An Interpretable Multimodal Visual Question Answering System using Attention-based Weighted Contextual Features

Extended Abstract

Yu Wang
Samsung Research America
Mountain View, CA
yu.wang1@samsung.com

Yilin Shen
Samsung Research America
Mountain View, CA
yilin.shen@samsung.com

Hongxia Jin
Samsung Research America
Mountain View, CA
hongxia.jin@samsung.com

ABSTRACT

Visual question answering (VQA) is a challenging task that requires a deep understanding of language and images. Currently, most VQA algorithms focus on finding the correlations between basic question embeddings and image features by using an element-wise product or bilinear pooling between these two vectors. Some algorithms also use attention models to extract features. In this extended abstract, a novel interpretable multimodal system using attention-based weighted contextual features (MA-WCF) is proposed for VQA tasks. This multimodal system can assign adaptive weights to the features of questions and images themselves and to their contextual features based on their importance. Our new model yields state-of-the-art results on the MS COCO VQA datasets for open-ended question tasks.

ACM Reference Format:
Yu Wang, Yilin Shen, and Hongxia Jin. 2020. An Interpretable Multimodal Visual Question Answering System using Attention-based Weighted Contextual Features. In *Proc. of the 19th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2020), Y. Wang, Y. Shen, H. Jin, Auckland, New Zealand, May 2020,* IFAAMAS, 3 pages.

1 INTRODUCTION

Despite the decent results that have been obtained by using different attention mechanisms for VQA, the overall performance achieved is still not comparable to that of human beings [9]. One possible reason for this shortfall is that humans can use more contextual information in both the question and the image to infer the answer.

In the example shown in Figure 1, the orange-colored text is contextual information that should be de-emphasized by the model since the text in purple is the real question that needs to be addressed. Consequently, although the correct regions of the image are emphasized in the attention map, the model still cannot find the correct answer. To overcome these two potential difficulties, in this paper, a new interpretable multimodal structure for VQA is designed by considering the contextual information in questions and images. A weighted contextual feature (WCF) structure is also proposed to balance the essential information from the question/image and the contextual information by assigning appropriate ratios through learning. In this paper. we propose a framework in which contextual features extracted from multiple sources (image and query) are used to improve VQA performance by further considering the cross-impact of these features with different types of data. It is shown that our model achieves state-of-the-art results on two benchmark VQA datasets.

Proc. of the 19th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2020), B. An, N. Yorke-Smith, A. El Fallah Seghrouchni, G. Sukthankar (eds.), May 9–13, 2020, Auckland, New Zealand.

2 AN INTERPRETABLE VQA MULTIMODAL SYSTEM USING ATTENTION-BASED WEIGHTED CONTEXTUAL FEATURES (MA-WCF)

The system structure is as shown in Figure 2. We extract the semantic contextual features using an RNN-based encoder-decoder structure and image contextual features using an MDLSTM-based encoder-decoder structure. Specifically, the RNN structure in our system is chosen to be a bidirectional LSTM (BLSTM) structure [7]. Moreover, as demonstrated by the VQA example given in Figure 1, many instances of misinterpretation in VQA tasks can be attributed to a misunderstanding of the contextual information present in the question (which can be extremely important) or image. Inspired by these observations and model features, we propose an attention-based multimodal system that leverages contextual features of both questions and images for VQA tasks. Due to the page limitation, readers can refer to the full paper about the details of system structure [18]. Related works on multimodal VQA/QA, co-attention networks, multi-model neural networks for learning tasks can also be found in [4, 5, 8–12, 12, 14–17, 19–21].

Figure 1: An example in which an incorrect answer is obtained using a VQA model

3 EXPERIMENT

In this section, we evaluate our new model's performance on two VQA datasets: MS COCO VQA dataset (v1 [3] and v2 [6])

3.1 Dataset and Different Model Configurations

We report our evaluation results obtained using the test-standard dataset. Moreover, the results obtained on open-ended tasks (on both VQA v1 and VQA v2) are reported. The model was trained on the training and validation sets, and the results are compared with those of the current state-of-the-art models for each category and the whole dataset. We tested our MA-WCF model with several different configurations:

MA-WCF model without question contextual features: In this model configuration, we removed the question contextual features

