StackOverflowVQA: Stack Overflow Visual Question Answering Dataset

Motahhare Mirzaei, Mohammad Javad Pirhadi and **Sauleh Eetemadi**
Iran University of Science and Technology at Tehran, Iran
{m_mirzaei96, mohammad_pirhadi}@comp.iust.ac.ir, sauleh@iust.ac.ir

Abstract

In recent years, people have increasingly used AI to help them with their problems by asking questions on different topics. One of these topics can be software-related and programming questions. In this work, we focus on the questions which need the understanding of images in addition to the question itself. We introduce the StackOverflowVQA dataset, which includes questions from StackOverflow that have one or more accompanying images. This is the first VQA dataset that focuses on software-related questions and contains multiple human-generated full-sentence answers. Additionally, we provide a baseline for answering the questions with respect to images in the introduced dataset using the GIT model. The dataset and codes will be publicly available.

1 Introduction

Visual Question Answering (Antol et al., 2015) (VQA) aims to answer a question based on an image. Multi-modal generative models and in particular VQA has received a lot of attention from the research community in recent years and has improved significantly (Chen et al., 2023; Li et al., 2022; Sammani and Deligiannis, 2023; Wang et al., 2021). On the other hand, assisting programmers has also received a lot of attention (e.g., Friedman, 2021 introduced GitHub Copilot). Using VQA models, a computer assistant can use the screenshot as the context for a question and better help users. We introduce the StackOverflowVQA dataset which contains the questions from StackOverflow which has at least one image in their question and one answer. Although single or multi-word answers can be useful in some scenarios, for a computer assistant full sentences or even multiple sentences are often required. However, to the best of our knowledge, datasets with full sentence answers for VQA (Shin et al., 2016; Sheikhi et al., 2022) have only been artificially synthesized using single or multi word answer datasets such as VQA 2.0 (Goyal et al., 2017). StackOverflowVQA is the first large scale human generated full sentence VQA dataset. Training and evaluation of generative tasks such as Machine Translation, Question Answering and VQA is challenging since the same meaning can be expressed in various ways but only a single answer is often available for each question. Evaluation metrics such as BLEU are also much more accurate when multiple answers are provided. StackOverflowVQA is the first VQA dataset that provides multiple full sentence answers for most questions. Hence this dataset is the best dataset available for training and evaluation of a visually aware computer assistant. While the main contribution of this work is construction of the dataset, we use GIT (Wang et al., 2022) as a state of the art multi-modal generative model which can be used for image/video captioning/question-answering or even image classification. We fine-tune this model on our proposed dataset as a baseline and report the results.

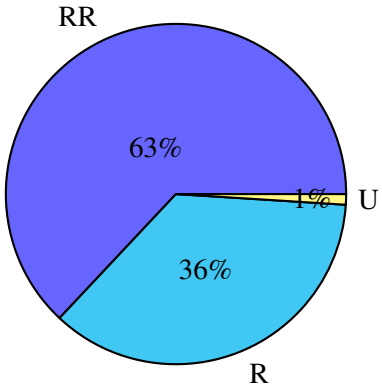


Figure 1: Distribution of images with respect to relatedness and requirement. (RR=Related and required, R=Related but not required, U=Unrelated)

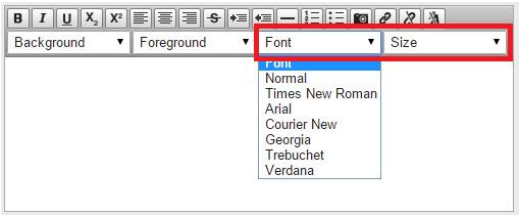
2 Dataset

We develop StackOverflowVQA by processing and filtering the Mike (2023) dataset which is available

Data	B-1	B-2	B-3	B-4	R-1	R-2	R-L	R-LSUM
StackOverflowVQA-filtered-small	33	18	11	6.8	31	5.8	23	28

Table 1: All the results are calculated using the test set of the StackOverflowVQA-filtered-small dataset. (B=BLEU, R=ROUGE)

(a) Q: I want to have a box in HTML such as this one: <IMAGE> [rest of question]



(b) Q: I have been trying to add custom fonts and custom sizes in select boxes provided with the 'RichTextArea' of Vaadin. <IMAGE> How do I do this?

Figure 2: Examples of a required image (left) and a related image that is not required to answer the question (right).

Indicator	All	F	FS
#Q	451,394	204,041	20,458
#A	799,270	204,041	20,458
#QwAA	287,050	204,041	20,458
#WpQ-min	8	3	3
#WpQ-max	26,948	10,380	425
#WpQ-mean	299.5	267.0	122.1
#WpQ-std	374.4	341.9	69.7
#WpA-min	0	0	0
#WpA-max	8,389	8,377	425
#WpA-mean	151.3	167.2	87.7
#WpA-std	208.3	227.0	62.7
#ApQ-min	1	1	1
#ApQ-max	132	1	1
#ApQ-mean	1.8	1	1
#ApQ-std	1.5	0	0
#PApQ-min	0	1	1
#PApQ-max	127	1	1
#PApQ-mean	1.1	1	1
#PApQ-std	1.3	0	0
#IpQ-min	1	1	1
#IpQ-max	33	1	1
#IpQ-mean	1.4	1	1
#IpQ-std	0.8	0	0

Table 2: Datasets' statistics (F=Filtered, FS=Filtered-Small, Q=Question, A=Answer, QwAA=Questions with Accepted Answer, WpQ=Word per Question, WpA=Word per Answer, ApQ=Answers per Question, PApQ=Positive-scored Answers per Question, IpQ=Images per Question)

on Hugging Face. The dataset has 58.3M rows including questions, answers, and some wiki-related posts submitted to StackOverflow before June 14th of 2023 formatted as Markdown text. The data is sourced from StackExchangeCommunity (2023). The total number of questions in this dataset is 23,536,500.

