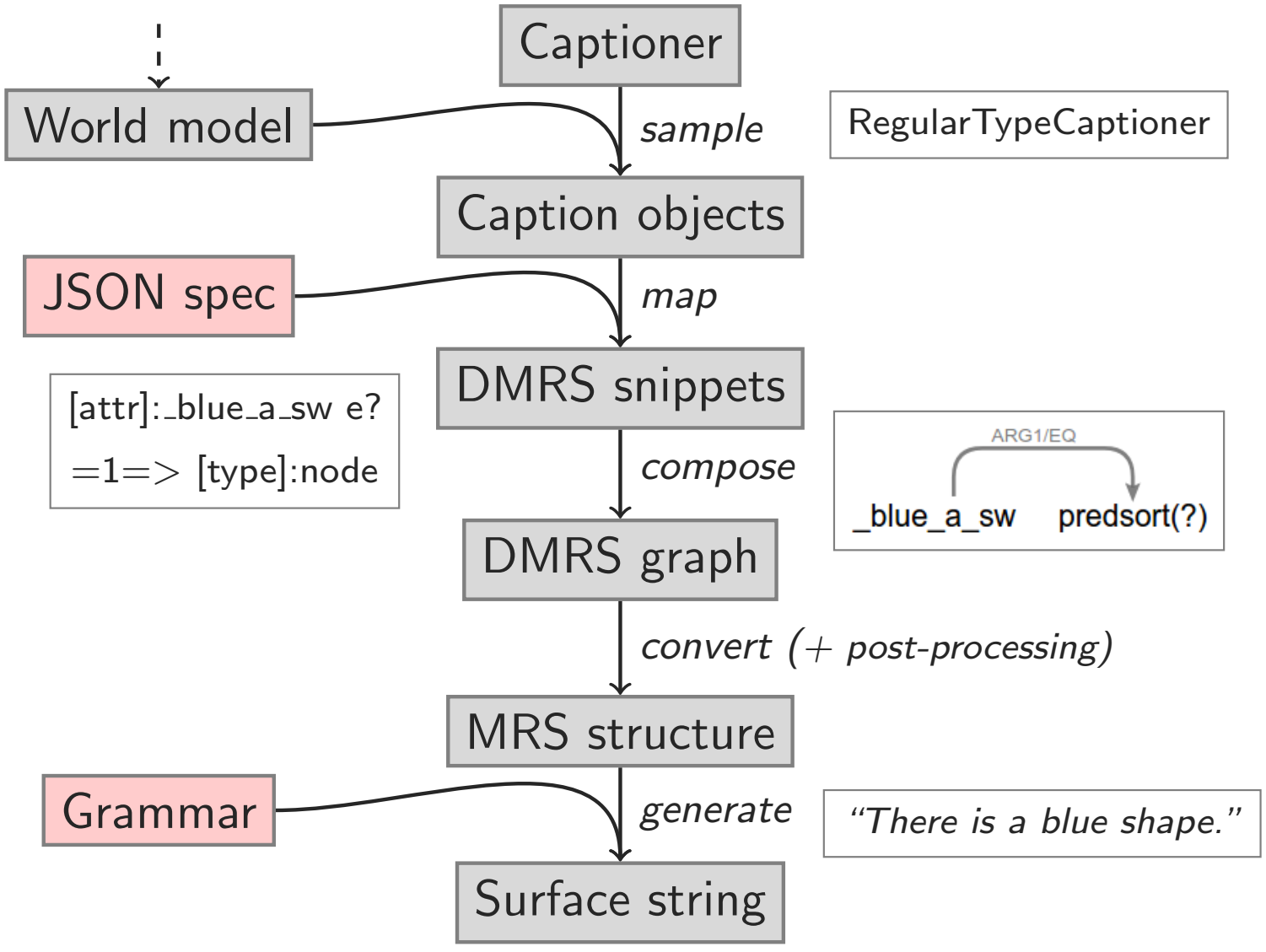
"Most crosses are yellow."

"A pentagon is below a cross."

Agreement?

