ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-6327-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICASSP49357.2023.10097017

IMAGE GENERATION IS MAY ALL YOU NEED FOR VQA

Kyungho Kim, Junseo Lee, Jihwa Lee

ActionPower Corporation, South Korea

ABSTRACT

Visual Question Answering (VQA) stands to benefit from the boost of increasingly sophisticated Pretrained Language Model (PLM) and Computer Vision-based models. In particular, many language modality studies have been conducted using image captioning or question generation with the knowledge ground of PLM in terms of data augmentation. However, image generation of VQA has been implemented in a limited way to modify only certain parts of the original image in order to control the quality and uncertainty. In this paper, to address this gap, we propose a method that utilizes the diffusion model, pre-trained with various tasks and images, to inject the prior knowledge base into generated images and secure diversity without losing generality about the answer. In addition, we design an effective training strategy by considering the difficulty of questions to address the multiple images per QA pair and to compensate for the weakness of the diffusion model. VQA model trained on our strategy improves significant performance on the dataset that requires factual knowledge without any knowledge information in language modality.

Index Terms— Visual question answering, image generation, diffusion model, knowledge base, data augmentation

1. INTRODUCTION

In recent years, the VQA [1, 2] task is receiving tremendous attention as a multimodal task that requires understanding both language and vision. Such VQA problems require a massive amount of image-question-answer triplet because it is necessary to understand not only each modality but also the relationship between different modalities [3, 4]. From this perspective, the research has put many endeavors into collecting various types and more significant amounts of data for the VQA dataset [5, 6] . However, making triplet data manually is expensive and time-consuming. Also, even if humans try to collect as many question types and images as possible in a balanced manner, human bias is eventually involved, resulting in an unbalanced dataset [7]. To address such data construction problems, methods for automatically transforming data and creating a dataset that requires an external knowledge base and much deeper reasoning have been proposed [8, 9].

With the advances in Natural Language Process (NLP), Language Models (LM) are used to automate the VQA dataset construction and obtain a larger number of VQA triplets. [10] uses translation-pair to increase the number of annotations. [11] takes advantage of image alt-text annotation to collect image-caption pairs. VQ^{2}A [12] uses question generation from image caption to increase the VQA triplets. However, since these methods increase the number of triplets by creating a new question, there are disadvantages of equally treating easy questions, such as true/false binary choice, and difficult questions, such as ‘how’ and ‘why’ which require deeper reasoning. To bridge the gap between the difficulty of questions, researchers leverage the prior knowledge of PLM by injecting additional information into the VQA model to clarify the information about facts. [13, 14] infuse the external knowledge base as language embedding of PLM to solve the fact-based problem.

In the field of vision, many attempts exist to augment the VQA dataset to overcome a limited volume of an image that is costly to collect. [15, 16] apply the basic data augment to images like flipping, rotation, and random masking. [17] utilizes the Generative Adversarial Network (GAN) [18] to generate the image related to the answer. However, this generation framework proceeds in a way that changes only the fine-grained portion of the original image due to the instability and model collapse of the GAN. Consequently, existing methods using GAN have focused on generating counterfactual images through minimal editing, which modifies only the part related to the correct answer in the original image, like the color of an object. Through this augmentation, the model can grasp where is the critical part of an image to solve the question. However, due to the nature of minimal editing, poor performance is shown in cases such as relatively large deformations like shape change. In addition, the GAN framework is difficult to apply to QA when abstract nouns are answer-words because it targets the specific details of the image. Moreover, augmented dataset consists of high image fidelity that leads to degrading the generality of the model.

In the domain of image generation, the diffusion model [19] has received a huge amount of attention in the research area. The diffusion model defines a Markov chain of diffusion steps to slowly add random noise to data and then learn to reverse the diffusion process to construct desired data samples from the noise. This process is mathematically traceable and can be flexibly applied to various datasets, unlike the existing image generation model like GAN, having a trade-off between fidelity and diversity. Therefore, the diffusion model is spotlighted as a new framework that satisfies tractability and flexibility. DDPM [20] improves performance by adding a denoising step. DDIM [21] solves the disadvantage that the diffusion model takes a long time. DALLE-2 [22] and IMAGEN propose text-to-image models by using the diffusion model with the large pre-trained language model. Stable-diffusion [23] shows that the text-guided image generation model can be extended to diverse domains guaranteeing a certain level of quality.

