Deep learning evaluation using deep linguistic processing
Alexander Kuhnle & Ann Copestake
University of Cambridge
{aok28, 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.
- ‘Zifan’ 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.

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.

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.