Weakly Supervised Visual Question Answer Generation

Charani Alampalle
AlphaICs
Bengaluru, India
nagasai.charani@alphaics.com

Shamanthak Hegde
KLE Technological University
Hubballi, India
01fe19bcs233@kletech.ac.in

Soumya Jahagirdar
CVIT, IIIT Hyderabad
Hyderabad, India
soumya.jahagirdar@research.iiit.ac.in

Shankar Gangisetty
IIIT Hyderabad
Hyderabad, India
shankar.gangisetty@ihub-data.iiit.ac.in

Abstract

Growing interest in conversational agents promote two-way human-computer communications involving asking and answering visual questions have become an active area of research in AI. Thus, generation of visual question-answer pair(s) becomes an important and challenging task. To address this issue, we propose a weakly-supervised visual question answer generation method that generates a relevant question-answer pairs for a given input image and associated caption. Most of the prior works are supervised and depend on the annotated question-answer datasets. In our work, we present a weakly supervised method that synthetically generates question-answer pairs procedurally from visual information and captions. The proposed method initially extracts list of answer words, then does nearest question generation that uses the caption and answer word to generate synthetic question. Next, the relevant question generator converts the nearest question to relevant language question by dependency parsing and in-order tree traversal, finally, fine-tune a ViLBERT model with the question-answer pair(s) generated at end. We perform an exhaustive experimental analysis on VQA dataset and see that our model significantly outperform SOTA methods on BLEU scores. We also show the results wrt baseline models and ablation study.

1. Introduction

Conversational agents that can communicate with a human have been an active area of research and becoming popular due to artificial intelligence (AI). In conversational agent communication, visual question answering (VQA) is not only the desired characteristic where the agent answers a natural language question about the image, but also an intelligent agent should have the ability to ask a meaningful and relevant question with respect to its current visual perception. Keeping these points in mind, we can say that the best way to establish communication between humans and machines is by making machines how to ask meaningful and relevant question and at the same time, answer it properly when asked. Recently more focus is given to understanding the scene, asking and answering the questions from images and videos.

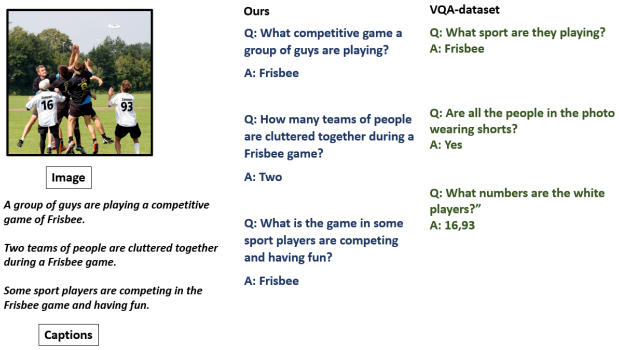

Figure 1. Our proposed weakly supervised VQA generation method gives importance to visual properties relevant to the given input image in order to generate specific question-answer pairs. Our results are automatically generated compared to VQA dataset [1] i.e., manually generated question-answer pairs.

Works in various domains are mainly focusing on VQA [1] and visual question generation [5,6]. Given an image, generating meaningful questions, also known as visual question generation (VQG) is an essential component of a conversational agent [16] along with answering the questions. Visual question answer generation (VQAG) is a precursor of visual dialogue systems and might help in building large-scale VQA datasets automatically with less human

effort. All the previous works [1, 3, 21] on question answer pair generation are dependent largely on datasets which was manual and tedious task to train. With less human effort to build VQA datasets needed for training, we build a VQAG model that can generate meaningful question-answer pairs for a given image and associated captions. Works in the direction of question-answer generation for image, given the caption focuses on generating generic questions or category based answers [2, 18]. Unlike these we focus on category based questions (As categories we consider six types of question words like "*how many*", "*who*", "*what*", "*which*", "*how much*", and "*where*") which help the conversational agents to communicate properly in understanding the context. Good question-answer pair is the one that has a tightly focused purpose and must be relevant to the image content. In this work, we fill-up the gap prevailing in the literature by introducing a method to solve the problem of generating question-answer pairs for the given image and associated captions with question being categorised. We refer our work as VQAG.

