Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 4, No. 2, 2023

A Survey of Language Priors for Visual Question Answering

Hantao Xu, Xia Ye, Zhangping Yang, Pujie Zhao

Academy of Combat Support, Rocket Force University of Engineer, Xi'an 710025, China

Abstract: In recent years, with the development of deep learning technology, visual question answering tasks have gradually attracted the attention of scientific researchers. Due to the continuous improvement of relevant large-scale standard data sets, a large number of visual questions answering research results have been released one after another, and the accuracy rate of the visual question answering model based on deep learning on the data set has been continuously improved. Recent studies have found that the previously proposed visual question answering model has different degrees of data set language prior problems, that is, the model is overly dependent on the strong phase between the question and the answer in the training process. Many articles briefly describe various research methods, and look forward to the future development direction of alleviating the prior problem of visual question answering based on the existing research.

Keywords: Visual Question Answering; Multimodality; Language Priors.

1. Introduction

Visual question answering (VQA) is one of the representative problems in the intersection of computer vision and natural language processing, and is an important research task of artificial intelligence. The goal of this task is that the question answering model takes the given image and text questions as input and gives the correct answer through the visual question answering model. Thanks to the development of deep learning, computer vision, natural language processing and other technologies, as well as the continuous improvement of related data sets, a large number of visual questions answering research results have emerged in recent years.

Due to the enthusiasm of researchers for visual question answering tasks, more and more complex models achieve higher and higher accuracy rates on larger and larger data sets. However, recent studies have shown that the current SOTA (state-of-the-art) visual question answering model lacks a good image foundation and answers questions by making extensive use of the superficial correlation between questions and answers in the training data, that is, language priors. question. This leads to the fact that the model trained on the dataset cannot be well applied to the real world. The training model affected by the language prior question will blindly output the answer according to the first few predicted word distributions of the question. If the answer to the question asking the number of questions in the training data is the highest proportion of "2", the model will use "2" As the question "How many ... in the graph?" predicts the answer; answering "yes/no" questions also give affirmative answers due to the large number of questions answered "yes" during training. During training, the model focuses too much on the superficial association between the question and the answer, while ignoring the more information contained in the image. This is very unreasonable. The generalization and robustness of the model trained in this way are severely limited; the same unreasonable problem is that there is a strong language prior in the test data with the same distribution as the training data, which leads to the performance of the model cannot be accurately evaluated under the existing evaluation criteria. Usually tested on datasets with different question-answer distributions, most existing models experience significant performance degradation.

2. Related Work

2.1. Improvements for Dataset

Agrawal et al. proposed the VQA-CP dataset, which is a repartition of the existing VQA dataset, where the split of the answer distribution for each question type in the training and test sets is inverted. For example, in the training of VQA-CP v1, "tennis" was the most common answer to the question "What sport is...?". And "skiing" is uncommon; in the test split, this prior is reversed. The VQA-CP dataset is currently the main dataset for dealing with language priors.

In order to minimize the influence of language prior in the dataset on training, some scholars are committed to constructing a balanced dataset. Zhang et al. [1] collect fine-grained scene pairs for each question type asking "yes/no" such that for the exact same question, the answer to one scene is "yes" and the answer to another scene is "No"; Goyal et al. [2] also balance the VQA dataset by collecting complementary images, so that each question is not only related to one image, but to a pair of similar pictures. AGRAWAL et al. [3] re-divided the VQA-v2 dataset so that the distribution of each question type (such as "what color...", "how much...", etc.) and its answer is different on the training set and the test set is large, so models with poor image bases and overly reliant on language priors will perform poorly on this new split. The re-divided VQA-CP data set improves the strong correlation between the questions and answers of the original VQA data set to a certain extent. Currently, this dataset is mainly used to evaluate the generalization ability of the VQA model.

