“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.

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

<table>
<thead>
<tr>
<th>Models</th>
<th>familiar</th>
<th>unseen quantities</th>
<th>unseen colors</th>
</tr>
</thead>
<tbody>
<tr>
<td>RNN</td>
<td>65.7</td>
<td>62.0</td>
<td>49.7</td>
</tr>
<tr>
<td>Counting</td>
<td>86.5</td>
<td>78.4</td>
<td>32.8</td>
</tr>
<tr>
<td>qMN</td>
<td>88.8</td>
<td>97.0</td>
<td>54.9</td>
</tr>
<tr>
<td>-softmax</td>
<td>85.9</td>
<td>66.6</td>
<td>54.4</td>
</tr>
<tr>
<td>-softmax/gist</td>
<td>51.4</td>
<td>51.8</td>
<td>44.4</td>
</tr>
</tbody>
</table>

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