Towards Good Practices for Visual Question Answering

Zhe Wang, Xiaoyi Liu, Liangjian Chen, Limin Wang, Yu Qiao, Xiaohui Xie, and Charless Fowlkes

Abstract

Visual question answering (VQA) tasks are of significant interest due to their potential as a strong test of image understanding systems and probing the connection between language and vision. Despite much recent innovation, general VQA is far from a solved problem. In this paper, we focus on the VQA multiple-choice task, and provide some good practices for designing an effective VQA model that can capture language-vision interactions and perform joint reasoning. We explore mechanisms for incorporating part-of-speech (POS) tag guided attention, convolutional n-grams, and triplet attention interactions among the image, question and candidate answer. Our main contribution is a set of useful insights for guiding VQA system design. We evaluate our models on two popular datasets: Visual7W Telling and VQA Real Multiple Choice. Our final model achieves state-of-the-art performance of 68.0% on Visual7W Telling, and a competitive performance of 68.3% on the test-standard split of VQA Real Multiple Choice.

1. Introduction

Visual Question Answer (VQA) [1] tasks provide a natural framework for the development of techniques that jointly reason about computer vision and natural language processing [1, 10] and have attracted increasing attention. Existing VQA solutions range from symbolic approaches [7], to memory-approaches [8], to attention-based approaches [2, 4, 9]. Our approach builds on the architecture of [6] and is inspired by Jabri et al. [3], who demonstrated that simple averaging of word vectors yielded sentences that were competitive with more complex methods (e.g., LSTM).

2. Architecture

We propose a simple but effective VQA model (Pipeline in Fig 2) that achieves good performance on two popular datasets: Visual7W Telling and VQA Real Multiple-Choice. We start with the architecture in [6], which combines word features from the question and answer sentences as well as hierarchical CNN features from the input image. Our contributions are threefold: (i) To better capture the relevant semantics of questions and answers, we propose to exploit a part-of-speech (POS) tag-guided attention model that ignores less meaningful words (e.g., coordinating conjunctions such as "for", "and", "or") and places more emphasis on the important words such as nouns, verbs and adjectives. (ii) We leverage a convolutional n-gram model [4] to capture local context needed for phrase-level meaning in questions and answers. (iii) To integrate visual features (extracted from a pre-trained deep residual network (ResNet)), we introduce a triplet attention mechanism that measures compatability based an affinity matrix constructed by the inner product of vector representations of each word in the question (or answer) and each sub-region in the image. After pooling and normalization, we linearly combine the attention coefficients from questions and answers to produce a final weighting of relevant visual features.

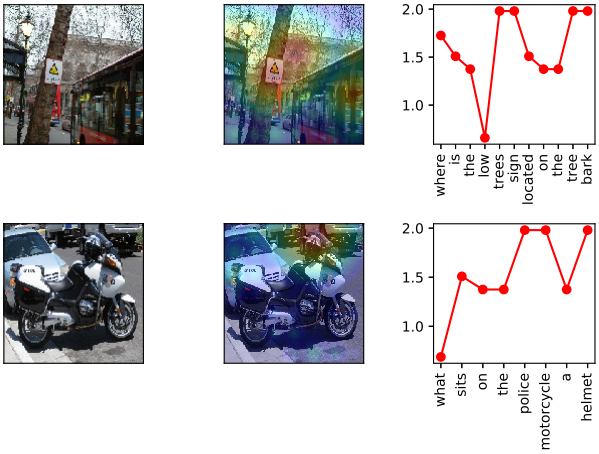

Figure 1. Illustration of triplet attention weights estimated for input image (second column) and POS tag-guided attention weights for corresponding question (third column).

3. Experiments

We compare our methods with the state-of-the-art performance in Table 3 and visualize the attention maps generated by triplet attention and POS tag in Fig 1. Our proposed methods achieve the state-of-the-art performance of 68.0% on Visual7W Telling benchmark, and competitive performance of 68.3% on the test-standard split of VQA Real Multiple Choice. We conclude that both our POS tag guided

