Artificial microworlds and deep linguistic processing for evaluating language understanding

Alexander Kuhnle & Ann Copestake

University of Cambridge
{aok25,aac10}@cam.ac.uk

Visual question answering datasets for evaluation

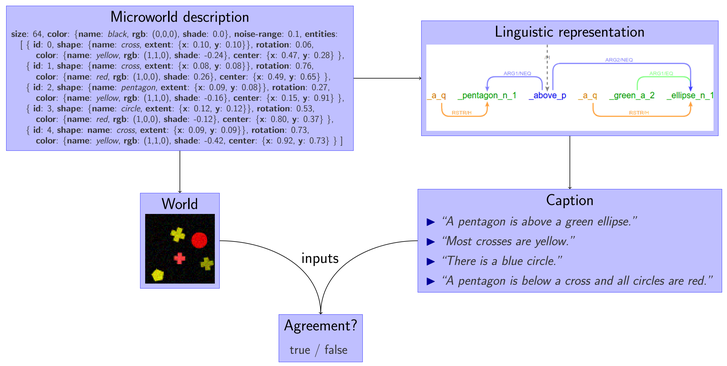
Properties and issues

- Photo data does not correspond to human perception of the world.
- Crowd-sourced language data tends to be simple (Zipf's law).
- "Clever Hans effect": Unexpected hidden correlations and biases, which do not relate to human-like language understanding.
- Deep neural networks are surprisingly good in fitting data, even mere noise (very different from "shallow" machine learning)

Three guiding evaluation principles

- Avoid training for multiple epochs on a fixed dataset.
- Focus on the true compositional generalization abilities required by dissimilar data distributions for training/evaluation.
- Do at least some experiments with clean data, which reduces the likelihood of hidden biases or correlations compared to more "realistic" and complex data.

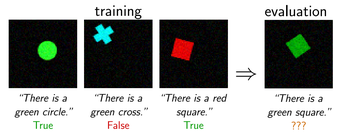
ShapeWorld: Generation of visually grounded language data



- Abstract world models are randomly sampled.
- These models can be visualized straightforwardly.
- A linguistic representation (Dependency Minimal Recursion Semantics) extracts relevant values.
- DMRS graphs can be realized as natural language.

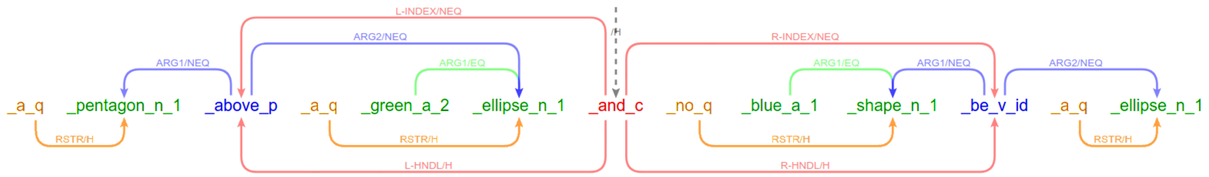
⇒ **Task: Image caption agreement (ICA)**

⇒ Evaluation data is different from training data, hence requiring ability to recombine/generalize.



Compositional ShapeWorld semantics

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



$\exists a$	a.shape = *pentagon*	a.y > b.y	$\exists b$	b.color = *green*	b.shape = *ellipse*	$\wedge$	$\neg\exists c$	c.color = *blue*	true	$c = d$	$\exists d$	d.shape = *ellipse*
$\exists a$: a.shape = *pentagon*		a.y > b.y	$\exists b$: b.color = *green* $\wedge$ b.shape = *ellipse*			$\wedge$	$\neg\exists c$: c.color = *blue*			$c = d$	$\exists d$: d.shape = *ellipse*	
$\exists a$: a.shape = *pentagon* $\wedge$ [$\exists b$: b.color = *green* $\wedge$ b.shape = *ellipse* $\wedge$ a.y > b.y]						$\wedge$	$\neg\exists c$: c.color = *blue* $\wedge$ [$\exists d$: d.shape = *ellipse* $\wedge$ $c = d$]					
($\exists a$: a.shape = *pentagon* $\wedge$ [$\exists b$: b.color = green $\wedge$ b.shape = *ellipse* $\wedge$ a.y > b.y]) $\wedge$ ($\neg\exists c$: c.color = *blue* $\wedge$ [$\exists d$: d.shape = *ellipse* $\wedge$ $c = d$])												