In this paper, we explore the diffusion model, which is pre-trained on diverse tasks and numerous images, to generate the image not only to maintain the core information about the answer but also to expand the diversity of images and to inject prior knowledge into the VQA model. In addition, we offer a Plug-and-Play modular design for generating images about VQA that make it easy to leverage rapid advances in the text-guided diffusion model. In other words, our framework is designed to use the diffusion model without fine-tuning with specific purposes. First, the diffusion model is pre-trained with a large amount of data and various distributions, allowing generated images to expand the knowledge base through various appearance concepts of a single image without any fine-tuning. In contrast, the generated images from GAN have a similar distribution to the original dataset due to minimal editing. Second, the diffusion model with text prompts can generate abstract images well for words like ‘romantic’, which can not be expressed as a small part of the image. We demonstrate that VQA models trained on such data, with no exposure to external language information at all, exhibit high performance on VQA tasks requiring a factual knowledge base. Also, we propose an efficient training method to determine how many images are used regarding the difficulty of the question . All our proposed models show better performance than competitive baselines on FVQA [8] and DAQUAR [2]. In particular, our best model obtains performance improvement more than twice compared to the strong baseline on FVQA.

Our contributions are as follows:

- We propose a framework that integrates dataset composition through image generation by diffusion model to a training strategy for efficient use of the corresponding augmented data.

- We design the training strategy with data augmentation for VQA, allowing the model to deal with the difficulty of questions.

- For the first time, we prove that image generation can inject the knowledge base into VQA dataset because the diffusion model can be guided by text and has diverse results with guaranteed quality. In particular, we improve the performance at FVQA more than twice, which requires factual knowledge.

- We conduct an extensive experiment on DAQUAR and FVQA datasets. Experimental results show that our proposed methods are effective.

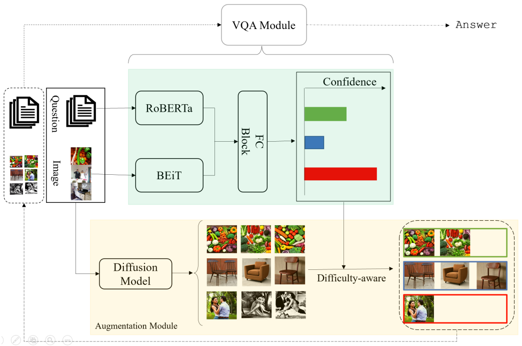


Fig. 1. overall framework of our proposed approach.

2. PRELIMINARY

2.1. Task: Visual Question Answering

We denote the VQA dataset as N triplets $\mathbb{D} = \{I_i, Q_i, A_i\}_{i=1}^{N}$, where for each image $I_i \in \mathbb{I}$, question $Q_i \in \mathbb{Q}$, and answer $A_i \in \mathbb{A}$. The goal of the VQA task is predicting the gold answer distribution $\mathbb{A}$ with given image $\mathbb{I}$ and question $\mathbb{Q}$:

$$P(A_i|I_i, q_i) = f_{vqa}(I_i, Q_i), \tag{1}$$

where f_{vqa} refers to the neural network model for VQA.

3. MODEL

We present our approach to utilizing the diffusion model for VQA in detail. We first present the overall framework and then introduce how to use the text-to-image diffusion model for generating VQA images and training strategy considering the difficulty of questions.

3.1. Framework

Fig. 1 provides an overview of our approach. Given a pair of questions and images, we first extract the embedding of each modality by a pre-trained model. And then, to integrate the different types of information, two embeddings are concatenated and processed by a subsequent FC block, consisting of several fully connected layers, dropout layers, and activation functions. The neural network, including FC block and pre-trained models, is trained to predict the golden answer of a given question-image pair. All the above steps are considered as the base VQA module.

