Multiview Language Bias Reduction for Visual Question Answering

Pengju Li, Zhiyi Tan, and Bing-Kun Bao, *Nanjing University of Posts and Telecommunications, Nanjing, 210049, China*

Current visual question answering models overly rely on language bias and fail to understand visual information sufficiently. Many recent works concentrate on mitigating the intraquestion type bias (bias in the distribution of answers to a question type) without taking the interquestion type bias (bias in distribution between question types) into consideration, causing the model to ignore the tail question types. In addition, they neglect the overall distribution bias of the answer set, leading to the model only focusing on the head answers. In this article, we propose the Multi-View Language Bias Reduction (MVBR) method to solve these problems. For the interquestion type bias, we introduce the Inter-Question Type Bias (IQTB) module. IQTB exploits the question type distribution of the training set to determine the question type bias, which is used to generate weighting factors to reshape the loss of each question type into a balanced form. For the overall distribution bias of the answer set, we utilize the Decoupled Training (DT) module. The DT module penalizes the weights of each class in the classifier and forces their distribution to be balanced. Experimental results demonstrate that MVBR is effective on VQA-CP v2 and effectively improves performance on the mainstream models.

The Visual Question Answering (VQA) task, which aims to utilize the given image and question to predict the correct answer, plays an important role in language and visual understanding. Existing models[1] have achieved well performance on several large-scale benchmarks.[2,3] However, previous models overly rely on strong correlations between questions and answers, ignoring the information of the visual modality during training.[3] For example, when the question is "How many...", "2" accounts for a high proportion of the answer set, so the model can still have an excellent performance only by answering "2." Such problem is known as "language bias," which has become one of the main bottleneck in VQA studies and it is reflected on out-of-domain datasets,[4] where model performance drops significantly.

To address the language bias problem in VQA, many approaches focus on minimizing language priors in VQA models. These methods can be roughly divided into two categories: 1) Enhancing understanding of visual features:[5,6,7] they aim to improve the model's ability to understand images. By enhancing the understanding of image information, the model reduces the reliance on language priors. But the bias caused by language prior can only be eliminated in language modality, which cannot be solved by enhanced visual information. 2) Reducing language modality bias[8,9,10,11]: they aim to mitigate the statistical bias introduced in language modalities. By constructing a question-only branch, the VQA model eliminates the language modality bias, thus the model can understand visual and textual information.

Although some studies have considered reducing the language modality bias, their elimination of bias is not complete. As shown in Figure 3, the existing methods only consider the reduction of the intra-question type bias: bias in the distribution of candidate answers corresponding to a question type. By analyzing VQA dataset, we find that the question type distribution of dataset is unbalanced, as shown in Figure 2, leading to model ignoring the "tail"question type in the training process. We view this bias as interquestion type bias: bias in distribution between question types. For example, the question types "what color is" and "How many" have

1070-986X © 2022 IEEE
Digital Object Identifier 10.1109/MMUL.2022.3217307
Date of publication 26 October 2022; date of current version 10 May 2023.

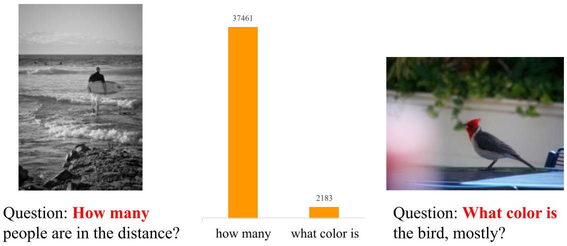

FIGURE 1. Example of question-type distribution in VQA-CP v2. Question type "How many" has 17 times as much data as question type "What color is." During the training process, the VQA model will pay too much attention to "How many" and "What color is" is far less concerned than "How many."

different numbers, as shown in Figure 1. Therefore, the model fails to pay sufficient attention to the question type "what color is" (tail question type) during training.

In addition, the overall distribution of the answer set is biased when the question type is not considered. We call this the overall distribution bias of the answer set, as shown in Figure 3. As a result, the model pays more attention to the head answers during training, which is detrimental to the robustness of the VQA model.

In this article, we propose a method named Multi-view language Bias Reduction (MVBR), which reduces the interquestion type bias and the overall distribution bias of the answer set effectively. MVBR[a] consists of two separate modules: Interquestion Type Bias module (IQTB) and Decoupled Training module (DT). First, IQTB is used to reduce the interquestion type bias. Specifically, IQTB extracts the distribution of the question types in training set to determine the interquestion type bias. Then, IQTB exploits this bias to generate a weighting factor for each question type, which reshapes the total loss. Compared with the loss before reshaping, the weights of the reshaped loss are more uniform for each question type, so the model will fairly consider each question type in the process of backpropagation. Second, inspired by Kang et al.[12] we introduce the Decoupled Training module (DT) to reduce the overall distribution bias in the answer set. In particular, we decoupled the training process into two stages: representation learning and classification. DT retrains the classifier by setting penalty weights for each class, which makes the weights of the classifier balanced.

