**ShapeWorld** - A new test methodology / environment

- Generate abstract microworlds of colored shapes
- Evaluate multimodal deep learning models
- Focus on “formal-semantics-style” tasks
- Test for multimodal language understanding and generalization abilities
- Analyze the learning process and basic capabilities of deep networks

![Diagram showing the process of generating and captioning images in ShapeWorld]
‘All dogs have four legs’: Learning Natural Language Quantifiers from Visual Experience

Bernardi, R.\textsuperscript{1}, Herbelot, A.\textsuperscript{1}, Kuhnle, A.\textsuperscript{2}, Pezzelle, S.\textsuperscript{1}, Sorodoc, I.\textsuperscript{1}

\textsuperscript{1}CIMeC, DISI - University of Trento
\textsuperscript{2}University of Cambridge
Overview

What can we infer from such experience?

➔ All dogs have *four legs*. / No dog has *two legs*.
➔ Most dogs are *brown*. / Some dogs are *black*. / No dog is *red*.
➔ All dogs have a *tail*.
➔ etc
Motivation

- From limited visual experience of objects, humans learn to generalize to rough frequency estimates of object attributes for a certain concept.
- Natural language quantifiers (*no*, *some*, *most*, *all*, etc) are used to express these frequency estimates.
- Hence they act as a proxy revealing the learned representation of this cognitive process (to some degree, at least).

Aim of this project

- Create a dataset resembling this learning setup
- Evaluate various deep learning models on this task
Creating the dataset

Two approaches:

1. Using quantified McRae’s feature norms (Herbelot and Vecchi, 2016):
   a. Extract images from existing resources (Visual Genome, MS COCO, etc) based on their provided annotations of objects and attributes.
   b. Query image search engines like Google, Bing, etc for “<concept> <attribute>”.
   c. Control for agreement between attribute frequency and associated quantifier.

2. Relying on annotations of image datasets (Visual Genome, MS COCO):
   a. Obtain relative frequencies of attributes for all concepts.
   b. Map these attribute frequencies to the corresponding quantifier, according to traditional formal semantic interpretation.
Ongoing issues

1. Problems with McRae’s feature norms:
   a. **Bad coverage** in existing resources (attributes do not match dataset annotations).
   b. Concepts are **too specific** and many attributes are **not visual**.

2. Problems with existing datasets:
   a. MS COCO: **only** 29 concepts, **only** few properties well-represented for each concept.
   b. Visual Genome: attributes **too sparse / too specific** (however, good concept coverage).

3. Images obtained from search engines:
   a. Results are **unreliable** and **inconsistent**, hence require **manual filtering**.
   b. **Vague concept boundaries** lead to many **“borderline” results**, which are hard to classify.
   c. Search results are sometimes **biased** towards some kind of **“prototypical” image structure**.
How we want to continue...

- Stick with McRae’s feature norms, since they are based on judgements of what humans consider “typical” for a concept.
- Focus on the Visual Genome dataset due to its good concept coverage.
- Include all attributes, even when covered only by a single image.

Dataset / Task

- Sample sequences of images (~100?, flexible) which consists of a certain attribute-per-concept frequency, amongst other attributes and concepts.
- Given such a sequence, target concept and attribute, the system has to decide which quantifier applies.