2.2. Improvements for Visual Question Answering Models

2.2.1. Methods based on Strengthening Visual Information

Selvaraju et al. [4] proposed the HINT method (Human Importance-aware Network Tuning), which is a general

framework for aligning network sensitivity with input regions that humans consider relevant to tasks. This method can effectively use human demonstrations to improve the visual basis, HINT encourages deep networks to be sensitive to the same input regions as humans. This method optimizes the alignment between human attention maps and gradient-based network importance to ensure that model training relies on task-relevant visual concepts for human predictions.

Li et al. [5] proposed a relation-aware graph attention network (Relation-Aware Graph Attention Network, ReGAT). This method constructs a relational graph for each image, and uses the attention mechanism to analyze the relationship between multiple types of objects. Modeling is performed to learn adaptive relational representations between problems. The relationship between objects is divided into explicit relationship and implicit relationship. The former is used to represent the geometric position relationship of objects and the semantic interaction relationship between objects, and the latter is used to capture implicit activities in image regions and hidden relationships between regions. Experiments show that on the VQA V2 and VQA-CP datasets, ReGAT's performance takes the lead; the article further shows that ReGAT can be used as a general encoder to embed today's cutting-edge visual question answering models and improve their performance in terms of robustness.

Hirota et al. [6] believe that it is difficult for traditional deep visual features to capture all the details in the image as humans do. At the same time, with the recent progress of natural language models, the authors propose to replace the "image-question" pair with the "description-question" pair as input and feed them into a language-only Transformer model. Experiments show that models based on deep visual features are not more competitive than pure language models.

Si et al. [7] proposed a select-and-rerank (SAR) framework based on Visual Entailment, in which the Visual Entailment task is used to judge the degree of correlation between the given text and the image. Specifically, the method first selects answers that are highly relevant to the image or question as candidates, and then uses Visual Entailment to verify whether the image semantically contains the synthetic statement of the question and the candidate answer. The SAR framework is a general framework, which can make full use of the interaction between images, questions and candidate answers, and can be combined with existing VQA models to improve its performance.

Table 1. Methods based on strengthening visual information [24]

Methods	Base	VQA-CP v2 test				VQA v2 val			
		All	Y/N	Num	Other	All	Y/N	Num	Other
AttAlign[4]	UpDn	39.37	43.02	11.89	45.00	63.24	80.99	42.55	55.22
HINT[4]	UpDn	46.73	67.27	10.61	45.88	63.38	81.18	42.99	55.56
SCR[25]	UpDn	49.45	72.36	10.93	48.02	62.20	78.80	41.60	54.50
ReGAT[5]	UpDn	40.42	-	-	-	67.18	-	-	-
ESR[26]	UpDn	48.90	69.80	11.30	47.80	62.60	-	-	-
VGQE[27]	UpDn	48.75	-	-	-	64.04	-	-	-
Picture[6]	UpDn	43.64	45.13	20.06	49.33	69.74	87.91	56.47	59.43
SAR[7]	UpDn	61.71	-	-	-	-	-	-	-
KAN[28]	UpDn	42.60	42.12	15.52	50.28	-	-	-	-

2.2.2. Methods based on Weakening Language Priors

Ramakrishnan et al. [8] introduced adversarial regularization (AdvReg) to the visual question answering task. It introduces a question-only model that uses the question encoding in the VQA model as input to capture as much language bias information as possible in the question encoding of the model, as shown in Figure 1. Then, the training process is regarded as an adversarial training between the main visual question answering model and the question-only model to prevent the main model from continuing to use the language bias information captured by the question-only model during the training process. At the same time, the article introduces the confidence measurement standard, which encourages the model to pay more attention to visual information by showing the rule to maximize the confidence difference between the two models. Subsequently, Grand et al. [9] studied the advantages and disadvantages of this adversarial regularization, and found some undesirable side effects of AdvReg, including the adversarial training method will bring a lot of noise to the gradient, resulting in unstable gradients and in-domain examples. (indomain example) Performance degrades dramatically. This study shows that introducing regularization during training can help alleviate the above problems but cannot completely solve them. Through error analysis, it is found that AdvReg improves the generalization ability of binary questions, but it will weaken the performance on questions with heterogeneous answer distribution; it also finds that the regularization model tends to rely too much on visual features, while ignoring the important in the question. language clues.