Ablation experiments prove that both modules of the MVBR are effective. Qualitative and quantitative analyses also demonstrate the advantages of MVBR.

The contributions of this article are as follows:

- We extract the priors of the question types in the VQA dataset and exploit it to construct weighting factors to reduce the interquestion type bias.
- We introduce decoupled training in the VQA task to reduce the overall distribution bias of the answer set.
- Our method is tested on the VQA-CP v2 dataset, and achieves the state-of-the-art results on the CSS[7] model.

RELATED WORK

As the VQA task has received extensive attention,[1,2,13] many studies[3,4] found that the VQA models suffer from language bias, causing the model to predict answers blindly due to the language shortcut. In addition, object hallucination[14,15,16] is also found in the image capture task. This shows that many existing models for solving multimodal tasks can not well deal with the gap between visual and linguistic modalities. To explore the impact of language bias on the VQA model, the VQA-CP dataset[4] is proposed, where the distributions of questions and answers in test set are different from the distributions in training set. Recently, some works have been proposed to eliminate the language bias and these methods are roughly divided into two categories: 1) enhancing visual reasoning ability and 2) reducing reliance on language prior.

Enhancing Visual Reasoning Ability

Enhancing visual reasoning ability methods are designed to allow the model to pay more attention to the visual information, thereby reducing the model's reliance on language bias. Hint[5] introduced the annotated attention map that guides the model to reason visual information by optimizing the consistency between the annotated attention map and the VQA model. By exploiting a self-supervised approach, SSL[6] divides the images into question relevant and question irrelevant, and increase the sensitivity of the model to visual content. CSS[7] used Grad-CAM to divide the regions relevant to the question and the regions irrelevant to the question in an image. However, all the above methods focus on assisting the model to learn visual information, but do not reduce the bias of language modality,[17] and the influence of language prior still exists.

[a] https://github.com/LSWXXBC/MVBR

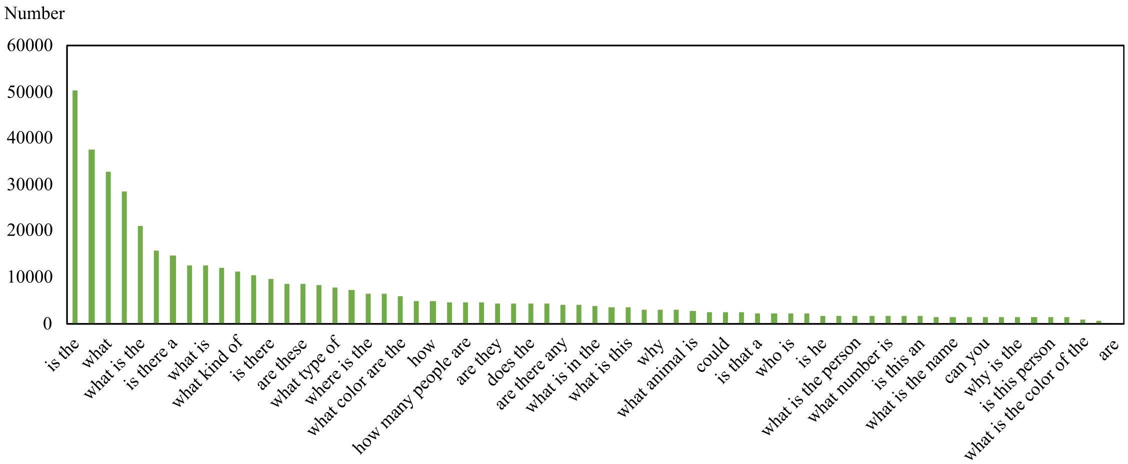

FIGURE 2. Distribution of question types in the training set. The question types distribution data presents a long tail distribution, which is unbalanced.

