Overcoming Language Priors with Counterfactual Inference for Visual Question Answering

Zhibo Ren, Huizhen Wang*, Muhua Zhu, Yichao Wang, Tong Xiao, Jingbo Zhu
NLP Lab, School of Computer Science and Engineering,
Northeastern University, Shenyang, China
rzb1998@qq.com
wanghuizhen@mail.neu.edu.cn

Abstract

Recent years have seen a lot of efforts in attacking the issue of language priors in the field of Visual Question Answering (VQA). Among the extensive efforts, causal inference is regarded as a promising direction to mitigate language bias by weakening the direct causal effect of questions on answers. In this paper, we follow the same direction and attack the issue of language priors by incorporating counterfactual data. Moreover, we propose a two-stage training strategy which is deemed to make better use of counterfactual data. Experiments on the widely used benchmark VQA-CP v2 demonstrate the effectiveness of the proposed approach, which improves the baseline by 21.21% and outperforms most of the previous systems.

1 Introduction

As an AI-complete task to answer questions about visual content, Visual Question Answering (VQA) has seen surging interest in recent years. The task is thought to be extremely challenging since a VQA system requires the capability of visual and language understanding and the capability of multi-modal reasoning. Recent researches in this field have paid increasing attention to the issue of language priors, aka language bias (Agrawal et al., 2018). The issue of language priors is caused by spurious correlation between the question pattern and the answer. See the example in Figure 1, "yellow" is the most likely answer to the question "what color are the bananas" in the training data. So a simple solution to answering the question is to give the answer "yellow" with no reference to visual content. Such a short cut can achieve an accuracy of 54.5% for the question.

To overcome language priors in VQA, previous works generally resort to data augmentation. In this direction, visual and textual explanations can be used as the data for augmentation (Das et al., 2017; Park et al., 2018). Besides, counterfactual training samples are also regarded as a valuable source for the purpose (Chen et al., 2020; Zhu et al., 2020; Gokhale et al., 2020; Liang et al., 2020). In the direction of causal effect for VQA, more recent work is counterfactual VQA that focuses on the inference instead of training phase (Niu et al., 2021), though, we still think of counterfactual data augmentation as an efficient and effective way to solve the issue of language priors. So in this paper we first design novel causal graphs specifically for the task of VQA, and then use the causal graphs to guide the generation of counterfactual data. Finally, to make better use of counterfactual data, we propose a two-stage training strategy. We evaluate the proposed approach on the widely used benchmark VQA-CP v2. Extensive experiments demonstrate the effectiveness of the approach, which improves over the baseline by 21.21% and outperforms most of previous systems. Moreover, to evaluate the generalization ability of the approach, we also experiment with VQA v2 and find that our approach achieves the best performance on the dataset.

The contributions of the paper are as follows.

- For the task of counterfactual VQA, we design a novel causal graph and methods to construct counterfactual data.

*Corresponding author.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 600-610, Harbin, China, August 3 - 5, 2023.

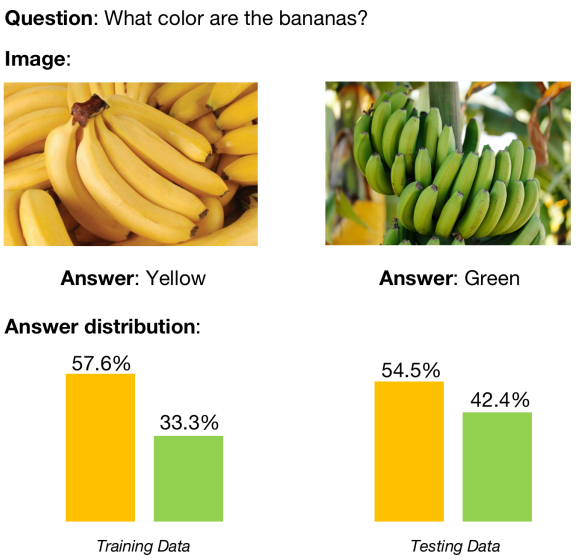

Figure 1: An example from VQA v2 which is used to illustrate 1) the task of visual question answering, and 2) the issue of language priors.

- Our approach achieves significant improvements over the baseline and is one of the best-performing systems on the benchmarks.

2 Methodology

In this section, we first describe the implementation of our baseline system. Then we introduce the design of VQA causal graphs which inspire us to come up with the proposed methods. Finally we describe the methods in detail. The system framework is presented in Figure 2.

