Evaluating multi-modal deep learning systems with micro-worlds

Alexander Kuhnle & Ann Copestake

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

Problems with recent multi-modal tasks

Impressive performance on visual tasks (Karpathy and Li, 2015; Antol et al., 2015), but:

- ▷ Opaqueness of the inner working of a deep neural network makes it difficult to comprehend the process and content of learning in such a network.
- ▷ *Unexpected, human-atypical erroneous behaviour* (Nguyen et al, 2015; Szegedy et al, 2013).
- ▷ Object recognition is still *far from being solved* (Hodosh and Hockenmaier, 2016).

⇒ **Current research largely considers deep networks as black-box approximators.**

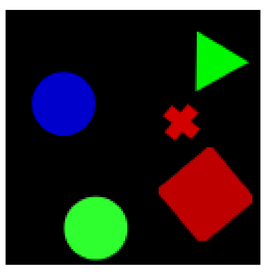
Proposed solution: Artificial toy data

- ▷ Artificial data can be *generated automatically in arbitrary quantities*.
- ▷ *Control over the content* of both individual instances and the entire training/test set split.
- ▷ Still, a *natural representation* of the data as image and natural language text.
- ▷ Toy data consists of *clear and structurally rich* instances which are *non-trivial* to handle (at least for deep learning approaches), hence *targeted evaluation for symbolic capabilities*.

⇒ **A first step towards understanding and consequently justifying the reasoning behind deep learning semantics.**

- ▷ In the spirit of traditional formal semantics and early AI research.
- ▷ Recently revived interest in artificial data for deep learning evaluation (see Sukhbaatar et al. (2015), Mikolov et al. (2015) or Weston et al. (2015)).

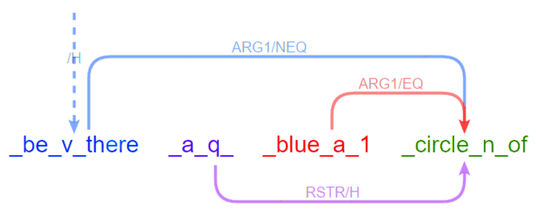
ShapeWorld – Abstract micro-worlds



Internal representation:

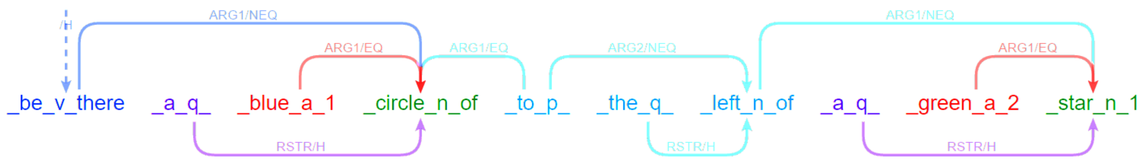
$$\begin{bmatrix} \{\text{shape} = \text{"triangle"}, \text{colour} = \text{"green"}, \text{x} = 0.9, \text{y} = 0.1, \text{etc.}\}, \\ \{\text{shape} = \text{"circle"}, \text{colour} = \text{"blue"}, \text{x} = 0.2, \text{y} = 0.3, \text{etc.}\}, \\ \{\text{shape} = \text{"cross"}, \text{colour} = \text{"red"}, \text{x} = 0.7, \text{y} = 0.5, \text{etc.}\}, \\ \{\text{shape} = \text{"square"}, \text{colour} = \text{"red"}, \text{x} = 0.8, \text{y} = 0.8, \text{etc.}\}, \\ \{\text{shape} = \text{"circle"}, \text{colour} = \text{"green"}, \text{x} = 0.3, \text{y} = 0.9, \text{etc.}\} \end{bmatrix}$$

- ▷ The internal representation of a shape world is *automatically generated by randomly sampling* shape, colour, number of entities, etc.
- ▷ The visual representation of a shape world is extracted from its internal representation.
- ▷ Captions are generated from an intermediate representation as DMRS semantic graph
- ▷ DMRS: Dependency Minimal Recursion Semantics, based on bi-directional DELPH-IN high-precision grammar capable of generating from a semantic graph structure (Copestake et al., 2016; Flickinger et al., 2014).



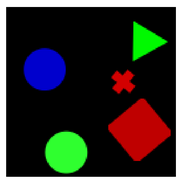
"There is a blue circle."

- ▷ The *compositionality of DMRS semantics* allows to combinatorially generate a wide range of captions with linguistic variety, for example:



"There is a blue circle which is to the left of a green star."

Examples requiring visual-linguistic understanding



- ▷ **Focus** *"There is a green circle."*
- ▷ **Quantification and counting**
  "Two shapes are green." *"All squares are red."*
  "Most circles are blue." *"Some green shapes are squares."*
- ▷ **Spatial relations**
  "All circles are to the left of a red cross." *"A blue circle is above a red square."*
  "Most red shapes are to the right of the green circle."
- ▷ **Comparative statements**
  "The left-most shape is a blue circle." *"The red cross is smaller than the red square."*
- ▷ **Underspecification**
  "The blue shape is above a green circle." *"The circle is to the left of the square."*
- ▷ **First-order logic**
  "There is a blue cross or all green triangles are not below the red square."
- ▷ **Syntactic/semantic re-combination**
  "At least one cross is blue, or all triangles which are green are above the red square."