At the same time, we generate the synthetic images by diffusion model with text prompt derived from question-answer pair. The probability of a golden answer predicted by the above-mentioned base VQA model is used to measure the difficulty of each question, as confidence. It determines how many generated images are used for each question. Such filtered images are appended to existing dataset as triplet $\{\mathbb{I}', \mathbb{Q}, \mathbb{A}\}$ to form new dataset $\mathbb{D}'$. Then, the final VQA module is trained by the new dataset $\mathbb{D}'$. The detailed prompt generation method and threshold of confidence for the difficulty-aware training strategy are described below subsection. The overall algorithm consists of five steps:

1. Train base VQA model M_{vqa} on raw dataset $\mathbb{D}$.
2. Generate synthetic image set $\mathbb{I}_{gen}$ by image generator M_{gen} with text-prompt from $\mathbb{Q} - \mathbb{A}$ pair.
3. Calculate the confidence of each question by M_{vqa}.
4. Filter $\mathbb{I}_{gen}$ with confidence to obtain $\mathbb{I}'$.
5. Train final VQA model on the new dataset $\mathbb{D}'$ where $\mathbb{D}$ is extended by $\mathbb{I}'$.

3.2. Text-to-Image Diffusion Model for VQA

Stable-diffusion [23] shows the outstanding result for text-guided image generation to the best of our knowledge. Therefore, we select it to make the synthetic images for VQA. The diffusion model is leveraged by only receiving text prompts without the original image to generate diverse images which share the core semantics but have a difference in superficial appearance from the original image. The text prompt used to induce the image is derived from the question-answer pair by replacing the interrogative word with an answer. If a question is not have any interrogative word, we empirically the first two words of the question with an answer token. Finally, the diffusion model can generate multiple images per question by text prompt. Formally, Each image can be derived as:

$$\mathbb{I}_{gen} = M_{gen}(prompt|\mathbb{Q}, \mathbb{A}), \tag{2}$$

where M_{gen} means the model of image generator and $\mathbb{I}_{gen}$ denotes the generated generated images.

3.3. Difficulty-aware training strategy

In this section, we discuss an effective training method to enable the model to achieve deeper reasoning by extending coverage of image examples of a difficult question when the question-image pair needs common sense to solve the problem. The base VQA model trained by the original dataset is utilized to predict the confidence of each question by calculating the probability of the golden answer. The lower confidence, the more difficult the model is to solve. Therefore, more image examples for difficult questions are included when constructing the new dataset. The dataset can be expanded more on difficult questions requiring the knowledge base by appending the generated images, which have the core concept of the golden answer. We divide the questions into N classes based on confidence and add a different number of image examples to the new dataset depending on the difficulty. We empirically set $N = 3$ and add 5, 3, and 1 images in the order of difficulty for each class. Formally, it can be described as below:

$$\begin{aligned} P(\mathbb{A}) &= M_{vqa}(\mathbb{Q}, \mathbb{A}), \\ \mathbb{I}' &= filter(\mathbb{I}_{gen}|P(\mathbb{A})), \end{aligned} \tag{3}$$

where M_{vqa} means the base VQA model trained by the original dataset, $P(\mathbb{A}))$ means the probability of a golden answer predicted by M_{vqa} and $\mathbb{I}'$ denotes the filtered image set by confidence for consisting of the new dataset.

4. EXPERIMENT

Approach	Evaluation Benchmark			
	DAQUAR		FVQA	
	Acc	F1	Acc	F1
Baseline	26.51	5.39	14.98	3.11
+ Diffusion Aug. 1	**27.55**	5.65	18.62	4.58
+ Diffusion Aug. 3	26.97	5.72	18.91	4.66
+ Diffusion Aug. 5	23.45	4.98	19.22	5.11
Diffusion Only. 5	13.24	2.79	16.53	4.63
Difficulty-aware Aug.	27.25	**6.01**	**19.61**	**5.31**

Table 1. Performance comparison on DAQUAR and FVQA.

Dataset	Image Type	Inception Score (IS)
DAQUAR	Base	1.923
	+ Generated	**2.483**
FVQA	Base	1.942
	+ Generated	**2.279**

Table 2. Inception score on the original and augmented dataset.

We empirically validate the effectiveness of difficulty-aware image generation for VQA based on diffusion model, using two widely adopted benchmark datasets - DAtaset for QUestion Answering on Real-world images(**DAQUAR**) and Fact-based Visual Question Answering(**FVQA**).

4.1. Implement Details

We apply BEiT [1] [24] and RoBERTa [2] [25] with a default setting, which show outstanding performance among pre-trained models, to extract the features of each modality and consist of a strong baseline. The FC block consists of 2048 fully-connected layer and 1024 fully-connected layer with 0.5 dropout and GELU activation function. We use AdamW optimizer with 5e-05 learning rate for training, and batch size is set to 32. The stable diffusion [3] [23] is used, without any fine-tuning, to generate images by text prompt from question-answer pair. We follow the original setting of models mentioned above for the other settings. In our setting, the diffusion model generates 6 images in 10 seconds with one V100 GPU. Additionally, we follow widely used task setting for VQA on [12, 4, 5] that formulate VQA as a multi-label classification task.