2.1 The Baseline System

Following the conventional paradigm of VQA systems, we formalize the task as a multi-class classification problem. In general, a VQA dataset consists of N instances which are tuples of an image, a textual question, and the corresponding answer, denoted as $D = \{I_i, Q_i, A_i\}_{i=1}^{N}$. VQA models take an image-question pair (I, Q) as input, and predict an answer A by following

$$A^* = \arg\max_{A \in \mathcal{A}} P(A|I_i, Q_i), \tag{1}$$

where $P(A|I_i, Q_i)$ can be any model-based functions that map (I, Q) to produce a distribution over the answer space $\mathcal{A}$. Conventional VQA systems are generally composed of three components:

- **Feature Extraction**, which extracts the features of images and question as visual representation and text representation, respectively.
- **Multimodal Feature Fusion**, which fuses image and text features into the same vector space.
- **Answer Prediction**, which produces the answer prediction through a classifier.

We follow (Anderson et al., 2018) to implement our baseline system. The baseline system pays special attention to feature extraction by integrating a combined bottom-up and top-down attention mechanism to enable attention calculation at the fine-grained level of objects. Within the approach, the bottom-up attention proposes image regions while the top-down mechanism determines feature weightings.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 600-610, Harbin, China, August 3 - 5, 2023.

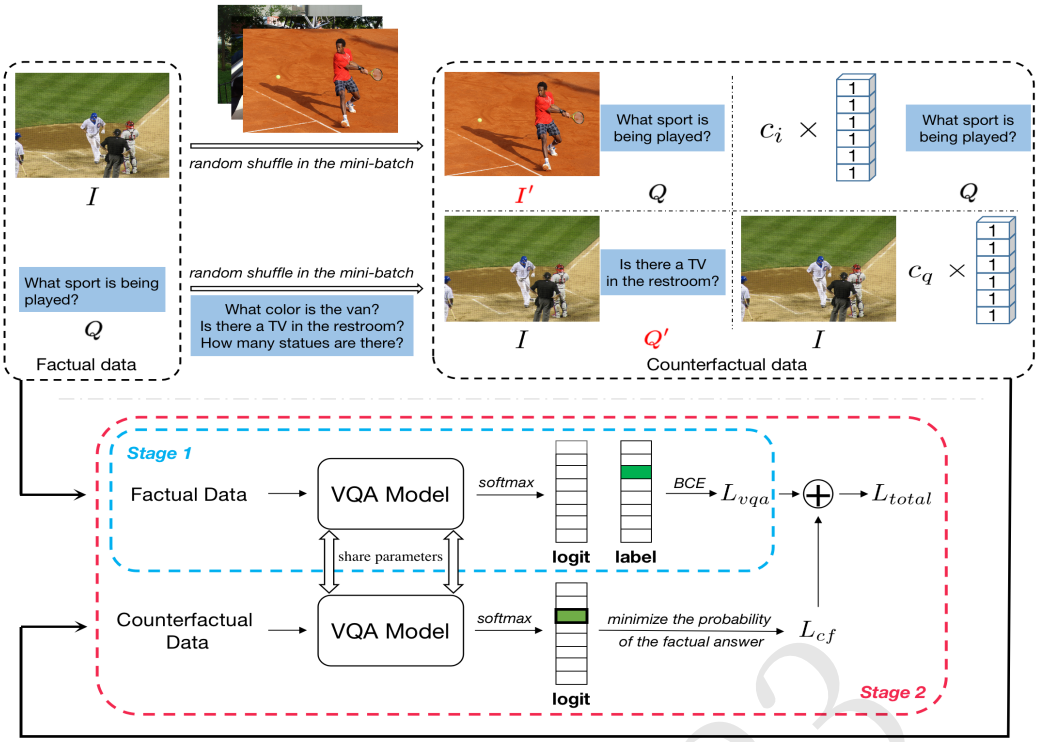

Figure 2: Illustration of our approach, where the upper half presents the process of counterfactual data generation and the bottom half represents the process of two-stage training.