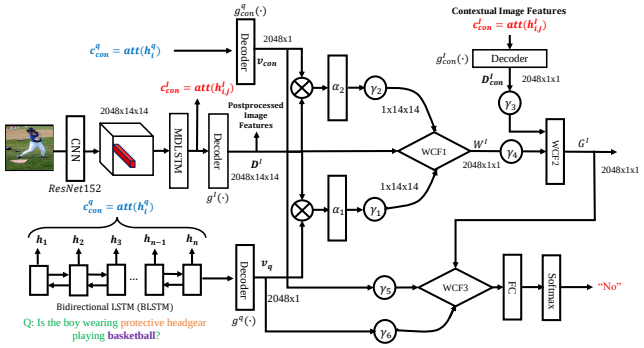

Figure 2: The general structure of the multimodal attention-based WCF (MA-WCF) model

Figure 3: Examples used for comparison between MA-WCF and state-of-the-art models

v_{con} shown in Figure 2. The entire model only contains the image contextual features D^I_{con}.

MA-WCF model without image contextual features: The second model configuration was created by removing the image contextual features from the MA-WCF model.

MA-WCF model with both question and image contextual features: This is the original model as shown in Figure 2.

3.2 Experiments

The three different configurations of our model were trained on the training and validation sets from the MS COCO VQA dataset for comparison with state-of-the-art models. The results are shown in Table 1.

3.2.1 Interpretablilty of MA-WCF. Figure 3 shows the results of applying the three different MA-WCF configurations to two different examples. In the first example, the noun "the person" is modified by the postpositional phrase "kicking the ball"; therefore, question-level contextual features must be considered for the question to be correctly understood. Otherwise, the model may not be able to focus on the correct subregions of the image since the question is not fully understood. In the second configuration, with the image-level contextual features removed, the model can locate the correct subregions of the image; however, it still cannot generate the correct answer since the contextual correlations between the masked sub-regions cannot be fully understood. Therefore, incorrect answers are generated when the image contextual features are not considered.

Table 1: Results for Open-ended Answers on the Test-standard Datasets in VQA v1 and VQA v2

	Test-standard (%)			
VQA v1	Y/N	Num.	Other	All Categories
Ensemble MCB [5]	83.2	39.5	58.0	66.5
Alpha VQA [1]	87.61	45.63	63.30	71.48
MA-WCF w/o Question Context	84.73	43.13	63.56	69.3
MA-WCF w/o Image Context	85.3	44.13	63.39	69.7
MA-WCF	**88.92**	**46.73**	**64.46**	**73.52**
VQA v2	Y/N	Num.	Other	All Categories
IL-QTA [13]	88.26	55.22	63.63	72.93
MIL@HDU [2]	90.36	59.17	**65.75**	75.23
MA-WCF w/o Question Context	89.45	55.46	62.63	73.35
MA-WCF w/o Image Context	90.35	57.86	63.94	74.98
MA-WCF	**91.45**	**59.65**	65.38	**76.33**

Similarly, in the second example, the question-level semantic contextual features help to locate the correct image subregion(s) to enable the identification of the "type of material". Then, the image contextual features further help to generate the correct answer by filtering out some noisy image information (such as the presence of 5 baseball bats) by giving them lower weights, hence generating the correct answer.

3.2.2 Experiment Results on VQA Datasets. One observation that can be drawn from Table 1 is that the performance of the model without question contextual features is far inferior to both that of the model without image contextual features and that of the model with both types of contextual features. This finding demonstrates the importance of question contextual features to our system. The MA-WCF model with both image and question contextual features outperforms the previous state-of-the-art results on each category of the test-standard sets in VQA v1 and on most categories in VQA v2. Excitingly, our model even shows better performance than ensemble/stacking-based models do. On VQA v1, the MA-WCF model outperforms the current state-of-the-art Alpha VQA model by 2.1% on the test-standard dataset. On VQA v2, the MA-WCF model outperforms the current state-of-the-art model by MIL@HDU by 1.1%.

