INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Enhancing Visual Question Answering via Deconstructing Questions and Explicating Answers

Feilong Chen[1,2], *Minglun Han*[1,3], *Jing Shi*[1*], *Shuang Xu*[1], *Bo Xu*[1,2,3]

[1]Institute of Automation, Chinese Academy of Sciences, Beijing, China
[2]School of Future Technology, University of Chinese Academy of Sciences, Beijing, China
[3]School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China
{chenfeilong2018, hanminglun2018, shijing2014, shuang.xu, xubo}@ia.ac.cn

Abstract

A compositional question refers to a question that involves multiple visual objects, as well as their attributes and relationships, which requires compositional reasoning to answer. Existing VQA models can well answer a compositional question, but few works can give the reasoning process and explain why this answer is given. In this paper, we propose a novel model (DEEX) to enhance visual question answering via **DE**constructing questions and **EX**plicating answers when answering compositional questions. Specifically, DEEX aims to accomplish three sub-tasks: (1) Compositional Question Answering (CQA), (2) Question Deconstructing (QD), and (3) Answer Explicating (AE). We utilize prompt-based multi-task learning to train the proposed DEEX to be able to answer questions and give explanations simultaneously. Experimental results on the GQA dataset demonstrate our method's effectiveness, which can enhance visual question answering by giving corresponding reasoning processes and explanations.

Index Terms: Visual question answering, compositional reasoning, deconstructing question

1. Introduction

Compositional visual question answering (VQA) aims to give an answer to a compositional question about an image. Compositional questions refers to questions that involves multiple visual objects, as well as their attributes and relationships. To answer compositional VQA well, VQA models are required to comprehensively understand the multi-modal inputs and perform compositional relational reasoning.

An important line of research to tackle compositional visual question answering is modular methods. Modular methods [1, 2, 3, 4, 5] execute several modules for reasoning which are assembled according to an input question. MAC [6] decomposes questions into their linguistic substructures and uses these structures to instantiate modular networks for reasoning dynamically. MMN [5] translates the question into programs, instantiates it into operation-specific modules, and executes the modules for reasoning. On the other hand, holistic methods [7, 6, 8, 9] use a single model for different inputs to achieve reasoning. ReaCon [10] integrates the sub-questions into the reasoning process for a compositional question like a dialogue task. MDETR [11], 12-in-1 [12], VLT5 [13] utilize vision-language pretraining [14, 15] for enhancing the understanding of multi-modality [16, 17]. In addition, although there are some works [18, 19, 20, 21, 22, 23] to explore self-explaining neural networks on NLP tasks and image captioning, we focus on constraining and improving pre-trained vision-language models for reasoning in compositional VQA.

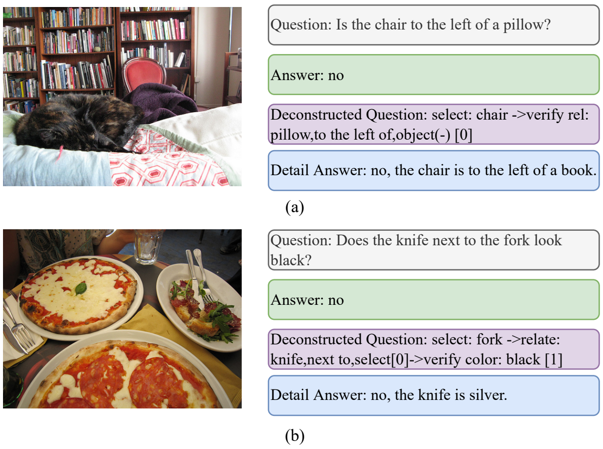

Question: Is the chair to the left of a pillow?
Answer: no
Deconstructed Question: select: chair ->verify rel: pillow,to the left of,object(-) [0]
Detail Answer: no, the chair is to the left of a book.
(a)
Question: Does the knife next to the fork look black?
Answer: no
Deconstructed Question: select: fork ->relate: knife,next to,select[0]->verify color: black [1]
Detail Answer: no, the knife is silver.
(b)

Figure 1: *A motivating example of enhancing visual question answering via deconstructing questions and explicating answers.*

As shown in Figure 1, we aim to deconstruct questions and explicate answers when answering compositional questions. We believe this is able to enhance VQA models and give explanations to some extent. As shown in Figure 1 (b), the reasoning process is "select forks", "judge the relation" and "verify color". The detailed answer needs to describe the real color of the knife, which can be regarded as an explanation.