Without depending on ground-truth QA pair(s) of images, the goal of VQAG is given a natural image and associated captions, list of objects are extracted from the image, generate a natural language question using caption, whose answer is one of the list of objects being identified. As we are not depending on ground-truth QA pair(s) and using captions instead we name our work as Weakly supervised Visual Question Answer Generation. As shown in Fig 2 given an image and associated caption(s), our aim is to generate a relevant language question such that the answer is in the list of the objects identified from the image. Here list of objects is [Fisbee, Person]. The problem is challenging as it requires in-depth semantic interpretation of the caption and from the image the visual content needs to generate meaningful and relevant questions. Question-answer pair generation has been a well-explored area in the language community [13], vision and language community [2, 18]. However, vision and language works often ignore the important visual information or objects appearing in the image, and only restrict themselves to the overall visual content while generating question and answer pairs.

It should be noted that objects in the image helps not only in asking semantically meaningful and relevant questions connecting visual and textual content, but also helps to avoid generic questions as well as generates detailed question-answer pair(s). Consider for example given an image with two teams of players playing frisbee in ground and enjoying the game shown in Fig 1. Our proposed method automatically generate questions that are meaningful, relevant, non-generic and detailed, such as "*What competitive game a group of guys are playing?*" ,"*How many teams of people are cluttered together during a frisbee game?*", "*What is the game in some sport players are competing and having fun?*". While the questions for the same image in VQA dataset [1] has generic questions like "*What sport are they playing?*", "*Are all the people in the photo wearing shorts?*", "*What numbers are the white players?*" respectively. We see that our proposed VQAG method generates more meaningful and detailed question-answer pairs unlike VQA. In our proposed approach, we first extract answer from the list of objects identified from the image. We use Faster RCNN [19] object detection technique to extract the objects from the image. Obtaining answers from the extracted objects, we then use these answer to generate meaningful question using associated caption (see Section 4). The proposed VQAG method significantly outperforms interms of question-answer generation compare to SOTA [10, 23]. We firmly believe that our work will boost ongoing research efforts [1, 6, 22] in the broader area of conversational AI and scene-text understanding. Our implementation will be made publicly available on acceptance of the work.

The major contributions of this paper are three folds:

- We draw the attention of the Document Analysis and Recognition community to the problem of visual question answer generation by leveraging image caption and semantically bridge the visual content with the associated caption. We are the first to explore the VQAG problem and is an important step in the development of conversational AI and useful in augmenting the training data of image-based question answering.

- We propose weakly supervised VQA generation method that creates nearest question using caption and visual information, that are then converted to relevant question using in-order traversal by dependency reconstruction method.

- Exhaustive experimental analysis are performed on proposed VQAG method. Our model significantly outperforms the existing works [10, 23] interms of qualitative and quantitative results. Extensive ablation study on proposed method is investigated.

2. Related works

Recently works started studying visual question answering [1, 3, 6, 9, 24, 26] and they are coming up with various ways to solve edge cases in VQA. However, there are only a few works that focus on generating question-answer pairs for given images [2, 7, 18]. Our goal is to increase the kind of questions that can be generated with correct answers provided they are relevant to the image. We carry out experiments in generating question-answer pairs in a weakly supervised manner that in turn helps in generating large datasets for solving various VQA tasks.

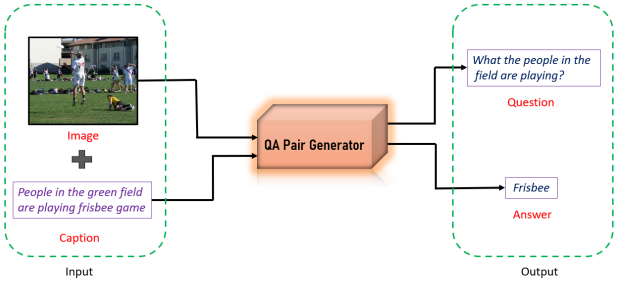

Figure 2. VQAG. We introduce a method of Visual Question Answer Generation. Given an image and its caption, our goal is to generate a meaningful natural language question whose answer is one of the objects from image.

2.1. Visual Question Generation (VQG)

VQG [10, 20, 25] is a well-studied problem in literature and is essential for developing conversational agents and visual dialogue systems [11, 16]. Further, ability to generate relevant and meaningful question by getting in-depth understanding of the visual content of the image is challenging. Few works in this direction are single question generator [16], multiple diverse questions generator [5, 27], goal-driven question generator [10, 23], and learn by context of VQA [12, 15]. In [10], the authors generate question based on category of answer. They proposed an information maximizing visual question generator that maximizes the mutual information between image, answer, and answer category during training that helps to generate questions based on the category of answer being asked without answer being given explicitly. But, some of the questions generated in [10] based on category of answer does not seem relevant to few images. For example, consider the image shown in Fig 3. Given an image of two street signs hanging over a pole with names over it as GREENWICH and VESEY respectively at intersection of roads in front of tall buildings. We aim to automatically generate questions that are meaningful, relevant, non-generic and detailed, such as "*Where VESEY street sign hanging on a pole?*" but, if given the category as color wrt same image information maximization VQG [10] might have generated question related to colour of the sky, building or clouds which seems very generic though it is category-based generator. In our method, we overcome this issue by first generating answer and depending on the answer, category-based questions are generated.

