Technical Report for CVPR 2024 VizWiz Challenge Track 1-Predict Answer to a Visual Question

Jinming Chai, Qin Ma, Kexin Zhang, Zhongjian Huang, Licheng Jiao, Xu Liu
Intelligent Perception Image Understanding Lab, Xidian University
{23171214444, 19200100144}@stu.xidian.edu.cn

Abstract

In this technical report, we briefly introduce the solution of our team ”chirmy” for VizWiz Challenge Track 1-Predict Answer to a Visual Question in CVPR 2024. In order to predict the answer to a visual question, we propose the following plan: Firstly, the training set and validation set are fused for training, and the dataset is deblurred. At the same time, motion blur data augmentation is performed on the data with anns-Type other; Classify the anns-Type in the test dataset by analyzing the frequency of prefix word combinations for train and val; Train and fine tune the ”yesno”, ”number”, ”other”, and ”unanswerable” parts of the dataset separately and perform inference tests, then fuse the inference results. Finally, adjust the fixed answer format of the results and integrate multiple versions of the reasoning results.

1 Introduction

This technical report describes our proposed solution for CVPR 2024 VizWiz Challenge Track 1-Predict Answer to a Visual Question. Predict Answer to a Visual Question in the Vizwiz Challenge at CVPR 2024 is part of the Vizwiz Challenge Workshop. The task is Given an image and question about it, the task is to predict an accurate answer. As shown in Figure 1, it shows a sample of Visual Question Answering task data, and each sample of training data is composed of images and corresponding questions and answers. Evaluation metric is the minimum between 1 and the number of people who provided the answer minus 1.

We propose the following plan: Firstly, the train- ing set and validation set are fused for training, and the dataset is deblurred. At the same time, motion blur data augmentation is performed on the data with anns-Type other; Classify the anns-Type in the test dataset by analyzing the frequency of prefix word combinations for train and val; Train and fine tune the ”yesno”, ”number”, ”other”, and ”unanswerable” parts of the dataset separately and perform inference tests, then fuse the inference results. Finally, adjust the fixed answer format of the results and integrate multiple versions of the reasoning results.

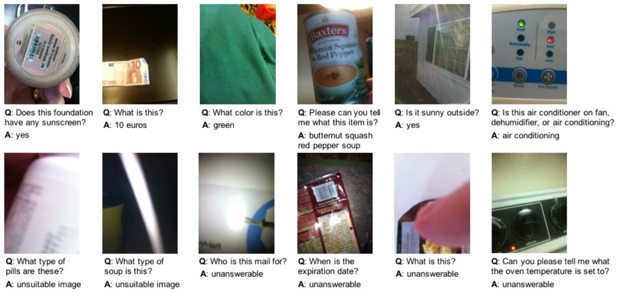
Q: Does this foundation have any sunscreen?
A: yes
Q: What is this?
A: 10 euros
Q: What color is this?
A: green
Q: Please can you tell me what this item is?
A: butternut squash red pepper soup
Q: Is it sunny outside?
A: yes
Q: Is this air conditioner on fan, dehumidifier, or air conditioning?
A: air conditioning
Q: What type of pills are these?
A: unsuitable image
Q: What type of soup is this?
A: unsuitable image
Q: Who is this mail for?
A: unanswerable
Q: When is the expiration date?
A: unanswerable
Q: What is this?
A: unanswerable
Q: Can you please tell me what the oven temperature is set to?
A: unanswerable

Figure 1: Visual question answering task data sample.

2 Method and Experiments

The proposed method contains four main steps, as follows:

A. Image pre-processing.
B. Annotations pre-processing.
C. Model train and inference.
D. Post-processing.

2.1 Image pre-processing

Firstly, using NAFNet-REDS to deblurring train/val/test. Through observation, it was found that the proportion of anns-Type=”other” class data in the dataset is close to 0.7. Therefore, data augmentation was performed on this type of data: the ASPDC[1] network was used to remove motion blur on the ”other” class data of train/val, and then these data were added to the training set.

2.2 Annotations pre-processing

The preprocessing of Annotations data can mainly be divided into the following two steps:

First, to statistically analyze the frequency of the first two keyword combinations (hereinafter referred to as prefix word combinations) in the question sentence of the train/val dataset, in order to discover the difference in the frequency of prefix word combinations between different anns-Types (”yesno”, ”number”, ”other”, ”unanswerable”).

Second, Using the difference in the frequency of prefix word combinations calculated in train/val to classify the test dataset into anns-Type: The anns-Type of the test is assigned the highest occurrence of its prefix word combination in the train/val dataset. If a prefix word combination appears in two or more train/val datasets with a frequency greater than 0.05, its anns-Type is assigned the value of ”unclassified”. By doing so, the test dataset will be assigned as ”yesno”, ”number”, ”other”, ”unanswerable”, ”unclassified” based on the question

2.3 Model train and inference

Inference uses the BLIP[2] network model to train and fine tune the ”yes no”, ”number”, ”other”, and ”unanswerable” parts of the dataset, and perform inference tests. Then, the inference results are fused. Perform BLIP2 model inference.

BLIP[2] introduces a multimodal mixture of encoder decoder (MED) structure, which can effectively perform multi task pre learning and transfer learning. MED consists of two single-mode encoders (lmage Encoder, Text Encoder), an image-grounded text encoder, and an image-grounded text decoder.

2.4 Post-processing

Format the fixed format answers in the reasoning results, such as replacing the time class ”22:34” with ”22:34”.

Result fusion: Integrate multiple inference results, prioritize using the result with the most occurrences of the same result but not the same, and use the result with the highest anns-Type score.

3 Conclusion

We propose the following plan: Firstly, the train- ing set and validation set are fused for training, and the dataset is deblurred. At the same time, motion blur data augmentation is performed on the data with anns-Type other; Classify the anns-Type in the test dataset by analyzing the frequency of prefix word combinations for train and val; Train and fine tune the ”yesno”, ”number”, ”other”, and ”unanswerable” parts of the dataset separately and perform inference tests, then fuse the inference results. Finally, adjust the fixed answer format of the results and integrate multiple versions of the reasoning results.However, there is still room for improvement in this method, such as VQA model and post-processing.

References

[1] D. Huo, A. Masoumzadeh, and Y.-H. Yang, “Blind Non-Uniform Motion Deblurring using Atrous Spatial Pyramid Deformable Convolution and Deblurring-Reblurring Consistency,” *arXiv e-prints*, p. arXiv:2106.14336, June 2021.

[2] J. Li, D. Li, C. Xiong, and S. Hoi, “BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation,” *arXiv e-prints*, p. arXiv:2201.12086, Jan. 2022.