Visual Question Answering with Multimodal Learning for VizWiz-VQA

Heegwang Kim[†,1] Chanyeong Park[†,1] Joonbo Jang[2] Jiyoon Lee[2] Jaehong Yoon[1] [*]Joonki Paik[1,2]
[1]Department of Image [2]Department of Artificial Intelligence
IPIS Lab, Chung-Ang University
{heegwang, chanyeong, junbojang, jylee67, jhyoon0101, paikj}@ipis.cau.ac.kr

Abstract

Visual Question Answering has emerged as a prominent task alongside the development of vision-language models. In this study, we utilize pre-trained image and text encoders from BeiT-3 to develop a VQA model for the VizWiz-VQA dataset. To address the dataset's characteristic of multiple answers per question, we created an answer vocabulary by consolidating frequently occurring answers across the dataset, resulting in a vocabulary of 6,484. On the VizWiz-VQA 2024 challenge we achieve 66.70 accuracy on Predict Answer to a Visual Question task.

1. Introduction

Vision-Language Pre-training (VLP) task has recently achieved significant success across various multi-modal downstream tasks, including image retrieval, image captioning, and visual question answering. Previous VLP approaches have heavily relied on image feature extraction, necessitating region supervision such as object detection and the use of CNN-based architectures. However, these methodologies are limited by the capabilities of the feature extractor and the predefined visual vocabulary, thereby impacting performance, processing speed, and feature representations. ViLT [4] markedly simplifies visual input processing in a convolution-free manner, thus unifying the treatment of visual and textual inputs. In BLIP [5], the issue of noisy captions in web data, prevalent in existing VLP models, is mitigated using CapFilt. Furthermore, a novel model architecture named MED is introduced, demonstrating proficiency in both text generation and image-text retrieval tasks. In BeiT-3 [6], a multiway transformer architecture is proposed and pre-trained using inputs of images, text, and image-text pairs.

The VizWiz dataset [2, 3], sourced from visually impaired individuals, deviates from conventional VQA datasets in format. It comprises videos directly captured by visually impaired individuals along with corresponding questions recorded as voice inputs. Consequently, instances exist where answers cannot be provided due to poor video quality or the absence of the inquired object. Additionally, some questions are improperly curated or invalid inquiries. Accordingly, our learning strategies are tailored to accommodate the unique characteristics of the VizWiz dataset. We adopt pre-trained image and text encoders from BeiT-3 and solely train a straightforward classification head.

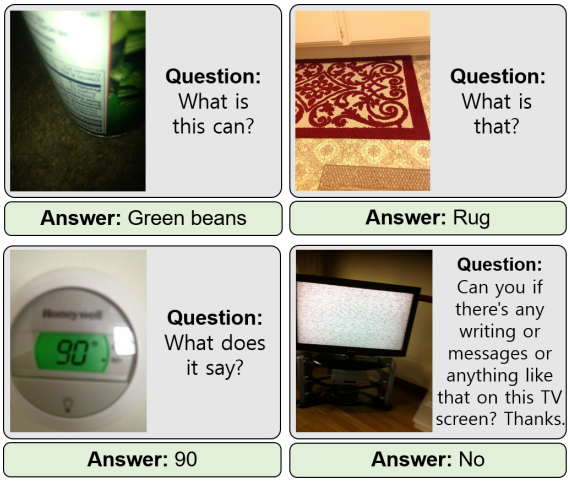

Figure 1. The proposed model results on VizWiz-VQA test set.

2. Methodology

The VizWiz-VQA dataset [3] consists of 20,523 train, 4,319 validation, and 8,000 test samples. Within the dataset annotations, an **answer_type** is defined, providing insight into the form of answers for respective questions. In the train set, categorized by **answer_type**, there are 957 yes/no, 5,532 unanswerable, 301 number, and 13,733 other questions. Similarly, in the validation set, there are 195 yes/no, 1,385 unanswerable, 195 number, and 2,691 other questions. Given the scarcity of question-answer pairs categorized as "number", only those categorized as such from

[*]† Co-first authors * Corresponding author

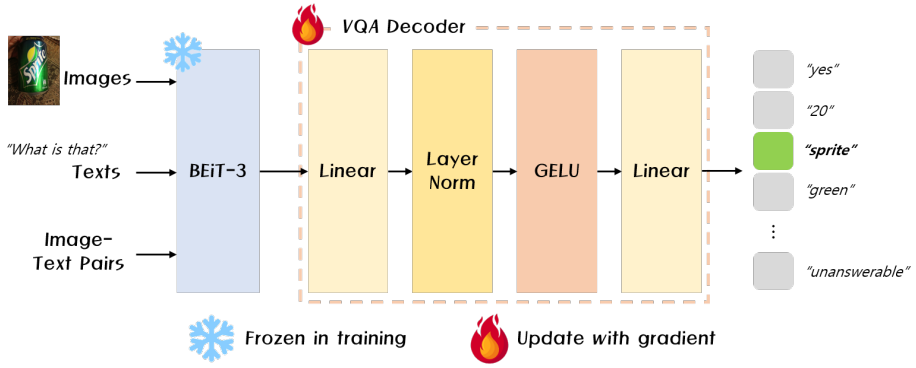

Figure 2. Our model for the VizWiz-VQA

the VQAv2 dataset [1] were included in the training data. Furthermore, due to the characteristic of the VizWiz-VQA dataset wherein multiple answers may correspond to a single question, potentially unrefined, experimentally, answers appearing six or more times across the entire dataset were organized into an answer vocabulary, resulting in a vocabulary size of 6,484.

3. Experimental Results

Table 1 shows the quantitative results of our method with other challenge models on the VizWiz test set. Our model shows good inference in answering yes/no questions and unanswerable questions. However, it can be observed that its performance drops in number questions and other types of questions. Our model achieved a ranking of 4th place on the challenge leaderboard with an overall result of 66.7.

4. Conclusion and Future Work

The results stems from setting an answer vocabulary and inferring responses within it. Hence, our model demonstrates good performance in tasks that require deriving predefined answers, such as yes/no or unanswerable questions. However, its performance decreases when faced with questions requiring inferences for answers not present in the vocabulary, such as number or other questions. Therefore, furhter research should be utilizing learnable parameters to generate answers, rather than classifying answers within a predefined closed answer vocabulary.

Table 1. VizWiz-VQA 2024 challenge results.

Accuracy	yes/no	number	other	unanswerable	overall
SLCV	90.54	75.20	70.62	98.02	78.91
v1olet	89.14	62.20	66.13	97.66	75.54
KNU	80.95	55.77	56.32	93.79	67.45
Ours (test-std)	86.93	44.23	53.86	96.87	66.70
Ours (test-dev)	89.78	47.62	54.52	97.15	67.77

Acknowledgments

This research was supported by Field-oriented Technology Development Project for Customs Administration through National Research Foundation of Korea(NRF) funded by the Ministry of Science & ICT and Korea Customs Service(2021M3I1A1097911).

References

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *International Conference on Computer Vision (ICCV)*, 2015. 2

[2] Danna Gurari, Qing Li, Abigale J. Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P. Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1

[3] Danna Gurari, Qing Li, Chi Lin, Yinan Zhao, Anhong Guo, Abigale Stangl, and Jeffrey P. Bigham. Vizwiz-priv: A dataset for recognizing the presence and purpose of private visual information in images taken by blind people. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1

[4] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International conference on machine learning*, pages 5583–5594. PMLR, 2021. 1

[5] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022. 1

[6] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. Image as a foreign language: Beit pretraining for all vision and vision-language tasks. *arXiv preprint arXiv:2208.10442*, 2022. 1