In [23], authors generate questions based on answer category but the approach taken was different where the mutual information between image, question, and answer category is maximized at latent space. The variational auto-encoder is used to reduce the level of supervision, but still they depend on manually created datasets for ground truth. In our method, we fill this gap with effective usage of visual (image) and textual (caption) meaningful information to generate question-answer pair in a weakly supervised manner.

2.2. Visual Question Answer Generation (VQAG)

Traditionally, VQAG methods focuses on category-based answer type [2, 18]. In [2, 18], authors generate category-wise synthetic question-answer pairs from captions using template-based learning. The categorization is based on answer that not only restricts the number of questions, but also the quality of question degrades. However, in our methods, we categorize based on the type of the question which leads to diversity of question-answer pairs that are being generated related to the input image. We consider six types of question words like "*how many*", "*who*", "*what*", "*which*", "*how much*", and "*where*". In [13], authors generate question-answer pairs from Wikipedia text and cited documents using context, question, and answer triplets. Thus, generating question-answer pairs from images is an ongoing research in AI that helps developing agents to connect easily with the scene and communicate with humans.

Q: Where VESEY street sign hanging on a pole?

Figure 3. Example question generated by our model which is meaningful, non generic, detailed and relevant to the image.

3. Proposed approach

Given an image and associated caption(s), our aim is to generate a relevant language question such that the answer is in the list of the objects identified from the image. The proposed method should be able to recognize the objects present in the image by properly understanding the visual content, get a clear idea about the caption associated with the image and semantically bridge the textual and visual content inorder to generate meaningful questions. Initially object detection is done using FasterRCNN [19] followed by a template based method to generate a question from caption and list of objects. After detecting objects, we extract the answer word from caption which is part of the list

of detected objects. We, then proceed with question generation. The question generation is a two step process, (i) nearest question generation and (ii) relevant question generation. The overall architecture of the proposed model is illustrated in Fig 4.

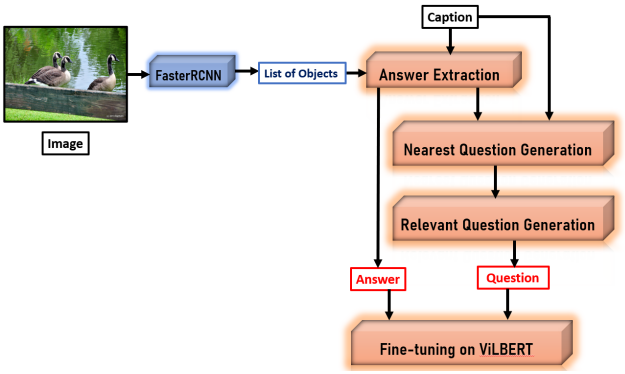

Figure 4. Proposed Visual Question Answer Generator (VQAG) architecture. VQAG has four modules, namely, (i) Answer Extraction, (ii) Nearest Question Generator, (iii) Relevant Question Generator, and (iv) Fine-tuning on ViLBERT.

3.1. Answer Extraction

We first extract the list of objects (Os) from the given input image (I) using pre-trained FasterRCNN [19] with ResNet-101 as backbone. Since we do not want to lose the information from the image that affects the objects being detected, we set the threshold of confidence to a lower value of -0.2. Once we detect objects $(O_1, O_2, ..., O_n)$, next extract the related answer word from the captions $(C_1, C_2, ..., C_n)$. We then check the words from the list of objects identified in last step from the caption. If the word from the list of objects is not found in caption, we then use noun chunkers or name entity recogniser (NER) from Spacy to extract answer words $(W_1, W_2, ..., W_n)$ from the captions. After extracting the answer words, we mask the answer word in the caption and call it as masked caption.

if $O_i == < C_1, ..., C_n >$ **then**
  $ans \leftarrow O_i$
else
  if $O_i \neq < C_1, ..., C_n >$ **then**
    $ans \leftarrow W_i$
  end if
end if

where W_i is answer words from $< W_1, W_2, ..., W_n >$

3.2. Question Generation

We generate nearest questions from the masked caption. We then introduce a rule-based method to rewrite the nearest questions to meaningful relevant questions, which utilizes the dependency structures. Depending on the answer word, Our model generates six types of questions like "*how many*", "*who*", "*what*", "*which*", "*how much*", and "*where*". The frequency of occurrence wrt each type of question word is shown in Fig 5. As question generation is governed by the caption and the detected objects, the distribution of questions are biased towards the count based category type that is "*how many*" and "*how much*" in which there is a scope for improvement.

