MoQA – A Multi-modal Question Answering Architecture

Monica Haurilet(✉), Ziad Al-Halah, and Rainer Stiefelhagen

Karlsruhe Institute of Technology, 76131 Karlsruhe, Germany
{haurilet,ziad.al-halah,rainer.stiefelhagen}@kit.edu

Abstract. Multi-Modal Machine Comprehension (M3C) deals with extracting knowledge from multiple modalities such as figures, diagrams and text. Particularly, Textbook Question Answering (TQA) focuses on questions based on the school curricula, where the text and diagrams are extracted from textbooks. A subset of questions cannot be answered solely based on diagrams, but requires external knowledge of the surrounding text. In this work, we propose a novel deep model that is able to handle different knowledge modalities in the context of the question answering task. We compare three different information representations encountered in TQA: a visual representation learned from images, a graph representation of diagrams and a language-based representation learned from accompanying text. We evaluate our model on the TQA dataset that contains text and diagrams from the sixth grade material. Even though our model obtains competing results compared to state-of-the-art, we still witness a significant gap in performance compared to humans. We discuss in this work the shortcomings of the model and show the reason behind the large gap to human performance, by exploring the distribution of the multiple classes of mistakes that the model makes.

1 Introduction

Answering questions based on natural images has received growing attention in the Computer Vision community for several years [14,15,18,20]. While at a very early age humans can answer basic question about their environment, we start to analyze and understand graphics at later time. In school years, children learn to analyze and understand complex illustrations, and are capable to extract important information and answer difficult questions about them.

The type and style of these illustrations have many different forms in terms of colors, structure types and complexity. While some illustrations in textbooks are easy, like simple drawings, we see in later school years more difficult types of figures like diagrams, plots and tables. Diagrams are especially challenging since we have different type of nodes like drawings, text, natural images etc.

Electronic supplementary material The online version of this chapter (https://doi.org/10.1007/978-3-030-11018-5_9) contains supplementary material, which is available to authorized users.

L. Leal-Taixé and S. Roth (Eds.): ECCV 2018 Workshops, LNCS 11132, pp. 106–113, 2019.
https://doi.org/10.1007/978-3-030-11018-5_9

Furthermore, we have various relationship types between nodes, e.g. textual description and nodes, and textual description and edges. We also have directed relations, usually represented with edges marked with an arrow sign, while some relations are not explicitly marked (see an example in Fig. 1).

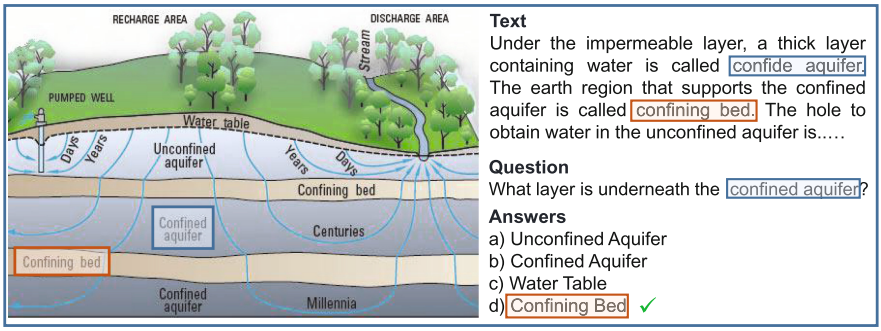


Fig. 1. Example diagram with corresponding question from the TQA dataset.

In this work, we compare different knowledge representations for our model: (1) the text-based model, where we use the surrounding text for answering the questions, (2) the image-based model uses the surrounding image by extracting the features of a pre-trained CNN and (3) the graph-based representation embeds the diagram as a graph where the nodes consist of the detected text and its location. We investigate the predictions of our model to find the reasons for the large gap to human performance, by analyzing a subset of incorrectly answered questions.

2 Related Work

VQA. Various topics join language with natural images like image captioning [19] and text-based image retrieval [5]. Visual Question Answering (VQA) obtains both an image and a question and produces an answer. In spite of a multitude of available datasets [1,11,17,21] and published models [14,15,18,20], VQA remains a hard task and the recognition rate remains far from human performance. Most VQA models do not consider the structure of the object instances in natural images, as most questions target single objects.

Textbook QA. In comparison to VQA, the Textbook Question Answering (TQA) task deals with different types of images: textbook illustrations like tables [4], plots [6,16] and diagrams [7,8,13]. Such figures are more structured than natural images, as the relations between the components have a higher importance for answering the questions. While tables are structural elements combining and ordering their entries - mostly text - in a specific way, diagrams

can have much more types of relations (e.g. location-based, 'eating' relation between animals). Furthermore, the nodes have various types like text, natural objects and drawings (as in Fig. 1). This makes the task of diagram question answering difficult to solve, as we see in the diagram QA models presented in [7,13]. Finally, the TQA dataset [8] contains questions about both diagrams and text. This makes the VQA task especially challenging, as the model has to decide from where to extract the relevant information to answer the question.