ShapeWorld

Language generation

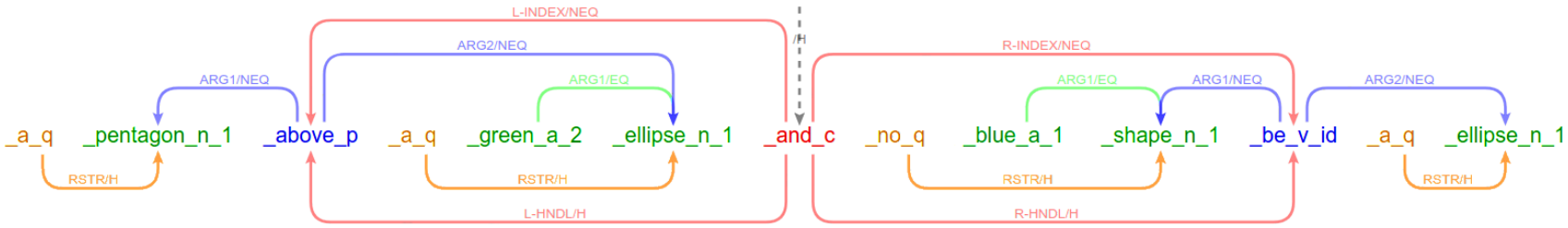


ShapeWorld

Compositionality

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

⇑ ERG + ACE realization ⇑



⇑ Internal DMRS mapping ⇑

$\exists a$	a.shape=pg	a.y>b.y	$\exists b$	b.color=gr	b.shape=el	$\wedge$	$\neg\exists c$	c.color=bl	true	$c{=}d$	$\exists d$	d.shape=el
$\exists a:$ a.shape=pg		a.y>b.y	$\exists b:$ b.color=gr $\wedge$ b.shape=el			$\wedge$	$\neg\exists c:$ c.color=bl			$c{=}d$	$\exists d:$ d.shape=el	
$\exists a:$ 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 $\wedge$ $c{=}d$]					
($\exists a:$ 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 $\wedge$ $c{=}d$])												

ShapeWorld

Design choices

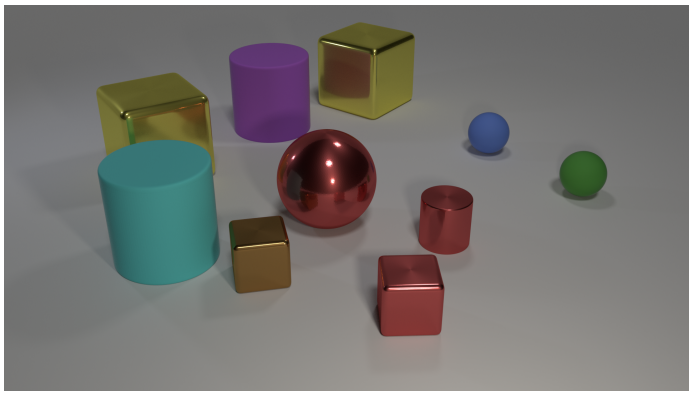
- Caption is extracted from image, i.e. world model
- Incorrect caption via minimal modification of correct one
- Three agreement values to avoid ambiguous cases
- Initialize generator/captioner values before sampling
- Various tautology/contradiction checks

Other synthetic datasets

Comparison to other VQA datasets

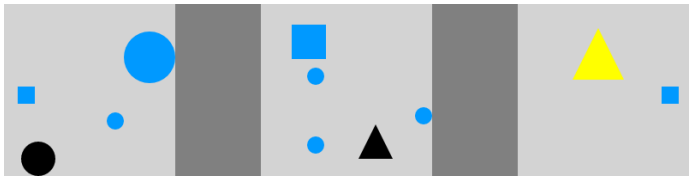
CLEVR

(Compositional Language and Elementary Visual Reasoning)



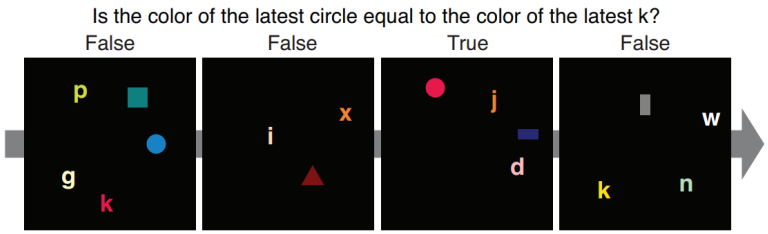
NLVR

(Natural Language Visual Reasoning)

