Alleviating shortcut learning behavior of VQA model with context augmentation and adaptive loss adjustment

Zerong Zeng, Ruifang Liu*, Huan Wang
School of Artificial Intelligence, Beijing University of Posts and Telecommunications Beijing, China

ABSTRACT

Despite the impressive improvements of Visual Question Answer (VQA), it still remains a challenge of how to avoid the suffering of spurious correlations from textual content to answer. Previous researches have shown that due to the existence of language bias in the VQA dataset, VQA models may tend to capture superficial statistical correlation and suffer from the poor generalization capability in the out-of-distribution data. To alleviate the biases caused by language modality, we propose a method of context augmentation and adaptive loss adjustment, which can alleviate shortcut learning behavior of VQA models. Specifically, the existence of language bias is due to the high co-occurrence frequency of categories and the words in "Question", therefore, we propose to use "Paraphrase Generation" to produce paraphrases with diverse contexts, so as to mitigate such correlation. Secondly, we use adaptive loss adjustment to adjust the importance of samples, that is, reduce the importance of bias-aligned samples and improve the importance of bias-conflicting samples, so as to guide the model to capture the intrinsic attributes that are beneficial to generalization. The experiments have demonstrated the feasibility and validity of our method on a variety of VQA models.

Keywords: Visual question answering, context augmentation, adaptive loss adjustment

1. INTRODUCTION

Visual Question Answering[1] is one of the fundamental tasks of multi-modal learning, which requires the AI systems to perceive the features of an image and a question, and answer the questions according to the visual clues. With the increasing interest, impressive improvements have been achieved in VQA. However, due to the existence of the language bias[2,3], VQA models tend to capture the superficial statistical correlation, rather than intrinsic attributes that have better generalization. Therefore, VQA models may suffer from the poor generalization capability in the out-of-distribution data, such as the VQA-CP[4], whose priors are quite different in training and test sets[2]. Language bias is manifested in that for some types of questions, certain answers occupy the majority, which leads to good performance of the VQA model using only unimodal information for prediction. For example, the answer to about 40% of the questions beginning with "what sport" is "tennis"[5], and the answer to about 90% of the questions beginning with "Do you see a ..." is "yes"[2].

There have been efforts to tackle the language bias issue, such as those methods based on annotated data[6] or counterfactual method[2]. Those methods based on annotated data, would encourage the model to focus on the regions (annotated by human annotators) on the picture that are most relevant to the question. Those counterfactual-based methods would construct a structural causal model to formulate VQA, and use the method of causal intervention to remove the spurious correlation between the question and the answer. However, the annotation-based method requires expensive labor, and the ensemble-based method cannot ensure that the "language bias" is fully captured. Therefore, the "bias" may still exist in the multi-modal information, which means that the model may still have the behavior of shortcut learning.

In this paper, we consider the unimodal shortcuts that come from language bias and the multimodal shortcuts that involve both visual and textual contents. For language bias, we consider that it originates from two aspects, i.e., keyword bias[7] and label bias. Keyword bias issue occurs in such a scenario that the categories and the words in "Question" have a high co-occurrence frequency, e.g., "White" and "red" are common answers to questions beginning with "what color" in the VQA-CP v1 training set[4]. Besides, when some categories appear significantly more frequently than others in a training set, the problem of label bias issue occurs. In addition, the textual and visual attributes exhibited by most of the samples in the dataset are not inherently intrinsic, but their high co-occurrence frequency with specific labels leads the model to

* lrf@bupt.edu.cn

Third International Conference on Computer Science and Communication Technology (ICCSCT 2022)
edited by Yingfa Lu, Changbo Cheng, Proc. of SPIE Vol. 12506, 125062A
 · 0277-786X · doi: 10.1117/12.2661996

incorrectly treat such bias attributes as an intrinsic attribute of the samples. For example, in the training set, the answer to a "what sport" type question has a high probability of being "tennis" when there is a racket in the corresponding picture, but in the validation set, this phenomenon does not necessarily hold[8]. Following Reference[9], we define the *bias-aligned* samples as data items that contain strong correlations between the bias attributes and the labels, and the *bias-conflicting* samples as other rarely occurring cases. Therefore, we suggest using "Paraphrase Generation"[10] to generate "Paraphrase" as new "Questions" with more diverse contexts, then we then follow the contrastive framework to introduce diverse information into the model. In this way, we can alleviate the "co-occurrence relationship" the categories and the words in "Question" to a certain extent. In addition, to guide the model to focus more on learning the intrinsic properties of the samples, we propose to use the mean classification score as the indicator of the model's learning status, then we use the mean classification score to re-weight different samples so as to emphasize bias-conflicting samples and de-emphasize the bias-aligned ones.