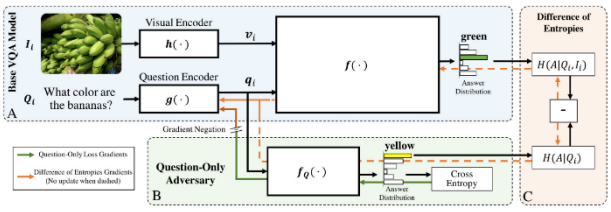

Figure 1. Diagram of introducing confrontational regularization visual question answering model

When training a deep network classifier on a multimodal dataset, multimodal data is exploited at different scales, all modalities contribute to better classification, and some modalities are better than others in actual training. states are more likely to contribute to the classification results. This is not ideal for training because the classifier is inherently biased towards a subset of modalities causing the model to ignore data from one or more other modalities. In order to alleviate this shortcoming, Gat [10] et al. proposed a new

regularization term based on function entropy, which encourages to balance the contribution of each modality to the classification results; and designed a method based on the "log-Sobolev" inequality method, binding the function entropy with its Fisher information to maximize the amount of information contributed by the mode.

Cadene et al. [11] proposed the RUBi (Reducing Unimodal Biases) training strategy to reduce the bias in the visual question answering model. This strategy reduces the most biased samples, that is, samples that can be answered correctly without image information. The visual question answering model designed in this article uses two input modules to replace the dependence on the surface correlation between questions and answers, and captures language bias during training through a question-only model, through which the loss is dynamically adjusted to compensate for the effects of prejudice. The specific process refers to treating the prediction result of the problem model as a mask between 0 and 1, and merging the prediction result distribution of the main model with the prediction distribution of only the problem model before using the prediction result to calculate the loss, and correcting the prediction result, the mask is used to dynamically change the loss, that is, the mask is multiplied by the original prediction result to generate a new prediction result. It can be seen that this strategy increases the score of the correct answer through the mask output by the question model and reduces the score of other answers, which ultimately leads to a reduction in the loss caused by biased samples and reduces the importance of biased samples. Clark et al. [12] designed a new method based on this. The first step is still to train a question-only model that captures bias; the second step is to train a main model that integrates the question-only model. When the article constructs the model, it presupposes a strong assumption that the bias factors in the known samples are independent of other factors except the bias factors. Because this assumption is too strong, the article introduces a new learning factor $g(x^i)$ in subsequent experiments to show the correlation between the two. And for the case where $g(x^i)$ is zero, a regularization method that adds entropy penalties to the loss is designed, through which the model's attention to the answers of highly biased samples is reduced.

Han [13] et al. analyzed several robust visual question answering models through experiments, and proposed that language bias in VQA can be divided into distribution bias and shortcut bias, and based on this, a greedy gradient integration (Greedy Gradient Ensemble, GGE) strategy to eliminate these two biases. GGE combines multiple biased models for integrated training to achieve an unbiased model: a greedy strategy is used to force the biased model to overfit the biased data from the beginning, so that the basic model can pay more attention to the biased model that cannot be solved of samples. Experiments show that this method can make better use of visual information and achieve good performance on the VQA-CP dataset without using additional annotations.