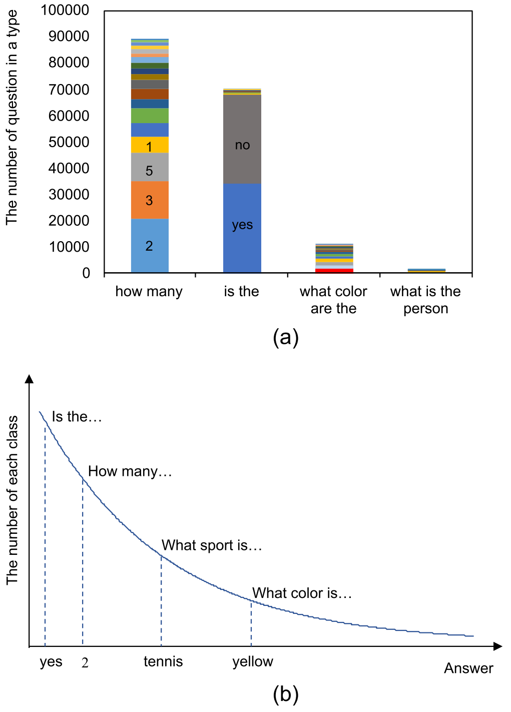

FIGURE 3. (a) We extracted four types of questions to explain the difference between **the intraquestion type bias** and **the interquestion type bias**. It can be seen that intraquestion type bias appears in a single question type that the answers have unbalanced distribution. However, the interquestion type bias appears between question types. (b) **The overall distribution bias of the answer set**. There are more than 2000 answer classes for the VQA-CP v2 dataset, but the answer classes with the top 200 distribution account for 70% of the data.

Reducing Reliance on Language Prior

At present, there are some effective solutions to reduce language modality bias. AReg[8] trained a question-only branch that shares encodings with the VQA question branch to capture language bias, but it affects the stability of the model. To solve this problem, Rubi[9] proposed a fusion training method that fuses the question-only branch's prediction with the answer predicted by the VQA model. LPF[10] directly used the prediction result of the question-only branch as the weighting factors to reshape the VQA loss. On the other hand, LMH[11] extracted training set priors and modified the biased model in the training process. Nevertheless, these methods only reduce intraquestion type bias, while ignoring the interquestion types bias and the overall distribution bias of the answer set. Motivated by those issues, we propose our MVBR method to solve these two problems.

METHODS

In this section, we briefly outline VQA models and detail the mechanisms of MVBR for reducing language bias. Figure 4 describes the structural framework of MVBR, which consists of two modules. The detail of each section is described in the following.

Fundamentals of VQA

VQA task is often viewed as a multiclassification task. The VQA model predicts the correct answer based on a given question Q and image I, which is expressed as follows:

$$a = \arg\ \max P(a \mid Q, I) \tag{1}$$

where P is the predicted result of the VQA model.

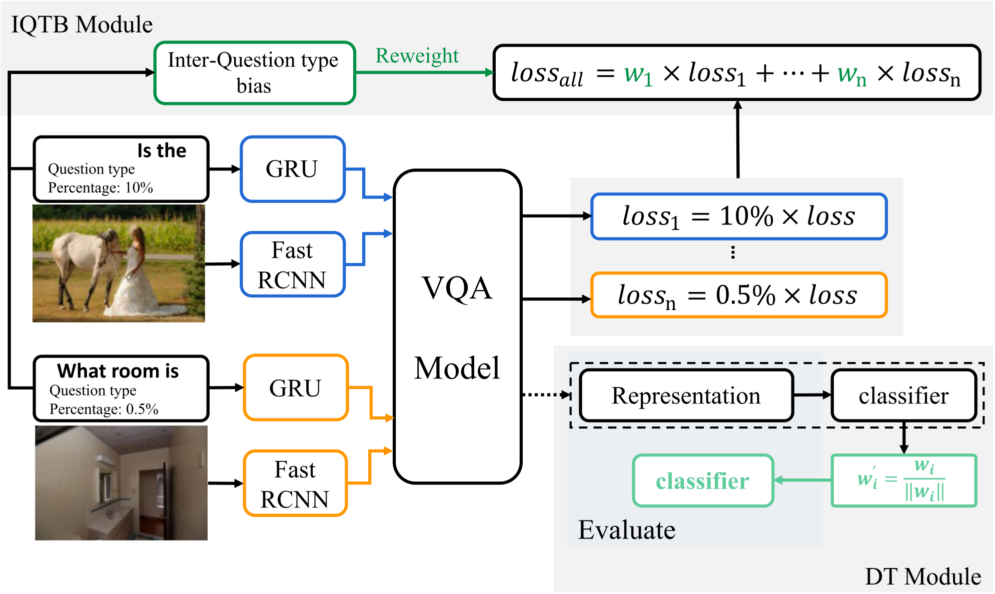

FIGURE 4. Overview of the MVBR structure. It consists of two parts: the IQTB module and the DT module. First, MVBR exploits the IQTB module to extract the bias in the training set for generating weighting factors to penalize the loss. Second, after the training is complete, MVBR utilizes the DT module to retrain the VQA model classifier to adjust the classifier weights.

