Int. J. Appl. Math. Comput. Sci., 2024, Vol. 34, No. 3, 453–466
DOI: 10.61822/amcs-2024-0032

DCF–VQA: COUNTERFACTUAL STRUCTURE BASED ON MULTI–FEATURE ENHANCEMENT

GUAN YANG [a,b,*], CHENG JI [a,b], XIAOMING LIU [a,b], ZIMING ZHANG [a,b], CHEN WANG [a,b]

[a]School of Computer Science
Zhongyuan University of Technology
No. 41, Zhongyuan Middle Road, 450007 Zhengzhou, Henan, China
e-mail: 6172@zut.edu.cn

[b]Henan Key Laboratory on Public Opinion Intelligent Analysis
Zhongyuan University of Technology
No. 41, Zhongyuan Middle Road, 450007 Zhengzhou, Henan, China

Visual question answering (VQA) is a pivotal topic at the intersection of computer vision and natural language processing. This paper addresses the challenges of linguistic bias and bias fusion within invalid regions encountered in existing VQA models due to insufficient representation of multi-modal features. To overcome those issues, we propose a multi-feature enhancement scheme. This scheme involves the fusion of one or more features with the original ones, incorporating discrete cosine transform (DCT) features into the counterfactual reasoning framework. This approach harnesses fine-grained information and spatial relationships within images and questions, enabling a more refined understanding of the indirect relationship between images and questions. Consequently, it effectively mitigates linguistic bias and bias fusion within invalid regions in the model. Extensive experiments are conducted on multiple datasets, including VQA2 and VQA-CP2, employing various baseline models and fusion techniques, resulting in promising and robust performance.

Keywords: visual question answering, multi-feature enhancement, counterfactual, discrete cosine transform.

1. Introduction

Machine learning has found extensive applications across various domains (Yang *et al.*, 2023a; 2023b; Surówka and Ogorzałek, 2022). Among them, visual question answering (VQA) (Antol *et al.*, 2015) has emerged as a fundamental component underpinning numerous cutting-edge interactive artificial intelligence systems, including visual dialogue (Das *et al.*, 2017b), visual language navigation (Anderson *et al.*, 2018), and visual common sense reasoning (Zellers *et al.*, 2019), to name a few. VQA systems are tasked with the complex challenge of performing visual analysis, comprehending natural language, and engaging in multi-modal reasoning.

Recent investigations (Goyal *et al.*, 2017; Agrawal *et al.*, 2018; Kafle and Kanan, 2017b; 2017a) have illuminated a noteworthy issue in VQA models—they may exhibit a tendency to rely on superficial linguistic correlations rather than robust multi-modal reasoning. This issue can be attributed to two primary explanations. Firstly, there exists a prevalent linguistic bias, where the question–answer pair tends to have a strong linguistic correlation, while the image–answer relationship often appears superficial. For instance, in the VQAv1 dataset, a significant portion of questions pertaining to the color of a banana yielded the answer "yellow". Secondly, another facet contributing to this challenge is the "visual priming bias". Here, the question content tends to align closely with the presence of objects in the image. As evidenced in the VQAv1 dataset, questions framed as "Did you see…?" resulted in a "yes" response approximately 90% of the time. In both of these interpretations, the model's focus is predominantly on the question itself, often neglecting the crucial visual context provided by the image.

Achieving high accuracy solely based on linguistic cues can be deceptive. When the test scenario diverges from the training data, language priors alone prove

*Corresponding author

inadequate, leading to the limitations of VQA models when applied in real-world scenarios.

One immediate strategy to alleviate language bias is to enrich the training data through additional annotations or data enhancements. Notably, both visual (Das *et al.*, 2017a) and textual (Park *et al.*, 2018) interpretations have been employed to enhance the foundational aspects of visual understanding (Selvaraju *et al.*, 2019; Wu and Mooney, 2019). Furthermore, the generation of counterfactual training samples (Chen *et al.*, 2023; Abbasnejad *et al.*, 2020; Zhu *et al.*, 2020; Gokhale *et al.*, 2020; Liang *et al.*, 2020) has proven to be an effective approach for balancing the training data. It stands out as a superior debiasing method, particularly on VQA-CP (Agrawal *et al.*, 2018). These methodologies demonstrate that debiasing during training can significantly enhance the generalization capabilities of VQA models. However, VQA-CP's primary objective is to assess the model's capacity to disentangle learned visual knowledge from memorized linguistic priors (Agrawal *et al.*, 2018). Therefore, mastering unbiased reasoning amidst biased training remains a substantial challenge in the field of VQA.