Based on the assumption that the model can be enhanced by predicting its logical flow of reasoning and detailed answers, we propose a novel model (DEEX) to enhance visual question answering via deconstructing questions and explicating answers when answering compositional questions. Specifically, DEEX aims to accomplish three sub-tasks: (1) Compositional Question Answering (CQA), (2) Question Deconstructing (QD) and (3) Answer Explicating (AE). We utilize prompt-based multi-task learning to train the proposed DEEX with shared modules to be able to answer questions and give explanations simultaneously. In addition, DEEX maintains the consistency of CQA and AE to avoid error prediction. We conduct experiments on the compositional VQA dataset, GQA [24], to verify our method. The contributions of this paper are two-fold:

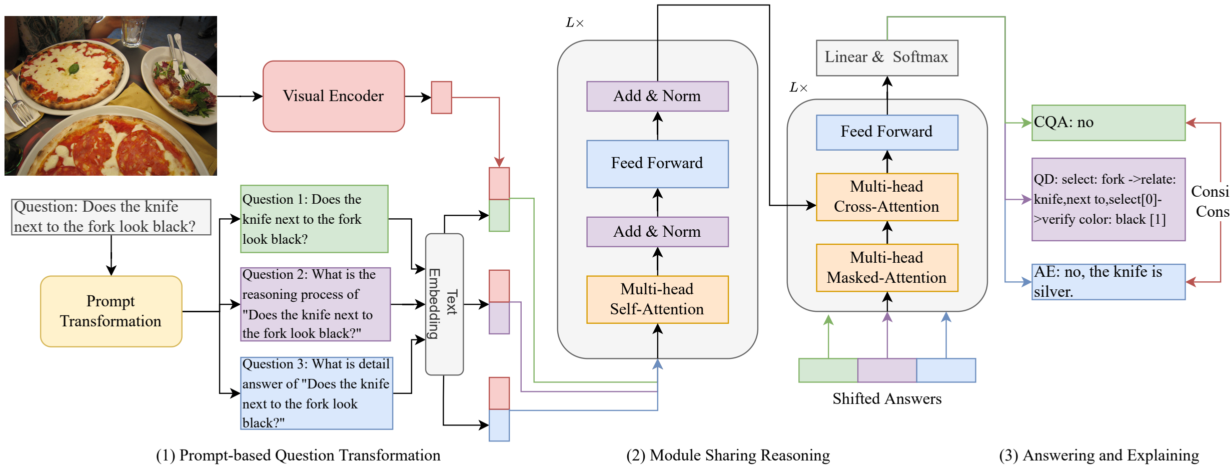
1. We propose prompt-based multi-task learning and maintain consistency constraints to enhance visual question answering via deconstructing questions and explicating answers.

2. We explore how language constraints (Structured questions and detailed answers) affect the performance of Vision-Language pre-trained VQA models

3. Experimental results show that our approach improves

[*] Corresponding author.

10.21437/Interspeech.2023-539


Visual Encoder
Question: Does the knife next to the fork look black?
Prompt Transformation
Question 1: Does the knife next to the fork look black?
Question 2: What is the reasoning process of "Does the knife next to the fork look black?"
Question 3: What is detail answer of "Does the knife next to the fork look black?"
Text Embedding
L×
Add & Norm
Feed Forward
Add & Norm
Multi-head Self-Attention
Linear & Softmax
L×
Feed Forward
Multi-head Cross-Attention
Multi-head Masked-Attention
Shifted Answers
CQA: no
QD: select: fork ->relate: knife,next to,select[0]->verify color: black [1]
AE: no, the knife is silver.
Consistency Constraints

(1) Prompt-based Question Transformation (2) Module Sharing Reasoning (3) Answering and Explaining

Figure 2: *The framework of DEEX. DEEX utilizes prompt-based multi-task learning and maintains consistency constraints to enhance visual question answering. We feed these three questions with the image to the model separately, but add their losses for optimization. We omit some modules and connections in Prompt-based Question Transformation and Module Sharing Reasoning for simplicity.*

VQA models and gives reasoning processes and explanations when answering compositional questions.