The major contributions of this paper can be summarized as follows: (1) Through the contrastive framework, we introduce "paraphrases" with diverse con- texts into the training process of the model, reduce the frequency of co-occurrence of the categories and the words in "Question", thereby alleviating the model's memory of the priors in dataset. (2) We introduce mean classification score to indicate the learning status of the model, and adaptively adjust the importance of different samples.

2. RELATED WORK

2.1 Paraphrase generation

Paraphrase generation[10] is a fundamental task in natural language processing, which aims to generate a new, semantically identical sentence given a source sentence, which has a different structure or choice of words than the source sentence, i.e., has a different context. An important property of paraphrase generation is diversity, which makes it useful for data enhancement, improving model robustness, etc. Reference[11] introduces Diverse Paraphraser using Sub-modularity to obtain diverse paraphrases for data augmentation on multiple down-stream tasks. Reference[12] combines pretrained Language Models (LMs) with a novel retrieval-based target syntactic parse selection module to augment the training data, which can effectively improve the robustness of the classification models against syntactic attacks. Based on the diversity of paraphrase generation, we obtain more contextually diverse "Questions" through paraphrase generation, thus reducing the co-occurrence frequency of categories and the words in "Questions" in the VQA dataset to a certain extent.

2.2 Contrastive learning

As a representation learning method, contrastive learning has been widely used in computer vision[13,14]. Recently, there have been several efforts to apply contrast learning to the field of natural language processing, such as sentence representation learning[15,16], and unsupervised clustering[17]. The goal of contrastive learning is to draw similar samples closer and dissimilar samples farther. Given a sample, and its augmented sample, the contrast loss will pull them closer together, while pulling the original sample and the augmented samples of other samples farther apart. In this way, the different instances in the representation space can be sufficiently separated, while the local invariance of each sample can be retained[17]. When using more contextually diverse "Questions" generated by "Paraphrase Generation" for context augmentation, we want to introduce richer contexts while avoiding introducing additional noise, so we use contrast learning to allow the VQA model to learn sentence representations that are more contextually rich and robust to the linguistic variations.

2.3 Alleviating shortcut learning behavior

When biases exist in the dataset, neural networks tend to capture spurious associations and exploit dataset biases as shortcuts to obtain higher evaluation performance instead of truly understanding the language and images, i.e., the models may not learn the intrinsic attributes of better generalization and the ability of reasoning. Recently, several research works[18,19] have shown that intrinsic attributes are usually more difficult to learn than bias attributes, and that in the early training phase of the model, the model first learns shortcut features for fast loss reduction, and only after that the model gradually learns intrinsic attributes for further loss reduction. In short, in the early training stage, the model will first learn to fit the bias-aligned samples, and then gradually fit the bias-conflicting samples. A straightforward approach to alleviate the shortcut learning behavior of the model is to increase the diversity of bias-conflicting samples in the dataset[9]. For instance, to improve the generalization of the domain, Reference[20] mixed the style of different source domains. However,

this method requires knowledge of the type of bias and additional manual collection of samples. Another approach of alleviating the shortcut learning behavior is model design. Reference[21] introduced a question-only branch to capture superficial statistical correlation from the textual modality and re-weight samples. Reference[22] introduced multiple biased models and make the biased models preferentially over-fit the biased data distribution by a greedy strategy, allowing the base model to learn those examples that are difficult for the biased models to solve, resulting in unbiased predictions.

Table 1. An example of "paraphrase generation".

Original question	**What color pants is the man wearing?**
Paraphrases	Which color is the man's pants? Can you name the color of the pants you're wearing? Tell me the color of pants that the man is wearing?

In Table 1, "Original Question" denotes the "Question" from the VQA-CP v2 dataset and the "Paraphrases" are generated by the Paraphraser based on the original question.

3. METHODOLOGY

3.1 Task definition

VQA is one of multimodal learning task that can be formulated as a multi-class classification problem. Given a multi-modal dataset $D = \{I_i, Q_i, A_i\}_i^N$ where each sample is a triplet, including a picture $I_i \in I$, a question $Q_i \in Q$, and an answer $A_i \in A$, the task of VQA model is to learn a mapping function $fvqa : I \times Q \rightarrow R^e$. Specifically, given a picture and a question, VQA model will output a probability distribution on the candidate answers according to the input.

