CLIC Lab

“Look, some green circles!”: Learning to quantify from images

Ionut-Teodor Sorodoc, Angeliki Lazaridou, Gemma Boleda
Aurélie Herbelot, Sandro Pezzelle, Raffaella Bernardi
{firstname.lastname}@unitn.it

Motivation

Word meaning can be modelled in a cognitively plausible way by learning representations from both **linguistic** and **visual** contexts

Current models are very effective in representing content words, but fail with **function words** like natural language quantifiers

In grounded contexts, children can provide quantification estimates before learn to count via Approximate Number System (**ANS**) [1,2]

Research question

We investigate whether a neural network can learn the meaning of quantifiers (*no*, *some*, *all*) from utterances grounded in vision

Consistently with human ANS, we hypothesize that counting is neither sufficient not necessary for the acquisition of quantifiers

Task

Given a set of objects (circles) with different properties (colors), the model learns to apply the **correct quantifier** to the scenario
→ e.g., *no/some/all circles are green*

In formal semantics terms, we focus on the **scope** of quantification, since the domain restrictor is fixed (objects are all circles)

Visual quantification dataset

We experiment with an **artificial dataset** including 5K datapoints <image, query, quantifier>

Images

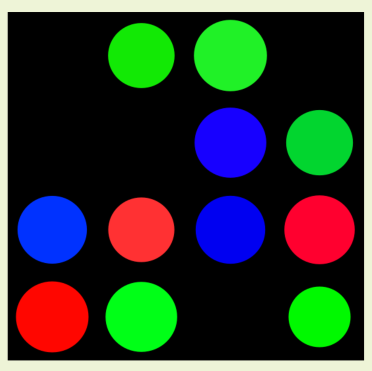
Each image contains 1 to 16 circles of 15 different colors; all the possible combinations wrt number of circles and colors are built

Image representation

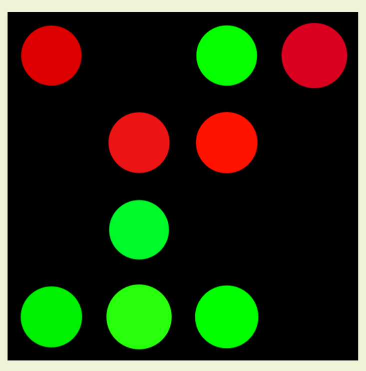
- Each circle is represented by a real-valued, 20-dim vector that is normalized to unit norm and has pairwise similarity < 0.7
- Empty cells are represented by orthogonal vectors
- Gaussian noise is added to all object vectors to simulate shades

Queries

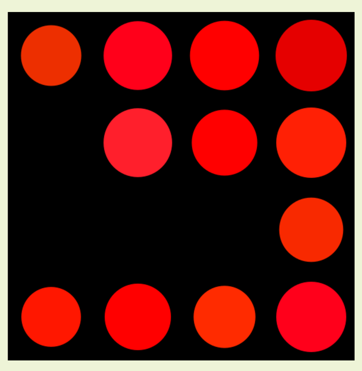
Each image is associated with a query, i.e. the property (*green*), and the correct quantifier for that property, e.g. *some*