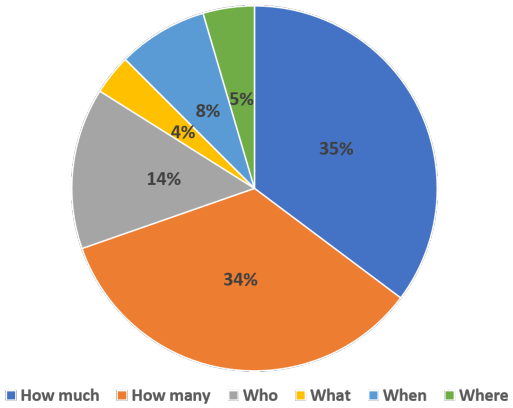

Figure 5. Distribution of different types of question categories generated by our VQAG model.

3.2.1 Nearest Question Generation

This is the first step in generating questions from the masked caption. Here, we replace the answer words in the statements with a special mask token, depending on its answer category. Using the masked caption and the answer (with a type label ANIMAL in Fig 6 as masked word is animal in this case), we replace the masked word with special category word which gives us with the Nearest question.

3.2.2 Translate Nearest to Relevant Questions

To generate the relevant questions from nearest questions, we perform a dependency reconstruction. First, we create the tree from the words taken from nearest question. Then, we move answer-related words in the dependency tree to the front of the question, as answer-related words are crucial in framing the question. We do this procedure of moving answer related words to the front of the question keeping the intuition that relevant questions usually start with question words. Dependency parsing is applied to the nearest questions, we follow three steps to translate them to relevant questions, (i) keep the right child nodes of the answer-related word and prune its lefts, (ii) for each node in the parsing tree, if the sub tree of its child node contains the answer node, we move the child node to the first child node, (iii) finally, do in-order traversal on the reconstructed tree to

Table 1. Quantitative evaluation of our method against other weakly supervised SOTA model using standard metrics.

Models	BLEU	METEOR	ROUGE-L
IA2Q [2]	30.42	9.42	-
V-IA2Q [2]	35.40	13.35	-
IMVQG [10]	31.2	12.11	40.27
C3VQG [23]	41.87	13.60	42.34
Ours	**47.78**	**27.61**	18.89

obtain the relevant question. Using rule-based method mapping, which replaces each answer category with the most appropriate *wh** word. For example, the LOCATION category is mapped to "*WHERE*" and the COUNT category is replaced by "*HOW MANY*". Fig 6 shows the detailed explanation of Visual QA generation following the above mentioned steps.

3.3. Fine-tuning on ViLBERT

Generated questions along with the object words that are identified as most appropriate answers are taken as question-answer pairs with their corresponding images and are fine-tuned on popular and state-of-art work in visual question answering [14]. Based on the question-answer pairs generated, we create a new vocabulary and then fine-tune on the new vocabulary being appended to the VQA's vocabulary. After fine-tuning with our questions and answers, we test it on VQA test dataset and obtain a score of 49.2.The reason for less score is, during the testing time, some of the answer words might be missing that are part of new vocabulary which are not found in [1] vocabulary as it is up-streaming task.

4. Experiments and Results

In this section, we experimentally validate our proposed method for question-answer pair generation. The quantitative and qualitative experimental results of the proposed method and the comparative analysis with SOTA is provided.

4.1. Datasets

Since we consider captions along with images to generate question-answer pair, there is no dedicated dataset available for this task containing question-answer pair with captions for the images. We, therefore, make use of the two popular datasets, namely, MSCOCO [4] and VQA [1]. Note that unlike these datasets where originally the task is to generate caption for the image [13] [8]and answer a question about the image [1] [6] respectively, our aim is to generate question-answer pairs using captions for the images. In other words, given an image and associated captions (taken from [4]), our proposed method learns to automatically generate question-answer pairs similar to the manually curated VQA dataset.