3 Method

We define the multi-modal comprehension task in the context of question answering. That is, given knowledge K from a textbook lesson (a set of sentences S, a set of nodes N or a global image representation I) and an embedding of the question Q, choose the correct answer from a set of answers $A = \{A_i\}$.

Approach. Since in case of the text-based and graph-based networks we receive a large amount of data, we filter out unrelated sentences and nodes. Our approach relies on the basic intuition that for each question Q, there is a set of supporting sentences/nodes $K^Q = \{K_j\}$ in K that would help in verifying the correctness of each (Q, A_i) pair. The text-based approach consists of two main steps: (1) Selecting k supporting sentences/nodes from K for a given question Q. (2) Based on (Q, K^Q), verify the correctness of each answer $A_i \in A$.

Supporting Nodes and Supporting Sentences. To select the set of supporting knowledge for a certain question Q, we measure the similarity of all K_j in the provided text and diagram to the question in an embedding space. That is, for each $K_j \in K$ we calculate $f_s(f_v(Q), f_v(K_j))$, where f_v is a sentence encoding function (e.g. recurrent neural network) and f_s is a similarity metric (e.g. cosine similarity). Then, the top k most similar knowledge information are selected to be in K^Q. Given the supporting sentences/nodes and the question, we use the deep neural network presented in Fig. 2 to verify each of the available answers.

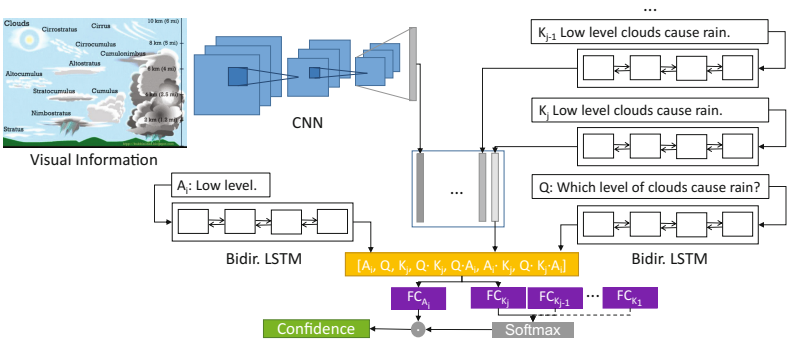


Fig. 2. Architecture of the image+text deep neural model

Neural Network. We start by encoding the triplet $(fc(Q), fc(K_j), fc(A_i))$ separately using fully connected layers fc. Then, the new embeddings are concatenated with the pairwise and triple-wise similarity of embeddings using element-wise multiplication, for each answer and knowledge: $mapping(A_i, K_j) = [K_j, Q, A_i, K_j \cdot Q, K_j \cdot A_i, Q \cdot A_i, Q \cdot A_i \cdot K_j]$.

Next, we split the output of this layer into two streams. The first stream captures the confidence of the answer A_i to be the correct one, while the second stream weights the model confidence in the knowledge subset K_j for being suitable to verify (Q, A_i). We calculate this confidence using an attention module using a softmax layer. The input of the softmax layer is the output of the fc_s layer for each of $K_j \in K^Q$ encoded by the same neural model. Finally, the two streams are fused using element-wise multiplication. In testing, the answer with highest confidence is selected as the correct one.

Text-Based Network. Our text-only model uses solely the surrounding text to generate the answers to the question ($K = T$). In case of the Text+Image Network, we include the visual information as another vector in the supporting sentences set: $K = T \cup I$ (see Fig. 2).

Graph-Based Network. In a similar manner, as we have a high number of nodes in the diagrams, we select a set of supporting nodes based on the question. In this case, k nodes are selected that have the highest similarity to the question, where the similarity is $f_s(f_v(Q), f_v(N_i))$. However, the difference to the text-based model lies in the representation of the nodes for the neural network, as instead of using the representation of the supporting nodes, we use an edge representation. For each node N_j in the set of k supporting nodes, we use the source node N_j concatenated with the nearest node, i.e. $[N_j, N_{nearest_j}]$, as the knowledge representation K_j.

Graph Baseline. In the first step in the baseline model, we take the top-1 supporting node and calculate its nearest neighbor. The answer is chosen based on the similarity of the nearest node and the answers.

Image-Based Network. The image-based network receives in addition to the question and answer pair, solely a global representation of the diagram I using features extracted from a pre-trained CNN.