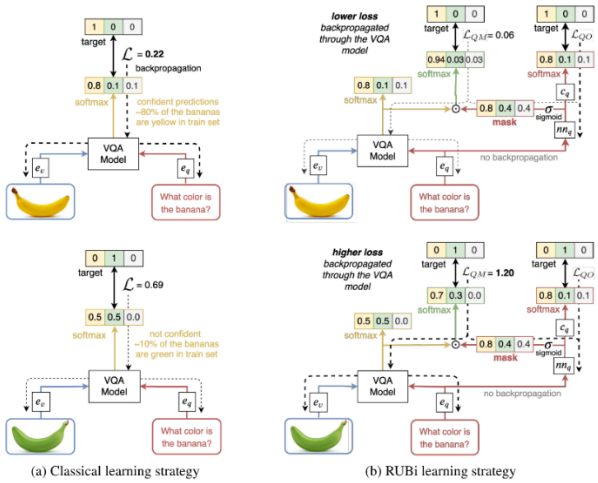

Figure 2. RUBi training strategy diagram

Shrestha et al. [23] pointed out in their article that the performance improvement of the method of alleviating VQA language bias is not the result of improving the visual basis, but often the result of regularization to prevent over-fitting language priors, even if random, irrelevant the visual cues of the model can also have a similar improvement effect on the model. The article further demonstrates that there is no statistically significant change in model predictions for correlated, unrelated, or random regions. Therefore, the article proposes a regularization method under the assumption that degrading the model on the training set will reduce its dependence on language priors and improve test accuracy. This method always penalizes the model regardless of whether its predictions are correct or not, in order to achieve the purpose of model degradation. It is worth mentioning that the article questioned whether visual cues were actually used.

Table 2. Methods based on weakening language priors [24]

Methods	Base	VQA-CP v2 test				VQA v2 val			
		All	Y/N	Num	Other	All	Y/N	Num	Other
AdvReg[8]	UpDn	41.17	65.49	15.48	34.48	62.75	79.84	42.35	55.16
GRL[9]	UpDn	42.33	59.74	14.78	40.76	-	-	-	-
RUBi[11]	UpDn	44.23	67.05	17.48	39.61	-	-	-	-
LM[12]	UpDn	48.78	72.78	14.61	45.58	63.26	81.16	42.22	55.22
LM+H[12]	UpDn	52.01	72.58	31.12	46.97	56.35	65.06	37.63	54.69
Semantic[21]	UpDn	47.50	-	-	-	-	-	-	-
RMFE[10]	UpDn	54.55	74.03	49.16	45.82	-	-	-	-
GGE-DQ[13]	UpDn	57.32	87.04	27.75	49.59	59.11	73.27	39.99	54.39
LPF[29]	UpDn	55.34	88.61	23.78	46.57	55.01	64.87	37.45	52.08

2.2.3. Methods Based on Data Argumentation

Hirota [6] et al. use descriptive text to replace traditional deep visual features, while increasing the diversity of training data to avoid learning language bias. Experiments show that

most data augmentation techniques improve model performance, especially question-based back-translation performs very well.

Chen et al [15] proposed a cross-modal training strategy for model-independent Counterfactual Samples Synthesizing and Training (CSST), where CSS (Counterfactual Samples Synthesizing) consists of two types of sample synthesis: V-CSS and Q-CSS. For V-CSS, a counterfactual image is synthesized by masking key objects in the original image, and a new image-question (VQ) pair is composed of the generated counterfactual image and the original text image; for Q-CSS, by using a special the counterfactual question is synthesized by replacing keywords in the original question with the tag "MASK", and a new VQ pair is composed of the generated counterfactual question and the original image. For newly generated samples, a dynamic answer allocation mechanism is used to form new triplet samples. By using this method for sample amplification, the model can pay more attention to the visual area related to the question, that is, to answer the correct area; and be sensitive to the language change of the question, that is, when the sensitive word in the question changes. The answers to should also change accordingly, improving the visual interpretability and question sensitivity of the model. The dynamic answer assignment mechanism designed in this paper approximates the answers of all synthesized vision-question pairs, avoiding expensive human annotations.