2.2 Causal Graph for VQA

To better understand the casual graphs we propose for the VQA task, we need to revisit the procedure of VQA data annotation. Specifically, when curating a dataset, annotators are required to produce a question regarding visual content of a presented image and give a correct answer. Therefore, we can construct a casual graph to exhibit the relationship between three variables: the image I, the question Q, and the answer A. Figure 3(a) illustrates the casual graph, where I indirectly and directly affects A through $I \rightarrow Q \rightarrow A$ and $I \rightarrow A$, respectively. In the chain of $I \rightarrow Q \rightarrow A$, the question Q acts as a mediator to influence A. If we control the mediator Q, the causal association between I and A in the chain $I \rightarrow Q \rightarrow A$ will be blocked, that is, when the association between I and A is not well learned through $I \rightarrow A$ (the middle and right graph in Figure 3(a)), the model will give the answer based on the question only but ignore the content of the image. This phenomenon corresponds exactly to the language prior problem in VQA. Therefore, we propose to introduce counterfactual data to weaken the effect that comes from the chain $I \rightarrow Q \rightarrow A$, which is shown in Figure 3(b).

2.3 Automatic Generation of Counterfactual Data

We propose two methods to construct counterfactual data, corresponding to multimodal counterfactual data and unimodal counterfactual data, respectively.

Multimodal Counterfactual Data. First of all, we realize that the issue of language priors is caused by the chain $I \rightarrow Q \rightarrow A$, so we need to mitigate the influence of this branch on the selection of the answer. Inspired by (Zhu et al., 2020), for each pair(I_i, Q_i) in factual data, we construct counterfactual data (I_i', Q_i) by shuffling image I_i in the same mini-batch, such that the image and the question in counterfactual data are mismatched. The causal graph of counterfactual image data is shown in Figure 4(a). Following the same idea, we also propose to construct counterfactual question data by shuffling questions in the same mini-batch. The corresponding causal graph is illustrated in Figure 4(b). Subsequent experiments show that incorporation of multimodal counterfactual question data is also beneficial to the performance, which demonstrates the presence of vision bias in the VQA task, a phenomenon not often

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 600-610, Harbin, China, August 3 - 5, 2023.

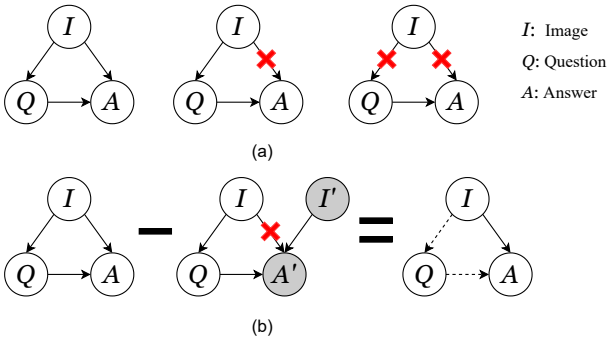

Figure 3: (a) Casual graph for VQA. (b) Overcome language priors with counterfactual data.

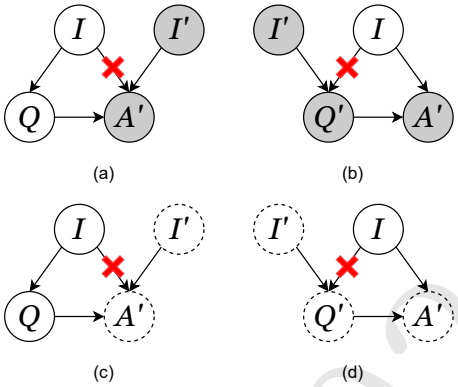

Figure 4: Causal graph demonstrating the methods for generating counterfactual data.

mentioned before.

It is worth noting that we do not resort to any extra human annotations during the construction of the multimodal counterfactual data, but simply make use of the factual data itself. The underlying idea is quite different from the methods proposed in previous works for the construction of counterfactual data (Chen et al., 2020; Liang et al., 2020; Gokhale et al., 2020).

Unimodal Counterfactual Data. We further consider to construct unimodal counterfactual data. We hope the model to accept information from only one modality as input. Concretely, we construct unimodal counterfactual data by passing only images$(I_i, \emptyset)$ or questions$(\emptyset, Q_i)$ into the model, which the causal graph is illustrated in Figure 4(c)(d). However, the model cannot handle the case where the input is empty during implementation, so we choose to use a learnable parameter c multiplied by a matrix whose elements are all ones and the shape is same as image representation or text representation as the null modal information. Finally, the unimodal counterfactual data can be represented as (I_i, c_q) and (c_i, Q_i).