Another prevalent approach (Cadene *et al.*, 2019; Clark *et al.*, 2019) is the utilization of distinct problem branches to learn language during training. During the testing phase, priors are mitigated by excluding additional branches. However, linguistic priors encompass both "undesirable" linguistic biases (e.g., linking "orange" to the predominant color "orange") and "beneficial" linguistic contexts (e.g., narrowing the answer space based on the question type, such as "what color"). Simply excluding extra branches does not fully leverage the contextual information. Simultaneously, existing models grapple with single-mode bias in the training data, along with the challenge of bias fusion within invalid regions of learning, impacting the model's overall generalization.

To harness image features more effectively within the model, this article introduces the DCF-VQA model. It leverages the concept of multi-feature enhancement to refine the visual question and answer causal model. It employs features transformed using the discrete cosine transform (DCT) (AlFawwaz *et al.*, 2022), which are mapped into the frequency domain in the causal model, enhancing the representation of image features. This enables a more distinct separation of influences from textual features. The DCT is a widely-used transformation method in image signal processing, recognized for its effectiveness in filtering and noise reduction. DCT transforms features into frequency domain signals through a linear transformation. It maps high-dimensional features into low-dimensional signal space, efficiently compressing redundant information into a few low-frequency coefficients while reducing the impact of image noise by eliminating high-frequency components. In image processing, the discrete cosine transform can enhance certain texture and shape information to a considerable extent, aiding in object correlation and distinction. Additionally, is also explored the discrete sine transform (DST) (Metwaly *et al.*, 2017) as another prevalent transformation method in image processing, which shares similar characteristics with the discrete cosine transform. In this section, the DST discrete sine transform serves as a supplementary method for enhancing image features, providing an alternative reference.

As illustrated in Fig. 1, the basic model can identify the image region pertinent to the question, yet it falls short in fully leveraging the image information, primarily due to semantic bias. This limitation results in a superficial correlation between the image and the answer. For instance, consider the following question–answer pair: "The woman is smiling". In this scenario, the model, constrained by semantic bias, prioritizes the question-answer relationship, leading to an answer that closely mirrors the question itself. On the other hand, the DCF-VQA model, devoid of semantic bias, excels in harnessing information from the interplay between the image and the question. It demonstrates superior focus on the image area relevant to the question, resulting in more accurate answers. For instance, when posed with the same question, "The woman is smiling", the DCF-VQA model correctly identifies the image context and provides the accurate response: "The woman is eating".

This study proposes a DCF-VQA model that utilizes a multi-feature enhancement strategy to improve visual question–answering models, which integrates features extracted through discrete cosine transform into a counterfactual causal framework. The proposed model enriches the subtle connection between image and text queries, and significantly reduces the impact of language bias. S-MRL and UpDn models are used as the base structure for the model. Experimental evaluations performed on the VQAv2 and VQA-CP datasets yielded satisfactory results, demonstrating the effectiveness of the method.

2. Related work

The task of visual question answering necessitates the extraction of features from images and text, followed by the prediction of answers through the comprehension of the semantic information conveyed by these features. However, visual question answering models often struggle with linguistic bias, due to the insufficient representation of multi-modal features. This bias manifests as the model heavily relying on text features while underutilizing visual features, leading to a strong correlation between questions and answers, with only a superficial connection between images and answers. Semantic bias diminishes

Fig. 1. Example of the DCF-VQA model. The proposed model makes full use of the interaction information between the image and the problem, and shows great attention to the regions of the image that are relevant to the problem.

the robustness of visual question answering models and hampers their performance across multiple datasets, significantly impacting their real-world applicability. This section introduces various methods and approaches aimed at mitigating language bias in current visual question answering tasks. These methods can be categorized into three main strategies: enhancing visual information, reducing language priors, and data augmentation.