We hope that VQA model can make the correct choice among the candidate answers by comprehensively considering the visual and textual information. Unfortunately, VQA model may tend to capture the superficial statistical correlation, instead of the intrinsic attributes that have better generalization. Therefore, in this section, we will introduce the Context Augmentation and Adaptive Loss Adjustment to alleviate shortcut learning behavior of VQA models.

3.2 Context augmentation

The existence of data bias leads to models that can achieve higher evaluation metrics by using bias as shortcut. As discussed in Section 1, data bias in VQA dataset is due in part to the high frequency co-occurrence relationship. Such co-occurrence relationship occurs not only in unimodal information, that is, the category and the words in the "question" have high co-occurrence, but also in multimodal information. For example, for "what sport" type questions, when "racket" appears in the picture, the answer is often "tennis". It is because of this that VQA models tend to capture such spurious statistical correlations rather than intrinsic features with better generalization, which then leads to shortcut learning behavior of the model. Therefore, we propose to alleviate the shortcut learning behavior of the VQA model by reducing such spurious co-occurrence relationship. Specifically, we want to introduce diverse information, and considering that paraphrase generation can yield semantically identical but more contextually rich sentences, we decided to use paraphrase generation to generate "Paraphrase" as a new "Question" with more diverse contexts to mitigate such correlation.

Paraphrases refer to text (often sentences) that share the same meaning but use different choices of words and their ordering. For each "question" in the VQA dataset, we used the "Paraphraser"[23] to generate multiple sentences with different contexts. In Table 1, we selected a sample of question type "What color" as an example. Compared with the original "question", the paraphrases have more context words and different ways of expression. After obtaining the paraphrases to the "questions" of all samples in the VQA training set, we selected two paraphrases (p_{i0} and p_{i1}) for each sample xi as the enhanced views of the "Questions". Then we follow the contrastive framework in Reference[13] and use the paraphrases to construct (x_i, x_i^+) pairs, where $x_i = (I_i, Q_i)$ denotes the original sample, which consists of a picture $I_i \in I$ and a question $Q_i \in Q$, and $x_i^+ = (I_i, p_i)$ denotes the positive in- stance, which consists of a picture $I_i \in I$ and a

paraphrase p_i. For a mini-batch with N pairs, the training objective of context augmentation is:

$$L_{i,k} = \log \frac{e^{sim(\mathrm{h}^{\mathrm{i}}, \mathrm{h}^{+}_{i,k})/\tau}}{\sum_{j=1}^{N} e^{sim(\mathrm{h}^{\mathrm{i}}, \mathrm{h}^{+}_{j,k})/\tau}}, \quad k \in (0,1) \tag{1}$$

where h_i denotes the representation of x_i, and τ is a temperature hyper-parameter and $sim\left(h_i, h^{+}_{j,k}\right)$ indicates cosine similarity. Finally, we calculate the contrastive loss through the following formula:

$$L_{cl} = L_i = \frac{L_{i,0} + L_{i,1}}{2} \tag{2}$$

3.3 Adaptive loss adjustment

The shortcut learning behavior of the model occurs not only because of the bias in the dataset itself, but also because the model over-fits the bias-aligned samples during the training process, and thus learns some of the bias features of the bias-aligned samples, especially when the bias features are easier to learn than the intrinsic features. As discussed in Section 2.3, increasing the diversity of bias-conflicting samples during the training of the model is crucial to improve the robustness of the model. Considering that increasing the diversity of bias-conflicting samples may change the distribution of the dataset itself, as well as being labor-intensive, we adopted the adaptive loss adjustment method. Based on this intuition, the model tends to learn some features that are better learned but not needed in the biased alignment sample, so the model may ignore some features that are harder to learn and that we want the model to learn. Therefore, we propose to guide the model to focus more on learning the intrinsic properties during training. We implement this idea by emphasizing bias-conflicting samples and de-emphasizing bias-aligned samples during the training process. It is unrealistic to let the model focus on the bias-conflicting samples by artificially labeling the data, and it is even more unreasonable to directly emphasize those tail data, so we hope to let the model adaptively complete the weight adjustment by monitoring the learning state of the model.