2.4 Two-stage Training Strategy

In the real world, we can only give the right answer when we see the right factual image-question pair. Conversely, we often cannot give the correct answer when we see a counterfactual image-question pair. But usually in this case the correct answer will change and the previously correct answer will often become the wrong answer, which is the only thing we know for sure. We hope to solve language prior problems by using counterfactual image data in the manner shown in Figure 3(b). Specifically, when the VQA model takes the counterfactual image data as input, we construct the loss function by minimizing the probability of the ground truth answer:

$$\begin{aligned} P(A'|I'_i, Q_i) &= softmax(F(I'_i, Q_i)) \\ L_{mm\_cf\_i} &= P(A'|I'_i, Q_i)[k] \end{aligned} \tag{2}$$

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 600-610, Harbin, China, August 3 - 5, 2023.

Systems	VQA-CP v2 test(%)				VQA v2 val(%)			
	All	Y/N	Num	Other	All	Y/N	Num	Other
UpDn	39.74	42.27	11.93	46.05	63.48	81.18	42.14	55.66
GVQA	31.3	57.99	13.68	22.14	48.24	72.03	31.17	34.65
SAN	24.96	38.35	11.14	21.74	52.41	70.06	39.28	47.84
Systems without counterfactual inference								
DLR	48.87	70.99	18.72	45.57	57.96	76.82	39.33	48.54
VGQE	48.75	-	-	-	64.04	-	-	-
AdvReg	41.17	65.49	15.48	35.48	62.75	79.84	42.35	55.16
RUBi	44.23	67.05	17.48	39.61	-	-	-	-
LMH	52.01	72.58	31.12	46.97	56.35	65.06	37.63	54.69
CVL	42.12	45.72	12.45	48.34	-	-	-	-
Unshuffling	42.39	47.72	14.43	47.24	61.08	78.32	42.16	52.81
RandImg	55.37	83.89	41.6	44.2	57.24	76.53	33.87	48.57
SSL	57.59	86.53	29.87	50.03	63.73	-	-	-
Systems with counterfactual inference								
CSS	58.95	84.37	49.42	48.21	59.91	73.25	39.77	55.11
CSS+CL	59.18	86.99	49.89	47.16	57.29	67.29	38.40	54.71
CF-VQA	53.55	**91.15**	13.03	44.97	63.54	**82.51**	43.96	54.30
MUTANT	**61.72**	88.90	49.68	**50.78**	62.56	82.07	42.52	53.28
This Paper	60.95	87.95	**50.41**	49.70	**64.11**	82.23	**44.09**	**56.75**

Table 1: Comparison with the state-of-the-art methods on the VQA-CP v2 test set and VQA v2 validation set. The evaluation metric is accuracy, and the backbone of all models is UpDn. Overall best scores are **bold** and the second best of scores are underlined.

where k denotes the index of the ground truth in the answer set A. For the counterfactual question data, the corresponding loss function is similar to equation(2): , which can be defined as:

$$\begin{aligned} P(A'|I_i, Q_i') &= softmax(F(I_i, Q_i')) \\ L_{mm\_cf\_q} &= P(A'|I_i, Q_i')[k] \end{aligned} \tag{3}$$

Finally, the loss of the multimodal counterfactual data is defined as:

$$L_{mm\_cf} = \lambda_i^{mm} L_{mm\_cf\_i} + \lambda_q^{mm} L_{mm\_cf\_q}, \tag{4}$$

where λ_i and λ_q are hyperparameters.
Similar to multimodal counterfactual data, the unimodal counterfactual loss function can be defined as:

$$\begin{aligned} P(A'|c_i, Q_i) &= softmax(F(c_i, Q_i)) \\ L_{um\_cf\_i} &= P(A'|c_i, Q_i)[k] \end{aligned} \tag{5}$$

$$\begin{aligned} P(A'|I_i, c_q) &= softmax(F(I_i, c_q)) \\ L_{um\_cf\_q} &= P(A'|I_i, c_q)[k] \end{aligned} \tag{6}$$

The total loss of unimodal counterfactual data is defined as:

$$L_{um\_cf} = \lambda_i^{um} L_{um\_cf\_i} + \lambda_q^{um} L_{um\_cf\_q} \tag{7}$$

Simply combining counterfactual and factual data together as training data may render these two types of data interfere with each other. For this reason, we adopt a two-stage training strategy, which utilize factual data and the normal VQA loss function for training in the first stage and utilize counterfactual data and counterfactual loss functions in the second stage. are introduced on top of the first stage to alleviate the problem of the language priors of the VQA model:

$$L_{total} = L_{vqa} + \lambda^{mm} L_{mm\_cf} + \lambda^{um} L_{um\_cf} \tag{8}$$

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 600-610, Harbin, China, August 3 - 5, 2023.

