Image Question Answering using Convolutional Neural Network with Dynamic Parameter Prediction II

Liangjie Cao

Jun 28, 2018

1. Final datasets

The authors evaluate the proposed network on several public ImageQA benchmark datasets such as DAQUAR [3], COCOQA [4] and VQA [1]. They collected question-answer pairs from existing image datasets and most of the answers are single words or short phrases.

DAQUAR is based on NYUDv2 dataset, and pro- vides two benchmarks. DAQUAR-all consists of 6,795 and 5,673 questions for training and testing respectively, and includes 894 categories in answer. DAQUAR-reduced in- cludes only 37 answer categories for 3,876 training and 297 testing questions. Some questions in this dataset are associated with a set of multiple answers.

VQA [1], which is also based on MS COCO dataset [2], contains the largest number of questions: 248,349 for training, 121,512 for validation, and 244,302 for testing, where the testing data is split into test-dev, test-standard, test-challenge and test-reserve. Each question is associated with 10 answers annotated by different people. About 90% of answers have single words and 98% of answers do not exceed three words.

	Open-Ended				Multiple-Choice			
	All	Y/N	Num	Others	All	Y/N	Num	Others
Question	48.09	75.66	36.70	27.14	53.68	75.71	37.05	38.64
Image	28.13	64.01	00.42	03.77	30.53	69.87	00.45	03.76
Q+I	52.64	75.55	33.67	37.37	58.97	75.59	34.35	50.33
LSTM Q	48.76	78.20	35.68	26.59	54.75	78.22	36.82	38.78
LSTM Q+I	53.74	78.94	35.24	36.42	57.17	78.95	35.80	43.41
CONCAT	54.70	77.09	36.62	39.67	59.92	77.10	37.48	50.31
RAND-GRU	55.46	79.58	36.20	39.23	61.18	79.64	38.07	50.36
CNN-FIXED	56.74	80.48	37.20	40.90	61.95	80.56	38.32	51.40
DPPnet	57.22	80.71	37.24	41.69	62.48	80.79	38.94	52.16

Table 1. performances of different methods on Shanghaitech dataset

2. Evaluation Metrics

VQA dataset provides open-ended task and multiple-choice task for evaluation. For open-ended task, the answer can be any word or phrase while an answer should be chosen out of 18 candidate answers in the multiple-choice task. In both cases, answers are evaluated by accuracy reflecting human consensus. For predicted answer a_i and target answer set T_i of the i^{th} example, the accuracy is given by

$$Acc_{VQA} = \frac{1}{N}\sum_{i=1}^{N}\min\{\frac{\Sigma_{t\in T^i}\Pi[a_i = t]}{3}, 1\} \qquad (1)$$

where $\Pi[\cdot]$ denotes an indicator function. In other words, a predicted answer is regarded as a correct one if at least three annotators agree, and the score depends on the number of agreements if the predicted answer is not correct.

3. Results

The authors test three independent datasets, VQA, COCO-QA, and DAQUAR, and first present the results for VQA dataset in Table 1. The proposed Dynamic Parameter Prediction network (DPPnet) outperforms all existing methods non- trivially. We performed controlled experiments to ana- lyze the contribution of individual components in the pro- posed algorithmdynamic parameter prediction, use of pre-trained GRU and CNN fine-tuning, and trained 3 additional models, CONCAT, RAND-GRU, and CNN-FIXED. The qualitative results of the proposed algorithm are presented in Figure 1. In general, the proposed network is successful to handle various types of questions that need different levels of semantic understanding. Figure 1(a) shows that the network is able to adapt recognition tasks depending on questions. However, it often fails in the questions asking the number of occurrences since these questions involve the difficult tasks (*e.g.*, object detection) to learn only with image level annotations. On the other hand, the proposed network is effective to find the answers for the same question on different images fairly well as illustrated in Figure 1(b). Refer to our project website 2 for more detailed results.

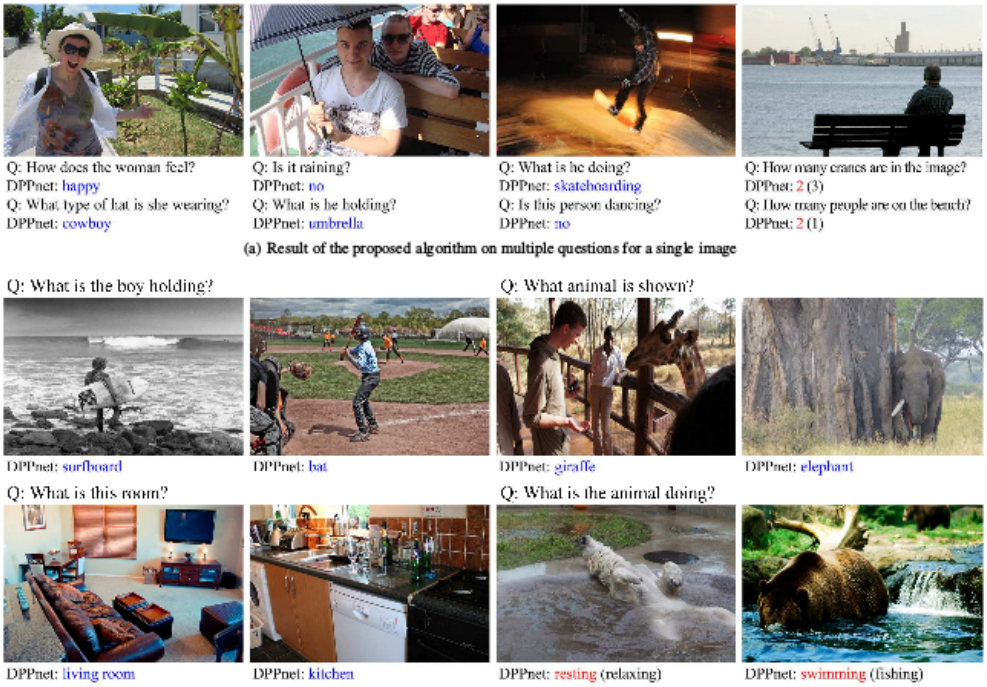

Figure 1. **Sample images and questions in VQA dataset [1]. Each question requires a different type and/or level of understanding of the corresponding input image to find correct answer. Answers in blue are correct while answers in red are incorrect. For the incorrect answers, ground-truth answers are provided within the parentheses**

References

[1] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh. Vqa: visual question answering. In *ICCV*, 2015. 1, 2

[2] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 1

[3] M. Malinowski and M. Fritz. A multi-world approach to question answering about real-world scenes based on uncertain input. In *NIPS*, 2014. 1

[4] M. Ren, R. Kiros, and R. S. Zemel. Exploring models and data for image question answering. In *NIPS*, 2015. 1