To adaptively adjust importance of the samples, we attempt to track the learning state (i.e., classification accuracy) of the model, then adjust the importance of the samples during the training process. There was some work[24,25], which tried to balance the classification loss qualitatively based on the number of samples in each category of the training set, without caring about the real learning state of the model for different types of samples. Such methods can not make adaptive adjustment according to the learning state of the model. Therefore, similar but different from Reference[26], we propose to use the dynamic mean classification score S RC (C denotes the numbers of the categories) as the indicator of the learning status of the model for each category. We hope that the model can quickly fit those bias attributes first, and then we can distinguish the bias-aligned samples with more bias attributes through the average classification score. Then, we make the model adaptively adjust the importance of different samples according to the cur- rent learning state in the training process, so as to pay more attention to the bias-conflicting samples that are difficult to distinguish.

Let $p_j^{ik} \in R^{1*C}$ denotes the predicted probability of the k-th sample in the j-th iteration and the i-th epoch of the model. At the i-th epoch, the dynamic mean classification score of the model can be calculated as:

$$S^i = \frac{\sum_{j=0}^{N-1} \sum_{k=0}^{bs-1} p_{jk}^{i}}{\sum_{j=0}^{N-1} m_j} \tag{3}$$

where $S^i \in S$, $S^i = [s_0^i, s_1^i, ..., s_n^i, ..., s_{c-1}^i]$, s_n^i denotes the mean classification score corresponding to category n in the i-th epoch, bs denotes the batch size, N denotes the total number of iterations in one epoch and m_j denotes the number of samples in the j-th iteration.

After obtaining the dynamic mean classification score Si for each category in the training process, we introduce a dynamically loss adjustment to dynamically adjust the importance of the sample. As discussed in Section 2.3, in the early training stage, the model will first learn to fit the bias-aligned samples, and then gradually fit the bias-conflicting

samples. Therefore, we do not adjust the sample weights in the early training phase of the model, thus allowing the model to quickly fit the bias-aligned samples. At this time, we consider that samples with high mean classification score are more likely to correspond to bias-aligned samples. Later, we adjust the weights of the bias-aligned and bias-conflicted samples according to the mean classification score, so as to guide the model to focus more on the bias-conflicted samples. To achieve this, we formulate the adaptive loss adjustment as:

$$L_{al} = L_{vqa}(g(p); y) \tag{4}$$

$$g(p^i) = \begin{cases} p^i, & i \le b \\ \alpha p^i + (1-\alpha) S^{i-1} p^i, & otherwise \end{cases} \tag{5}$$

where L_{vqa} is the cross-entropy, p^i denotes the predicted probability of the model in the i-th epoch, b de- notes the beginning epoch to start adjusting the importance of the sample, and α is a regulating factor, indicating to what extent the model adjusts the importanceof samples according to "learning status".

3.4 Objective function

Finally, our training paradigm optimizes both (L_{cl} and L_{al}) losses together, and the overall objective function is in the form of:

$$L_{total} = \beta L_{cl} + L_{al} \tag{6}$$

where β is a hyper-parameter.

4. EXPERIMENTS

4.1 Experiment settings

Datasets: We use VQA-CP v2[4] dataset to evaluate the performance of our proposed approach. VQA-CP v2 dataset was built by reorganizing VQA v2, which is designed to test the robustness of the VQA models. The answer distribution of its training set and test set is quite different, so it can usually be used to evaluate the robustness of the VQA models. We follow the protocol of[3,21] to train and evaluate our model, and use the standard VQA evaluation metrics to evaluate the performance of our model.

Implementation details: The weight of the contrastive loss to the total loss, namely the hyper- parameter β in equation (6), is set to 0.4. The value of the hyper-parameter α in equation 5 is set to 0.5. In addition, the setting of "start epoch", namely the b in equation 5, needs to be determined according to the training effect of the first few epochs of the model, which is related to the model itself. In the experiment, we set it to 4, which means that in the first four epochs, we did not use the *mean classification score* to adjust the im- portance of the samples. After four epochs of training, the model tends to be stable, and we begin to adjust the loss. Other experimental settings, such as dimension of the "Question" features and the batch size, are the sameas[21].