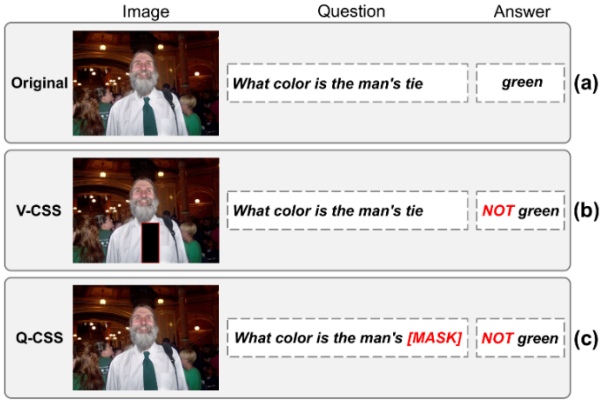

Figure 3. Counterfactual Samples Synthesizing diagram

Zhu et al. [16] also adopted a strategy-assisted model with enlarged training samples for training, and introduced a self-supervised learning framework to solve the language prior problem in the model. Specifically, the method first automatically generates labeled data and balances the biased data through a self-supervised auxiliary task, which does not require the introduction of external annotations. Gokhale et al. [17] proposed an input mutation training strategy, which is assisted by multi-angle sample expansion. The method includes covering, replacing or negating keywords, color inversion and removal of key areas and key objects in the image, its essence is still a way of data augmentation. A large number of samples are formed by pairing new images and questions generated by mutations of images and questions; different from traditional classification methods, this paper uses a noise-contrastive method to predict the correct answer, and a regularized loss function of pairwise consistency to narrow the real answer and the distance to predict the answer. This idea reaches the highest level in the VQA-CP v2 dataset.

2.2.4. Methods based on Training Strategies

Mahabadi et al. [18] introduced three new strategies to reduce bias: the first is an ensemble-based approach that combines multiple probabilistic models of the same data by multiplying the probabilities together and then renormalizing them. The idea is still to combine the probability distributions of the problem-only model and the main model, enabling them to make predictions based on different characteristics of the input. The second method proposes two variants on the basis of RUBi, namely "RUBi+logarithmic space" method and "RUBi+standardization". The third method improves the method of improving a single classifier by reducing the weight of well-classified samples proposed by Lin et al. [19], and designs a new loss function "Debiased Focal Loss" to reduce the importance of biased samples, and make the model pay more attention to samples that require visual information to answer.

Wu et al. [19] found through the research that the existing research robust visual question answering model is encouraged to pay more attention to the image area that people think is important, even if the image area will eventually cause the model to produce wrong answers, and the so-called influential area Often identified automatically through human visual/textual annotations or more key words in questions and answers. In response to this phenomenon, the paper proposes a "self-criticism" method, which reduces the sensitivity to the area of the image by criticizing the image area that the wrong answer focuses on, so as to ensure that the visual interpretation of the correct answer can be compared with other competing answer candidates. More closely matches the most influential image areas. For the training of each sample, first determine the area that most affects the correctness of the model, and when the model answers the question wrongly through this image area, then punish the attention to this area, so as to ensure that the correct answer can be more important than other answers in the image There is stronger competitiveness in the region.

Generally, the visual question answering task is regarded as a classification problem. Although this method is easy to handle, each answer is independent of each other, and the similarity between answers cannot be calculated, so the semantic relationship between them cannot be considered. Kervadec et al. [21] believe that directly defining the loss function will lead to damage to the generalization of the model, and if there is content that is not in the training set during the test, it will also lead to incorrect prediction of the model answer. Aiming at this problem, the author designed a new loss function semantic-loss. By establishing the semantic space embedded in the answer, the distance function is defined to measure the similarity between the answers, so as to guide the model to adjust the loss value reasonably. Guo et al. [22] believe that the existing work on alleviating language priors for visual question answering models cannot explain the reasons for language bias well. It believes that the reason for the obvious error of the model during the test is caused by the sparseness of the answer. Therefore, Guo et al. [22] proposed to assign different weights to each answer, so that the model will have different losses when predicting the answer. In this way, the dependence of the model on the problem is adjusted. However, visual question answering datasets are often complex, large and interconnected, so it is difficult to assign reasonable weights to each answer.

