Deep learning evaluation using deep linguistic processing

Alexander Kuhnle & Ann Copestake

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

Evaluation methodology and real-world datasets

Properties and issues

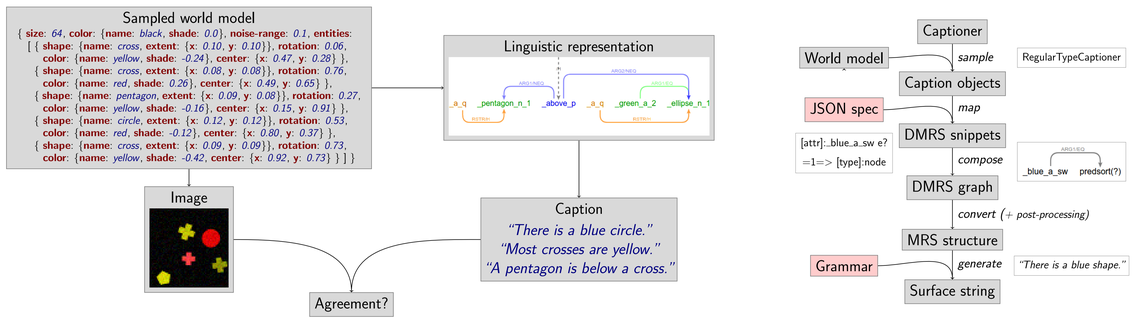
- Natural or repurposed? — photo data does not correspond to human perception of the world.
- 'Zipfian' tendency to simplicity — crowd-sourced language (and image) data tends to be simple.
- "Clever Hans effect" — unintended biases and correlations confounding experimental results.
- Adversarial examples — surprisingly odd system behavior on minimally modified data.

Three guiding principles

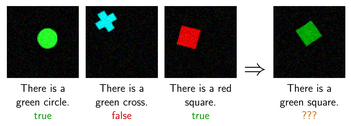
- No single general evaluation benchmark, but investigation tailored to the model.
- Dissimilar train/test distributions, requiring compositional generalization.
- Clean data with clear image/text relationship, instead of uncontrolled content.

⇒ Synthetic data as targeted 'unit-testing' evaluation, complementing general real-world benchmarks.

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
- Evaluation data is different from training data, hence requiring ability to recombine/generalize:



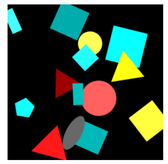
Conclusion: why use artificial data?

- **Challenging test data:** nontrivial multimodal reasoning required, more complex than what crowd-sourcing would plausibly produce.
- **Avoid Clever Hans effect:** data is comparatively unbiased, data space is covered relatively uniformly and exhaustively.
- **Flexibility & reusability:** data generation system is easily reusable, even for unforeseen use cases or changes in evaluation focus.
- **Rich evaluation:** unit-testing deep neural networks on specific instance types which are configurable in detail, hence a better way to establish trust in a model's understanding abilities than a monolithic dataset.

ShapeWorld GitHub: `https://github.com/AlexKuhnle/ShapeWorld` — ShapeWorld arXiv: `https://arxiv.org/abs/1704.04517`

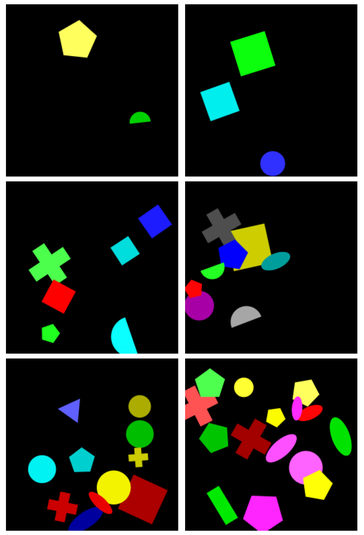
Examples: relations and quantifiers

- 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.

Coverage and configurability of generation system



Generator modules and configurability

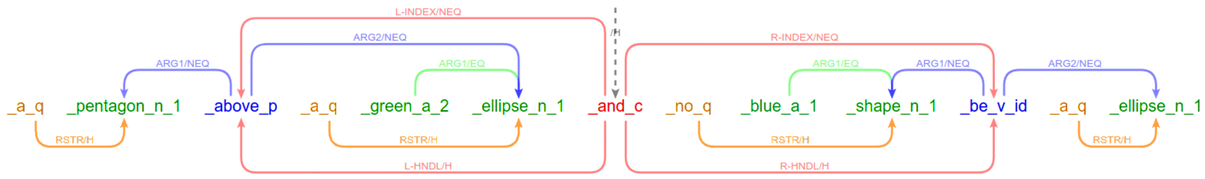
- **World attributes:** number of objects, available attributes, withheld combinations
- **Primary object attributes:** location, size, shade
- **Secondary object attributes:** rotation, distortion, collision tolerance
- **Attribute choice:** random, reinforced, limited subset

Captioner modules and configurability

- **Object(s) description:** red square, square, red shape, shape
- **Spatial relations:** left, right, above, below, in front of, behind, closer, farther
- **Attribute relations:** same/different shape/color as, bigger, smaller, lighter, darker
- **Relation variants:** negation, comparative, superlative
- **Numbers:** zero, one, two, three, four, five
- **Quantifiers:** no, a quarter, a third, half, two thirds, three quarters, all
- **Number/quantifier modifiers:** less than, at most, exactly, at least, more than, not
- **Number bounds:** of the two/. . . /eight
- **Comparative quantifiers:** one/. . . /five less/more than, as many, half/twice as many
- **Logical connectives:** and, or, if, if and only if

Compositional grounded semantics of captions

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



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