This paper gives several insights as follows: (1) We can utilize structured questions to improve models' performance, which can be obtained by using NLP tools given textual questions. (2) Informative answers can be also used to improve models' performance. (3) Language constraints benefit Vision-Language pre-trained VQA models.

2. Approach

2.1. Overview

The compositional VQA task is to answer a natural language compositional question Q about an Image I. In this paper, in addition to giving the brief answer A to the question Q, our model also needs to give the deconstructed question Q_D and the detailed answer A_D.

As shown in Figure 2, DEEX consists of three main parts: (1) Prompt-based Question Transformation transforms the original question into three questions to accomplish the proposed three sub-tasks. (2) Module Sharing Reasoning takes image features and three question features as inputs, concatenates image features with each question feature, and performs reasoning using the shared modules. (3) Answering and Explaining give the answers to the three questions and maintain the consistency of CQA and AE to avoid error prediction.

2.2. Feature Encoding

2.2.1. *Image Features*

Given an Image I, we represent it into region features with k regions using Faster RCNN [25] trained on Visual Genome [26] for object and attribute classification [27]. We obtain each region features v_i by summing appearance features, spatial features, and region id embeddings. Appearance features can be extracted directly from Faster RCNN. We encode the bounding box coordinates with a linear layer to obtain spatial features. Moreover, we give each region an id ($0 \sim k$), and encode it with learned embeddings [28]. Thus, we obtain the image features $\mathbf{v} = \{v_0, v_1, \ldots, v_k\}$, where v_0 denotes the global representation of the whole image.

2.2.2. *Question Features*

Given a compositional question Q, we first transform the original question Q into another two questions Q_{QD} and Q_{AE} with prompts "What is the reasoning process of " and "What is detail answer of ", respectively [1]. Thus we obtain three questions: the original question Q_{CQA}, the prompt-based transformed question Q_{QD}, and Q_{AE}. We tokenize the input question and encode it as learned question features $\mathbf{q} = \{q_1, q_2, \ldots, q_{|q|}\}$ following T5 [13]. We also add learned positional embeddings to the question features.

2.3. Module Sharing Reasoning

We here utilize shared modules to accomplish the three-sub tasks because we have used prompts to distinguish different questions. Given the question features $\mathbf{q}$ and the image features $\mathbf{v}$, we obtain the input embedding $\mathbf{x} = \mathbf{W_x}[\mathbf{q}; \mathbf{v}]$ by concatenating the two features and projecting the concatenated feature to a common space, where $\mathbf{W_x}$ is a learnable matrix. Note that we use question features $\mathbf{q}$ to denote each of three questions features $\mathbf{q}_{CQA}$, $\mathbf{q}_{QD}$, $\mathbf{q}_{AE}$ for simplicity.

We use transformer encoder-decoder architecture to encode visual and text inputs and generate target text. Our bidirectional multimodal encoder is a stack of L transformer blocks, which consist of a self-attention layer and a fully-connected layer with residual connections with normalization layers as shown in Figure 2. The encoder takes the concatenated embeddings $\mathbf{x}$ and outputs their contextualized joint representations $\mathbf{H}$ as follows:

$$\mathbf{Q} = \mathbf{H}^{l-1}\mathbf{W}_l^Q, \mathbf{K} = \mathbf{H}^{l-1}\mathbf{W}_l^K, \mathbf{V} = \mathbf{H}^{l-1}\mathbf{W}_l^V, \tag{1}$$

$$\mathbf{M}_{ij} = \begin{cases} 0, & \text{allow to attend,} \\ -\infty, & \text{prevent from attending,} \end{cases} \tag{2}$$