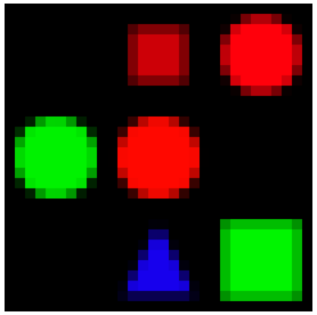


COG

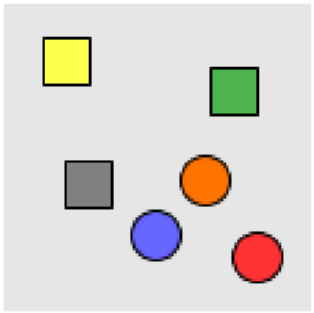
("visual reasoning in time, in parallel with human cognitive experiments")



SHAPES



Sort-of-CLEVR



Images: `https://github.com/facebookresearch/clevr-dataset-gen`, `https://arxiv.org/abs/1511.02799`, `https://arxiv.org/abs/1706.01427`, `http://lic.nlp.cornell.edu/nlvr/`, `https://arxiv.org/abs/1803.06092`

Other synthetic datasets

Artificial text-only datasets

bAbI Tasks

Task 13: Compound Coreference

1. Daniel and Sandra journeyed to the office.
2. Then they went to the garden.
3. Sandra and John travelled to the kitchen.
4. After that they moved to the hallway.

Q: *Where is Daniel?*

⇒ garden

FraCaS Test Suite

Monotonicity (upwards on second argument)

1. Every European has the right to live in Europe.
2. Every European is a person.
3. Every person who has the right to live in Europe can travel freely within Europe.

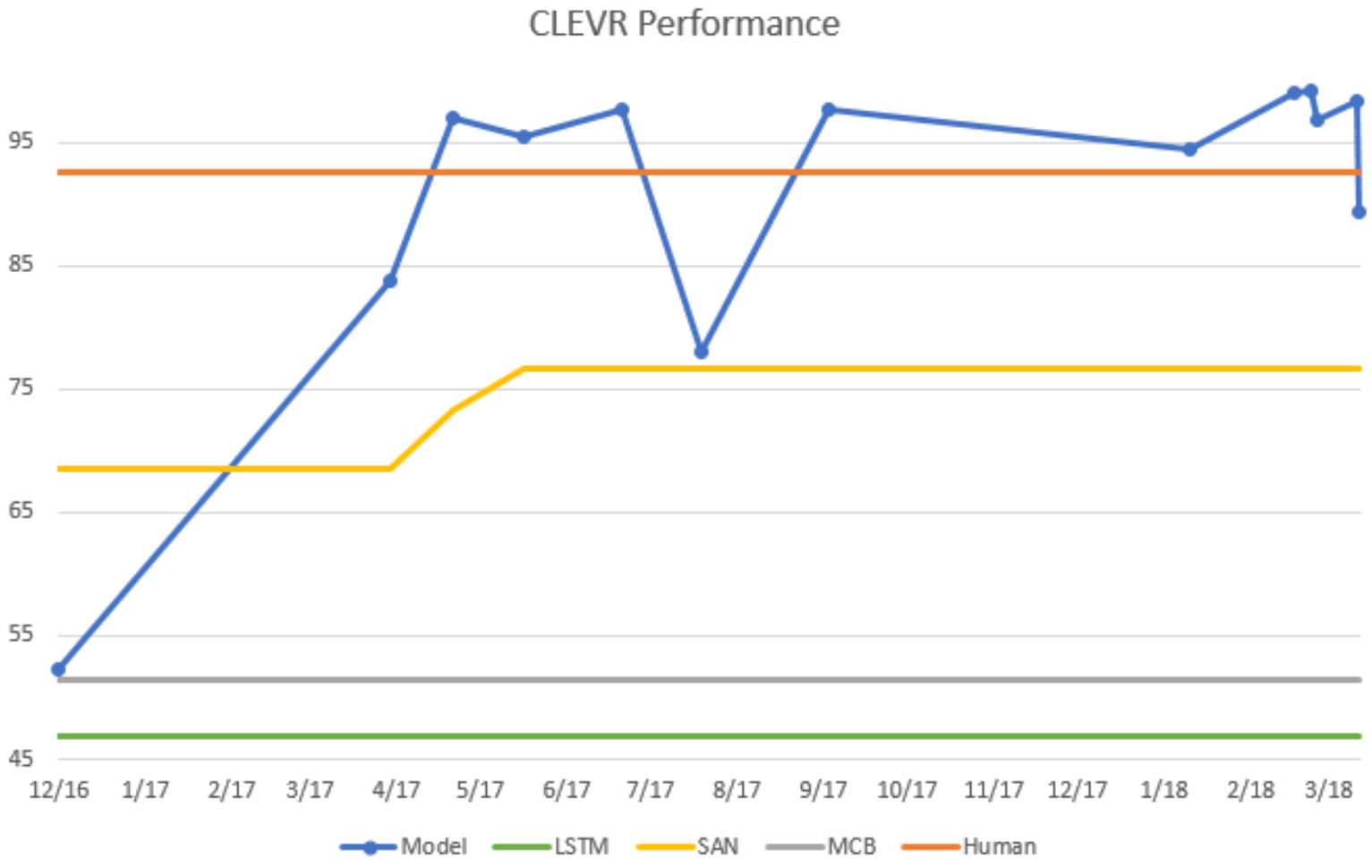
Q: *Can every European travel freely within Europe?*