Table 3. Methods based on data argumentation and training strategies [24]

Methods	Base	VQA-CP v2 test				VQA v2 val			
		All	Y/N	Num	Other	All	Y/N	Num	Other
ActSeek[30]	UpDn	46.00	58.24	29.49	44.33	-	-	-	-
A1C-WS[31]	UpDn	39.60	42.70	12.90	45.3	-	-	-	-
CSS[15]	UpDn	58.95	84.37	49.42	48.21	59.91	73.25	39.77	55.11
CL-VQA[32]	UpDn	59.18	86.99	49.89	47.16	57.29	67.27	38.40	54.71
GradSup[33]	UpDn	46.80	64.50	15.30	45.90	-	-	-	-
Mutant[34]	UpDn	61.72	88.90	49.68	50.78	62.56	82.07	42.52	53.28
RandImg[35]	UpDn	55.37	83.89	41.60	44.20	57.24	76.53	33.87	48.57
SSL[16]	UpDn	57.59	86.53	29.87	50.03	63.73	-	-	-
Unshuffling[36]	UpDn	42.39	47.72	14.43	47.24	61.08	78.32	42.16	52.81
LP-Focal[37]	UpDn	58.45	88.34	34.67	49.32	62.45	-	-	-
ADA-VQA[38]	UpDn	54.67	72.47	53.81	45.58	-	-	-	-
CCB-VQA[39]	UpDn	59.12	89.12	51.04	45.62	59.17	77.28	33.71	52.14
SBS[40]	UpDn	59.57	87.44	52.96	46.79	61.97	78.80	42.17	54.41
WeaQA[41]	UpDn	41.20	68.50	29.80	30.00	-	-	-	-
X-GGM[42]	UpDn	45.71	43.48	27.65	52.65	-	-	-	-
CFT-VQA[43]	UpDn	59.37	87.95	52.42	46.30	59.82	74.91	38.64	53.97

3. Conclusion

In the robustness research method of the visual question answering model, there are still problems such as poor generalization of the model, uninterpretable answers, and low accuracy of counting questions. Among them, the visual question answering model is affected by the surface correlation of the training data and lacks image foundation, that is, it is affected by the language prior problem and has become a popular research direction in the field of visual question answering in the past two years, and has become a landmark branch in visual question answering. This research direction attempts to solve the problem of unbalanced data distribution in multimodal information deep learning models, and unbalanced data distribution is also a common problem faced by machine learning.

The current VQA task is generally to predict the correct answer from the preset answers, and one of the sources of language bias makes the data distribution uneven. Therefore, enabling VQA models to answer questions using external databases and external datasets or combining knowledge graphs may effectively alleviate the problem of language bias. Auxiliary methods through external knowledge may become one of the research directions to address language bias. Due to the uneven distribution of question answers in the data set, there is a long-tail distribution problem, and the lack of finer-grained labels for question answers makes it difficult to transform the language bias problem into a long-tail problem. Due to the limitation of human resources and other practical factors, it is difficult to classify the data at a very fine-grained level, which leads to the imbalance of fine-grained labels in the data, and then the problem of language bias in the training process. How to solve the problem of language bias in the VQA task being transformed into a long-tail distribution, deeply mining language and visual information, and extracting fine-grained labels may become the research direction to effectively solve the problem of language bias in the future. Some of the existing methods for mitigating language priors and data augmentation can be categorized as causal inference methods. The problem of language bias in VQA can be analyzed from the perspective of causality, including the construction of causality graphs, counterfactual data augmentation, etc. Inferring models at multiple levels through targeted data augmentation and constructing counterfactual examples can analyze the source of language bias from a causal perspective and mitigate it to some extent. Therefore, allowing machine learning data to see profound causal relationships is an effective way to solve language bias.

To achieve true artificial intelligence, there is still a long way to go. In future research work, dealing with language prior issues is still a direction worthy of further research.

References

[1] Zhang P, Goyal Y, Summers-Stay D, et al. Yin and yang: Balancing and answering binary visual questions [C]// Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 5014-5022.

