Graph representation of scenes and questions

Chaonan Song

May 19, 2018

1 Evaluation

1.1 Evaluation on the "balanced" dataset

During training, They take care to keep pairs of complementary scenes together when forming mini-batches. This has a significant positive effect on the stability of the optimization. They did not notice any tendency toward overfitting when training on balanced scenes. They hypothesize that the pairs of complementary scenes have a strong regularizing effect that force the learned model to focus on relevant details of the scenes. In Fig. 2 (and in the supplementary material), they visualize the matching weights between question words and scene objects (Eq. 1). As expected, these tend to be larger between semantically related elements (*e.g.* daytime↔sun, dog↔puppy, boy↔human) although some are more difficult to interpret. Our best performance of about 39% is still low in absolute terms, which is understandable from the wide range of concepts involved in the questions (see examples in Fig. 2 and in the supplementary material). It seems unlikely that these concepts could be learned from training question/answers alone, and they suggest that any further significant improvement in performance will require external sources of information at training and/or test time.

$$[H]a_{ij} = \sigma(W_5(\frac{x_i^{'Q}}{\|x_i^{'Q}\|} \circ \frac{x_j^{'S}}{\|x_j^{'S}\|}) + b_5) \qquad (1)$$

Ablative evaluation They evaluated variants of Their model to measure the impact of various design choices (see numbered rows in Table 1). On the question side, they evaluate (row 1) their graph approach without syntactic parsing, building question graphs with only two types of edges, *previous/next* and linking consecutive nodes. This shows the advantage of using the graph method and syntactic parsing. The optimization starts from scratch (row 2) because starting from the pre-trained Glove vector [3] results in a significant decrease in performance. On the scene side, they removed the edge features (row 3) by setting $e_{ij}^S = 1$ It confirms the spatial relationship between the model's edge-coded objects. In rows 4C6, they disabled the recurrent graph processing ($x_i = x_i$) for the either the question, the scene, or both. They finally tested the model with uniform matching weights ($a_{ij} = 1$, row 10). It performed badly. Their weights are similar to those in other models (*e.g.* [1, 2, 4, 5, 7]), and their observations confirm that these mechanisms are critical to good performance.

Precision/recall They are interested in assessing the confidence of their model in its predicted answers. Most existing VQA methods treat the answer as a hard classification of the candidate answer, and almost all reported results consist of a single accuracy metric. To provide more insight, they produce *precision/recall* curves for predicted answers. A precision/recall point (p, r) is obtained by setting a threshold t on predicted scores such that

$$p = \frac{\sum_{i,j} 1\left(\mathbf{S}(i,j) > t\right) s(i,j)}{\sum_{i,j} 1\left(\mathbf{S}(i,j) > t\right)} \qquad (2)$$

$$r = \frac{\sum_{i,j} 1\left(\mathbf{S}(i,j) > t\right) s(i,j)}{\sum_{i,j}} \qquad (3)$$

where $1(\bullet)$ is the 0/1 indicator function. They plot

precision/ recall curves in Fig. 3 for both datasets2. The predicted score proves to be a reliable indicator of the model confidence, as a low threshold can achieve near-perfect accuracy (Fig. 3, left and middle) by filtering out harder and/or ambiguous test cases.

Effect of training set size They trained our model with limited subsets of the training data (see Fig. 1). Unsurprisingly, the performance grows steadily with the amount of training data, which suggests that larger datasets would improve performance. Their use of parsing and word embeddings is a small step in that direction. Both techniques clearly improve generalization(Fig. 1).

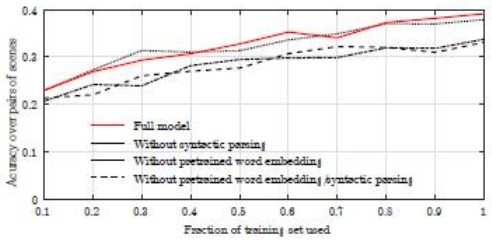

Figure 1: Impact of the amount of training data on performance(accuracy over pairs on the balanced dataset). Language preprocessing always improve generalization: pre-parsing and pretrained word embeddings both have a positive impact individually, and their effects are complementary to each other.

References

[1] Kan Chen, Jiang Wang, Liang Chieh Chen, Haoyuan Gao, Wei Xu, and Ram Nevatia. Abc-cnn: An attention based convolutional neural network for visual question answering. *Computer Science*, 2015.

[2] Aiwen Jiang, Fang Wang, Fatih Porikli, and Yi Li. Compositional memory for visual question answering. *Computer Science*, 2015.

[3] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Conference on Empirical Methods in Natural Language Processing*, pages 1532–1543, 2014.

[4] Huijuan Xu and Kate Saenko. Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. In *European Conference on Computer Vision*, pages 451–466, 2016.

[5] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. Stacked attention networks for image question answering. pages 21–29, 2015.

[6] Peng Zhang, Yash Goyal, Douglas Summersstay, Dhruv Batra, and Devi Parikh. Yin and yang: Balancing and answering binary visual questions. In *Computer Vision and Pattern Recognition*, pages 5014–5022, 2016.

[7] Yuke Zhu, Oliver Groth, Michael Bernstein, and Fei Fei Li. Visual7w: Grounded question answering in images. In *Computer Vision and Pattern Recognition*, pages 4995–5004, 2016.

Figure 2: 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.* [6] 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 [6] (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.

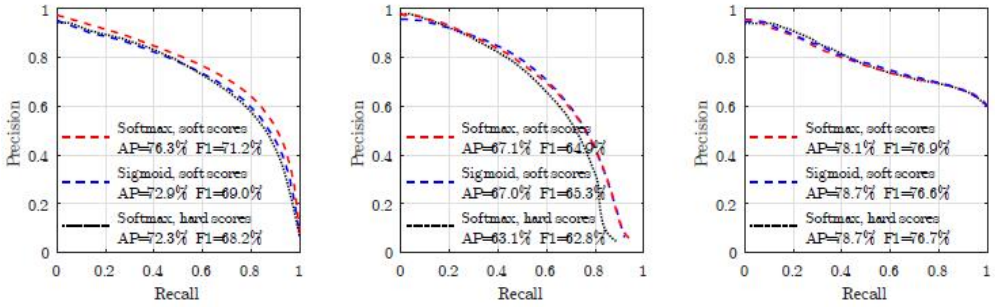

Figure 3: Precision/recall on the "abstract scenes" (left: multiple choice, middle: open-ended) and balanced datasets (right). The scores assigned by the model to predicted answers is a reliable measure of its certainty: a strict threshold (low recall) filters out incorrect answers and produces a very high precision. On the "abstract scenes" dataset (left and middle), a slight advantage is brought by training for soft target scores that capture ambiguities in the human-provided ground truth.