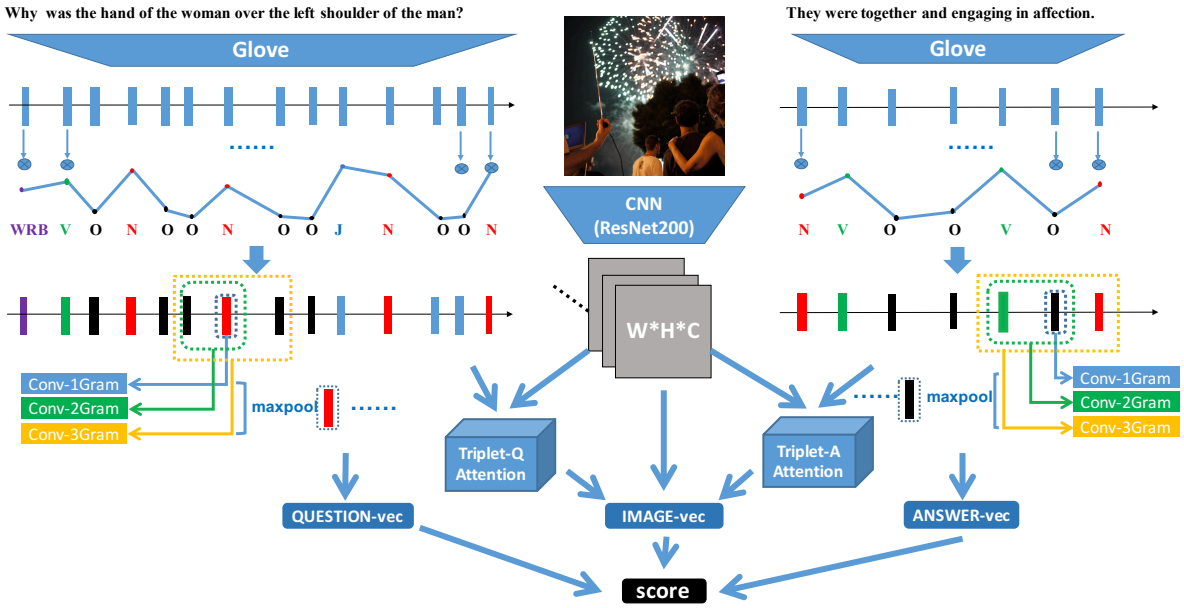

Figure 2. Illustration of our pipeline for VQA. We first extract GLOVE [5] vector representations of each word in the question and answer, which are weighted by a POS tag-guided attention for each word. We transform each sentence using a convolutional n-gram to encode contextual information and average to get QUESTION-vec. For visual features, we utilize a standard CNN model and conduct weighted pooling using triplet attention to produce IMAGE-vec. Finally we combine QUESTION-vec, IMAGE-vec and ANSWER-vec to score the quality of the proposed answer.

and triplet attention mechanisms are beneficial for VQA by helping the model focus on relevant inputs. Our approach offers some simple insights for effective practice building high performance VQA systems.

Method	V7W	VQA
Co-attention [4]	-	66.1
Attention-LSTM [10]	55.6	-
MCB [2]	62.2	65.4
Ensemble of 7 Att models [2]	-	70.1
MLP [3]	64.8	65.2
Full model (7×7 ResNet feature)	**67.3**	**68.3**
Full model (14×14 ResNet feature)	**68.0**	-

Table 1. Quantitative results on Visual7W Telling [10] and the test2015-standard split on VQA Real Multiple Choice [1]. Our full model outperforms the state-of-the-art performance by a large margin on Visual7W and obtains a competitive result to the state-of-the-art on VQA. Note that the best model in [2] uses an ensemble of 7 attention models with auxiliary training data and more word embeddings. Due to the memory constraints, we haven't yet evaluated the high-res Full Model (14×14 Resnet feature) on VQA.

References

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

[2] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. In *EMNLP*, 2016. 1, 2

[3] A. Jabri, A. Joulin, and L. van der Maaten. Revisiting visual question answering baselines. In *ECCV*, 2016. 1, 2

[4] J. Lu, J. Yang, D. Batra, and D. Parikh. Hierarchical question-image co-attention for visual question answering. In *NIPS*, 2016. 1, 2

[5] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *EMNLP*, 2014. 2

[6] D. Teney and A. van den Hengel. Zero-shot visual question answering. 2016. arXiv.1611.05546. 1

[7] Q. Wu, C. S. Peng Wang, A. Dick, and A. van den Hengel. Ask me anything: Free-form visual question answering based on knowledge from external sources. In *CVPR*, 2016. 1

[8] C. Xiong, S. Merity, and R. Socher. Dynamic memory networks for visual and textual question answering. In *ICML*, 2016. 1

[9] H. Xu and K. Saenko. Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. In *ECCV*, 2016. 1

[10] Y. Zhu, O. Groth, M. Bernstein, and L. FeiFei. Visual7W: Grounded Question Answering in Images. In *CVPR*, 2016. 1, 2