4.2. Performance metrics

We use popular evaluation measures such as bilingual evaluation understudy (BLEU) [17], recall-oriented understudy for listing evaluation-longest common sub-sequence (ROUGE-L), and metric for evaluation of translation with explicit ordering (METEOR). The BLEU score compares n-grams of the generated question with the n-grams of the reference question and counts the number of matches. The ROUGE-L metric indicates similarity between two sequences based on the length of the longest common subsequence even though the sequences are not contiguous. The METEOR is based on the harmonic mean of unigram precision and recall and is considered a better performance measure in the text generation. Higher values of all these performance measures imply better matching of generated questions with the reference questions.

4.3. Implementation Details

We fine-tune our generated QA pair(s) on [14] with 4 layer MLP using the Adam optimizer with initial learning rate of 1e-4, batch size of 64. The maximum length of the generated questions is set to 24. We fine-tune the model for 15 epochs. The model is fine-tuned on a single NVIDIA Quadro P5000.

4.4. Results and discussions

4.4.1 Quantitative Analysis

We evaluate the performance of our proposed method and compare it against three baseline approaches, "without list of objects as context - without filtering block, "without list of objects as context - with filtering block", "with list of objects as context - with filtering block".Explanation for these baselines is provided in section 4.4.2. This comparative result is shown in Table 2 using performance measures discussed in Section 4.2. Note that higher values for all these popularly used performance measure is considered superior. Among the three baseline approaches, "without list of objects as context - with filtering block" generates comparatively better questions. Our proposed final method significantly outperforms all the baseline models. For example our proposed method improves BLEU-score by 1.14 as compared to the most competitive baseline i.e., "without list of objects as context - with filtering block". It should be noted that under these performance measures these gains are considered significant. Further, by design, our proposed method tries to ensure that the answer being object detected in most of the cases and generated question is relevant and meaningful to answer and image. We also

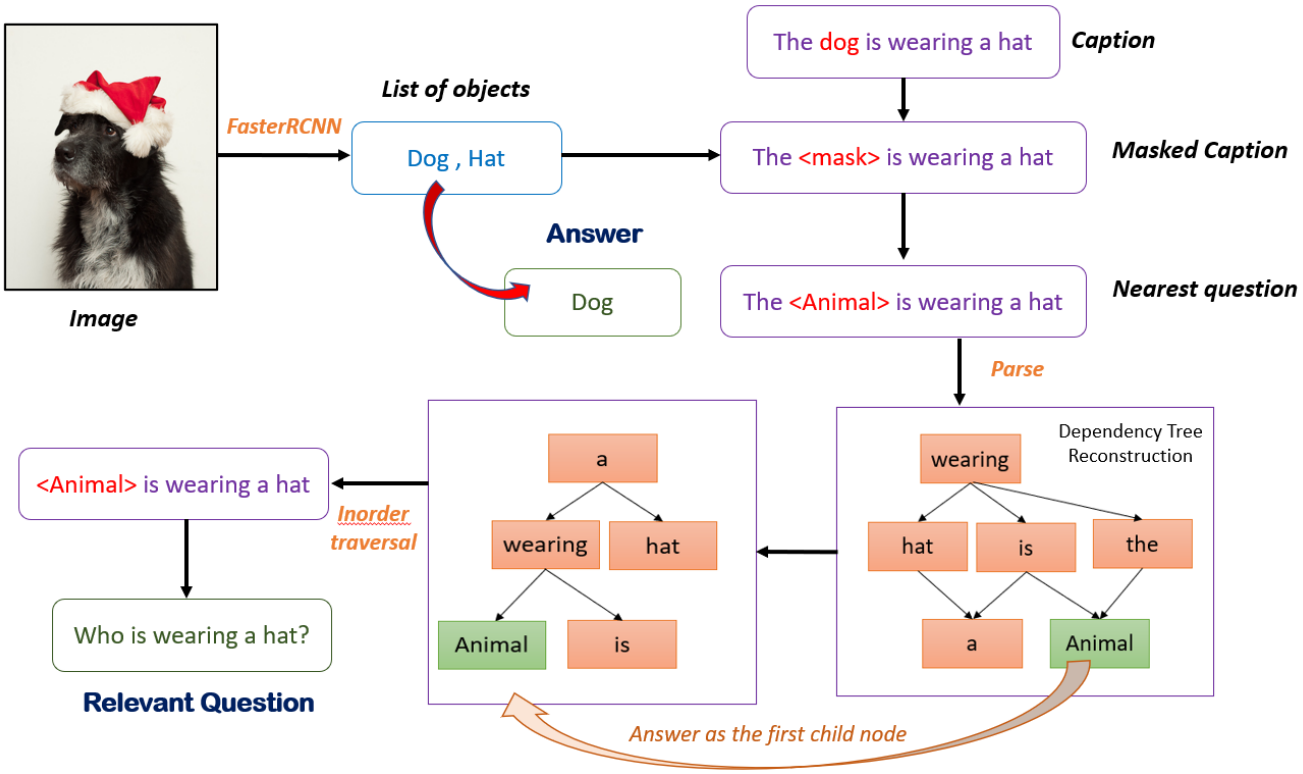

Figure 6. Illustration of the proposed VQAG with an example.