4.2 Results and discussion

In order to verify the effectiveness of our proposed method, we compare our method with the existing methods to solve the data bias issue, such as References[3,21,27], and we construct experiments based on the baseline "S-MRL+RUBi"[21] and "S-MRL+LPF"[3]. As listed in Table 2, the improvement on the VQA-CP v2, demonstrates the effectiveness of our method to reduce the impact of data bias on the model and alleviate shortcut learning behavior of VQA model. Among them, baseline "S-MRL+LPF" is based on baseline "S-MRL+RUBi", by using "language prior" to adjust the global training target, which has achieved a significant improvement (+11.11%). Our method outperforms baseline "S-MRL+RUBi" by 1.55%, and outperforms baseline "S-MRL+LPF" by 1.18%. This may seem a bit strange, because our method can achieve a 1.18% improvement on baseline "S-MRL+LPF", but the improvement on baseline "S-MRL+RUBi" is not obvious, only 1.55%. After analysis, we think this is reasonable, because our method will have certain requirements for the learning ability of the model itself. In the introduction of Section 2.3, our "Adaptive Loss Adjustment" method depends on the learning state in the process of model training to a certain extent. Therefore, this just shows the effectiveness and applicability of our method in different models.

Table 2. Comparison on VQA-CP v2 test set.

Model	Overall	Yes/no	Number	Other
S-MRL15	38.46	42.85	12.81	43.20
S-MRL+RUBi15	47.11	68.65	20.28	43.18
S-MRL+VGQE21	50.11	66.35	**27.08**	**46.77**
S-MRL+LPF14	53.38	88.06	25.00	42.99
S-MRL+RUBi+CAALA (ours)	48.66	73.04	22.42	43.08
S-MRL+LPF+CAALA (ours)	**54.56**	**88.19**	26.07	44.76

Note: All metric values of all the baselines are taken directly from the original paper.

4.3 Ablation studies

In this section, we conduct ablation experiments to prove the effectiveness of Context Augmentation and Adaptive Loss Adjustment. The results are reported in Table 3, the "base1" indicates the "S-MRL+RUBi" model and the "base2" indicates the "S-MRL+LPF" model. The fact that "base1+CA" outperforms "base1" and "base2+CA" outperforms "base2" shows that in- creasing the diversity of "language information" can relieve the model's memory of "co-occurrence relations" to a certain extent. In addition, by adding "CAL" to the "base1+CA" and "base2+CA" model, the performance is improved, which illustrates that adjusting the importance of samples according to the learning state of the model during the training process can prevent the model from over learning some bias attributes.

Table 3. Ablation studies on the VQA-CP v2 test set.

Model	Overall	Yes/no	Number	Other
Base 1	47.11	68.65	20.28	43.18
Base 2	53.38	88.06	25.00	42.08
Base 1+CA	47.74	70.94	20.83	42.96
Base 2+CA	54.21	87.79	25.45	44.51
Base 1+CA+ALA	48.66	73.04	22.42	43.08
Base 2+CA+ALA	54.56	88.19	26.07	44.76

Note: The "base 1" indicates the "S-MRL+RUBi" model and the "base 2" indicates the "S-MRL+LPF" model. "base+CA" indicate the "base" model trained with our proposed Context Augmentation, "base+CA+ALA" indicate the "base" model trained with our proposed Context Augmentation and Adaptive Loss Adjustment.

5. CONCLUSION

In this work, we propose to use Context Augmentation and Adaptive Loss Adjustment to alleviate short-cut learning behavior of VQA model. We increase the diversity of language information through para- phrase generation and introduce adaptive loss regulation to adjust the importance of samples according to the learning state of the model. The experimental results demonstrate the feasibility and validity of our pro-posed method.

ACKNOWLEDGMENTS

This research was supported by the National Natural Science Foundation of China (Grant No. 61872338).

REFERENCES

[1] Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L. and Parikh, D., "Vqa: Visual question answering," Proc. of the IEEE Inter. Conf. on Computer Vision, 2425-2433 (2015).

[2] Niu, Y., Tang, K., Zhang, H., Lu, Z., Hua, X. S. and Wen, J. R., "Counterfactual vqa: A cause- effect look at language bias," Proc. of the IEEE/CVF Conf. on Computer Vision and Pat tern Recognition, 12700-12710 (2021).

[3] Liang, Z., Hu, H. and Zhu, J., "Lpf: A language-prior feedback objective function for debiased visual question answering," arXiv preprint arXiv:2105.14300, (2021).

[4] Agrawal, A., Batra, D., Parikh, D. and Kembhavi, A., "Don't just assume; look and answer: Overcoming priors for visual question answering," Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, 4971-4980 (2018).

[5] Wu, J. and Mooney, R. J., "Self-critical reasoning for robust visual question answering," arXiv preprint arXiv:1905.09998, (2019).

[6] Das, A., Agrawal, H., Zitnick, L., Parikh, D. and Batra, D., "Human attention in visual question answering: Do humans and deep networks look at the same regions?" Computer Vision and Image Understanding 163, 90-100 (2017).