2.1. Visual information enhancement. HINT offered a generic approach (Selvaraju *et al*., 2019) that revolves around enhancing the model's visual foundation. This optimization involves refining the human attention map and bolstering the alignment between gradient-based network inputs. The objective is to ensure that the model not only learns to perceive but also relies on visual concepts that align with human understanding and relevance for the task. By providing interpretive information that is comprehensible and verifiable by humans, the reliability and interpretability of visual question answering models can be significantly enhanced.

The ReGAT model (Li *et al*., 2019) offers an effective means of improving model accuracy by capturing the relationship between images and questions. It has demonstrated impressive results in verifying both linguistic bias datasets and commonly used datasets. The ESR model (Shrestha *et al*., 2020), founded on a regularization scheme, introduces a straightforward regularization approach that does not necessitate external annotations. It has achieved favorable results on datasets designed to assess language bias. VGQE (Kv and Mittal, 2020) introduces a novel approach involving a problem encoder and an image encoder. During the encoding process, it equally leverages both visual and linguistic modes to furnish ample visual underpinning for problem features. This approach effectively diminishes the model's dependence on language priors through visual feature supplementation.

The progressive model SAR (Si *et al*., 2021), centered around visual entailment selection and ranking, fully exploits the relationships between images, questions, and candidate answers. This enhances the utilization of image information within the model. The Kan model (Zhang *et al*., 2020), based on a knowledge enhancement network, introduces richer visual information. For different types of problems, it adaptively balances the significance of visual information and external knowledge. The adaptive scoring attention module aids the model in automatically selecting suitable information sources based on problem types.

2.2. Weakening language prior. AdvReg introduced a novel regularization scheme that takes the question encoding of the visual question answering model as input Ramakrishnan *et al*. (2018). It employs adversarial training to confront the visual question answering model with another model containing only questions. This process enables the visual question answering model to recognize and rectify language bias within its question encoding.

RUBI (Cadene *et al*., 2019) presents learning strategies that compel the model to provide answers using information from both modalities, rather than solely relying on the correlation between the question and the answer. This is achieved by reducing instances where a correct answer can be given solely based on textual features, even when image features are not utilized. The learn-mixin method (Clark *et al*., 2019) trains both a biased model, which makes predictions based on dataset bias, and a robust model, serving as a reference. This encourages the model to focus on other patterns within the data that are more likely to generalize, rather than relying solely on the correlation between the question and the answer.

RMFE (Gat *et al*., 2020) introduces a new regularization method based on functional entropy. It aims to balance the contribution of each modality to the prediction answer, maximizing the information provided by each modality and promoting more equitable utilization of information from each mode. CF-VQA (Niu *et al*., 2021) incorporates counterfactual causality thinking to extract the direct influence of language on the answer, differentiating it from the impact of multi-modal information on the answer. It retains both the direct and the indirect influence of vision and language, thereby compelling the model to allocate more attention to image information.

2.3. Data enhancement. ActSeek (Teney and Hengel, 2019) introduces the concept of dynamically utilizing external data. It defines a set of specific weights tailored to a given question, enabling the retrieval of specific information for the visual question answering model. This allows the visual question answering system to reason and respond effectively beyond the confines of its training set. The counterfactual sample synthesis training scheme CSS (Chen *et al*., 2023) generates corresponding correct answers as counterfactual training samples by obscuring key objects and keywords within images. This compels the visual question-answering model to focus on these crucial elements.

CL-VQA (Liang *et al*., 2020) introduces a novel self-supervised contrastive learning mechanism. It aims to capture the relationship between original, factual, and counterfactual samples. GradSup (Teney *et al*., 2020) puts forth an auxiliary training objective designed to enhance the generalization capabilities of neural networks. This is achieved by leveraging overlooked supervisory signals within existing datasets. MUTANT (Gokhale *et al*., 2020) represents a novel approach in the realm of visual question answering. It constrains information input by tailoring the training target. This method effectively balances the influence of questions and semantic alterations in images on answer prediction.