4 Evaluation

Dataset. TQA [8] is a dataset for multi-modal machine comprehension, which contains lessons and exercises from the sixth grade curricula. In total, the dataset contains 1 K lessons from Life Science, Earth Science and Physical Science textbooks with 26 K corresponding multi-modal questions. Around half of the questions have corresponding text (*text questions*), while the other ones also have an accompanying diagram (*diagram questions*). The text questions are further split into true/false questions, where the only possible answers are true and false, and multiple-choice, where we can have different answers.

Parameters. As a similarity metric (f_s) for selecting the supporting sentences we use the cosine similarity. We empirically set $k = 4$ for the multiple choice model and $k = 2$ for the true/false model. A sentence embedding (f_v), if not otherwise specified we use the SkipThought [10] encoding, however we also provide results for InferSent [2]. We represent the images using a Residual Network [3] trained on ImageNet [12]. Our model is trained using Adam [9] for stochastic optimization with an initial learning rate of 0.01.

Comparison to State-of-the-Art. We are able to outperform state-of-the-art in the true/false questions and obtain competitive results in the entire text-only task (see Table 1). In the case of diagrams, our model has a lower performance, but is able to outperform complex models such as BiDAF and Memory Networks. We notice that InferSent obtains a higher accuracy in the true/false questions than SkipThought. InferSent was trained in a supervised setting in a similar scenario as the true/false task, namely, to find the relation between a pair of sentences (i.e. no relation, contradiction and entailment).

Table 1. Validation accuracy of our model compared to state-of-the-art (left) and comparison of different variations of our model (right).

	T/F	MC	Text	Diag.
Random	50.1	22.9	33.6	25.0
MemN + VQA [8]	50.5	31.1	38.7	31.8
MemN + VQA + HT [13]	50.3	28.1	36.9	29.8
MemN + DPG [8]	50.5	30.1	38.7	32.8
BiDAF + DPG [8]	50.4	30.5	38.7	32.7
Challenge	–	–	45.6	35.9
IGMN [13]	**57.4**	**40.0**	**46.9**	**36.4**
Ours [InferSent]	61.9	36.2	46.4	33.4
Ours [SkipThought]	60.2	36.4	45.6	34.0

Modality	#S	#N	Diag.
Image-only	–	–	33.2
Text-only	3	–	33.8
	4	–	33.9
	5	–	33.4
Graph-baseline	–	1	29.2
	–	2	28.3
	–	4	25.8
Graph-only	–	4	33.3
Text+Image	4	–	34.0

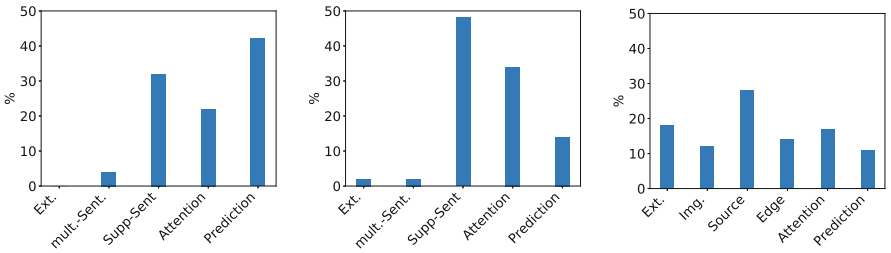
Different Knowledge Representations. In Table 1 (right) we show the performance of the model for the three different knowledge modalities and varying number of supporting sentences S and nodes N. The image-only model obtained the worst accuracy, which however, can be explained with the use of a CNN pretrained on natural images and not diagrams. Furthermore, we note that the text may play a significant role for many questions, which is not taken into account in this approach.

5 In-Depth Analysis

In this section we explore the properties of our model and attempt to find the cause behind the existing gap between the model and human performance.

Text-Based Task. To have a better overview of the common problems, we categorize them into the following groups: (1) necessity of external knowledge to answer the question (*ext.*), (2) the required information spreads over more than one sentence (*mult.-Sent.*), (3) the supporting sentences selected by our model do not contain the correct one (*Supp.-Sent.*), (4) the attention module failed to attend to the correct sentence (*Attention*), and finally (5) the prediction module was not able to provide the correct answer, even though all other modules were correct (*Prediction*).

We show in Fig. 3, the distribution of the problem types for true/false and multiple choice questions for 100 randomly selected questions in the dataset. For the true/false case, most of our mistakes are due to the prediction module, followed by the supporting sentence and the attention module. Deciding if two sentences are contradictory or have the same statement is a hard task, especially when a sentence consists of multiple statements. Furthermore, finding the correct supporting sentence is the reason for around 30% of the mistakes of our T/F model, which is less than in the case of multiple choice. This is surprising as the the true/false models have two supporting sentences and thus the probability of the sentence being in the set is lower compared to multiple-choice case.