Colors: true, unclear, false

Comparison to formal semantics approach

- ▷ *"There is a green circle."*

$$\exists s \in W : \text{circle}(s.\text{shape}) \wedge \text{green}(s.\text{colour})$$

- ▷ *"Two shapes are green."*

$$\exists s_1, s_2 \in W : s_1 \neq s_2 \wedge \text{green}(s_1.\text{colour}) \wedge \text{green}(s_2.\text{colour})$$

- ▷ *"All circles are to the left of a red cross."*

$$\forall s_1 \in W : \text{circle}(s_1.\text{shape}) \Rightarrow \left(\exists s_2 \in W : \text{cross}(s_2.\text{shape}) \wedge \text{red}(s_2.\text{colour}) \wedge s_1.\text{x} < s_2.\text{x}\right)$$

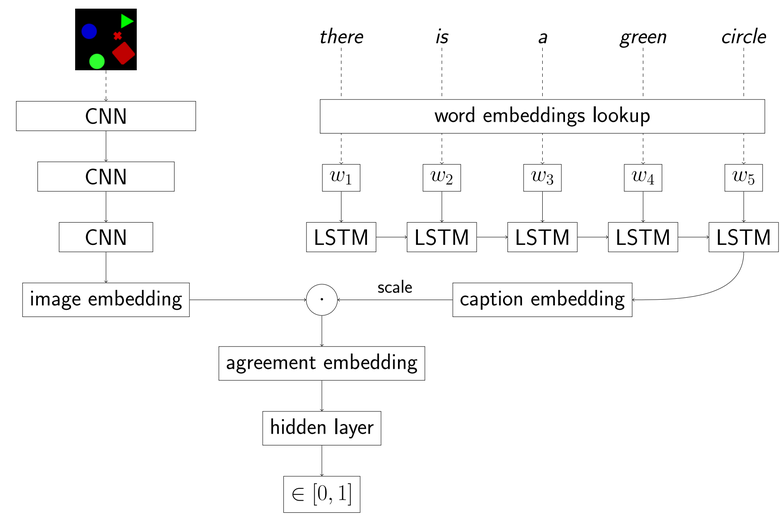
- ▷ *"The left-most shape is a blue circle."*

$$\forall s_1 \in W : \left(\forall s_2 \in W : s_1.\text{x} \leq s_2.\text{x}\right) \Rightarrow \text{circle}(s_1.\text{shape}) \wedge \text{blue}(s_1.\text{colour})$$

⇒ **Given a symbolic world model, formal semantics can handle such statements.**

Multi-modal deep learning architecture

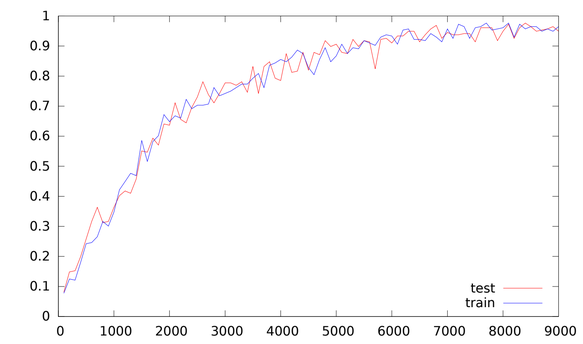
Task: Decide whether the caption is in agreement with the image.



- ▷ Potential baselines: Bag of words, general-purpose object recognition network
- ▷ Below first "sanity check" results:

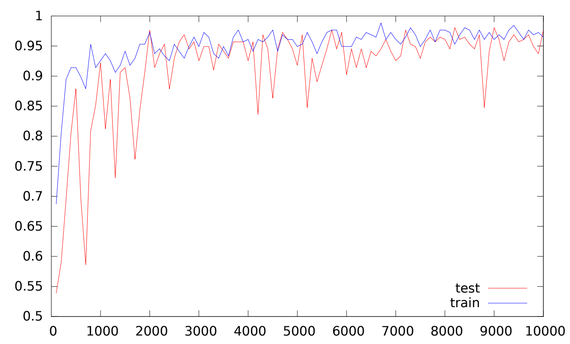
Object recognition

- ▷ Single-object images only
- ▷ Picked from 8 shapes and 7 colors
- ▷ Consequently, 56 distinct objects



Existential statements

- ▷ Images: Only a single object
- ▷ Captions: *"There is a [colour] [shape]."*
- ▷ Test: Unseen combination *"red triangle"*



Coming soon: Analysing the learnability of statements involving quantifiers

What's next? – Ideas for the PhD project

- ▷ **Noise** Visual variation can be increased by introducing pixel value noise, diversifying colour shades and shape sizes, (potentially) applying photo filters and effects.
- ▷ **Cliparts** Similar to Zitnick et al. (2016), the visual scenario of the micro-worlds can easily be changed to more diverse and more detailed objects.
- ▷ **Paraphrasing** Earlier work on paraphrase rules as DMRS graph mapping (Copestake et al., 2016) can be applied to introduce more linguistic variety.
- ▷ **Human captions** Humans can be asked to give various (true or false) statements about the images of our system. These can be compared to the automatically generated captions to inspire increasing linguistic variety, or used as test set for a more realistic evaluation.
- ▷ **Different language** The captions can easily be generated in another language by switching to a different DELPH-IN grammar, e.g. for German or Japanese or for a more exotic language in the Grammar Matrix project (Bender et al., 2002).
- ▷ **Multi-agent experiments** Our image generation system can be used to develop symbolic tasks with the aim of *evaluating language evolution, pragmatics and psycholinguistics* in a multi-agent setup, similar to Lazaridou et al. (2016). In contrast to their approach, however, the micro-worlds of our system are more diverse and cover a larger configuration space, so they are likely to avoid the issue of the agents developing a specialised language which exploits hidden regularities in the data.