The SSL approach (Zhu *et al*., 2020) introduces a self-supervised learning framework that operates independently of external annotations. It helps mitigate data bias by achieving a balanced dataset. ADA-VQA (Guo *et al*., 2021) employs feature space learning to address language bias issues. It designs an adaptive cosine loss to differentiate between the frequency and sparsity of answers based on different question types, thus reducing limitations imposed by language patterns and diminishing language priors. CCB-VQA (Yang *et al*., 2021) bolsters the model's ability to learn contextual priors. This is achieved by establishing content and context branches and incorporating local critical context and global effective context.

Inspired by the idea of counterfactual causation, a multi-feature counterfactual causation structure is proposed in order to solve the problem of language bias in existing VQA models. It aims to better utilize the image information and reduce the influence of linguistic bias.

3. Methods

The proposed model uses a multi-feature enhancement approach to incorporate a set of features transformed by a discrete cosine transform into a counterfactual causality framework. By averaging the computations, it spreads the predictive distribution of incorrect answers across answers to increase the probability of a correct answer, thus improving the performance of the model. The specific implementation process is as follows.

3.1. Causal structure in visual question answering. The causal relationship in visual question answering is shown in Fig. 2, where $v \in V$ represents picture data, $q \in Q$ stands for question language data related to the image, and $a \in A$ represents the answer corresponding to the question image. The influence of V and Q on A can be divided into single-mode influence and multi-mode influence. The direct effects of V on A and Q on A are obtained through $V \rightarrow A$ and $Q \rightarrow A$, respectively. Multi-model influence M after multi-model fusion affects A; M is further divided into M_1 (effective fusion) and M_2 (ineffective fusion). Effective fusion refers to the fusion that matches the object in the image with the text content through attention, while ineffective fusion refers to the fusion that does not match the object and the text exactly. What the visual Q&A task wants is $V, Q \rightarrow M_1 \rightarrow A$, but in the actual experiment, $Q \rightarrow A$ and $V, Q \rightarrow M_2 \rightarrow A$ interfere with the experiment. Therefore, it is necessary to exclude the effect of pure language on the experiment and the effect of mismatching on the experiment.

Firstly, the relationship between question A, answer A and image V is established. If indirect effects are not considered, the relationship between them can be simply

expressed as follows:

$$A_{v,q} = A(V = v, Q = q). \tag{1}$$

In practice, there is feature fusion M, expressed as

$$YA_{q,v,m} = A(V = v, Q = q, M = m), \tag{2}$$

where $m = M_1(V = v, Q = q) \cup M_2(V = v, Q = q)$.

According to the definition of causality, the total effect TE of $V \to v$ and $Q \to q$ can be expressed as

$$\text{TE} = A_{q,v,m} - A_{q^*,v^*,m^*} \tag{3}$$

with

$$m^* = M_1\left(M = m^*, Q = q^*\right) \cup M_2\left(V = v^*, Q = q^*\right).$$

Here v^* and q^* mean that no conditions for v and q are given. The VQA model may have false answers trained between simple questions, thus skipping the multi-modal reasoning process. Therefore, it is hoped that the model will exclude the case where the answer is derived from the question alone. First, the natural direct effect NDE of Q on A is obtained,

$$\text{NDE} = A_{q,v^*,m^*} - A_{q^*,v^*,m^*}. \tag{4}$$

Because the influence of q on the intermediate quantity M is blocked, the direct effect of NDE captures language bias. What the model wants to obtain is the total indirect effect, that is, the effect of M on A, so the total indirect effect is obtained by subtracting the total direct effect from the total effect. The appropriate formula is

$$\text{TIE} = \text{TE} - \text{ND}E = A_{q,v,m} - A_{q,v^*,m^*}. \tag{5}$$

The total indirect effects at this time include the total indirect effects of M_1 and M_2, and M_2 is the non-effective fusion part, which needs to be removed. Again, in a similar way to causation, we do the same operation in frequency space to eliminate linguistic bias. The overall diagram of the model is shown in Fig. 2.

