CIMeC

UNIVERSITÀ DEGLI STUDI DI TRENTO

CLIC Lab

Be *Precise* or *Fuzzy*: Learning the Meaning of Cardinals and Quantifiers from Vision

S. Pezzelle[1], M. Marelli[2], R. Bernardi[1]

[1]CIMeC, DISI - University of Trento
[2]Department of Psychology, University of Ghent

Overview

How many of the animals are **dogs**? **‘Three’** / **‘Most’**

Motivation and Goal

Motivation

- People can refer to quantities in a visual scene by using either exact cardinals (Cs) or natural language quantifiers (Qs)
- In humans, processes underlying different cognitive and neural mechanisms
- Meaning of both Cs and Qs is learned in multimodality

Goal

- Single, computational architecture for learning the meaning of Cs and Qs capitalizing on 2 different functions (**cosine** and **dot product**)

Dataset

We build a dataset of **synthetic scenes** by join together 1-9 real images from ImageNet (each image depicting **one** object)

Properties:

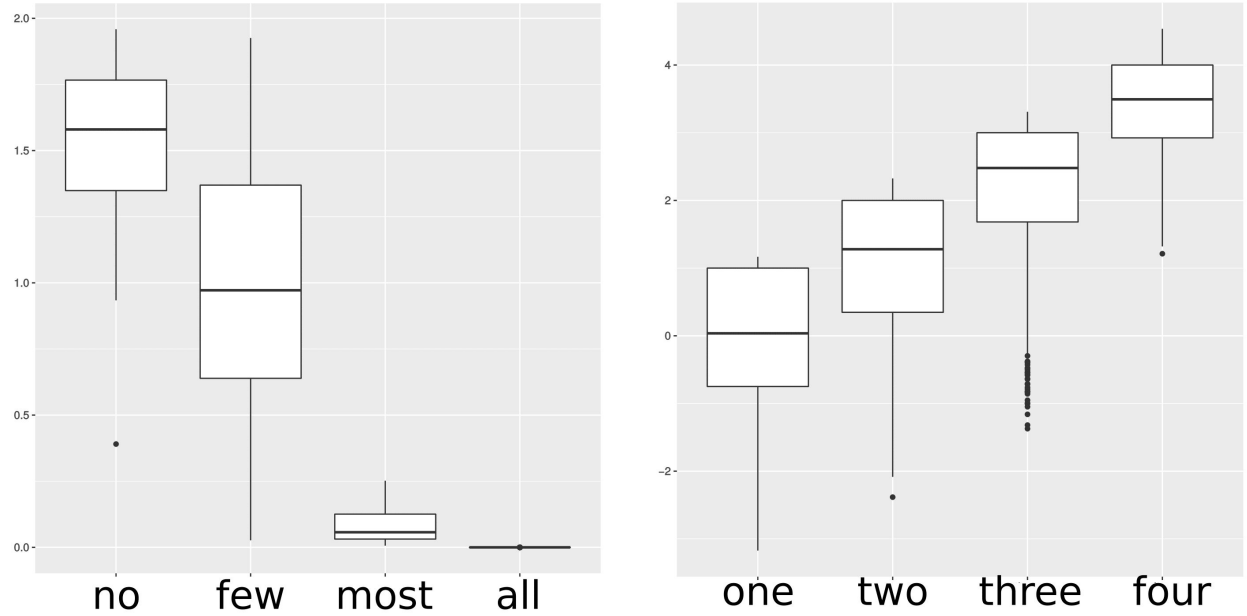
1. Balanced number of scenarios depicting *no, few, most, all* (Qs); 1,2,3,4 (Cs)
2. Qs' percentages defined 'a priori' (0%, 1-49%, 51-99%, 100%, resp.)
3. Train, Test differing w.r.t. different combination targets-distractors

Combinations

Train-q				**Train-c**			
no	few	most	all	one	two	three	four
0/1	1/6	2/3	1/1	1/1	2/2	3/3	4/4
0/2	2/5	3/4	2/2	1/3	2/3	3/4	4/5
0/3	2/7	3/5	3/3	1/4	2/5	3/5	4/6
0/4	3/8	4/5	4/4	1/6	2/7	3/8	4/7
Test-q				**Test-c**			
no	few	most	all	one	two	three	four
0/5	1/7	4/6	5/5	1/2	2/4	3/7	4/8
0/8	4/9	6/8	9/9	1/7	2/9	3/9	4/9

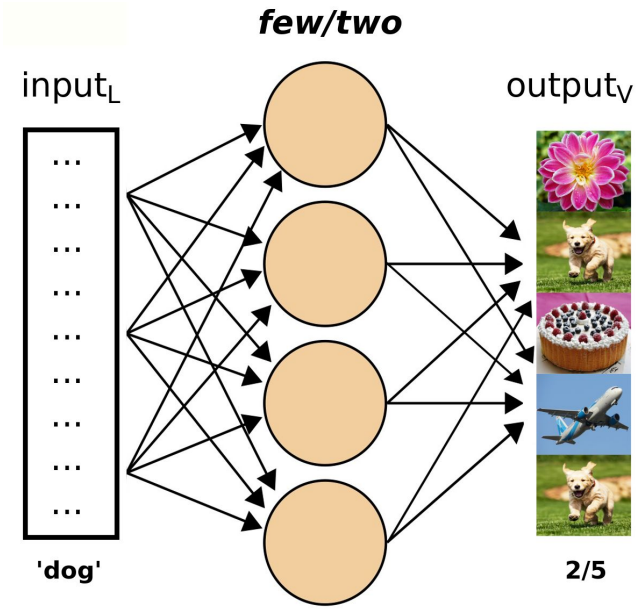
Combinations in Train, Test. Numerator: n of target objects. Denominator: n of targets+distractors