$$\mathbf{A}_l = \text{softmax}(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}} + \mathbf{M})\mathbf{V}, \tag{3}$$

where $\mathbf{W}_l^Q, \mathbf{W}_l^K, \mathbf{W}_l^V \in \mathbb{R}^{d_h \times d_k}$ are learnable weights for computing the queries, keys, and values respectively, and $\mathbf{M} \in \mathbb{R}^{|\mathbf{x}| \times |\mathbf{x}|}$ is the self-attention mask that determines whether tokens from two sources can attend each other, and $\mathbf{H}^0 = \mathbf{x} =$

[1] Note that since we use simple prompts, which is likely that engineering in text prompts [29]. As this is not the focus of this paper, we leave it as future works.

$[x_1, ..., x_{|\mathbf{x}|}]$. Then $\mathbf{A}_l$ is passed into a feedforward layer to compute $\mathbf{H}^l$ for the next layer:

$$\mathbf{H}^l = \mathrm{FFN}(\mathbf{A}_l) \tag{4}$$

We omit the operation for layer normalization.

Our decoder is another stack of L transformer blocks similar to the multimodal encoder, where each block has an additional cross-attention layer to attend to the multimodal input representation and then predicts the probability of target text tokens, as shown in Figure 2. We get three answers to the three questions by performing three times inference, respectively.

2.4. Answering and Explaining

With prompting of different prompts, DEEX iteratively attends to previously generated tokens and the encoder outputs, then predicts the probability of future text tokens.

To avoid error prediction, we maintain the consistency of answers for CQA and AE. Given the predicted answer features $\hat{y}_{CQA}$ and $\hat{y}_{AE}$ of CQA and AE, which are the average of the hidden states of the decoder's last layer. The consistency constraint enforces the semantic similarity between $\hat{y}_{CQA}$ and $\hat{y}_{AE}$ as:

$$\mathcal{L}_{cons} = 1 - \frac{\hat{y}_{CQA} \cdot \hat{y}_{AE}}{|\hat{y}_{CQA}||\hat{y}_{AE}|} \tag{5}$$

DEEX generates corresponding answers by minimizing the negative log-likelihood of label target Y_{CQA}, Y_{QD}, Y_{AE} tokens given corresponding input text Q_{CQA}, Q_{QD}, Q_{AE} and the image I, respectively:

$$\mathcal{L}_{CQA} = -\sum_{j=1}^{|Y_{CQA}|} \mathrm{log}P(Y_{CQA_j}|Y_{CQA_{<j}}, Q_{CQA}, I), \tag{6}$$

$$\mathcal{L}_{QD} = -\sum_{j=1}^{|Y_{QD}|} \mathrm{log}P(Y_{QD_j}|Y_{QD_{<j}}, Q_{QD}, I), \tag{7}$$

$$\mathcal{L}_{AE} = -\sum_{j=1}^{|Y_{AE}|} \mathrm{log}P(Y_{AE_j}|Y_{AE_{<j}}, Q_{AE}, I). \tag{8}$$

We use prompt-based multi-task learning to train DEEX and minimize the total loss $\mathcal{L}$ as follows:

$$\mathcal{L} = \mathcal{L}_{CQA} + \mathcal{L}_{QD} + \mathcal{L}_{AE} + \lambda_{cons}\mathcal{L}_{cons}, \tag{9}$$

Considering that at the beginning of the training stage, the predictions of the model may not be accurate enough for consistency constraint to be effective, we first set λ_{cons} in Eq. 9 as 0 and train the model for several epochs, and then train it with the full objective.

3. Experiments

3.1. Dataset

We evaluate our model on the GQA dataset [24]. GQA is a large-scale visual question answering dataset with real images from the Visual Genome dataset [26]. We train our model using the train split in GQA balanced version with 943,000 questions. Following [10], we test our model on the validation split and the test split of the GQA dataset, respectively. We report the accuracy of question answering following previous works [1]. The GQA dataset provides corresponding images, questions, answers, structured questions, and detailed answers. So we can use these annotations for training our model via deconstructing questions and explicating answers.