4 CONCLUSION

In this paper, we have proposed a novel interpretable multimodal system using attention-based weighted contextual features (MA-WCF) to address the visual question answering (VQA) problem. By using adaptively weighted contextual features extracted from both questions and images, our system gains the advantageous ability to pinpoint the most important parts of both questions and images while de-emphasizing less important features. We have achieved new state-of-the-art results on the MS COCO VQA dataset for open-ended question tasks. As a relatively general technique, our MA-WCF approach can be further extended to other text- or image-related tasks, such as question answering, text summarization or visual grounding. The interpretability of the model also endows it with great potential for application to more complex VQA tasks; we are currently working on these problems and will report our progress in future works.

REFERENCES

[1] 2016. AlphaVQA. *URL: https://visualqa.org/roe.html* (2016).

[2] 2019. State-of-the-art VQA model in 2019 VQA Challenge. *URL: https://visualqa.org/roe.html* (2019).

[3] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*.

[4] Kan Chen, Jiang Wang, Liang-Chieh Chen, Haoyuan Gao, Wei Xu, and Ram Nevatia. 2015. ABC-CNN: An attention based convolutional neural network for visual question answering. *arXiv preprint arXiv:1511.05960* (2015).

[5] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. 2016. Multimodal Compact Bilinear Pooling for Visual Question Answering and Visual Grounding. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. 457–468.

[6] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

[7] Alex Graves and Jürgen Schmidhuber. 2005. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks* 18, 5-6 (2005), 602–610.

[8] Laurent Itti and Christof Koch. 2001. Computational modelling of visual attention. *Nature reviews neuroscience* 2, 3 (2001), 194.

[9] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2016. Hierarchical question-image co-attention for visual question answering. In *Advances In Neural Information Processing Systems*. 289–297.

[10] Kumpati S Narendra, Yu Wang, and Wei Chen. 2014. Stability, robustness, and performance issues in second level adaptation. In *American Control Conference (ACC), 2014*. IEEE, 2377–2382.

[11] Kumpati S Narendra, Yu Wang, and Snehasis Mukhopadhay. 2016. Fast Reinforcement Learning using multiple models. In *Decision and Control (CDC), 2016 IEEE 55th Conference on*. IEEE, 7183–7188.

[12] Kevin J Shih, Saurabh Singh, and Derek Hoiem. 2016. Where to look: Focus regions for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4613–4621.

[13] Do Tuong, Tran Huy, Do Thanh-Toan, Tjiputra Erman, and D. Tran Quang. 2019. Interaction Learning with Question-type Awareness for Visual Question Answering. In *Computer Vision and Pattern Recognition, 2019. CVPR 2019. IEEE Conference on*. IEEE.

[14] Yu Wang. 2017. A new concept using LSTM Neural Networks for dynamic system identification. In *American Control Conference (ACC), 2017*. IEEE, 5324–5329.

[15] Yu Wang, Yue Deng, Yilin Shen, and Hongxia Jin. 2020. A New Concept of Multiple Neural Networks Structure Using Convex Combination. *IEEE Transactions on Neural Networks and Learning Systems* (2020), 1–12.

[16] Yu Wang and Hongxia Jin. 2019. A Deep Reinforcement Learning based Multi-Step Coarse to Fine Question Answering (MSCQA) System. In *AAAI, 2019*. AAAI.

[17] Yu Wang, Abhishek Patel, Yilin Shen, and Hongxia Jin. 2018. A Deep Reinforcement Learning Based Multimodal Coaching Model (DCM) for Slot Filling in Spoken Language Understanding (SLU). *Proc. Interspeech 2018* (2018), 3444–3448.

[18] Yu Wang, Abhishek Patel, Yilin Shen, Hongxia Jin, and Larry Heck. 2018. An Interpretable (Conversational) VQA model using Attention based Weighted Contextual Features. *The 2nd Conversational AI Worshop, NeurIPS* (2018).

[19] Yu Wang, Yilin Shen, and Hongxia Jin. 2018. A Bi-model based RNN Semantic Frame Parsing Model for Intent Detection and Slot Filling. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2*, Vol. 2. 309–314.

[20] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning*. 2048–2057.

[21] Zhou Yu, Jun Yu, Chenchao Xiang, Jianping Fan, and Dacheng Tao. 2018. Beyond Bilinear: Generalized Multimodal Factorized High-Order Pooling for Visual Question Answering. *IEEE Transactions on Neural Networks and Learning Systems* (2018).