Given a dataset D consisting of N triples $\{Q_i, I_i, A_i\}$, where $Q_i \in Q$ is the question, $I_i \in I$ is the image, and $A_i \in A$ is the answer. Each image I_i will be encoded into visual vectors $V \in \{v_1, v_2, \ldots, v_n\}$, where v_i is the i th object feature. Each question Q_i is encoded into question vectors $E \in \{e_1, e_2, \ldots, e_j\}$, where e_j is the jth word feature.

The VQA model fuses the visual vector and the question vector into R^d as the representation of the answer, then the model maps R^d to the answer space $R^{\|A\|}$ and selects the most probable answer. The process is represented as follows:

$$p_i(A \mid I_i, Q_i) = f_{vqa}(f_c(V, E)) \tag{2}$$

where f_c represents the function that convert V, E into R^d, and d represents the dimension of the joint representation. f_{vqa} represents the function that maps R^d to $R^{\|A\|}$, and $\|A\|$ represents the dimension of the answer space.

In the training stage, the VQA task usually uses the cross entropy loss function to train the model, which can be expressed as

$$\text{Loss} = \sum_{i}^{N} y_i \log(p_i) + (1 - y_i)\log(1 - p_i) \tag{3}$$

where y_i is the label corresponding to each answer a_i, $y_i \in \{0, 1\}$, p_i is the prediction result of the model, and N is the number of samples.

Multiview Language Bias Reducing

In this part, we introduce the interquestion type bias module and the decoupled training module of MVBR model.

The Interquestion Type Bias Module

In order to solve the problem caused by interquestion type bias, we propose the IQTB module to make the VQA model pay sufficient attention to each question type. As shown in Figure 5, this module consists of a question type bias extractor and a weighting factors generator, respectively.

In order to measure the interquestion type bias, one effective approach is to extract the distribution of each question type that can be written as

$$b_i = \frac{q_i}{\sum_i^n q_i} \tag{4}$$

where q_i represents the number of questions in a type and n represents the number of question types.

The role of the weighting factors generator is to generate a penalty factor for each question type. As the number of questions in a question type increases, the penalty should increase, so the weighting factors can be expressed as

$$w_i = (1 - a \times b_i)^r \tag{5}$$

where a and r are hyperparameters and b_i is the interquestion type bias.

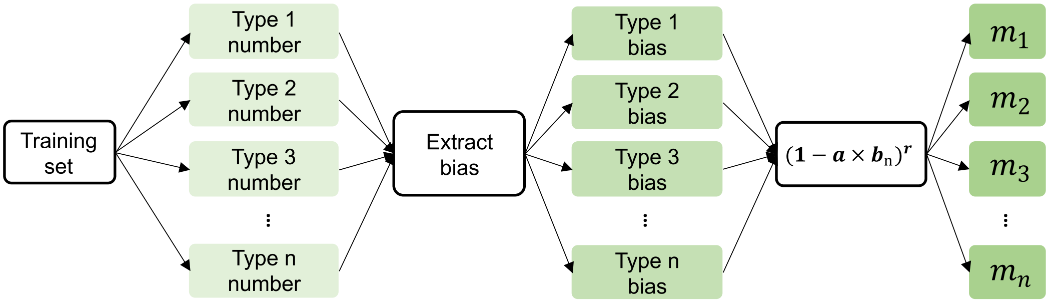

FIGURE 5. The architecture of the Interquestion Type Bias module. It consists of a question type bias extractor and a weighting factors generator, respectively.

The loss of the VQA model is reshaped by the penalty of the weighting factors and becomes more balanced. Specifically, the use of IQTB is very simple and does not require too many changes to the original model. The generated penalty factors are directly weighted to the loss function of each question type, which can be written as

$$\text{loss} = \sum_{i}^{n} \sum_{j}^{m} w_i \times [y_j \log(p_j) + (1 - y_j) \log(1 - p_j)] \qquad (6)$$

where n represents the number of question types and m represents the number of questions in a type.

The Decoupled Training module

The VQA task also suffers from the overall distribution bias of the answer set. Due to the unbalanced distribution of the answer set, the weights of the classifier are unbalanced during the training process, which affects the final classification decision.

Inspired by Kang et al.[12] we introduce the decoupled training module to reduce the overall distribution bias of the answer set by automatically balancing the weights of the classifier. Specifically, DT sets a penalty factor for each class of the classifier, the specific formula is as follows:

$$w_i' = \frac{w_i}{\|w_i\|^x} \qquad (7)$$

where w_i represents each class of the classifier, $\|w_i\|$ represents l_2 norm of w_i, and x is the hyperparameter. DT set a penalty weight $\|w_i\|$ for each class w_i to balance the classifier.