⇒ yes

Examples: `https://arxiv.org/abs/1502.05698`, `https://nlp.stanford.edu/~wcmac/downloads/fracas.xml`

Other synthetic datasets

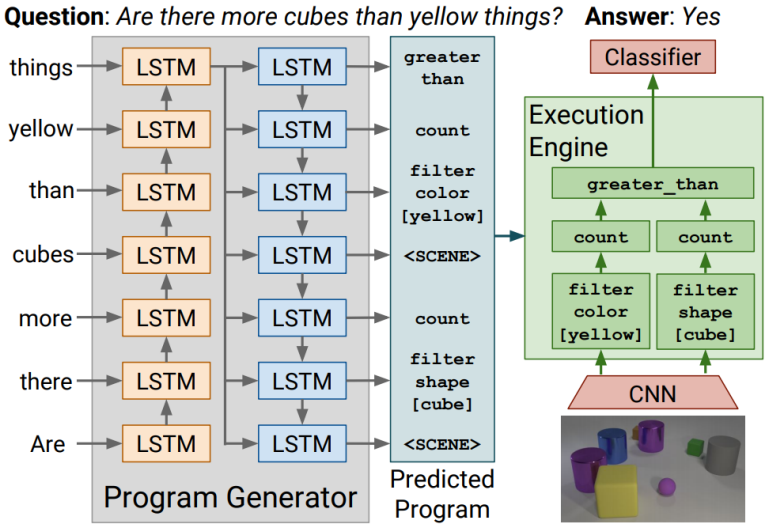
Performance on CLEVR



Other synthetic datasets

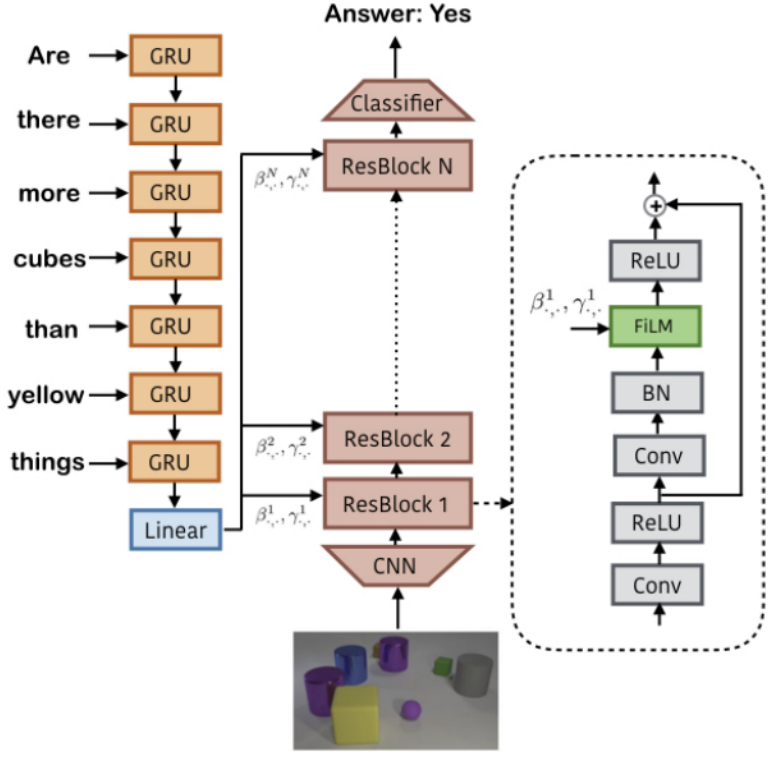
CLEVR-inspired systems

Seq2Seq Module Network



FiLM Model

(Feature-wise Linear Modulation)



Images: `https://arxiv.org/abs/1705.03633`, `https://arxiv.org/abs/1709.07871`

Experiments

Early experimental results

Dataset configuration	LSTM-only	CNN-LSTM	HCA-par	HCA-alt
OneShape	51 / 46 / 50	81 / 70 / 66	90 / 77 / 78	**92 / 81 / 77**
C: no hypernyms	90 / 70 / 100	95 / 64 / 57	98 / 71 / 73	97 / 68 / 66
C: only hypernyms	100 / 100 / 100	52 / 34 / 30	96 / 78 / 82	95 / 75 / 73
I: changed shape	6 / 5 / 7	70 / 81 / 82	60 / 63 / 58	73 / 78 / 78
I: changed color	8 / 15 / 0	100 / 100 / 99	100 / 92 / 96	100 / 97 / 89
I: changed both	7 / 5 / 6	96 / 97 / 98	87 / 85 / 84	93 / 92 / 89
MultiShape	62 / 67 / 67	**72 / 71 / 72**	**72 / 71 / 69**	71 / 68 / 68
correct instances	48 / 49 / 50	76 / 64 / 54	81 / 68 / 65	71 / 59 / 53
I: random attr.	58 / 63 / 68	67 / 74 / 79	64 / 67 / 68	70 / 73 / 78