Table 2. Comparison of our method with various baselines techniques based on context and filtering block.

Models	BLEU
w/o list of objects as context, w/o filtering block	45.63
w/ list of objects as context, w/ filtering block	46.29
w/o list of objects as context, w/ filtering block	45.13
w/ list of objects as context, w/o filtering block (Ours)	**47.78**

demonstrate our results with and without fine-tuning on [14]. It is also seen that fine-tuning on ViLBERT plays a major role in achieving better results which in turn prove that our generated QA pair are meaningful.

After fine-tuning on [14], we evaluate our generated questions on VQA questions and our scores are compared with [2, 10, 23]. Our model outperforms all the models as shown in Table 1. The ROUGE-L scores are low compared to the other works as the process of generating questions varies from their approach to our approach. Question generation in their approach follows category based on answers or directly via image features which generates questions similar to that of VQA. Unlike those works we generate questions via captions and detected objects which generates questions way different from VQA.

4.4.2 Experiments

We experimented with and without using filtering in our proposed architecture (see Fig 4). In case of "with filtering block" we filter the captions mean we consider the captions only if the particular caption has one of the word from list of detected objects for that image. In this, we filter out certain captions and filtered captions are only considered for further question generation. But this approach removed nearly 23 percent of the captions which in turn lead to lose of good quality questions.

4.4.3 Context as secondary input

We used context as secondary input to the nearest question generation. We experimented using "list of objects and caption as context" and only caption as context. Using of context with "list of objects and caption" helped our model to

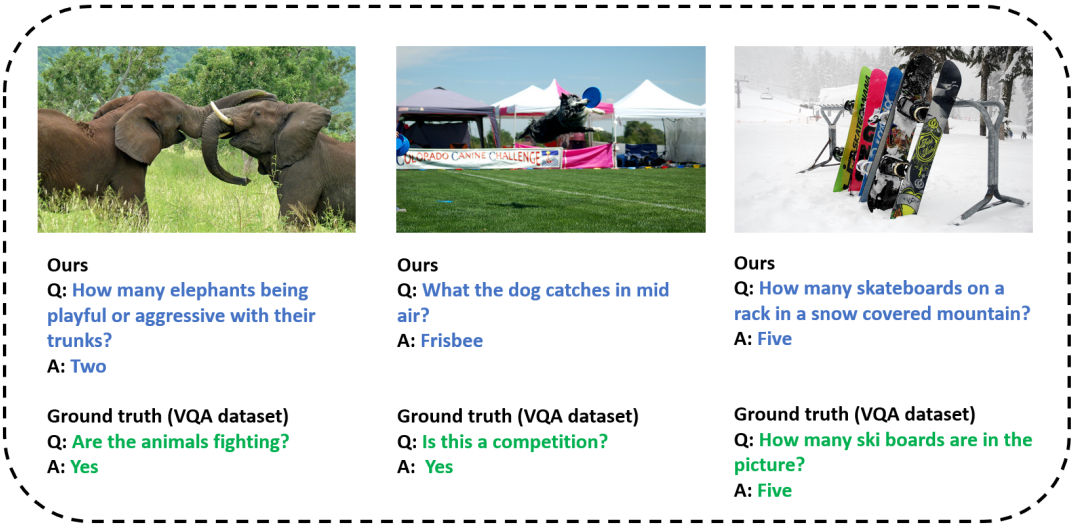

Figure 7. Generated question-answer pairs on VQA dataset [1]. Results showing from our method and ground truth from VQA dataset respectively. **[Best viewed in color]**

get better questions which are more relevant to the image as it uses list of objects too along with caption.

4.4.4 Problem with words that are identified as objects

In one of the baseline models (w/ list of objects as context, w/ filtering block) we had a problem with the certain object words which are identified by object detector (Faster-RCNN) that are not found in captions and hence we may lose few good questions there. We tackled this problem by appending the word identified to the similar words having the same meaning for the detected objects. For example, we replaced person by adult, man, woman, boy, girl. However we removed the filtering block as the scores are low compared to the model without filtering block as shown in Table 2. We removed the filtering block from the methodology diagram Fig 4 as it is not considered in the finalized model.