[7] Qian, C., Feng, F., Wen, L., Ma, C. and Xie, P., "Counterfactual inference for text classification debiasing," Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th Inter. Joint Conf. on Natural Language Processing (Volume 1: Long Papers), 5434-5445 (2021).

[8] Dancette, C., Cadene, R., Teney, D. and Cord, M., "Beyond question-based biases: Assessing multimodal shortcut learning in visual question answering," arXiv preprint arXiv:2104.03149, (2021).

[9] Lee, J., Kim, E., Lee, J., Lee, J. and Choo, J., "Learning debiased representation via disentangled feature augmentation," Advances in Neural Informa tion Processing Systems 34, (2021).

[10] Zhou, J. and Bhat, S., "Paraphrase generation: A survey of the state of the art," Proc. of the 2021 Conf. on Empirical Methods in Natural Language Processing, 5075-5086 (2021).

[11] Kumar, A., Bhattamishra, S., Bhandari, M. and Talukdar, P., "Submodular optimization-based diverse paraphrasing and its effectiveness in data augmentation," Proc. of the 2019 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 3609-3619 (2019).

[12] Sun, J., Ma, X. and Peng, N., "Aesop: Paraphrase generation with adaptive syntactic control," Proc. of the 2021 Conf. on Empirical Methods in Natural Language Processing, 5176-5189 (2021).

[13] Chen, T., Kornblith, S., Norouzi, M. and Hinton, G., "A simple framework for contrastive learning of visual representations Inter. Conf. on Machine Learning, 1597-1607 (2020).

[14] Chen, X. and He, K., "Exploring simple Siamese representation learning," Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition, 15750-15758 (2021).

[15] Yan, Y., Li, R., Wang, S., Zhang, F., Wu, W. and Xu, W., "Consert: A contrastive framework for self-supervised sentence representation transfer," arXiv preprint arXiv:2105.11741, (2021).

[16] Liu, C., Wang, R., Liu, J., Sun, J., Huang, F. and Si, L., "Dialoguecse: Dialogue-based contrastive learning of sentence embeddings," arXiv preprint arXiv:2109.12599, (2021).

[17] Zhang, D., Nan, F., Wei, X., Li, S., Zhu, H., McKeown, K., Nallapati, R., Arnold, A., and Xiang, B., "Supporting clustering with contrastive learning," arXiv preprint arXiv:2103.12953, (2021).

[18] Du, M., Manjunatha, V., Jain, R., Deshpande, R., Dernoncourt, F., Gu, J., Sun, T. and Hu, X., "Towards interpreting and mitigating shortcut learning behavior of nlu models," arXiv preprint arXiv:2103.06922, (2021).

[19] Nam, J., Cha, H., Ahn, S., Lee, J. and Shin, J., "Learning from failure: Training debiased classifier from biased classifier," arXiv preprint arXiv:2007.02561, (2020).

[20] Zhou, K., Yang, Y., Qiao, Y. and Xiang, T., "Domain generalization with mixstyle," arXiv preprint arXiv:2104.02008, (2021).

[21] Cadene, R., Dancette, C., Cord, M., Parikh, D., et al., "Rubi: Reducing unimodal biases for visual question answering," Advances in Neural Information Processing Systems 32, 841-852 (2019).

[22] Han, X., Wang, S., Su, C., et al., "Greedy gradient ensemble for robust visual question answering," Proc. of the IEEE/CVF Inter. Conf. on Computer Vision, 1584-1593 (2021).

[23] Damodaran, P., "Parrot: Paraphrase generation for nlu," (2021).

[24] Menon, A. K., Jayasumana, S., Rawat, A. S., Jain, H., Veit, A. and Kumar, S., "Long-tail learning via logit adjustment," arXiv preprint arXiv:2007.07314, (2020).

[25] Tan, J., Wang, C., Li, B., Li, Q., Ouyang, W., Yin, C. and Yan, J., "Equalization loss for long-tailed object recognition," Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition, 11662-11671 (2020).

[26] Feng, C., Zhong, Y. and Huang, W., "Exploring classification equilibrium in long-tailed object detection," Proc. of the IEEE/CVF Inter. Conf. on Computer Vision (ICCV), 3417-3426 (2021).

[27] KV, G. and Mittal, A., "Reducing language biases in visual question answering with visually-grounded question encoder," 16th European Conference on Computer Vision, 18-34 (2020).