3 Experiments

3.1 Datasets and Comparative Systems

Datasets. We conducted extensive experiments on the most widely used benchmark VQA-CP v2 (Agrawal et al., 2018) adopting the standard evaluation metric. Because the dataset of VQA v2 (Goyal et al., 2017) has the language prior problem, (Agrawal et al., 2018) reorganized the data splitting of VQA v2 to construct VQA-CP v2 where answers have different distributions in the training and validation set. Thus, VQA-CP v2 is an appropriate benchmark for evaluating the generalization ability of VQA models. Briefly, the training set of VQA-CP v2 contains approximately 121k images and 245k questions, and the test set consists of approximately 98k images and 220k questions.

Comparative Systems. System participating in the comparison against our approach can be categorized into two groups: 1) systems without counterfactual inference, including **DLR** (Jing et al., 2020), **VGQE** (KV and Mittal, 2020), **AdvReg** (Ramakrishnan et al., 2018), **RUBi** (Cadène et al., 2019), **LMH** (Clark et al., 2019), **Unshuffling** (Teney et al., 2021), **RandImg** (Teney et al., 2020), **SSL** (Zhu et al., 2020), and 2) systems with counterfactual inference, including **CF-VQA** (Niu et al., 2021), **CSS** (Chen et al., 2020), **CL** (Liang et al., 2020), and **MUTANT** (Gokhale et al., 2020).

3.2 Implementation Details

As mentioned above, our VQA system builds on the base of UpDn (Anderson et al., 2018). Following previous researches, we use the Faster-RCNN (Ren et al., 2015) model previously trained by (Anderson et al., 2018) to extract image features. We extract 36 region features for each image and the dimension of each region feature is set to 2048. Moreover, each question is padded so as to have the same length of 14 tokens, and each token in questions is encoded by the pretrained language model BERT (Devlin et al., 2019) with a dimension of 768. Then word embeddings are fed into GRUs to obtain the question representation with a dimension of 1280. Inspired by SSL (Zhu et al., 2020), we also add a BatchNorm layer before the MLP classifier of UpDn. We train our model for 25 epochs every time. We adopt the Adam optimizer to update model parameters, whose learning rate is set to 0.001 and the learning rate decreases by half every 5 epochs after 10 epochs. The batch size is set to 256. We implement our system using PyTorch, and we train our model with one Nvidia 2080Ti card.

3.3 Main Experimental Results

Table 1 presents the comparison results between our approach and previous systems on both VQA-CP v2. From the results, we can see that our approach significantly improves the baseline UpDn by $+21.21\%$ on VQA-CP v2. The improvement demonstrates the effectiveness of our approach on mitigating the issue of language prior. Moreover, our approach outperforms all the comparative systems on VQA-CP v2 except for MUTANT which requires additional human annotations of key objects in images. Moreover, we can see our approach achieves stable performance on VQA v2 with the best performance over all the previous systems. To demonstrate the generality of our approach, we also experiment with VQA v2, and the results show that our approach achieves the best performance among all the participating systems.

3.4 Experiment Analysis

Impact of Counterfactual Data Combination

We proposed several types of counterfactual data, so we conducted a study on the effect of each type of counterfactual data and the effect of their combinations. From the results shown in Table 2, we have the following observations:

- Both counterfactual image data (I_i', Q_i) and counterfactual question data (I_i, Q_i') are able to improve the performance. The use of counterfactual image data achieves significant improvements, while the counterfactual question data achieves relatively limited improvements. This suggests that the main cause of the language prior problem is the superficial correlation between questions and answers, but there are also some vision bias that cannot be ignored.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 600-610, Harbin, China, August 3 - 5, 2023.