4.4.5 Qualitative Analysis

We perform a detailed qualitative analysis of the baselines as well as our proposed method. We first show a comparison of generated QA pair using all the three baselines versus proposed method in Fig 9 which is differentiated by color. We observe that the baselines are capable of generating almost linguistically meaningful QA pair with minute chances in sentence formation but not as meaningful as our proposed method. For the given image in Fig 9, the expected question is " What has a yellow cartoon dog on it?" which is generated by our proposed method. The baseline approaches have generated questions which are appropriate

Figure 8. Failure case examples of visual question generation by our model on images from VQA dataset.

to their corresponding answers but they are not as meaningful as our proposed method. Though the QA pairs generated by baselines are relevant to the input image, as the complexity of the image and the caption increases, the baseline models fail to generate more appropriate and meaningful questions than the proposed model. The proper utilization of visual and textual information generated better questions in our proposed method.

The best baseline model "without list of objects as context - with filtering block" generates question "What cartoon dog?" which is more meaningful than rest baseline approaches but it fails to specify the important information in the image as it is not using list of objects as context and not using filtering block that has removed the caption which dont have one of the words from list of objects, in this case

the important information can be Frisbee which is missing. The answer here ("Yellow") is generated by noun chunkers and NER toolkit. Whereas the proposed model generates "What has a yellow cartoon dog on it?" which is more relevant to the image and meaningful than baseline approaches.

Further, more results of our model are shown in Fig 7 questions and answers generated by our model are in blue, and corresponding [1] questions and answers are in green. Here our model successfully generates meaningful and relevant questions to the image.

The failure of our model pronounced when VQAG misunderstood the scene and output incorrect objects during answer extraction or there is need of generating questions in which objects from image may not act as answer or fails to build the semantic relation between caption and the answer. Fig 9 shows failure case examples generated by our model. Though the failure case examples does not sense meaningful yet they are relevant to the image.

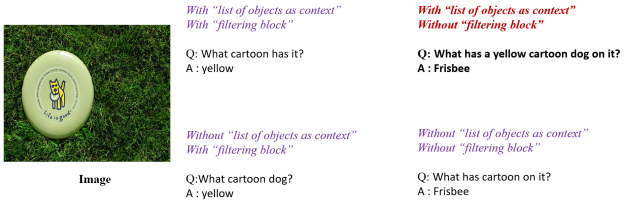

Figure 9. Baseline models and our model qualitative comparison on image from VQA dataset.

5. Conclusions

We proposed a visual question answer generation method in weakly supervised manner for a given image and associated caption. Our proposed method properly utilizes the visual properties and generates the question-answer pairs that are meaningful and relevant to the image. Our method has outperformed on SOTA question generating models with a BLEU score value increased by **6%**. Ours is the first work towards developing a visual question-answer pair generation model which considers answer as the one of the object from the image, we restrict our scope to generating questions whose answer is the object from image. Our question-answer pair generator can be used in generating large datasets with no human effort and can also be used in task related to meta-learning and self-supervised learning. Future directions include complex, specific and realistic question-answer pair generation that require deeper semantic reasoning using transformers in understanding image and text together.

References

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: visual question answering. In *ICCV*, pages 2425–2433. IEEE Computer Society, 2015. 1, 2, 5, 7, 8

[2] Pratyay Banerjee, Tejas Gokhale, Yezhou Yang, and Chitta Baral. Weaqa: Weak supervision via captions for visual question answering. In *ACL/IJCNLP (Findings)*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 3420–3435. Association for Computational Linguistics, 2021. 2, 3, 5, 6

[3] Ali Furkan Biten, Rubèn Tito, Andrés Mafla, Lluís Gómez i Bigorda, Marçal Rusiñol, C. V. Jawahar, Ernest Valveny, and Dimosthenis Karatzas. Scene text visual question answering. In *ICCV*, pages 4290–4300. IEEE, 2019. 2

[4] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO captions: Data collection and evaluation server. *CoRR*, abs/1504.00325, 2015. 5

[5] Zhihao Fan, Zhongyu Wei, Piji Li, Yanyan Lan, and Xuanjing Huang. A question type driven framework to diversify visual question generation. In *IJCAI*, pages 4048–4054. ijcai.org, 2018. 1, 3