3.2. Model structure and its application. The UpDn model (Anderson *et al.*, 2018) combined with DCF-VQA structure is shown in Fig. 3. By using the word embedding glove method based on matrix decomposition, the problem statement uses global information to capture the relationship and semantic features between words and generate word vectors. Then a long short-term memory network (LSTM) is used to capture the dependencies between words and generate a usable problem feature vector Q. In the image, effective regions are identified to extract regional features. The regional features are taken as the preliminary feature V based on the linear

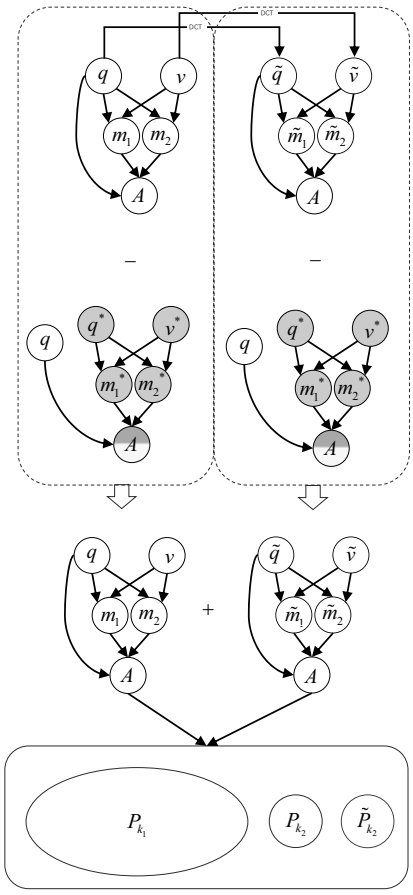

Fig. 2. Multi-feature enhanced structure. The proposed method averages the DCT-transformed features with the original ones to obtain a higher distribution of correct answers.

layer, and the preliminary feature V_1 is obtained by the discrete cosine transform of V. V and V_1 are guided by the attention of the problem characteristics respectively to get the final V and V_1 involved in causality. Image features V and V_1 are taken as the input of the VA neural network model, image features V, V_1 and problem features Q are taken as the input of the VQA neural network model, and problem Q is taken as the input of the QA neural network model, where the QA* model is the one that blocks image V, V_1 and indirect influence to obtain the influence of problem Q under the ideal state. The combination of the VA model, the VQA model and the QA model is a conventional visual question-answering model. Subtracting it from the QA* model, we can obtain an answer distribution free of linguistic bias.

In order to make the model use more features from images, the DCF-VQA model employs the idea of multi-feature enhancement to take the features transformed by the discrete cosine transform into frequency domain as the images in the causal model,

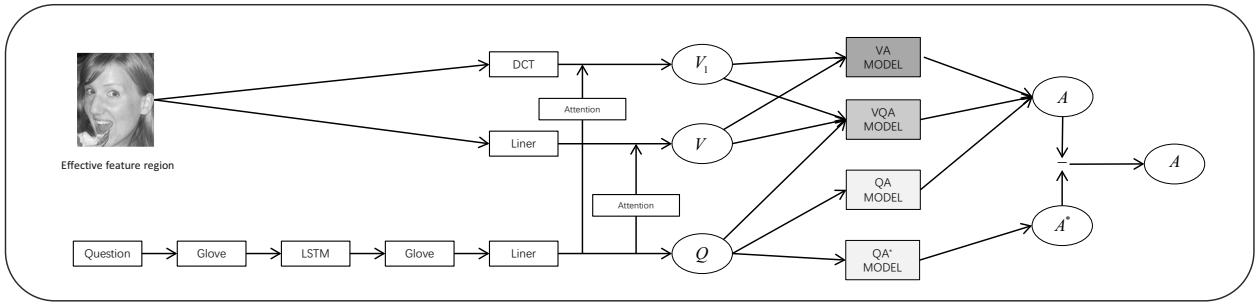

Fig. 3. UpDn model and the multi-feature causal combination process.

enhance the representation of image features, and better strip the influence of text features. The model derivation is completed with the formula below. First, v and q are mapped to the frequency space by the DCT, and the formula is

$$\begin{aligned} \tilde{v} &= \mathrm{DCT}(v), \quad \tilde{q} = \mathrm{DCT}(q), \\ \mathrm{DCT}(v) &= f_n \\ &= \sum_{i=0}^{L-1} v_i \sqrt{\frac{2}{L}} \cos\left(\frac{\pi n}{L}\left(i+\frac{1}{2}\right)\right), \\ & n \in \{0,1,\ldots,L-1\}. \end{aligned} \tag{6}$$