	Counterfactual Data (I_i', Q_i)	(I_i, Q_i')	(c_i, Q_i)	(I_i, c_q)	Acc.
			-		41.52
MM	✔	-	-	-	57.59
	-	✔	-	-	41.87
	✔	✔	-	-	**59.05**
UM	-	-	✔	-	41.83
	-	-	-	✔	41.70
	-	-	✔	✔	**41.88**
Total	✔	✔	✔	✔	**60.95**

Table 2: Impact of different types of counterfactual data, evaluated on VQA-CP v2 test set. MM refers to multimodal counterfactual data and UM refers to unimodal counterfactual data, respectively

λ	Ratio	VQA-CP v2 test(%)
λ_i^{mm}:λ_q^{mm}	1:0.5	58.06
	1:0.7	**59.46**
	1:1	59.32
	1:2	59.15
	1:3	58.76
λ_i^{um}:λ_q^{um}	1:0.5	60.03
	1:0.7	60.29
	1:1	**60.34**
	1:2	59.51
	1:3	58.07
λ^{mm}:λ^{um}	1:0.5	60.17
	1:0.7	60.53
	1:1	**60.95**
	1:2	58.21
	1:3	60.29

Table 3: Impact of different ratio between λ. We divide λ into three groups($\lambda_i^{mm} : \lambda_q^{mm}$), ($\lambda_i^{um} : \lambda_q^{um}$),($\lambda^{mm} : \lambda^{um}$) according to the counterfactual data used, with the latter group realized on the best results of the previous group's experiment. The evaluation metric is accuracy(%).

- Both multimodal counterfactual data and unimodal counterfactual data can improve the model performance, which demonstrates that these data can prompt the generalization ability of model.

In summary, the above experimental results verify the validity of the counterfactual data.

Impact of Varying Settings of λ

As we can see from the results in Table 2, different types of counterfactual data have diverse effect on the performance. So we need to evaluate the effect of varying settings of the hyperparameters λ in the loss functions. We divide λ into three groups for comparison and conducted extensive experiments with different λ values. From results in Table 3, we can observe that the model gets the best performance when $\lambda_i^{mm} : \lambda_q^{mm}$ is 1:0.7, $\lambda_i^{um} : \lambda_q^{um}$ is 1:1, and $\lambda^{mm} : \lambda^{um}$ is 1:1.

Impact of Varying Starting Points of the Second Stage Training

In the process of two-stage training, different starting points of the second stage tend to achieve different results. So we conducted an experiment to show the effect of varying starting points. As can be seen in Figure 5, starting the training on counterfactual data too early or too late will bring negative effect on

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 600-610, Harbin, China, August 3 - 5, 2023.

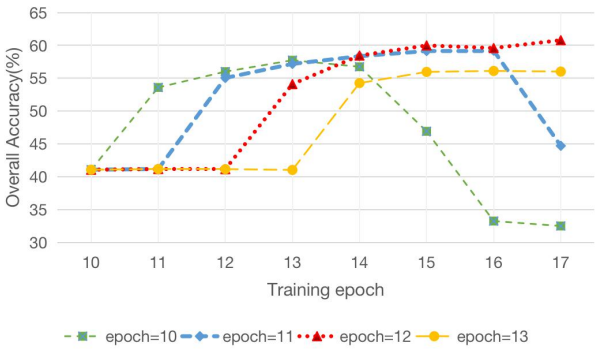

Figure 5: Impact of different starting points of the second stage training, evaluated on the VQA-CP v2 test set.

Methods	Overall(%)	Gap Δ ↑
UpDn	39.74	**+21.21**
UpDn + counterfactual data	**60.95**	
SAN	24.96	**+27.46**
SAN + counterfactual data	**52.42**	

Table 4: Performance of different backbones on VQA-CP v2 test set.

the performance. Empirically, we find the second stage can start its training at the 12th epoch.

Impact of Different Backbones

We also conducted experiments on another backbones SAN (Yang et al., 2016) to verify that our approach is model agnostic. From the results in Table 4, we can observe that our approach can achieve significant improvements no matter what backbone is used.

4 Related Work

Visual Question Answering

Visual Question Answering aims to answer the question according to the given image, which involves both natural language processing and computer vision techniques. At present, the dominant methods are attention-based models. (Anderson et al., 2018; Yu et al., 2019; Yang et al., 2016) use attentions mechanisms to capture the alignment between images and natural language in order to learn the intrinsic interactions between image regions and words. (Antol et al., 2015) maps two modal features(visual and textual features) into a common feature space and then passes the joint embedding into the classifier to obtain the answer of the question. Another methods including that compositional models that (Andreas et al., 2016) applies neural module network to the VQA task, which is a combination of several modular networks. The neural module network is dynamically generated according to the linguistic structure of the question. (Wu et al., 2016) introduces external knowledge to help model with answering the questions.