I: random existing attr.	100 / 100 / 100	78 / 86 / 95	55 / 71 / 79	72 / 87 / 95
Spatial	52 / 51 / 50	57 / 52 / 54	**63 / 65 / 64**	54 / 52 / 55
C: no hypernyms	85 / 85 / 69	45 / 44 / 41	83 / 83 / 86	92 / 62 / 100
C: only hypernyms	95 / 95 / 97	4 / 6 / 4	60 / 59 / 65	49 / 40 / 52
I: swapped direction	11 / 13 / 16	98 / 97 / 98	36 / 39 / 30	50 / 61 / 47
I: object random attr.	15 / 12 / 16	88 / 88 / 91	69 / 68 / 68	63 / 66 / 60
I: subject random attr.	13 / 12 / 17	87 / 88 / 89	69 / 71 / 70	61 / 64 / 56
Quantification	57 / 57 / 56	56 / 56 / 58	**76 / 77 / 78**	**74 / 77 / 78**
correct instances	23 / 22 / 18	25 / 30 / 26	74 / 71 / 72	70 / 71 / 75
incorrect instances	94 / 93 / 93	88 / 90 / 88	81 / 83 / 88	78 / 82 / 82
instances with "no"	52 / 51 / 48	61 / 60 / 61	56 / 56 / 51	55 / 55 / 58
instances with "the" (=1)	53 / 58 / 61	55 / 59 / 58	59 / 59 / 55	63 / 63 / 63
instances with "a" ($\geq$1)	34 / 35 / 36	34 / 36 / 37	49 / 50 / 51	48 / 52 / 50
instances with "two" ($\geq$2)	53 / 48 / 48	50 / 50 / 49	70 / 69 / 62	72 / 67 / 58
instances with "most"	49 / 50 / 49	48 / 48 / 49	69 / 68 / 60	60 / 52 / 51
instances with "all"	52 / 54 / 50	48 / 50 / 51	47 / 52 / 51	49 / 50 / 51