[2] Goyal Y, Khot T, Summers-Stay D, et al. Making the v in vqa matter: Elevating the role of image understanding in visual question answering [C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 6904-6913.

[3] Agrawal A, Batra D, Parikh D, et al. Don't just assume; look and answer: Overcoming priors for visual question answering [C]// Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 4971-4980.

[4] Selvaraju R R, Lee S, Shen Y, et al. Taking a hint: Leveraging explanations to make vision and language models more grounded[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2019: 2591-2600.

[5] Li L, Gan Z, Cheng Y, et al. Relation-aware graph attention network for visual question answering[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2019: 10313-10322.

[6] Hirota Y, Garcia N, Otani M, et al. A picture may be worth a hundred words for visual question answering [J]. https://doi. org/10. 48550/arXiv.2106.13445,2021-06-25.

[7] Si Q, Lin Z, yu Zheng M, et al. Check It Again: Progressive Visual Question Answering via Visual Entailment [C]//

Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). 2021: 4101-4110.

[8] Ramakrishnan S, Agrawal A, Lee S. Overcoming language priors in visual question answering with adversarial regularization[C]//Proceedings of the 32nd International Conference on Neural Information Processing Systems. 2018: 1548-1558.

[9] Grand G, Belinkov Y. Adversarial regularization for visual question answering: Strengths, shortcomings, and side effects. In: Proc. of the 57th Conf. on Computational Natural Language Learning.ACL, 2019. 1–13.

[10] Gat I, Schwartz I, Schwing A, et al. Removing bias in multi-modal classifiers: Regularization by maximizing functional entropies[J]. Advances in Neural Information Processing Systems, 2020, 33: 3197-3208.

[11] Cadene R, Dancette C, Ben-Younes H, et al. RUBi: Reducing Unimodal Biases for Visual Question Answering [C]//Neural Information Processing Systems. Curran Associates, Inc., 2019, 32: 841-852.

[12] Clark C, Yatskar M, Zettlemoyer L. Don't Take the Easy Way Out: Ensemble Based Methods for Avoiding Known Dataset Biases[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 2019: 4069-4082.

[13] Han X, Wang S, Su C, et al. Greedy gradient ensemble for robust visual question answering [C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 1584-1593.

[14] Selvaraju R R, Lee S, Shen Y, et al. Taking a hint: Leveraging explanations to make vision and language models more grounded [C]//Proceedings of the IEEE/CVF international conference on computer vision. 2019: 2591-2600.

[15] Chen L, Zheng Y, Niu Y, et al. Counterfactual samples synthesizing and training for robust visual question answering [J].https: //doi.org/10.48550/arXiv.2110.01013, 2021-10-03.

[16] Zhu X, Mao Z, Liu C, et al. Overcoming language priors with self-supervised learning for visual question answering [C]// Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence. 2021: 1083-1089.

[17] Gokhale T, Banerjee P, Baral C, et al. MUTANT: A training paradigm for out-of-distribution generalization in visual question answering[C]//2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020. Association for Computational Linguistics (ACL), 2020: 878-892.

[18] Mahabadi R K, Henderson J. Simple but effective techniques to reduce biases. [J]https:// doi.org/ 10. 48550/ arXiv. 1909. 06321, 2020-04-23.

[19] Lin T Y, Goyal P, Girshick R, et al. Focal loss for dense object detection[C]//Proceedings of the IEEE international conference on computer vision. 2017: 2980-2988.

[20] Wu J, Mooney R J. Self-critical reasoning for robust visual question answering[C]//Proceedings of the 33rd International Conference on Neural Information Processing Systems. 2019: 8604-8614.

[21] Kervadec C, Antipov G, Baccouche M, et al. Estimating semantic structure for the VQA answer space [J]. https: // doi. org/ 10.48550/arXiv.2006.05726, 2021-04-08.

[22] Guo Y, Nie L, Cheng Z, et al. Loss re-scaling VQA: revisiting the language prior problem from a class-imbalance view[J]. IEEE Transactions on Image Processing, 2021, 31: 227-238.