Attacking Language Priors in VQA

Despite the progress made in the field of VQA, recent researches have found that VQA systems tend to exploit superficial correlations between question patterns and answers to achieve state-of-the-art performance (Agrawal et al., 2016; Kafle and Kanan, 2017). To help build a robust VQA system, (Agrawal et al., 2018) propose a new benchmark named VQA-CP whose training and testing data have vast distributions. Recent solutions to overcome the language priors can be grouped into two categories as without counterfactual inference (Clark et al., 2019; Zhu et al., 2020; Teney et al., 2021) and with counterfactual inference (Agrawal et al., 2019; Pan et al., 2019; Chen et al., 2020; Liang et al., 2020;

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 600-610, Harbin, China, August 3 - 5, 2023.

Gokhale et al., 2020).
For the methods that without counterfactual inference, RUBi (Cadène et al., 2019) proposes to dynamically adjust the weights of samples according to the effect of the bias, LMH (Clark et al., 2019) ensembles a question-only branch to discriminates which questions can be answered without utilizing image and then penalizes these questions. Unshuffling (Teney et al., 2021) describes a training procedure to capture the patterns that are stable across environments while discarding spurious ones. SSL (Zhu et al., 2020) proposes a self-supervised framework that generates labeled data to balance the biased data.
For the methods that with counterfactual inference, One solution is is to modify model architecture that implement counterfactual inference to reduce the language bias (Niu et al., 2021). The other one is to synthesize counterfactual samples to improve the robustness of VQA systems (Agrawal et al., 2019; Pan et al., 2019; Chen et al., 2020; Liang et al., 2020; Gokhale et al., 2020). CSS (Chen et al., 2020) generates the counterfactual samples by masking objects in the image or some keywords in the question. Based on CSS, CL (Liang et al., 2020) introduces a contrastive learning mechanism to force the model to learn the relationship between original samples, factual samples and counterfactual samples. MUTANT (Gokhale et al., 2020) utilizes the extra object-name annotations to locates critical objects in the image and critical words in the question and then mutates these critical elements to generate counterfactual samples.

5 Conclusion and Future Work

To mitigate the effect of language priors in the VQA task, we proposed a causal inference approach that automatically generates counterfactual data and utilize the data in a two-stage training strategy. We also designed several causal graphs to guide the generation of counterfactual data. Extensive experiments on the benchmark VQA-CP v2 shows that our system achieves significant improvements over the baselines and outperforms most of previous works. Moreover, our system achieves the best performance on VQA v2 which demonstrates the capability of generalization.

The starting point of the the second stage training is critical to the performance, in our future work, we would like to determine the starting point in an automatic way. Moreover, it is interesting to evaluate the performance when other networks such as SAN are used as the backbone. We will also study this problem in our future work.

Acknowledgements

This work was supported in part by the National Science Foundation of China (No. 62276056), the National Key RD Program of China, the China HTRD Center Project (No. 2020AAA0107904), the Natural Science Foundation of Liaoning Province of China (2022-KF-16-01), the Yunnan Provincial Major Science and Technology Special Plan Projects (No. 202103AA080015), the Fundamental Research Funds for the Central Universities (Nos. N2216016, N2216001, and N2216002), and the Program of Introducing Talents of Discipline to Universities, Plan 111 (No. B16009).

References

Aishwarya Agrawal, Dhruv Batra, and Devi Parikh. 2016. Analyzing the behavior of visual question answering models. In *Proceedings of EMNLP*, pages 1955–1960.

Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. 2018. Don't just assume; look and answer: Overcoming priors for visual question answering. In *Proceedings of CVPR*, pages 4971–4980.

Vedika Agrawal, Rakshith Shetty, and Mario Fritz. 2019. Towards causal VQA: Revealing and reducing spurious correlations by invariant and covariant semantic editing. In *Proceedings of CVPR*, pages 9690–9698.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of CVPR*, pages 6077–6086.

Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. 2016. Neural module networks. In *Proceedings of CVPR*, pages 39–48.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: visual question answering. In *Proceedings of ICCV*, pages 2425–2433.