EXPERIMENTS

In this section, we comprehensively evaluate the MVBR method in reducing language bias. First, MVBR is tested on the VQA-CP and VQA-CP v2 datasets and applied on three baselines. Further, we conduct ablation studies to demonstrate the performance of each module. The experimental results and ablation studies are shown below.

Setup

VQA-CP v1 [4] and VQA-CP v2 are used to evaluate the robustness of VQA models when the answer distribution of training and test splits is significantly different. For fair comparisons, we follow the evaluation metrics of VQA.[2] We use the UpDn model [1] for feature preprocessing and conduct experiments with three baselines: UpDn,[1] LMH,[11] and CSS.[7]

Implementation Details

In this article, we use Fast R-CNN to extract the object from the image. The question representation uses the attention mechanism in the UpDn model to guide the distribution of visual features. For each image, the UpDn generates a set of 36 objects with a 2048-D feature. Questions are trimmed to a maximum of 14 words for computational efficiency. We use Glove to initialize the word embedding, and the embedding size is set to 300. The training epochs and batch size is set to 30 and 512, respectively. In order to compare LHM and CSS models fairly, we follow the original paper in parameter settings and reproduce the results.[b] We set the hyperparameters $a = 2$ and $r = 5$ in IQTB module, and x is set to 2 in the DT module. For the selection of hyperparameters, we put more detailed analysis on the ablation studies section. To support the running of the code, we used a NVIDIA Tesla V100 to meet the performance requirements.

Experimental Results and Analysis

As shown in Table 1, we first compare the methods for reducing reliance on language prior in VQA-CP v2. Our MVBR method improves the accuracy from 52.45% to

[b] https://github.com/yanxinzju/CSS-VQA

TABLE 1. Accuracies on VQA-CP v2 test set of the state-of-The-Art models. the part above the dividing line in the table focuses on the method of reducing reliance on language prior, and the part below the dividing line in the table is the method for enhancing visual reasoning ability.

Model	VQA-CP v2 test			
	Overall	Yes/ No	Number	Other
GVQA[4]	31.30	57.99	13.68	22.14
UpDn[1]	39.84	41.96	12.36	46.26
AReg[8]	41.17	65.49	15.48	35.48
Rubi[9]	47.11	68.65	20.28	43.18
SCR	48.47	70.41	10.42	47.29
DLR	48.87	70.99	18.72	45.57
LMH[11]	52.45	69.81	44.46	45.54
GGE[17]	57.32	**87.04**	27.75	**49.59**
LMH+CCB[18]	**57.99**	86.41	45.63	48.76
LMH+MVBR (Ours)	56.72	74.11	**57.29**	47.45
HAN[19]	28.65	52.25	13.79	20.33
HINT[5]	47.70	70.04	10.68	46.31
SCR	49.17	71.55	10.72	47.49
SSL[6]	57.59	86.53	29.87	50.53
CSS†	58.91	85.03	51.18	47.34
CSS+CCB[18]	59.12	89.12	51.04	45.62
CSS+IntroD[20]	60.17	**89.17**	46.91	48.62
CSS+MVBR(Ours)	**60.21**	84.53	**55.21**	**48.83**

† is the result of our reimplementation.

56.72% on LHM. Moreover, we compare the approaches to enhancing visual reasoning ability. MVBR improves the accuracy from 58.91% to 60.21% on the CSS model and achieves the state-of-the-art result. MVBR improves the accuracy of "Number" questions from 51.18% to 55.21%, and the accuracy of "Other" questions from 47.34% to 48.83%. In Table 2, our method also achieved competitive results on VQA-CP v1. In this dataset, MVBR achieves the accuracy of 86.74% and 44.60% on the "Yes/No" and "Number" questions, respectively.

Through the above experimental analysis, we draw the following two conclusions: 1) Experiments on different datasets prove that MVBR can effectively alleviate the language bias caused by the long tail distribution. 2) MVBR can improve accuracy of the tail data with little loss of accuracy of the head data.

TABLE 2. Performance on VQA-CP v1 test set.