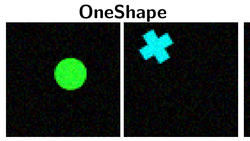
GitHub project & arXiv preprints

Project: https://github.com/AlexKuhnle/ShapeWorld

Preprints: https://arxiv.org/abs/1704.04517 https://arxiv.org/abs/1706.01322

Examples of ShapeWorld datasets/generators

OneShape



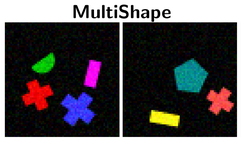
- *"There is a blue rectangle."*
- *"There is a triangle."*
- *"There is a yellow shape."*

Training combinations
- 50 combinations

Validation combinations
- *"red square"*, *"green triangle"*, *"blue circle"*

Test combinations
- *"yellow rectangle"*, *"cyan ellipse"*, *"magenta cross"*

MultiShape

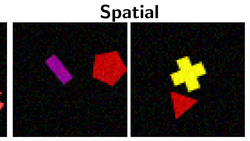


- *"There is a magenta semicircle."*
- *"There is a pentagon."*
- *"There is a cyan shape."*

(Same as for OneShape)

Number of objects
- Training: 1, 2, 3, 5
- Validation: 4
- Testing: 6

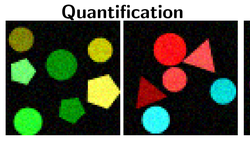
Spatial



- *"A red circle is to the left of a cyan semicircle."*
- *"A white circle is above a pentagon."*
- *"A red shape is to the right of a green triangle."*
- *"A shape is below a cross."*

Training / validation / test combinations:
- Same as for OneShape

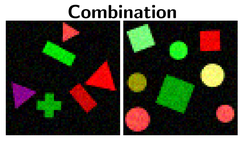
Quantification



- *"The shape is green."*
- *"Most shapes are rectangles."*
- *"No shape is a red triangle."*
- *"All triangles are green."*
- *"Two blue shapes are pentagons."*

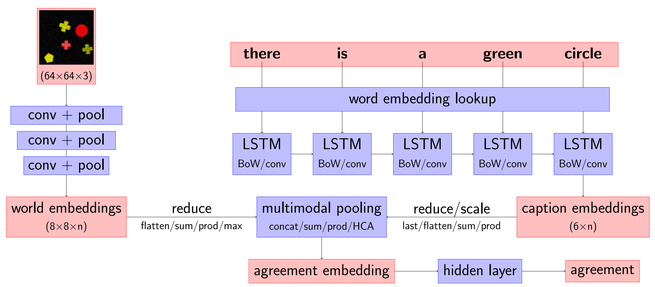
Number of objects
- Training: 3, 4, 5, 7
- Validation: 6
- Testing: 8

Combination



- *"A cross is above an ellipse, and most shapes are green."*
- *"A triangle is below a yellow circle, or two circles are yellow."*
- *"All triangles are white, or all blue shapes are circles."*
- *"Most green shapes are semicircles, or a rectangle is above a rectangle."*
- *"Two red shapes are rectangles, and two shapes are red."*

Neural network architecture for ICA



- Corresponds to popular VQA architectures.
- Entire architecture trained end-to-end on the task.
- Both the image-processing CNN and the word embeddings are learned from scratch.
- We train for 5000 iterations with a batch size of 128.
- Important choices (according to our experiments):
  - Multimodal pooling operation
  - Reduction operation for world/caption embeddings

A new evaluation methodology for language understanding

Dataset configuration	LSTM-only	CNN+LSTM:Mult	CNN+CNN:HCA-par	CNN+CNN: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"* (≥1)	34 / 35 / 36	34 / 36 / 37	49 / 50 / 51	48 / 52 / 50
instances with *"two"* (≥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

General properties and methodology

- ShapeWorld datasets are generators, not fixed datasets, hence an instance is unlikely to be seen twice during training.
- Datasets can be configured to focus on specific instance types for an in-depth evaluation, yielding detailed insights.
- Datasets and their components can be recombined in mixer datasets for a combinatorially large number of instance types.
- Generated language can be more challenging than human annotations, e.g. syntax, requiring good understanding abilities.
- Abstract data and random sampling reduces risk of "Clever Hans effect", i.e. unexpected hidden biases in the data.

Uncovered shortcomings in experiments

- Existential statements in the context of multiple shapes.
- Spatial relations and quantification examples.