This dataset can be challenging because the model has to have programming knowledge as well as an understanding of the image contents. Also, the questions and the answers can be very long (See #WpQ and #WpA in Table 2).

2.1 StackOverflowVQA

StackOverflowVQA includes all questions that have at least one image as well as their answers (at least one answer) which has 451,394 questions with their corresponding answers (799,270 answers in total).

To understand whether images are essential to answering questions, we select 100 random samples from StackOverflowVQA and manually examine whether the accompanying image is required to answer the question. One percent of the images are memes or similar unrelated images while 63 percent of the samples have images that are required to answer the question correctly. While the remaining 36 percent are related images but the questions can be answered correctly without them (Figure 1 shows the distribution and figure 2 shows examples of an image that is required for answering the question and one that is not required).

2.2 StackOverflowVQA-filtered

This version of the dataset only includes the questions which have only one image and have an accepted answer. Also, we download the images, and if the image of a question can not be downloaded, we omit that question as well. This dataset contains 204,574 question-answer pairs which have been split to train and test using 90% and 10% of the samples respectively.

2.3 StackOverflowVQA-filtered-small

Due to the unavailability of resources, we use a subset of the filtered version of the dataset to fine-tune the GIT model. This subset has only 10% of the filtered dataset and the question + answer tokens of it are not more than 512. This dataset has 18,412 and 2,046 question-answer pairs for training and testing respectively.

3 Model

We use GIT as the baseline model. This model is trained on general images and is not ideal for images in our dataset because they are mostly computer screenshots. Also, the texts are programming questions and answers while the model is trained on general text. So, this dataset is out of domain data for the model.
Furthermore, It is also not suitable for the texts because it has a limit of 512 tokens but with our StackOverflowVQA-filtered-small dataset, this problem can be solved.
The model gets the image and the question as the input and generates the answer using them.

4 Training

We first split 10% of the training set as the validation set and use it to find the optimal number of epochs. Then, we use the whole training set to fine-tune the model and evaluate it using the test set. We use a gradient clip with the value of 12, an AdamW optimizer with default hyper-parameters and a learning rate of 5e-5, a cosine schedule without warm-up steps, and a batch size of 8 (due to limited resources).

5 Results

Table 1 shows the results. The results show that the introduced dataset is challenging for even a state-of-the-art model like GIT. The low scores for more n-grams with $n > 1$ show that the model does not have any programming knowledge to produce a coherent answer.
Also, these results can show the inefficiency of the evaluation metrics for such tasks and the need to design a better evaluation metric.
For sure, training the model with the whole dataset can improve the results.

6 Future Work

The purpose of this work is to attract researchers to a new application of visual question answering. One can do the following suggested actions to improve the results.

- The baseline accepts a fixed number of images and because of that we selected the samples which have only one image. This limitation reduces the accuracy of the model.
- The baseline has been trained on general images and text but the images of the proposed dataset are mostly screenshots and texts are programming questions and answers. A pre-training on such image-text pairs can significantly improve the results.
- Most of the StackOverflow questions have more than one answer which in most cases are relevant and can help the model to gain programming knowledge.
- Also, the text-only question-answer pairs can teach the model programming.
- The position of the images in the question can be important especially when we have more than one image.
- Some of the answers have images as well, future work can work on generating the images in answer.

7 Conclusion

In this work, we introduce a new dataset called StackOveflowVQA which consists of StackOverflow questions that have at least one image in their body. This can be used to train a model to answer the programming questions and use screenshots or other related images to help the model improve its accuracy. Also, we use the GIT model as a baseline on a small subset of this dataset.

References

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: visual question answering. *CoRR*, abs/1505.00468.

Chongyan Chen, Samreen Anjum, and Danna Gurari. 2023. Vqa therapy: Exploring answer differences by visually grounding answers.

Nat Friedman. 2021. Introducing github copilot: your ai pair programmer. *URL https://github.blog/2021-06-29-introducing-github-copilot-ai-pair-programmer*.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Chenliang Li, Haiyang Xu, Junfeng Tian, Wei Wang, Ming Yan, Bin Bi, Jiabo Ye, Hehong Chen, Guohai Xu, Zheng Cao, Ji Zhang, Songfang Huang, Fei Huang, Jingren Zhou, and Luo Si. 2022. mplug: Effective and efficient vision-language learning by cross-modal skip-connections.

Mike. 2023. Mikex86/stackoverflow-posts · datasets at hugging face.

Fawaz Sammani and Nikos Deligiannis. 2023. Uni-nlx: Unifying textual explanations for vision and vision-language tasks.

Hadi Sheikhi, Maryam Hashemi, and Sauleh Eetemadi. 2022. Generate answer to visual questions with pre-trained vision-and-language embeddings. *WiNLP Workshop at EMNLP*.

Andrew Shin, Yoshitaka Ushiku, and Tatsuya Harada. 2016. The color of the cat is gray: 1 million full-sentences visual question answering (fsvqa). *arXiv preprint arXiv:1609.06657*.

StackExchangeCommunity. 2023. Stack exchange data dump.

Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. 2022. GIT: A generative image-to-text transformer for vision and language. *Transactions on Machine Learning Research*.

Wenhui Wang, Hangbo Bao, Li Dong, and Furu Wei. 2021. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. *CoRR*, abs/2111.02358.