Model	VQA-CP v1 test			
	Overall	Yes/No	Number	Other
SAN	26.88	35.34	11.34	24.70
UpDn	39.74	42.27	11.93	46.05
GVQA	39.23	64.72	11.87	24.86
AReg	41.17	65.49	15.48	35.48
GRL	45.69	77.64	13.21	26.97
Rubi	50.90	80.83	13.84	36.02
LMH	55.27	76.47	26.66	**45.68**
CSS	60.95	85.60	40.57	44.62
CSS+MVBR	**62.47**	**86.74**	**44.60**	45.52

Comparison of Different Models

MVBR is a model-agnostic training method that can be integrated in other models. So we apply MVBR on three representative models. Specifically, the MVBR achieves a higher performance in UpDn and LHM that are the basic model for VQA tasks and an effective method in reducing intraquestion type bias. MVBR also improves the accuracy in CSS that is a de-bias method by enhancing visual reasoning ability and the result is shown in Table 3. Through the experiment on VQA v2 dataset, MVBR can maintain a competitive result when the distribution of training set and test set is the same. On the CSS model, MVBR can improve its accuracy.

By using different baselines and datasets for testing and comparison, we can draw the following conclusions: 1) MVBR is widely applicable and can be applied to many baselines; 2) MVBR has a competitive performance in different distributed datasets.

Ablation Studies

We conduct ablation experiments on VQA-CP v2 to investigate the validity of our method. The results are shown in Table 4. First, we set $x = 0$ in which the model is not affected by the DT module. Then, we set $a = 2$, $r = 5$ to investigate the effectiveness of the IQTB module. The results illustrate that our approach reduces the interquestion type bias effectively and the performance is improved from 58.91% to 59.55%.

We also compared the performance of the IQTB module with different parameters. The performance of the model begins to deteriorate when $a = 3$, $r = 5$. This indicates that as a increases, the performance of the model deteriorates due to excessive penalty. Compared with the CSS baseline, the accuracy of our method is improved from 85.03% to 85.35% on the

TABLE 3. Performance of different models on VQA-CP v2 and VQA v2 datasets.

Model	VQA-CP v2 test				VQA v2 test			
	Overall	Yes/No	Number	Other	Overall	Yes/No	Number	Other
UpDn	39.68	41.93	12.68	45.91	63.48	81.18	42.14	55.66
UpDn+MVBR	40.01	42.26	12.79	46.30	63.16	81.13	43.03	55.73
LMH	52.45	69.81	44.46	45.54	61.55	77.22	41.02	55.06
LMH+MVBR	56.72	74.11	57.29	47.45	61.50	77.53	39.55	55.12
CSS†	58.91	**85.03**	51.18	47.34	53.80	57.11	**39.07**	55.20
CSS+MVBR	**60.17**	84.53	**55.21**	**48.83**	**55.30**	**61.10**	37.51	**55.30**

† is the result of our reimplementation.

TABLE 4. Ablation studies on VQA-CP v2 test. † is the result of our reimplementation.

Model	a	r	x	Overall	Yes/No	Number	Other
CSS†	-	-	-	**58.91**	**85.03**	**51.18**	**47.34**
MVBR	2	5	0	59.55	85.35	51.18	48.33
	3	5	0	59.38	**85.41**	52.46	47.63
	0	0	1.0	59.53	84.98	53.41	47.86
	0	0	1.5	59.68	84.87	54.27	47.96
	0	0	2.0	59.76	84.68	55.08	47.98
	0	0	2.4	59.78	84.42	55.57	48.03
	0	0	3.0	59.69	83.75	**56.16**	48.05
	2	5	2.0	**60.21**	84.53	55.21	**48.83**

"Yes/No" question and from 47.34% to 48.33% on the "Other" question when $a = 2$, $r = 5$. This shows that by reshaping the loss of the VQA model, the model focuses on each question type balanced, which improves the robustness of the model.

Then, we set $a = 0$ and $r = 0$ to remove the effect of the IQTB module and explore the performance of the DT module. The hyperparameter x is set to 2.4 and the performance is improved from 58.91% to 59.78% based on the CSS model, proving that the DT module reduces the bias caused by data distribution. Figure 6 shows that the weights of the classifier after retraining become more balanced. Specifically, the accuracy of MVBR increases from 51.18% to 55.57% in the "Number" question, and from 47.34% to 48.83% in the "Other" question. Meanwhile, the "Yes/No" questions, i.e. the header data that make up the majority, have little loss of precision. This shows that our method can improve the performance of tail data with little loss of head data.

We also adjusted x to investigate its effect on the model. As can be seen from Table 4, as x increases, the model results do not always improve and the effect on the head data also increases.

Qualitative Analysis

Finally, to further validate the superiority of our method, we visualize the attention regions and output the top three candidate answers. The predicted probability is the confidence of the VQA model in the answer. As shown in Figure 7, for the question "What color are the balloons?", the baseline and MVBR are the same in the region of interest in the image. However, the baseline model predicts the wrong answer,

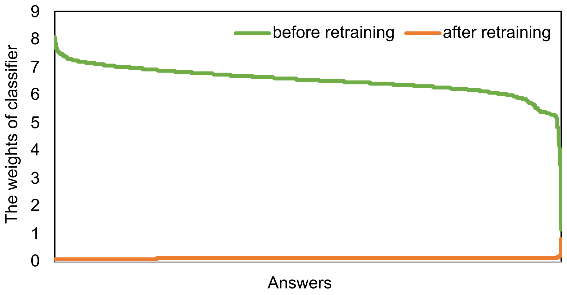

FIGURE 6. Distribution of classifier weights. The weights distribution is more balanced after retraining.