(a) T/F Questions (b) MC Questions (c) Diag. Questions

Fig. 3. Distribution of the problems of the model in the TQA task.

Diagram-Based Task. For the diagram questions, we additionally include the image information *Img.* that shows if visual information is necessary to answer the question. Furthermore, the *Source* shows if the supporting source nodes were correctly selected and the *Edge* shows if the target node is not the one that should be used to answer the question. We see that the model has the most difficulties selecting the source nodes, similar to the text-based questions where selecting the supporting sentences causes many mistakes. Extending the model with more nodes may be beneficial for this problem but leads to overfitting (see Table 1). Including visual information (Img.) has the potential to increase performance, however to attend to parts of the image without supervision and without a higher amount of data would probably lead also to overfitting. Overall, our text-based

model has shown very strong performance on the Diagram Task. As 20% of the mistakes are caused by the absence of external knowledge (e.g. surrounding text), we believe that including this information as a further knowledge source would lead to a significant improvement.

6 Conclusion

In this work we introduced a novel neural architecture for multi-modal question answering in the multiple choice setup. We compare the network for different knowledge modalities: text-, image- and graph-based, and show that the text-based model has the best performance in all tasks. Furthermore, we analyze the mistakes our model makes and show the difficulties that our model encountered.

References

1. Antol, S., et al.: Vqa: Visual question answering. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2425–2433 (2015)
2. Conneau, A., Kiela, D., Schwenk, H., Barrault, L., Bordes, A.: Supervised learning of universal sentence representations from natural language inference data. In: Conference on Empirical Methods in Natural Language Processing (EMNLP) (2017)
3. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778 (2016)
4. Jauhar, S.K., Turney, P., Hovy, E.: Tables as semi-structured knowledge for question answering. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), vol. 1, pp. 474–483 (2016)
5. Johnson, J., et al.: Image retrieval using scene graphs. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3668–3678 (2015)
6. Kahou, S.E., Atkinson, A., Michalski, V., Kádár, Á., Trischler, A., Bengio, Y.: Figureqa: an annotated figure dataset for visual reasoning. arXiv preprint arXiv:1710.07300 (2017)
7. Kembhavi, A., Salvato, M., Kolve, E., Seo, M., Hajishirzi, H., Farhadi, A.: A diagram is worth a dozen images. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016, Part IV. LNCS, vol. 9908, pp. 235–251. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46493-0_15
8. Kembhavi, A., Seo, M., Schwenk, D., Choi, J., Farhadi, A., Hajishirzi, H.: Are you smarter than a sixth grader? textbook question answering for multimodal machine comprehension. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
9. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. In: International Conference for Learning Representations (2014)
10. Kiros, R., et al.: Skip-thought vectors. In: Advances in Neural Information Processing Systems, pp. 3294–3302 (2015)
11. Krishna, R., et al.: Visual genome: connecting language and vision using crowdsourced dense image annotations. Int. J. Comput. Vis. **123**(1), 32–73 (2017)

12. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems, pp. 1097–1105 (2012)
13. Li, J., Su, H., Zhu, J., Wang, S., Zhang, B.: Textbook question answering under instructor guidance with memory networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3655–3663 (2018)
14. Lu, J., Yang, J., Batra, D., Parikh, D.: Hierarchical question-image co-attention for visual question answering. In: Advances In Neural Information Processing Systems, pp. 289–297 (2016)
15. Malinowski, M., Rohrbach, M., Fritz, M.: Ask your neurons: a neural-based approach to answering questions about images. In: Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), pp. 1–9. IEEE Computer Society (2015)
16. Reddy, R., Ramesh, R., Deshpande, A., Khapra, M.M.: A question-answering framework for plots using deep learning. arXiv preprint arXiv:1806.04655 (2018)
17. Ren, M., Kiros, R., Zemel, R.: Exploring models and data for image question answering. In: Advances in Neural Information Processing Systems, pp. 2953–2961 (2015)
18. Xu, H., Saenko, K.: Ask, attend and answer: exploring question-guided spatial attention for visual question answering. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016, Part VII. LNCS, vol. 9911, pp. 451–466. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46478-7_28
19. Xu, K., et al.: Show, attend and tell: neural image caption generation with visual attention. In: International Conference on Machine Learning, pp. 2048–2057 (2015)
20. Yang, Z., He, X., Gao, J., Deng, L., Smola, A.: Stacked attention networks for image question answering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 21–29 (2016)
21. Zhu, Y., Groth, O., Bernstein, M., Fei-Fei, L.: Visual7w: grounded question answering in images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4995–5004 (2016)