[1] https://huggingface.co/microsoft/beit-large-patch16-224-pt22k

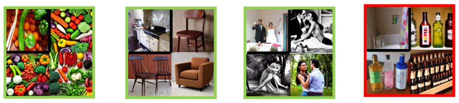
[2] https://huggingface.co/roberta-large

[3] https://huggingface.co/CompVis/stable-diffusion-v1-4

4.2. Performance Evaluation

The results are reported in Table 1.Each row means how many generated images from the diffusion model are appended to the dataset. All our proposed methods outperform the strong baseline except the one case. Especially the difficulty-aware data augmentation approach shows the best F1 score and accuracy on FVQA and the best F1 score on DAQUAR. The more generated images are included in the training data, the better the performance. It shows that generated images are adequate for VQA by giving diverse examples for each question. However, experiments that only such generated images resulted in poor performance. This is because the diffusion model is utilized without any fine-tuning to obtain different data distributions from existing dataset. In other words, the generated images can inject knowledge bases into the model with various visual information of the same image, but it has an independent data distribution that is not specific to the task. Therefore, the original image dataset is required to achieve good performance. Moreover, the worst performance is acquired on the DAQUAR dataset with five generated images. We analyze this phenomenon with qualitative examples in sec. 5.2.

Since the performance gap is much more significant on FVQA dataset, where the questions demand the knowledge base to solve, it implies that generated images, having core semantics on the golden answer but having different aspects, can expand the knowledge base. Additionally, this means that the knowledge base from diffusion model is conveyed in the form of an image where the model is pre-trained with a wide range of data. Also, the average number of augmented images of difficulty-aware augmentation is three, showing more remarkable performance improvement than the method of generating the same average number of images for all questions. It means our difficulty-aware training is important to increase performance.



Question	What sort of food can you see in this image?	What is the chair color?	What kind of property does the action?	How many bottles are there?
Answer	Vegetables	Brown	Romantic	3
Prediction	Vegetables	Brown	Romantic	2

Fig. 2. Qualitative examples of our proposed method.

5.1. Evaluating Image

To measure the quality and diversity of the generated dataset at once, Table 2 shows the Inception Score (IS) for each DAQUAR and FVQA dataset. In this experiment, we include five generated images in the augmented dataset to clearly compare the effect of images generated by the diffusion model to the existing dataset. On both benchmark datasets, higher IS is recorded in the augmented dataset. It means our proposed method increases the quality and diversity of the dataset by synthetic images.

5.2. Qualitative examples

Fig. 2 shows the qualitative examples of our proposed method. The green border refers to an example in which the prediction is wrong in a conventional way, but the prediction is changed to a golden answer in the way we proposed. The red example refers to an example in which the prediction is wrong due to our proposed method. The first image(left top) is the original image in the existing dataset. The others are generated images. In the first example, various vegetables that do not exist in the original image can be identified, like an onion cut in half. Second, we can get the diverse style of chairs with brown color. When creating an image about romantic, the styles of the generated images are entirely different from the original image. The original image is a wedding photo with a cake. However, the created images are related to romantic, such as kisses and hugs, which do not appear in the original image. This can be seen in the third example. The last example is the failure case of our proposed method. The generated images describe the various type of bottles but fail to draw exactly three bottles. The phenomenon of addressing specific object counts and location is a limitation of the diffusion model pointed out by [22]. This is why one of our proposed methods shows the worst performance on DAQUAR, where a lot of 'how many' questions exist. Therefore, when a large number of generated images are augmented, the number of images that do not match the golden answer increases, resulting in a decrease in performance. To handle above problem, our proposed method is based on the fact [12] that traditional VQA models solve well, unless it is a problem that requires a generalized knowledge base obtained through various appearances of objects.