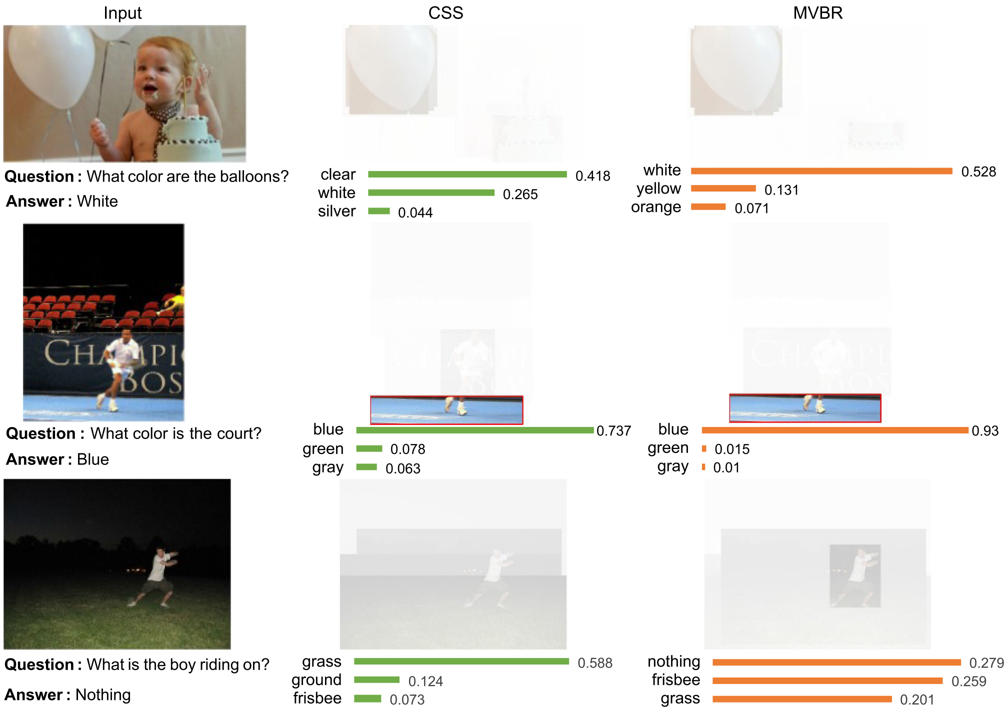

FIGURE 7. Qualitative comparison between our baseline CSS and our output Method MVR on the VQA-CP v2 test set. We output the attention level of each model to the picture, the clearer the image region, the more attention the model pays to the region. We also output the top-3 candidate answers of both models for comparison.

while MVBR is successful in predicting. It proves that when the VQA model finds the correct regions, MVBR allows the VQA model to understand the content of the image better. Although both the baseline model and MVBR predict the correct result for the question "What color is the court?," we have higher probability of correct answer and lower probability of wrong answer. It suggests that MVBR has higher accuracy in choosing answers. When the question is "What is the boy riding on?," MVBR finds the regions in the image more accurately than the baseline and predicts the correct result. It can be concluded that our model has a higher sensitivity of image understanding.

CONCLUSION

In this work, we propose a method named MVBR to reduce language bias in VQA tasks. Compared with the previous method that only reduces intraquestion type bias, MVBR solves the following two problems. First, we propose the concept of interquestion type bias, and design the IQTB module based on this bias to mitigate its negative impact on the VQA model. Second, we introduced decoupled training to reduce the negative effect of the overall distribution bias of the answer set. Experiment results show that MVBR achieves the state-of-the-art on VQA-CP v2 dataset. The proposed method effectively reduces the interquestion type bias and the overall distribution bias of the answer set. In addition, visual analysis also proves that MVBR can enhance the sensitivity of the model to visual content. The data of the visual question answering task is complex. The biases we currently solving are mainly generated by language modalities. In the future, we will consider the bias from visual modalities. Finally, we will also consider how to use the bias to help the model understand the semantic information of the data, and how to make the model better integrate the information of the two modalities.

