Graph-Structured Representations for Visual Question Answering

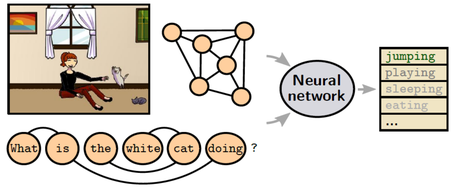
Damien Teney, Lingqiao Liu, Anton van den Hengel

THE UNIVERSITY of ADELAIDE

Overview

Definition of visual question answering
Input: image + text question
Output: text answer from predefined set of frequent ones (classification problem)

Contribution: inputs represented as graphs
Scene: structured description readily available
nodes = objects
edges = relative spatial relations (dense connectivity)
Question: leverage existing NLP tools for syntactic parsing
nodes = words
edges = syntactic relationships (sparse connectivity)



Technical details

Propagation of information over the graph from neighbours, over several iterations

$$h_i^0 = 0$$
$$n_i = \text{pool}_j(e'_{ij} \circ x'_j)$$
$$h_i^t = \text{GRU}(h_i^{t-1}, [x'_i; n_i]) \qquad t \in [1, T].$$

Matching the graphs of questions and image

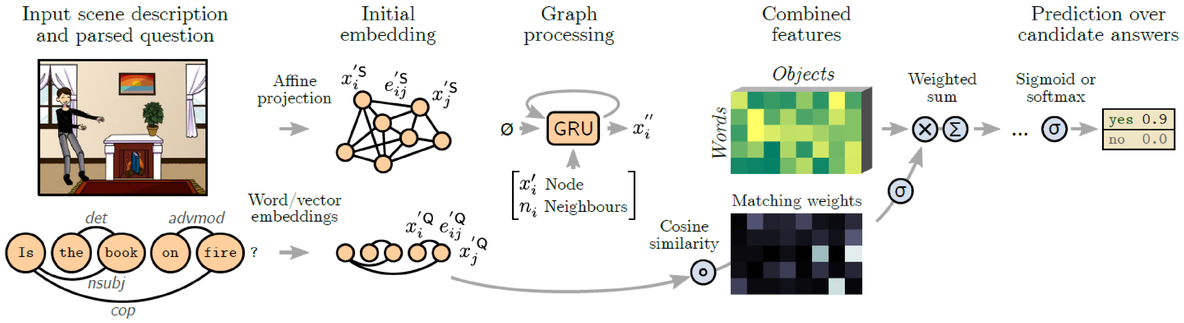
- Attention weights

$$a_{ij} = \sigma\left(W_5\left(\frac{x_i'^{\mathsf{Q}}}{\|x_i'^{\mathsf{Q}}\|} \circ \frac{x_j'^{\mathsf{S}}}{\|x_j'^{\mathsf{S}}\|}\right) + b_5\right)$$

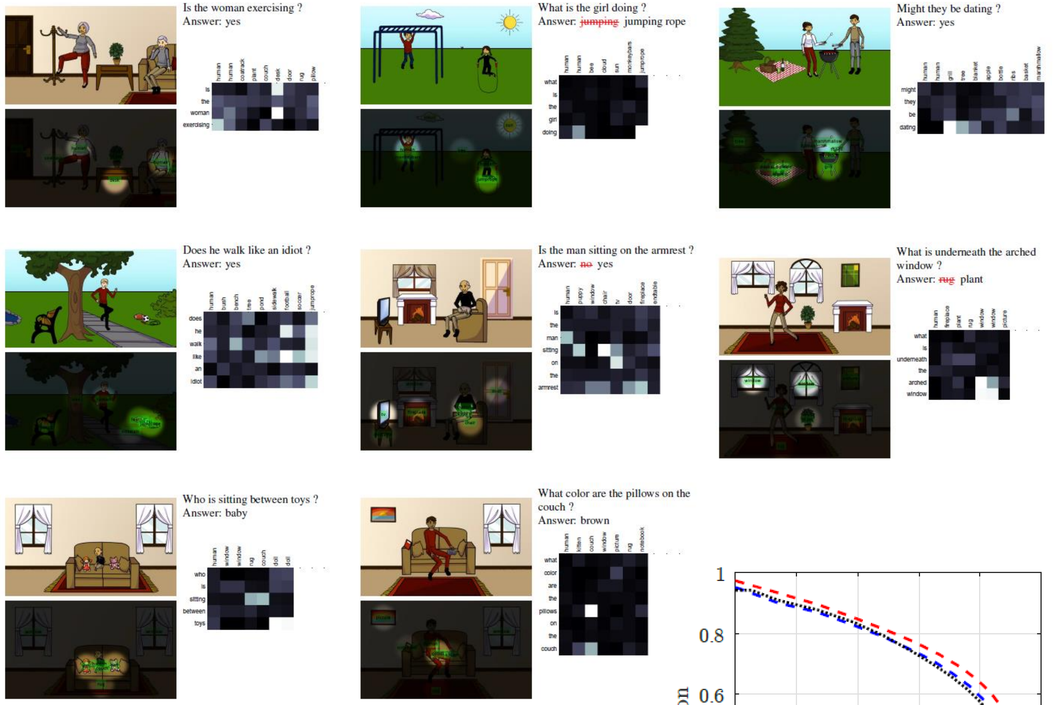
- Weighted sum of the features

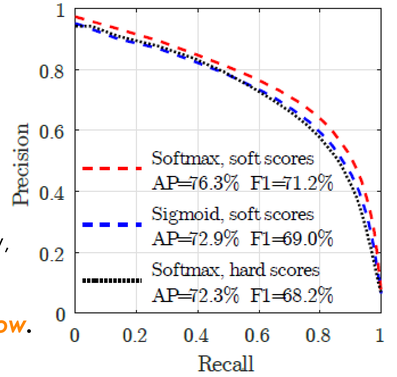
$$y_{ij} = a_{ij} \cdot [x_i''^{\mathsf{Q}}; x_j''^{\mathsf{S}}]$$

Network architecture



Results





As seen on P/R curve: the model's output (after softmax or sigmoid) is a good measure of its confidence/uncertainty, especially when trained with soft scores as targets.

Practically, this can be used to derive the answer *I don't know*.