Write P_{m_1} for the probability of each answer predicted by m_1, P_{m_2} for the probability of each answer predicted by m_2, $P_{\tilde{m}_1}$ for the probability of each answer predicted by $\tilde{m}_1$ and $P_{\tilde{m}_2}$ for the probability of each answer predicted by $\tilde{m}_2$. Because the correct fusion mode has uniqueness and consistency, and the wrong fusion mode has diversity, the distance between P_{m_1} and $P_{\tilde{m}_1}$ distributions is short, and the distance between P_{m_2} and $P_{\tilde{m}_2}$ distributions is great, the idea of substitution by approximation can be expressed as $P_{m_1} \approx P_{\tilde{m}_1}$. Do the following to these distributions to form a new distribution P_{new}. The formula is

$$\begin{aligned} P_{\text{new}} &= \frac{P_m + P_{\tilde{m}}}{2} = \frac{P_{m_1} + P_{m_2} + P_{\tilde{m}_1} + P_{\tilde{m}_2}}{2} \\ &\approx \frac{P_{m_1} + P_{m_2} + P_{m_1} + P_{\tilde{m}_2}}{2} \\ &= P_{m_1} + \frac{P_{m_2}}{2} + \frac{P_{\tilde{m}_2}}{2}. \end{aligned} \tag{7}$$

The new answer distribution is

$$P_{m_1} + \frac{P_{m_2}}{2} + \frac{P_{\hat{m}_2}}{2}$$

while to the original answer distribution is $P_{m_1} + P_{m_2}$. The probability of an incorrect prediction of the answer is spread over different answers, which is equivalent to the probability of an incorrect fusion prediction of the answer being reduced, because, in the end, the highest prediction probability of the answer is considered to be the correct answer, so the answer generated by m_1 will be substantially increased. Through repeated cycles, this effect is cumulative and strengthened, eventually getting closer to the point where all predicted answers are produced by m_1 fusion correctly, which greatly weakens the bias effect of incorrect fusion.

3.3. Model implementation details. In order to facilitate understanding, since image features V and V_1 are processed in the same way, this section introduces features V and V_1 collectively as V.

By combining the scores z_q, z_v and z_m of the three neural network models VQ, VA and VQA (the fusion function is h), the comprehensive score of $z_{v,q,m}$ is obtained as shown in (11). Among them, z_q, z_q and z_m are the language, visual and indirect influence branches, respectively:

$$Z_q = F_Q(q), \quad Z_v = F_V(v), \quad Z_m = F_{VQ}(v,q), \tag{8}$$

$$Z_{v,q,m} = h\left(Z_v, Z_q, Z_m\right). \tag{9}$$

Because the inputs to the neural model must be valid inputs, the model will take uniformly distributed learnable parameters as if no blocking treatment for v and q were given. In this case, the expressions for Z_q, Z_v and Z_m are

$$Z_q = \begin{cases} z_q = F_Q(q) & \text{if } Q = q, \\ z_q^* = c & \text{if } Q = \emptyset, \end{cases} \tag{10}$$

$$Z_v = \begin{cases} z_v = F_V(v) & \text{if } V = v, \\ z_v^* = c & \text{if } V = \emptyset, \end{cases} \tag{11}$$

$$Z_m = \begin{cases} z_m = F_{QV}(q,v) & \text{if } Q = q \text{ and } V = v, \\ z_m^* = c & \text{if } Q = \emptyset \text{ or } V = \emptyset. \end{cases} \tag{12}$$

For the fusion function h, there are two nonlinear fusion modes, harmonic (HM) and SUM:

$$\text{(HM)} \quad h\left(Z_v, Z_q, Z_m\right) = \log \frac{\sigma(Z_v)\times\sigma(Z_q)\times\sigma(Z_m)}{1+\sigma(Z_v)\times\sigma(Z_q)\times\sigma(Z_m)}, \tag{13}$$

$$\text{(SUM)} \quad h\left(Z_v, Z_q, Z_m\right) = \log \sigma\left(Z_v + Z_q + Z_m\right). \tag{14}$$