[23] Shrestha R, Kafle K, Kanan C. A negative case analysis of visual grounding methods for VQA [C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020: 8172-8181.

[24] Yuan D. Language bias in visual question answering: A survey and taxonomy[J]. https://doi. org/ 10. 48550/ arXiv. 2111. 08531, 2021-11-06.

[25] Wu J, Mooney R. Self-critical reasoning for robust visual question answering[J]. Advances in Neural Information Processing Systems, 2019, 32.

[26] Shrestha R, Kafle K, Kanan C. A negative case analysis of visual grounding methods for VQA[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020: 8172-8181.

[27] Kv G, Mittal A. Reducing language biases in visual question answering with visually-grounded question encoder[C]// Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16. Springer International Publishing, 2020: 18-34.

[28] Zhang L, Liu S, Liu D, et al. Rich visual knowledge-based augmentation network for visual question answering[J]. IEEE Transactions on Neural Networks and Learning Systems, 2020, 32(10): 4362-4373.

[29] Liang Z, Hu H, Zhu J. LPF: A language-prior feedback objective function for de-biased visual question answering [C]//Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval. 2021: 1955-1959.

[30] Teney D, Hengel A. Actively seeking and learning from live data [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 1940-1949.

[31] Zhou Y, Ji R, Sun X, et al. Plenty is plague: Fine-grained learning for visual question answering[J]. IEEE transactions on pattern analysis and machine intelligence, 2019, 44(2): 697-709.

[32] Liang Z, Jiang W, Hu H, et al. Learning to contrast the counterfactual samples for robust visual question answering [C]//Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP). 2020: 3285-3292.

[33] Teney D, Abbasnedjad E, van den Hengel A. Learning what makes a difference from counterfactual examples and gradient supervision[C]//Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part X 16. Springer International Publishing, 2020: 580-599.

[34] Gokhale T, Banerjee P, Baral C, et al. Mutant: A training paradigm for out-of-distribution generalization in visual question answering[J]. https://doi. org/10. 48550/ arXiv. 2009. 08566, 2020-10-16.

[35] Teney D, Abbasnejad E, Kafle K, et al. On the value of out-of-distribution testing: An example of goodhart's law[J]. Advances in neural information processing systems, 2020, 33: 407-417.

[36] Teney D, Abbasnejad E, van den Hengel A. Unshuffling data for improved generalization in visual question answering [C]// Proceedings of the IEEE/CVF international conference on computer vision. 2021: 1417-1427.

[37] Lao M, Guo Y, Liu Y, et al. A language prior based focal loss for visual question answering[C]//2021 IEEE International Conference on Multimedia and Expo (ICME). IEEE, 2021: 1-6.

[38] Guo Y, Nie L, Cheng Z, et al. Adavqa: Overcoming language priors with adapted margin cosine loss[J]. https:// doi.org/ 10. 48550/arXiv.2105.01993,2021-05-05.

[39] Yang C, Feng S, Li D, et al. Learning content and context with language bias for visual question answering[C]//2021 IEEE International Conference on Multimedia and Expo (ICME). IEEE, 2021: 1-6.

[40] Ouyang N, Huang Q, Li P, et al. Suppressing biased samples for robust vqa[J]. IEEE Transactions on Multimedia, 2021, 24: 3405-3415.

[41] Banerjee P, Gokhale T, Yang Y, et al. WeaQA: Weak Supervision via Captions for Visual Question Answering[J]. Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, 2021.

[42] Jiang J, Liu Z, Liu Y, et al. X-ggm: Graph generative modeling for out-of-distribution generalization in visual question answering[C]//Proceedings of the 29th ACM international conference on multimedia. 2021: 199-208.

[43] D. Yuan, X. Liu, Q. Wu, H. Li, F. Meng, K. N. Ngan, and L. Xu, “Empower counterfactual thinking via contrastive learning for robust visual question answering,” in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP). IEEE, 2022, p. under review.