EACL 2017
European Chapter of the Association for Computational Linguistics
Valencia, 3-7 April 2017
EACL 2017

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

Sandro Pezzelle[1], Marco Marelli[2], Raffaella Bernardi[3]

[1,3]CIMeC - Center for Mind/Brain Sciences, [3]DISI, University of Trento
[2]Department of of Psychology, University of Ghent

Overview

Quantities can be expressed by either 'exact' **cardinals** (*one*, *two*, *three*, *four*) or 'fuzzy' **quantifiers** (*no*, *few*, *most*, *all*)

In humans, different cognitive [1] and neural [2] mechanisms: numbers neither necessary nor sufficient for quantifiers

Meaning of both acquired in grounded contexts [3,4]

Our proposal:

Single computational architecture (**cross-modal mapping**) for learning meaning of cardinals (Cs) and quantifiers (Qs) from Vision

But, two different objective functions required:

1. **cosine** ('fuzzy') for quantifiers
2. **dot product** ('exact') for cardinals

@tweetabstract: #quantification as #crossmodalmapping from #language to #vision: learning #quantifiers and #cardinals @sandropezzelle @raffagbernardi @CIMeC_

How many pets are *dogs*? **Three** / **Most**

Dataset

Synthetic scenarios containing up to 9 objects representing either **targets** (*dogs*) or **distractors** (*cats*)

GT quantifiers defined by (simplified) **ratios** n_targets/n_total: no = 0%; few < 50%; most > 50%; all = 100%

Different combinations n_targets/n_total in training and testing (e.g. 'few' 1/6, 2/5, 2/7, 3/8 in training; 1/7, 4/9 in testing)

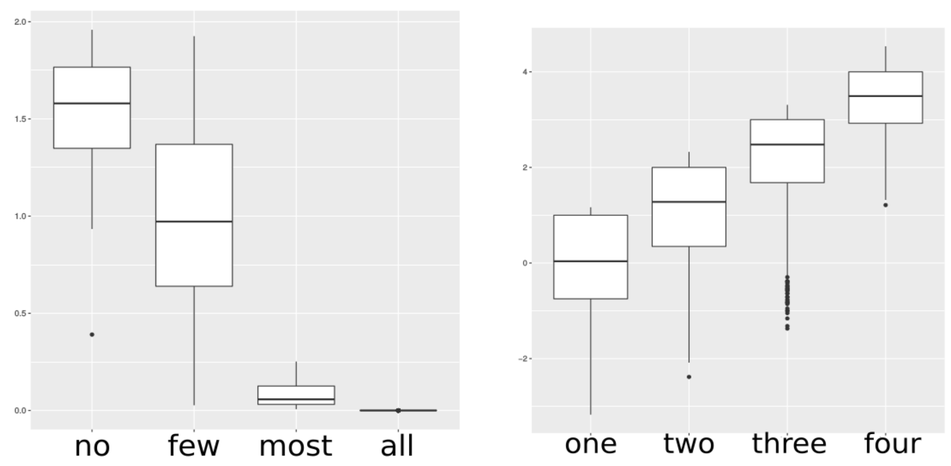
Scenarios: (a) 188 concepts from ImageNet ; (b) VGG-19 CNN feats extracted from all images; (c) 1 centroid vector per concept; (d) PCA; (e) normalization; (f) **sum**

<scenario,quantifier> datapoints: **3008 train**, **1504 test**

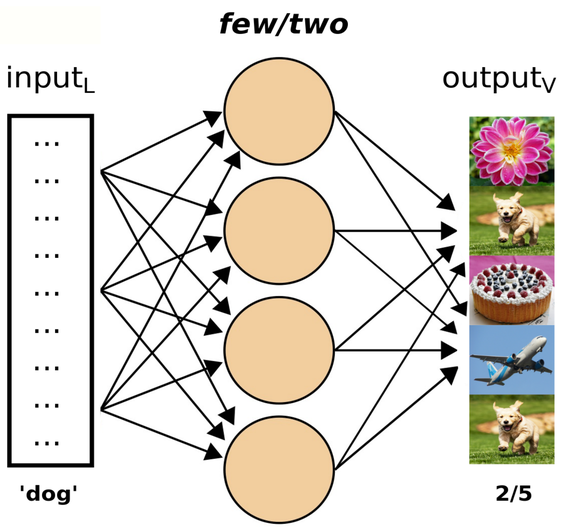
Only-vision Evaluation

Cosine distance and **dot product** between each scenario (e.g. 2/5 *dogs*) and corresponding target object (*dog*)

Cosine effective for Qs (left), dot product for Cs (right)



Cross-modal Mapping



Models:

lin: linear mapping

nn-cos: 1-layer NN maximizing cosine similarity (1) between in and out

nn-dot: 1-layer NN making dot product between in and out close to 1

Cross-modal mapping between linguistic representation of target object (word2vec) and visual scenarios representing a given Q/C

Each **Q/C** (e.g. 'few', 'two') learned as a **specific matrix/function**

Results & Discussion

	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

	no	few	most	all
no	**288**	88	0	0
few	141	**191**	38	6
most	0	0	111	265
all	0	0	0	**376**
	one	two	three	four
one	**168**	113	54	41
two	64	**136**	124	52
three	23	80	130	145
four	10	24	72	**272**

Retrieval task: as conjectured, nn-cos overall best model for Qs, nn-dot overall best model for Cs

Best-performing models make 'plausible' errors (ordered scale)

References:

[1] Pietroski et al. 2009. The meaning of 'most': Semantics, numerosity and psychology.
[2] Wei et al. 2014. Dissociated neural correlates of quantity processing of quantifiers ...
[3] Halberda et al. 2008. The development of 'most' comprehension and its potential ...
[4] Barner et al. 2009. Cross-linguistic relations between quantifiers and numerals in language ...