Table 1: *Results of our method and the state-of-the-art on the test split of the GQA dataset. Our approach improves the strong baseline significantly. "V" denotes the image. "L" denotes the question. "Program" denotes programs building by a set of syntax rules.*

Model	Required Inputs	ACC
Bottom-Up [27]	V+L	49.74
MAC [6]	V+L	54.06
NMN [1]	V+L+Program	55.70
BAN [31]	V+L	57.10
GRN [32]	V+L	57.04
LCGN [7]	V+L	57.07
RPR [8]	V+L	59.43
ReaCon [10]	V+L	59.58
LXMERT [33]	V+L	60.33
12-in-1 [12]	V+L	60.65
MMN [5]	V+L+Program	60.83
MDETR [11]	V+L	61.99
VLT5 [13]	V+L	60.20
DEEX (Ours)	V+L	**63.05**

Table 2: *Results of our method and the state-of-the-art on the validation split of the GQA dataset. "DA" means the data augmentation.*

Model	Required Inputs	ACC
Language-only [10]	L	43.86
Visual-only [10]	V	56.63
MAC [6]	V+L	62.08
MAC+DA [6]	V+L	61.04
LCGN [7]	V+L	64.16
LCGN+DA [7]	V+L	64.14
MMN [5]	V+L+Program	65.05
MMN+DA [5]	V+L+Program	65.65
ReaCon [10]	V+L	66.26
DEEX (Ours)	V+L	**67.86**

3.2. Implementation

We initialize our shared reasoning encoder and decoder with pre-trained VLT5 [13], which consists of 12 transformer blocks, each with 12 attention heads and hidden state dimensions of 768 for the encoder and decoder, respectively. We use Adam [30] with an initial learning rate of $5e-5$ and a batch size of 48 to train our model. A linear learning rate decay schedule with a warmup of 0.1 is employed. We train our model for 15 epochs on a cluster of 4 A100 GPUs with 40G memory. We first train the model for 5 epochs without the consistency constraint. The hyper-parameter λ_{cons} is set as 1.0. Note that modern approaches are discriminative models, where they tackle the VQA task as multi-label classification over a predefined set of answer candidates. We tackle the VQA task as a generative task for prompt-based multi-task learning.

3.3. Baselines

We compare our method with the several state-of-the-art compositional VQA models: (1) Modular method, NMN [1], MMN [5]. (2) Holistic methods: Bottom-Up [27], MAC [6], BAN [31], GRN [32], LCGN [7], PRP [8], LXMERT [33], 12-in-1 [12], MDETR [11]. All the methods mentioned above are discriminative models. We also compare our method with the generative model VLT5 [13].