some circles are green



no circles are blue



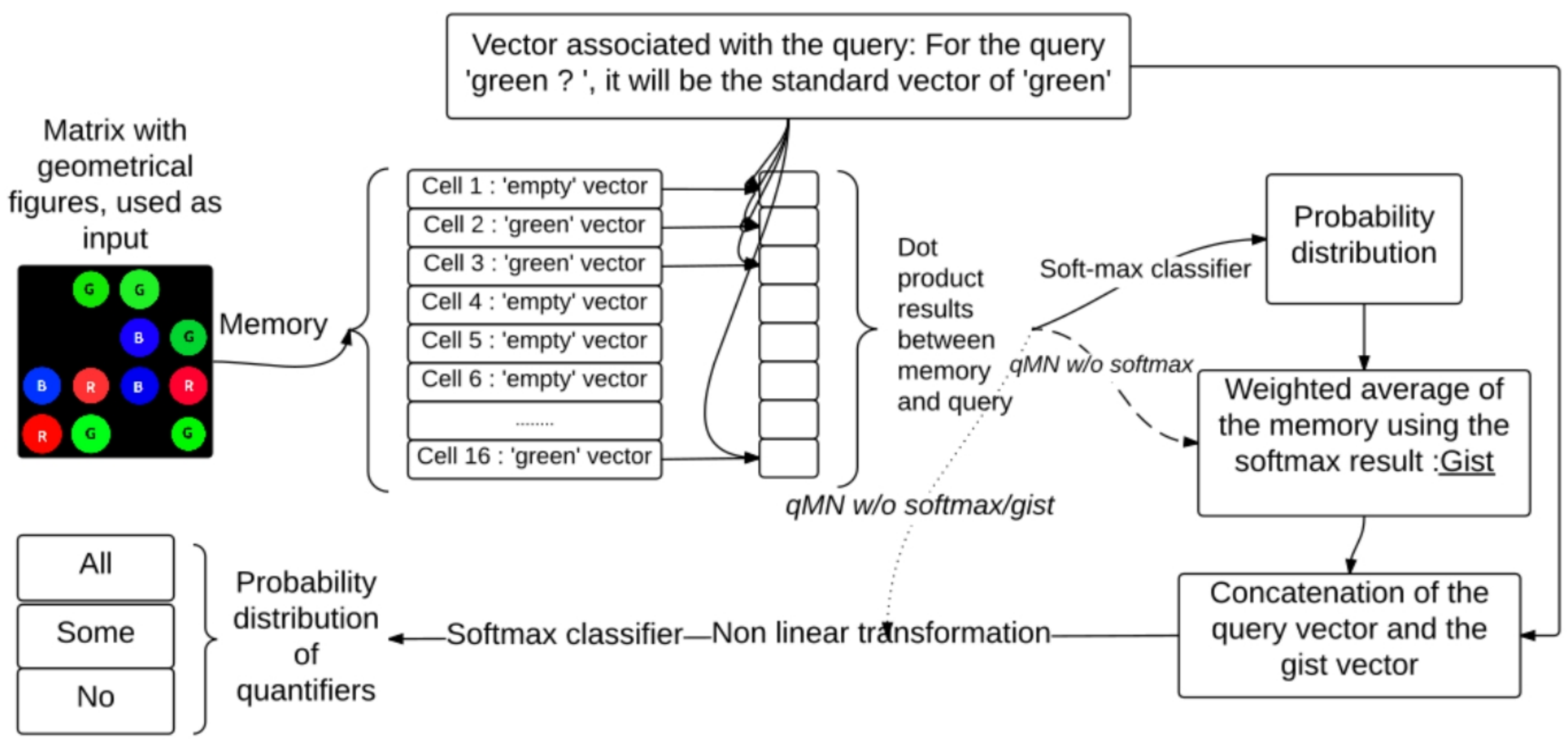
all circles are red

Models

1. Counting model

- each image is represented by a 16-d feature vector (1 for each color + 1 for empty cell) encoding the frequency of each color scaled by similarity with the query (one-hot 16-d vector)
- feature vector and query are concatenated + softmax classifier

2. Quantifier Memory Network (qMN) adapted from [3]



3. Recurrent Neural Network (RNN)

- uses hidden state to encode information about image gist
- at each timestep, the RNN receives as input first the query vector followed by each of the 16 object vectors
- last timestep: hidden layer linearly reduced to 3-d + softmax

Experimental setup

We test each model in 3 experimental setups:

1. ***familiar***: 5K datapoints randomly split in train/val/test set
2. ***unseen quantities***: no overlap train/test wrt number of objects in the image
3. ***unseen colors***: train with 10 colors and test w/ 5 unseen colors

Results

Models	familiar	unseen quantities	unseen colors
RNN	65.7	62.0	49.7
Counting	86.5	78.4	32.8
qMN	**88.8**	**97.0**	**54.9**
-softmax	85.9	66.6	54.4
-softmax/gist	51.4	51.8	44.4

Models accuracies (in %). Last 2 lines refer to qMN model versions without either softmax or softmax/gist (performance decreases)

Conclusion & current work

Counting is neither necessary nor sufficient to quantify over images

We are currently extending our investigation to other quantifiers as *few* and *most* and modeling the **restrictor** of the quantification

References

[1] L. Feigenson, S. Dehaene, and E. Spelke. 2004. Core system of number. Trends in cognitive sciences, 8(7):307-314
[2] M. M. Mazzocco, L. Feigenson, and J. Halberda. 2011. Preschoolers' precision of the approximate number system predicts later school mathematics performance. PLoS one, 6(9):e23749
[3] S. Sukhbaatar, A. Szlam, J. Weston, and R. Fergus. 2015. End-to-end memory networks. http://arxiv.org/abs/1503.08895