The training strategy is optimized by minimizing the cross-entropy losses of $z_{v,q,m}$, z_v, and z_q,

$$\mathcal{L}_{\text{cls}} = \mathcal{L}_{\text{VQA}}(v,q,a) + \mathcal{L}_{\text{QA}}(q,a) + \mathcal{L}_{\text{VA}}(v,a). \tag{15}$$

$\mathcal{L}_{\text{VQA}}$, $\mathcal{L}_{\text{VA}}$, and $\mathcal{L}_{\text{QA}}$ are obtained through $z_{v,q,m}$, z_v, and z_q. In the triplet (q,v,a), a is the correct answer to the problem image pair (q,v).

Ideally, the clarity of NDE should be similar to the total effect TE, and the KL divergence (Kullback–Leibler) should be used to calculate c, so that the total indirect effect TIE will not be dominated by the total effect TE or the natural direct effect NDE, as shown in

$$\mathcal{L}_{\text{kl}} = \frac{1}{|\mathcal{A}|}\sum_{a\in\mathcal{A}} -p(a \mid q,v,m)\log p\left(a \mid q,v^*,m^*\right), \tag{16}$$

where $p\left(a \mid q,v^*,m^*\right)$ is z_{v,q^*,m^*} obtained by the normalization of Softmax, and $p\left(a \mid q,v,m\right)$ is also obtained by $z_{v,q,m}$ via Softmax. Only c is updated when $\mathcal{L}_{\text{kl}}$ is minimized. The final loss function consists of $\mathcal{L}_{\text{kl}}$ and $\mathcal{L}_{\text{cls}}$, as shown in (17):

$$\mathcal{L} = \sum_{(v,q,a)\in D} \mathcal{L}_{\text{kl}} + \mathcal{L}_{\text{cls}}. \tag{17}$$

4. Experiment and analysis

4.1. Dataset used in the experiment. VQA-CP is a dataset for language bias in visual question-answering tasks (Agrawal *et al*., 2018). It aims to test the robustness of the visual question answering task and promote the development of the visual question answering task to multimodel deep association representation and inference learning. Many studies have found that the existing visual question-answering models, to a large extent, make insufficient use of the image basis when answering questions, and there is only superficial correlation with training data. Most models can achieve high accuracy on the VQAv2 dataset, but poor performance on the VQA-CP. For example, the stack attention based SAN model achieved a total accuracy of 52.41% on the VQAv2 dataset, but only 24.9% on VQA-CP. In the binary index of Yes/No, there is a decrease of about 31%.

The VQA-CP dataset consists of two parts, VQA-CPv1 and VQA-CPv2, each of which includes a training set, a validation set and a test set. Among them,

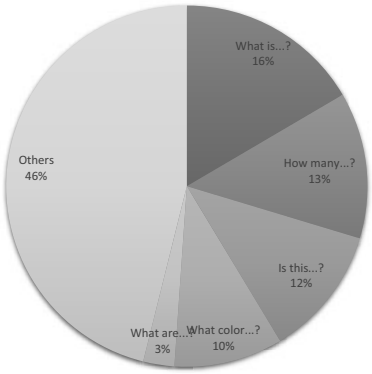

Fig. 4. Composition of the question types of the VQA-CP dataset.

VQA-CPv1 mainly examines the robustness of the model to some changes and disturbances in the scene. The questions and answers in the dataset are the same as those in VQAv2, but the images are randomly changed, such as adjusting brightness, contrast and tone, etc. These transformations may lead to different answers to the same questions on different images. The 4000 samples selected from VQAv1 are made up of 1000 samples from the training set, 1000 samples from the validation set, and 2000 samples from the test machine. These samples require the model to make full use of features from the images and problems, inferring more complex and deep connections between the two modes rather than simply providing the simple relationship in the training data of the VQAv1 dataset. For VQA-CPv2 (which mainly examines the model's ability to generalize to common sense reasoning), the questions and answers in the dataset are the same as those in VQAv2, but the images and questions are modified, requiring the model to be able to reason some additional information in order to give the correct answer. For example, for a problem regarding the color of an object, with the object being blocked, the model needs to deduce the relationship between the occlusion and the blocked object in order to give a correct answer. Then 8000 samples are selected from VQAv2, consisting of 2000 samples from the training set, 2000 samples from the verification set, and 4000 samples from the test set. Similarly to VQA-CPv1, the model is also required to have a stronger reasoning ability and a common sense inference ability.