6. CONCLUSION

In this paper, we show that image generation induced from question-answer pairs is useful for VQA training data. Especially the diffusion model is adopted as an image generator that generates diverse and high-fidelity data. Furthermore, we demonstrate that the knowledge of the large-scale pre-trained model can be transformed into the form of an image. Therefore, the model trained with the generated images can expand the knowledge about the question and achieve the best performance compared with a solid baseline. Also, we propose a difficulty-aware training strategy that differentially appends generated images to a new dataset, depending on the difficulty of the problem. From this, our best model obtains performance improvement more than twice compared to the strong baseline on FVQA where the dataset acquires the knowledge base to solve the problem. Overall, our work, handling data augmentation to training strategy, demonstrates that image generation can provide an efficient approach than raw dataset to build a more powerful VQA model with a more profound knowledge base and understanding.

7. REFERENCES

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh, "VQA: Visual Question Answering," in *ICCV*, 2015.

[2] Mateusz Malinowski and Mario Fritz, "A multi-world approach to question answering about real-world scenes based on uncertain input," in *NIPS*, 2014.

[3] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick, "Microsoft coco: Common objects in context," *ArXiv*, vol. abs/1405.0312, 2014.

[4] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi, "Ok-vqa: A visual question answering benchmark requiring external knowledge," in *CVPR*, 2019.

[5] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh, "Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering," in *CVPR*, 2017.

[6] Peng Zhang, Yash Goyal, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh, "Yin and Yang: Balancing and answering binary visual questions," in *CVPR*, 2016.

[7] Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, and Ji-Rong Wen, "Counterfactual VQA: A cause-effect look at language bias," *CoRR*, 2020.

[8] Peng Wang, Qi Wu, Chunhua Shen, Anton van den Hengel, and Anthony R. Dick, "FVQA: fact-based visual question answering," *CoRR*, vol. abs/1606.05433, 2016.

[9] Peng Wang, Qi Wu, Chunhua Shen, Anton Hengel, and Anthony Dick, "Explicit knowledge-based reasoning for visual question answering," *CoRR*, 2015.

[10] Mandy Guo, Qinlan Shen, Yinfei Yang, Heming Ge, Daniel Matthew Cer, Gustavo Hernández Ábrego, Keith Stevens, Noah Constant, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil, "Effective parallel corpus mining using bilingual sentence embeddings," in *WMT*, 2018.

[11] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut, "Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning," in *ACL*, 2018.

[12] Soravit Changpinyo, Doron Kukliansky, Idan Szpektor, Xi Chen, Nan Ding, and Radu Soricut, "All you may need for vqa are image captions," in *NAACL*, 2022.

[13] Jialin Wu, Jiasen Lu, Ashish Sabharwal, and Roozbeh Mottaghi, "Multi-modal answer validation for knowledge-based VQA," *CoRR*, 2021.

[14] Diego Garcia-Olano, Yasumasa Onoe, and Joydeep Ghosh, "Improving and diagnosing knowledge-based visual question answering via entity enhanced knowledge injection," WWW '22.

[15] Kailai Zhang, Zheng Cao, and Ji Wu, "Circular shift: An effective data augmentation method for convolutional neural network on image classification," *ICIP*, 2020.

[16] Aisha Al-Sadi, Hana' Al-Theiabat, and Mahmoud Al-Ayyoub, "The inception team at vqa-med 2020: Pretrained vgg with data augmentation for medical vqa and vqg," in *CLEF*, 2020.

[17] Zeyd Boukhers, Timo Hartmann, and Jan Jürjens, "Coin: Counterfactual image generation for vqa interpretation," *ArXiv*, vol. abs/2201.03342, 2022.

[18] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, "Generative adversarial networks," 2014.

[19] Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," *CoRR*, vol. abs/1503.03585, 2015.

[20] Jonathan Ho, Ajay Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *ArXiv*, vol. abs/2006.11239, 2020.

[21] Jiaming Song, Chenlin Meng, and Stefano Ermon, "Denoising diffusion implicit models," *ArXiv*, vol. abs/2010.02502, 2021.

[22] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen, "Hierarchical text-conditional image generation with clip latents," *ArXiv*, vol. abs/2204.06125, 2022.

[23] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and BjÃ¶rn Ommer, "High-resolution image synthesis with latent diffusion models," in *CVPR*, 2022.

[24] Hangbo Bao, Li Dong, and Furu Wei, "Beit: BERT pre-training of image transformers," *CoRR*, vol. abs/2106.08254, 2021.

[25] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov, "Roberta: A robustly optimized BERT pretraining approach," *CoRR*, vol. abs/1907.11692, 2019.