Rémi Cadène, Corentin Dancette, Hedi Ben-younes, Matthieu Cord, and Devi Parikh. 2019. Rubi: Reducing unimodal biases for visual question answering. In *Proceedings of NeurIPS*, pages 839–850.

Long Chen, Xin Yan, Jun Xiao, Hanwang Zhang, Shiliang Pu, and Yueting Zhuang. 2020. Counterfactual samples synthesizing for robust visual question answering. In *Proceedings of CVPR*, pages 10797–10806.

Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. 2019. Don't take the easy way out: Ensemble based methods for avoiding known dataset biases. In *Proceedings of EMNLP-IJCNLP*, pages 4067–4080.

Abhishek Das, Harsh Agrawal, Larry Zitnick, Devi Parikh, and Dhruv Batra. 2017. Human attention in visual question answering: Do humans and deep networks look at the same regions? *Computer Vision and Image Understanding*, 163:90–100.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Tejas Gokhale, Pratyay Banerjee, Chitta Baral, and Yezhou Yang. 2020. MUTANT: A training paradigm for out-of-distribution generalization in visual question answering. In *Proceedings of EMNLP*, pages 878–892.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *Proceedings of CVPR*, pages 6325–6334.

Chenchen Jing, Yuwei Wu, Xiaoxun Zhang, Yunde Jia, and Qi Wu. 2020. Overcoming language priors in VQA via decomposed linguistic representations. In *Proceedings of AAAI*, pages 11181–11188.

Kushal Kafle and Christopher Kanan. 2017. An analysis of visual question answering algorithms. In *Proceedings of CVPR*, pages 1983–1991.

Gouthaman KV and Anurag Mittal. 2020. Reducing language biases in visual question answering with visually-grounded question encoder. In *Proceedings of ECCV*, volume 12358, pages 18–34.

Zujie Liang, Weitao Jiang, Haifeng Hu, and Jiaying Zhu. 2020. Learning to contrast the counterfactual samples for robust visual question answering. In *Proceedings of EMNLP*, pages 3285–3292.

Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, and Ji-Rong Wen. 2021. Counterfactual VQA: A cause-effect look at language bias. In *Proceedings of CVPR*, pages 12700–12710.

Jingjing Pan, Yash Goyal, and Stefan Lee. 2019. Question-conditional counterfactual image generation for VQA. In *arXiv, preprint arXiv:1911.06352*.

Dong Huk Park, Lisa Anne Hendricks, Zeynep Akata, Anna Rohrbach, Bernt Schiele, Trevor Darrell, and Marcus Rohrbach. 2018. Multimodal explanations: Justifying decisions and pointing to the evidence. In *Proceedings of CVPR*, pages 8779—-8788.

Sainandan Ramakrishnan, Aishwarya Agrawal, and Stefan Lee. 2018. Overcoming language priors in visual question answering with adversarial regularization. In *Proceedings of NeurIPS*, pages 1548–1558.

Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. 2015. Faster R-CNN: towards real-time object detection with region proposal networks. In *Proceedings of NeurIPS*, pages 91–99.

Damien Teney, Ehsan Abbasnejad, Kushal Kafle, Robik Shrestha, Christopher Kanan, and Anton van den Hengel. 2020. On the value of out-of-distribution testing: An example of goodhart's law. In *Proceedings of NeurIPS*.

Damien Teney, Ehsan Abbasnejad, and Anton van den Hengel. 2021. Unshuffling data for improved generalization in visual question answering. In *Proceedings of ICCV*, pages 1397–1407.

Qi Wu, Peng Wang, Chunhua Shen, Anthony R. Dick, and Anton van den Hengel. 2016. Ask me anything: Free-form visual question answering based on knowledge from external sources. In *Proceedings of CVPR*, pages 4622–4630.

Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alexander J. Smola. 2016. Stacked attention networks for image question answering. In *Proceedings of CVPR*, pages 21–29.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 600-610, Harbin, China, August 3 - 5, 2023.

Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. 2019. Deep modular co-attention networks for visual question answering. In *Proceedings of CVPR*, pages 6281–6290.

Xi Zhu, Zhendong Mao, Chunxiao Liu, Peng Zhang, Bin Wang, and Yongdong Zhang. 2020. Overcoming language priors with self-supervised learning for visual question answering. In *Proceedings of IJCAI*, pages 1083–1089.