[6] Yash Goyal, Tejas Khot, Aishwarya Agrawal, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. *Int. J. Comput. Vis.*, 127(4):398–414, 2019. 1, 2, 5

[7] Soumya Jahagirdar, Shankar Gangisetty, and Anand Mishra. Look, read and ask: Learning to ask questions by reading text in images, 2022. 2

[8] Mahmoud Khademi and Oliver Schulte. Image caption generation with hierarchical contextual visual spatial attention. In *CVPR Workshops*, pages 1943–1951. Computer Vision Foundation / IEEE Computer Society, 2018. 5

[9] Alexey K. Kovalev, Makhmud Shaban, Evgeny Osipov, and Aleksandr I. Panov. Vector semiotic model for visual question answering. *Cogn. Syst. Res.*, 71:52–63, 2022. 2

[10] Ranjay Krishna, Michael S. Bernstein, and Li Fei-Fei. Information maximizing visual question generation. In *CVPR*, pages 2008–2018. Computer Vision Foundation / IEEE, 2019. 2, 3, 5, 6

[11] Yikang Li, Nan Duan, Bolei Zhou, Xiao Chu, Wanli Ouyang, Xiaogang Wang, and Ming Zhou. Visual question generation as dual task of visual question answering. In *CVPR*, pages 6116–6124. Computer Vision Foundation / IEEE Computer Society, 2018. 3

[12] Yikang Li, Nan Duan, Bolei Zhou, Xiao Chu, Wanli Ouyang, Xiaogang Wang, and Ming Zhou. Visual question generation as dual task of visual question answering. In *CVPR*, pages 6116–6124. Computer Vision Foundation / IEEE Computer Society, 2018. 3

[13] Zhongli Li, Wenhui Wang, Li Dong, Furu Wei, and Ke Xu. Harvesting and refining question-answer pairs for unsupervised QA. In *ACL*, pages 6719–6728. Association for Computational Linguistics, 2020. 2, 3, 5

[14] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *NeurIPS*, pages 13–23, 2019. 5, 6

[15] Ishan Misra, Ross B. Girshick, Rob Fergus, Martial Hebert, Abhinav Gupta, and Laurens van der Maaten. Learning by asking questions. In *CVPR*, pages 11–20. Computer Vision Foundation / IEEE Computer Society, 2018. 3

[16] Nasrin Mostafazadeh, Ishan Misra, Jacob Devlin, Margaret Mitchell, Xiaodong He, and Lucy Vanderwende. Generating natural questions about an image. In *ACL (1)*. The Association for Computer Linguistics, 2016. 1, 3

[17] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *ACL*, pages 311–318. ACL, 2002. 5

[18] Mengye Ren, Ryan Kiros, and Richard S. Zemel. Exploring models and data for image question answering. In *NIPS*, pages 2953–2961, 2015. 2, 3

[19] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(6):1137–1149, 2017. 2, 3, 4

[20] Mourad Sarrouti, Asma Ben Abacha, and Dina Demner-Fushman. Goal-driven visual question generation from radiology images. *Inf.*, 12(8):334, 2021. 3

[21] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards VQA models that can read. In *CVPR*, pages 8317–8326. Computer Vision Foundation / IEEE, 2019. 2

[22] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards VQA models that can read. In *CVPR*, pages 8317–8326. Computer Vision Foundation / IEEE, 2019. 2

[23] Shagun Uppal, Anish Madan, Sarthak Bhagat, Yi Yu, and Rajiv Ratn Shah. C3VQG: category consistent cyclic visual question generation. In *MMAsia*, pages 49:1–49:7. ACM, 2020. 2, 3, 5, 6

[24] Shaohua Wan, Chen Chen, and Alexandros Iosifidis. Editorial to special issue on cross-media learning for visual question answering. *Image Vis. Comput.*, 118:104355, 2022. 2

[25] Xing Xu, Tan Wang, Yang Yang, Alan Hanjalic, and Heng Tao Shen. Radial graph convolutional network for visual question generation. *IEEE Trans. Neural Networks Learn. Syst.*, 32(4):1654–1667, 2021. 3

[26] Huayi Zhan, Peixi Xiong, Xin Wang, Xin Wang, and Lan Yang. Visual question answering by pattern matching and reasoning. *Neurocomputing*, 467:323–336, 2022. 2

[27] Shijie Zhang, Lizhen Qu, Shaodi You, Zhenglu Yang, and Jiawan Zhang. Automatic generation of grounded visual questions. In *IJCAI*, pages 4235–4243. ijcai.org, 2017. 3