Experiments

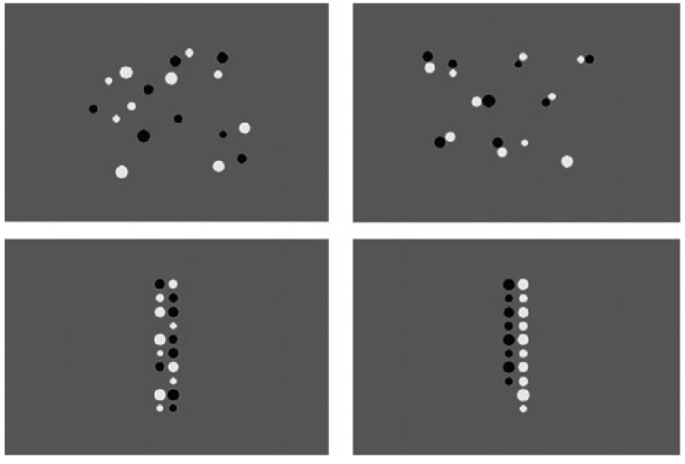
The meaning of "most" (Pietroski et al., 2009)

Cardinality-based mechanism

$$\text{most}(A, B) \Leftrightarrow |S_A \cap S_B| > \frac{1}{2}|A| \Leftrightarrow |S_A \cap S_B| > |S_A - S_B|$$

Pairing-based mechanism

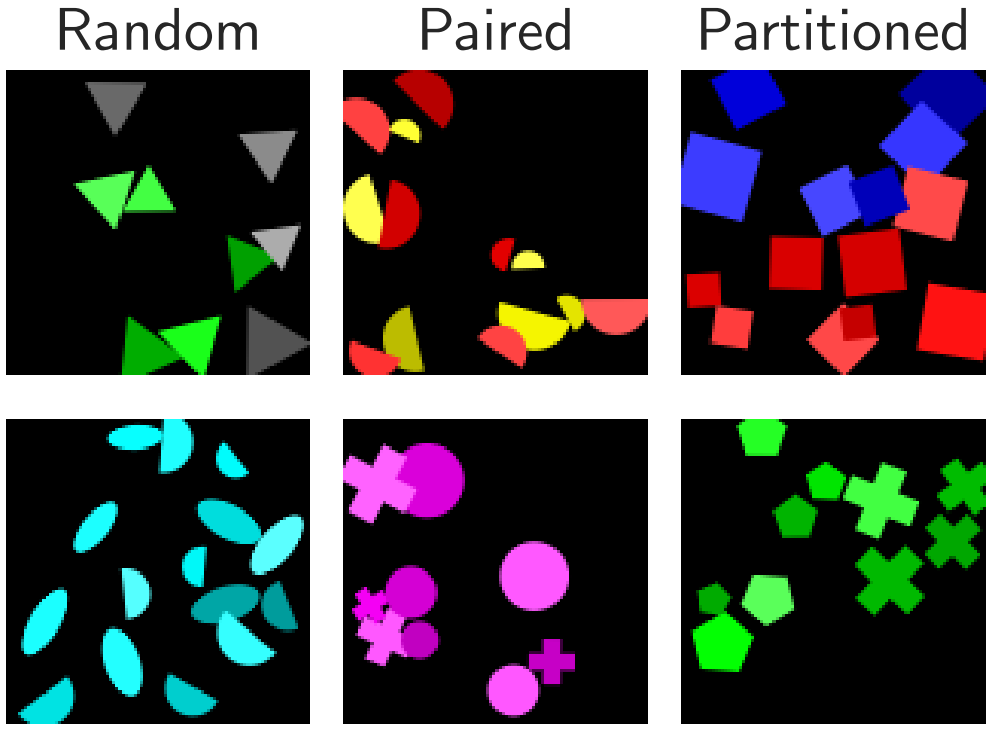
$$\text{most}(A, B) \Leftrightarrow \exists S\colon S \subset B \text{ and } \mathit{OneToOne}(A, S)$$



Images: https://doi.org/10.1111/j.1468-0017.2009.01374.x

Experiments

Replication with ShapeWorld data



+ *"More/less than half the shapes are X."*

Conclusion

Synthetic data. . .

- . . . can be generated in arbitrary quantities
- . . . avoids biases found in real-world data
- . . . provides challenging test data
- . . . can be tailored to the evaluation goals
- . . . is configurable, flexible and reusable
- . . . makes unit-testing deep learning models possible


Thank you for your attention!
Questions?