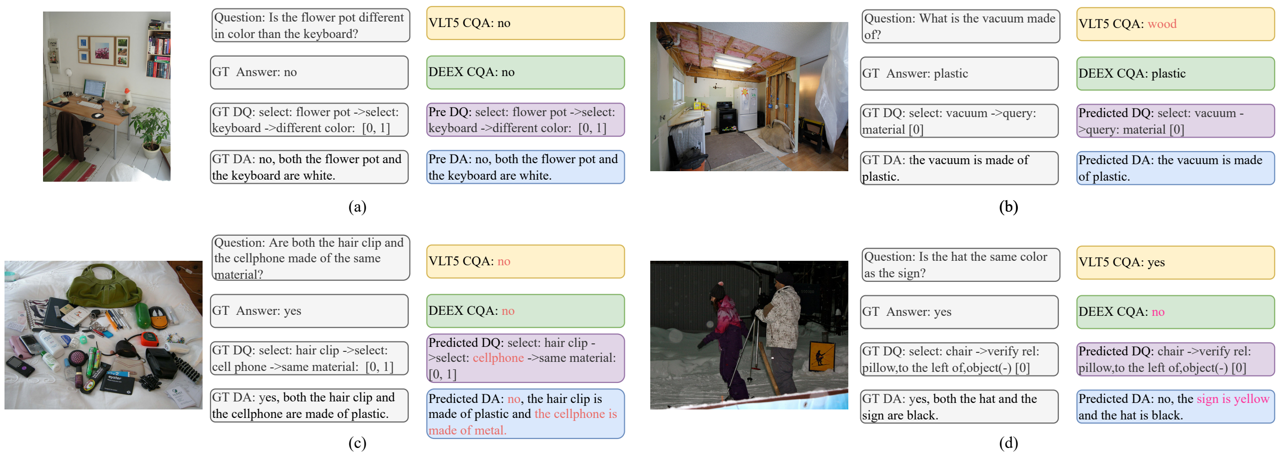

Question: Is the flower pot different in color than the keyboard?
VLT5 CQA: no
GT Answer: no
DEEX CQA: no
GT DQ: select: flower pot ->select: keyboard ->different color: [0, 1]
Pre DQ: select: flower pot ->select: keyboard ->different color: [0, 1]
GT DA: no, both the flower pot and the keyboard are white.
Pre DA: no, both the flower pot and the keyboard are white.
(a)
Question: What is the vacuum made of?
VLT5 CQA: wood
GT Answer: plastic
DEEX CQA: plastic
GT DQ: select: vacuum ->query: material [0]
Predicted DQ: select: vacuum ->query: material [0]
GT DA: the vacuum is made of plastic.
Predicted DA: the vacuum is made of plastic.
(b)
Question: Are both the hair clip and the cellphone made of the same material?
VLT5 CQA: no
GT Answer: yes
DEEX CQA: no
GT DQ: select: hair clip ->select: cell phone ->same material: [0, 1]
Predicted DQ: select: hair clip ->select: cellphone ->same material: [0, 1]
GT DA: yes, both the hair clip and the cellphone are made of plastic.
Predicted DA: no, the hair clip is made of plastic and the cellphone is made of metal.
(c)
Question: Is the hat the same color as the sign?
VLT5 CQA: yes
GT Answer: yes
DEEX CQA: no
GT DQ: select: chair ->verify rel: pillow,to the left of,object(-) [0]
Predicted DQ: chair ->verify rel: pillow,to the left of,object(-) [0]
GT DA: yes, both the hat and the sign are black.
Predicted DA: no, the sign is yellow and the hat is black.
(d)

Figure 3: *Examples of model prediction. The grey box denotes original questions, ground-truth answers, ground-truth deconstructed questions and ground-truth detail answers. The yellow, green, purple, and blue boxes indicate the predicted ones, respectively. The red texts indicate wrong predictions. "CQA" denotes compositional question answering. "DQ" denotes deconstructed questions. "DA" denotes detailed answers.*

Table 3: *Results of different variants of our model. "Cons" denotes consistency constraints.*

#	Model	ACC
0	DEEX (Ours)	**63.05**
1	DEEX w/o Question Deconstructing	61.17
2	DEEX w/o Answer Explicating	62.22
3	DEEX w/o Cons	62.54
4	DEEX w/o All (VLT5)	60.21

Table 4: *Performance on "DQ" and "DA" subtasks.*

Task	BLEU1	BLEU2	BLEU3	BLEU4	Rouge
DQ	87.60	85.32	83.15	80.61	82.92
DA	87.81	84.57	81.75	78.84	80.68

3.4. Compositional Reasoning Performance

As shown in Table 1 and 2, we compare our model DEEX with several SOTA models on the GQA test and validation split, respectively. Our method surpasses all methods, including modular methods and holistic methods, by +1.06 and +1.60 on the test split and the validation split, respectively. We also obverse that the proposed method outperforms VLT5 by a significant margin, which demonstrates the effectiveness of deconstructing questions and explicating answers. Deconstructing questions makes the model learn to reflect on how it is reasoning and explicitly express reasoning processes. Explicating answers enforces the model to learn detailed information instead of having a superficial knowledge of images. For example, to answer the question "Does the knife next to the fork look black?", explicating answers enforces the model to recognize the true color of the knife instead of just judging whether it is black or not. Moreover, we can judge whether the model really carries out correct reasoning through question reconstruction and refined answers to some extent.

