Graph representation of scenes and questions

Chaonan Song

May 17, 2018

1 Evaluation

Datasets Their evaluation use two datasets the original abstract scenes from Antol *et al.* [1] and its "balanced" extension from [2]. They all contain scenes created by humans for arranging clip art objects and graphics. The original dataset contains 20k/10k/20k scenes and 60k/30k/60k questions, each with 10 human provided ground-truth answers. Questions are categorized based on the type of the correct answer into *yes/no number*,and *other*, But all categories use the same method to test for unknown problems. The "balanced" version of the dataset contains only the subset of questions which have binary *yes/no* answers and, In addition, create complementary scenarios to elicit the opposite answer for each question. A pair of complementary scenes differs because only one or two objects are moved, removed or modified (see examples in Fig. 1, bottom rows). This makes the problem very challenging because of the need to take into account the subtle details of the scene.

Metrics The main indicator is the average "VQA score" [1], taking into account the variability of answers from multiple human annotators.

ground truth score $s(q, a) = 1.0$ if the answer a was provided by $m \geq 3$ annotators. Otherwise, $s(q, a) = m/3$ Their method outputs a predicted score $\mathbf{s}(q, a)$ for each question and answer and the overall accuracy is the average ground truth score of the highest prediction per question, *i.e.* $\frac{1}{M}\sum_q^M s\left(q, \arg\max_a \mathbf{s}(q, a)\right)$.

Their initial experiments confirmed that the performances of various algorithms on the balanced dataset were indeed better separated, and we used it for our ablative analysis. They also focus on the hardest evaluation setting [2], which measures the accuracy over pairs of complementary scenes. This is the only indicator of zero precision based on the guesswork. This setting also does not consider pairs of test scenes that are considered fuzzy due to inconsistencies between commentators. The metric is then a standard hard accuracy, *i.e.* all ground truth scores $s(i, j) \in \{0, 1\}$.

1.1 Evaluation on the "balanced" dataset

They compare their method against the three models proposed in [2]. The visual features in the three models are empty, global, or focused on two objects identified from the problem. These models are designed specifically for binary problems, and their models usually apply. Nevertheless, they obtain significantly better accuracy than all three (Table. 1). The difference in performance is mainly seen in the "pair" setting, which they believe is more reliable because it discards the fuzzing issue of the commenter disagreeing.

References

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. *In Proc. IEEE Int. Conf. Comp. Vis*, 2015.

[2] Peng Zhang, Yash Goyal, Douglas Summersstay, Dhruv Batra, and Devi Parikh. Yin and yang: Balancing and answering binary visual question-

Figure 1: Qualitative results on the abstract scenes dataset (top row) and on balanced pairs (middle and bottom row). We show the input scene, the question, the predicted answer, and the correct answer when the prediction is erroneous. We also visualize the matrices of matching weights (Eq. 6, brighter correspond to higher values) between question words (vertically) and scene objects (horizontally). The matching weights are also visualized over objects in the scene, after summation over words, giving an indication of their estimated relevance. The ground truth object labels are for reference only, and not used for training or inference.

Method	Avg. score over scenes	Avg. accuracy over pairs
Zhang *et al.* [2] blind	63.33	0.00
with global image features	71.03	23.13
with attention-based image features	74.65	34.73
Graph VQA(full model)	**74.94**	**39.1**
(1) Question: no parsing (graph with previous/next edges)		37.9
(2) Question: word embedding not pretrained		33.8
(3) Scene: no edge features ($e_{ij}^S = 1$)		36.8
(4) Graph processing: disabled for question ($x_i^Q = x_i^S$)		37.1
(5) Graph processing: disabled for scene ($x_i^S = x_i^Q$)		37.0
(6) Graph processing: disabled for question/scene		35.7
(7) Graph processing: only 1 iteration for question ($T^Q = 1$)		39.7
(8) Graph processing: only 1 iteration for scene ($T^S = 1$)		37.9
(9) Graph processing: only 1 iteration for question/scene		39.1
(10) Uniform matching weights ($a_{ij} = 1$)		39.1

Table 1: Results on the test set of the "balanced" dataset [2] (in percents , using balanced versions of both training and test sets). Numbered rows report accuracy over pairs of complementary scenes for ablated versions of our method.

s. In *Computer Vision and Pattern Recognition*,
pages 5014–5022, 2016.