Only-vision evaluation



Left: Qs against cosine distance. Right: Cs against dot product

Model



- **Cross-modal mapping** modelling each Q/C as a separate function
- **Cosine** (‘fuzzy’) is used for Qs, **dot product** (‘exact’) for Cs
- Single-layer neural network (criterion ReLU)

Evaluation & Results

Each mapping function is evaluated by means of **retrieval task** aimed at picking up the correct scenarios from Test combinations

	lin		**nn-cos**		**nn-dot**	
	mAP	*P2*	*mAP*	*P2*	*mAP*	*P2*
no	0.78	0.65	**0.87**	0.77	0.54	0.37
few	0.59	0.39	**0.68**	0.51	0.59	0.43
most	0.61	0.36	0.60	0.29	**0.62**	0.45
all	0.75	0.66	**1**	1	0.33	0.12
one	0.44	0.30	0.38	0.21	**0.61**	0.45
two	0.35	0.15	0.38	0.21	**0.57**	0.43
three	0.38	0.16	0.36	0.13	**0.56**	0.40
four	0.65	0.47	0.75	0.60	**0.76**	0.61

Discussion

- The two proposed objective functions turn out to best describe Cs and Qs
- Cosine is a ‘fuzzy’ measure evaluating the overall similarity target-scene; dot product includes information about the ‘exact’ number of instances

Open issues

- Cognitive plausibility of ‘one quantified expression, one function’ approach
- Is the approach feasible for numerosities exceeding ‘subitizing’ range?
- Do quantifiers lie on an ordered scale from ‘none’ to ‘all’?

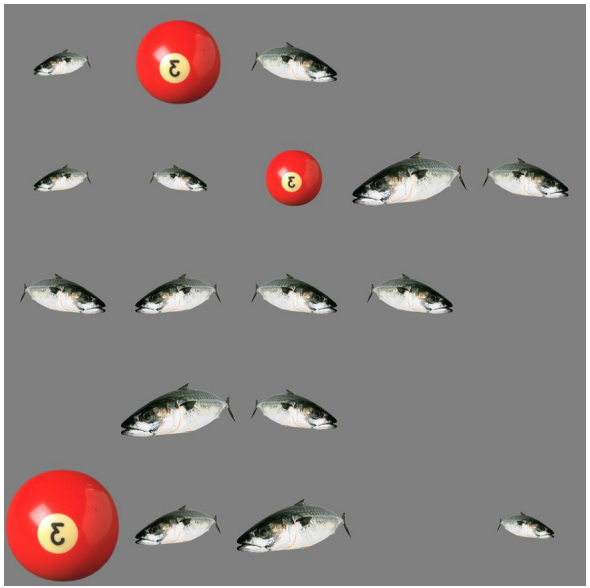
Ongoing work

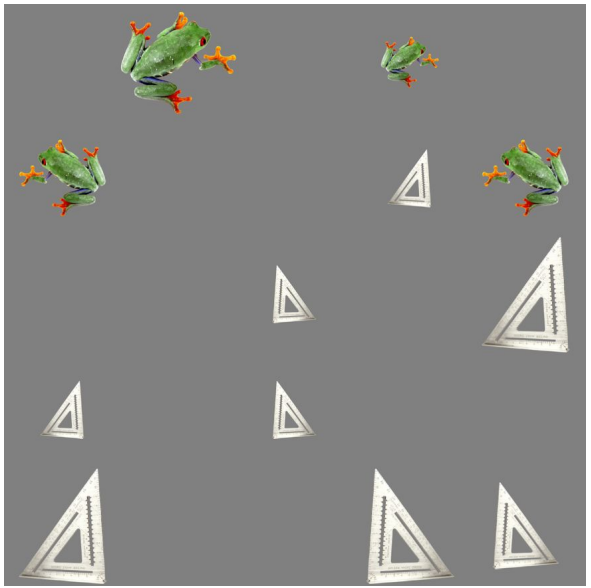
Two **behavioral studies**:

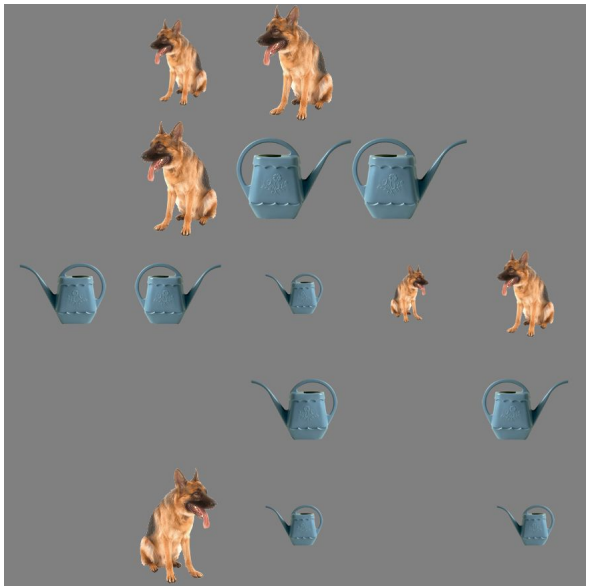
1. Only-language study investigating semantic similarity between Qs
   - aimed at empirically test the ‘ordered scale’ assumption for Qs
   - how similar is the meaning of, e.g. ‘none’ and ‘some’ in a 1-7 scale

2. Only-vision study investigating the meaning of Qs
   - given a visual scene containing animals and artifacts, provide correct Q out of 9 options: none, almost none, very few, few, some, many, most, almost all, all

How many of the objects are animals?







Thank you for your attention!