3.5. Ablation Study

As shown in Table 3, we propose several variants of our model to verify the effectiveness of each module. We observe that: (1) Comparing line 0 with line 1, 2, 3, all the proposed methods to improve the model performance. When one of them is removed, the performance of the model decreases. (2) Question Deconstructing has the most important among the three methods, which greatly improves the model's performance. (3) Comparing line 0 with line 4, the three methods can work together to improve the model performance further. Note that results in line 4 is the same as VLT5, our approach significantly improves the performance of pretrained VLT5 by deconstructing questions and explicating answer. As shown in Table 4, we utilize BLEU [34] and Rouge [35] to test the performance of question deconstructing and answer explicating. The high BLEU and Rouge scores show that DEEX gives deconstructed questions and detail answers well, which demonstrates that DEEX has good understanding of the question and the image, thus enhancing the performance of question answering.

3.6. Quantitative Analysis

As shown in Figure 3, we provide some examples of the model prediction. Our DEEX answers the compositional questions accurately and gives explanations simultaneously. The explanations give the reasoning why the model answers questions correctly or gives wrong answers. DEEX also corrects error predictions due to training with proposed methods. As shown in Figure 3 (b), DEEX corrects the wrong answer predicted by VLT5 and gives the right answer. As shown in Figure 3 (c), VLT5 and DEEX all predict wrong answers, but DEEX gives the reason why it predicts a wrong answer. As shown in Figure 3 (d), when the answer to the question is ambiguous, DEEX does not simply fit the real answer but gives its own opinion.

4. Conclusion

In this paper, we propose a novel model (DEEX) to enhance visual question answering via deconstructing questions and explicating answers when answering questions. Specifically, DEEX utilizes prompt-based multi-task learning to answer questions and give explanations simultaneously. Experimental results on the GQA dataset verify our assumption and demonstrate the effectiveness of our method. In the future, we will explore how to introduce our method into datasets which do not have such annotations. This work was supported by the National Key Research and Development Program of China (2018AAA0100400), the Strategic Priority Research Program of the Chinese Academy of Sciences (No. XDA27030300) and the National Natural Science Foundation of China (62206294).

5. References

[1] J. Andreas, M. Rohrbach, T. Darrell, and D. Klein, "Neural module networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 39–48.

[2] D. Hudson and C. D. Manning, "Learning by abstraction: The neural state machine," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[3] R. Hu, J. Andreas, M. Rohrbach, T. Darrell, and K. Saenko, "Learning to reason: End-to-end module networks for visual question answering," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 804–813.

[4] J. Shi, H. Zhang, and J. Li, "Explainable and explicit visual reasoning over scene graphs," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8376–8384.

[5] W. Chen, Z. Gan, L. Li, Y. Cheng, W. Wang, and J. Liu, "Meta module network for compositional visual reasoning," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 655–664.

[6] D. A. Hudson and C. D. Manning, "Compositional attention networks for machine reasoning," *arXiv preprint arXiv:1803.03067*, 2018.

[7] R. Hu, A. Rohrbach, T. Darrell, and K. Saenko, "Language-conditioned graph networks for relational reasoning," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 10 294–10 303.

[8] C. Jing, Y. Jia, Y. Wu, C. Li, and Q. Wu, "Learning the dynamics of visual relational reasoning via reinforced path routing," in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, vol. 2, 2022, p. 7.

[9] P. Zhang, X. Li, X. Hu, J. Yang, L. Zhang, L. Wang, Y. Choi, and J. Gao, "Vinvl: Revisiting visual representations in vision-language models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 5579–5588.

[10] C. Jing, Y. Jia, Y. Wu, X. Liu, and Q. Wu, "Maintaining reasoning consistency in compositional visual question answering," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 5099–5108.

[11] A. Kamath, M. Singh, Y. LeCun, G. Synnaeve, I. Misra, and N. Carion, "Mdetr-modulated detection for end-to-end multi-modal understanding," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 1780–1790.