ACKNOWLEDGMENTS

This work was supported in part by National Key Research and Development Project under Grant 2020AAA0106200, in part by the National Nature Science Foundation of China under Grants 61936005 and 61872424, and in part by the Natural Science Foundation of Jiangsu Province under Grants BK20200037 and BK20210595.

REFERENCES

1. P. Anderson et al., “Bottom-up and top-down attention for image captioning and visual question answering,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 6077–6086.
2. S. Antol et al., “VQA: Visual question answering,” in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 2425–2433.
3. Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh, “Making the V in VQA matter: Elevating the role of image understanding in visual question answering,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6904–6913.
4. A. Agrawal, D. Batra, D. Parikh, and A. Kembhavi, “Don’t just assume; look and answer: Overcoming priors for visual question answering,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 4971–4980.
5. R. R. Selvaraju et al., “Taking a hint: Leveraging explanations to make vision and language models more grounded,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 2591–2600.
6. X. Zhu, Z. Mao, C. Liu, P. Zhang, B. Wang, and Y. Zhang, “Overcoming language priors with self-supervised learning for visual question answering,” in *Proc. Int. Joint Conf. Artif. Intell.*, 2020.
7. L. Chen, X. Yan, J. Xiao, H. Zhang, S. Pu, and Y. Zhuang, “Counterfactual samples synthesizing for robust visual question answering,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 10800–10809.
8. S. Ramakrishnan, A. Agrawal, and S. Lee, “Overcoming language priors in visual question answering with adversarial regularization,” *Proc. Adv. Neural Inform. Process. Syst.*, vol. 31, 2018.
9. R. Cadene et al., “Rubi: Reducing unimodal biases for visual question answering,” *Adv. Neural Inf. Process. Syst.*, vol. 32, pp. 841–852, 2019.
10. Z. Liang, H. Hu, and J. Zhu, “LPF: A language-prior feedback objective function for de-biased visual question answering,” in *Proc. 44th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2021, pp. 1955–1959.
11. C. Clark, M. Yatskar, and L. Zettlemoyer, “Don’t take the easy way out: Ensemble based methods for avoiding known dataset biases,” in *Proc. Conf. Emp. Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process.*, 2019.
12. B. Kang et al., “Decoupling representation and classifier for long-tailed recognition,” in *Proc. Int. Conf. Learn. Representations*, 2019.
13. J. Wang, B. K. Bao, and C. Xu, “Dualvgr: A dual-visual graph reasoning unit for video question answering,” *IEEE Trans. Multimedia*, vol. 24, pp. 3369–3380, 2021.
14. A. Rohrbach, L. A. Hendricks, K. Burns, T. Darrell, K. Saenko, and U. Berkeley, “Object hallucination in image captioning,” in *Proc. Conf. Emp. Methods Natural Lang. Process.*, 2018.
15. X. Yang, C. Gao, H. Zhang, and J. Cai, “Auto-parsing network for image captioning and visual question answering,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 2197–2207.
16. M. Yuan, B.-K. Bao, Z. Tan, and C. Xu, “Adaptive text denoising network for image caption editing,” *ACM Trans. Multimedia Comput.*, 2022.
17. X. Han, S. Wang, C. Su, Q. Huang, and Q. Tian, “Greedy gradient ensemble for robust visual question answering,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 1584–1593.
18. C. Yang, S. Feng, D. Li, H. Shen, G. Wang, and B. Jiang, “Learning content and context with language bias for visual question answering,” in *Proc. IEEE Int. Conf. Multimedia Expo*, 2021, pp. 1–6.
19. M. Malinowski, C. Doersch, A. Santoro, and P. Battaglia, “Learning visual question answering by bootstrapping hard attention,” in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 3–20.
20. Y. Niu and H. Zhang, “Introspective distillation for robust question answering,” *Adv. Neural Inf. Process. Syst.*, vol. 34, pp. 16292–16304, 2021.

PENGJU LI is a student at Nanjing University of Posts and Telecommunications, Nanjing, 210049, China. His research interests include artificial intelligence, multimedia recognition, and cross-modal understanding. Li received his M.E. degree in signal and information processing from the Nanjing University of Posts and Telecommunications. Contact him at 1020010507@njupt.edu.cn.

ZHIYI TAN is a lecturer at Nanjing University of Posts and Telecommunications, Nanjing, 210049, China. His research interest include sequential modeling and prediction, multimedia recognition and cross-modal understanding. Tan received his Ph.D. degree in communication and information engineering from Shanghai Jiao Tong University. Contact him at tzy@njupt.edu.cn

BING-KUN BAO is currently a professor with College of Telecommunications & Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing, 210049, China. Her current research interests include pattern recognition, multimedia recognition, and cross-modal understanding. Contact her at bingkunbao@njupt.edu.cn.