[12] J. Lu, V. Goswami, M. Rohrbach, D. Parikh, and S. Lee, "12-in-1: Multi-task vision and language representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10 437–10 446.

[13] J. Cho, J. Lei, H. Tan, and M. Bansal, "Unifying vision-and-language tasks via text generation," in *International Conference on Machine Learning*. PMLR, 2021, pp. 1931–1942.

[14] F.-L. Chen, D.-Z. Zhang, M.-L. Han, X.-Y. Chen, J. Shi, S. Xu, and B. Xu, "VLP: A survey on vision-language pre-training," *Machine Intelligence Research*, vol. 20, no. 1, pp. 38–56, 2023.

[15] F. Chen, D. Zhang, X. Chen, J. Shi, S. Xu, and B. XU, "Unsupervised and pseudo-supervised vision-language alignment in visual dialog," in *Proceedings of the 30th ACM International Conference on Multimedia*, ser. MM '22. New York, NY, USA: Association for Computing Machinery, 2022, p. 4142–4153.

[16] F. Chen, X. Chen, S. Xu, and B. Xu, "Improving cross-modal understanding in visual dialog via contrastive learning," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 7937–7941.

[17] F. Chen, F. Meng, J. Xu, P. Li, B. Xu, and J. Zhou, "DMRM: A dual-channel multi-hop reasoning model for visual dialog," in *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020*. AAAI Press, 2020, pp. 7504–7511.

[18] A. Ross, M. E. Peters, and A. Marasović, "Does self-rationalization improve robustness to spurious correlations?" *arXiv preprint arXiv:2210.13575*, 2022.

[19] S. Palaskar, A. Bhagia, Y. Bisk, F. Metze, A. W. Black, and A. Marasovic, "On advances in text generation from images beyond captioning: A case study in self-rationalization," *arXiv preprint arXiv:2205.11686*, 2022.

[20] A. Marasović, I. Beltagy, D. Downey, and M. E. Peters, "Few-shot self-rationalization with natural language prompts," *arXiv preprint arXiv:2111.08284*, 2021.

[21] S. Wiegreffe and A. Marasovic, "Teach me to explain: A review of datasets for explainable nlp," *arXiv preprint arXiv:2102.12060*, vol. 3, 2021.

[22] D. Alvarez Melis and T. Jaakkola, "Towards robust interpretability with self-explaining neural networks," *Advances in neural information processing systems*, vol. 31, 2018.

[23] W. Peng, Y. Hu, J. Yu, L. Xing, and Y. Xie, "APER: adaptive evidence-driven reasoning network for machine reading comprehension with unanswerable questions," *Knowl. Based Syst.*, vol. 229, p. 107364, 2021. [Online]. Available: https://doi.org/10.1016/j.knosys.2021.107364

[24] D. A. Hudson and C. D. Manning, "Gqa: A new dataset for real-world visual reasoning and compositional question answering," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 6700–6709.

[25] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems*, vol. 28, 2015.

[26] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma *et al.*, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *International journal of computer vision*, vol. 123, no. 1, pp. 32–73, 2017.

[27] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6077–6086.

[28] J. D. M.-W. C. Kenton and L. K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of NAACL-HLT*, 2019, pp. 4171–4186.

[29] T. Gao, A. Fisch, and D. Chen, "Making pre-trained language models better few-shot learners," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2021, pp. 3816–3830.

[30] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. [Online]. Available: http://arxiv.org/abs/1412.6980

[31] J.-H. Kim, J. Jun, and B.-T. Zhang, "Bilinear attention networks," *Advances in neural information processing systems*, vol. 31, 2018.

[32] D. Guo, C. Xu, and D. Tao, "Graph reasoning networks for visual question answering," 2019.

[33] H. Tan and M. Bansal, "Lxmert: Learning cross-modality encoder representations from transformers," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019, pp. 5100–5111.

[34] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2002, pp. 311–318. [Online]. Available: https://www.aclweb.org/anthology/P02-1040

[35] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Text Summarization Branches Out*, 2004, pp. 74–81. [Online]. Available: https://www.aclweb.org/anthology/W04-1013