The optimized model in this section mainly constructs the model based on causal reasoning. Therefore, compared with the conventional data set VQAv2, the VQA-CP data set for language bias is more suitable for examining the effectiveness of the improved causal relationship and can directly reflect the effect of the model. Figure 4 shows the distribution of problems in the VQA-CPv2 test set.

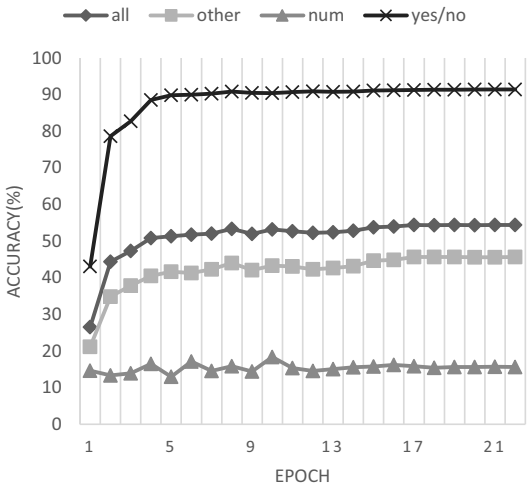

Fig. 5. DCF-VQA (UpDn) model training rounds and accuracy (%).

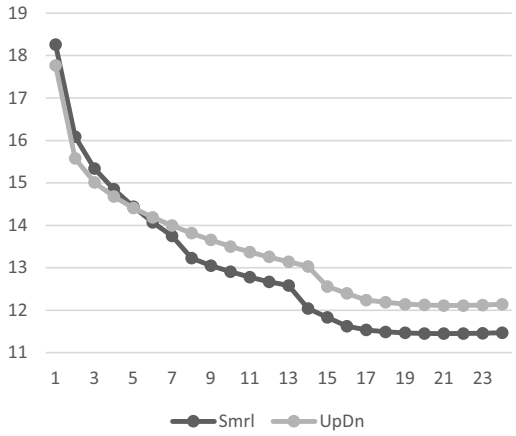

Fig. 6. DCF-VQA model training rounds and the loss.

Table 1. Comparison of the DCF-VQA and UpDn models on the VQA-CP dataset.

Model	All	Yes&No	Num	Other
Updn	37.69	43.17	12.53	41.72
DCF-VQA(HM)	50.08	76.15	**16.77**	45.56
DCF-VQA(SUM)	**54.40**	**91.38**	15.67	**45.64**

Table 2. Comparison of DCF-VQA and S-MRL models on the VQA-CP dataset.

Model	All	Yes&No	Num	Other
S-MRL	38.36	42.85	12.81	43.20
DCF-VQA(HM)	50.39	75.75	19.48	**45.59**
DCF-VQA(SUM)	**55.10**	**88.99**	**27.60**	44.88

This experiment was carried out under the framework of Torch 1.10.1 supported by the NVIDIA RTX A5000 graphics card, the Ubuntu 18.04 operating system and the CUDA 11.4 version. Using the block.bootstrap.pytorch framework, the file loads the options contained in the yaml file, creates the corresponding experiment directory, and begins the training process. The experimental parameters are as follows: BatchSize is set to 256, the problem input dimension is 4800, the image input dimension is 2048, the dropout is set to 0.25, and the learning rate LR is set to 0.0003. Here, the Adam optimizer (Kingma and Ba, 2014) is used to train the DCF-VQA model. We set the fusion mode to Block, the fusion module input dimension to 4800 and 2048, and the output dimension to 2048. For comparison, the train subset of the VQA-CP dataset is used here for training and the val subset for validation.

As shown in Fig. 5, the overall indicators All, Num and Yes/No gradually improve, achieving the best effect in the 22nd round; however, the Num index achieved the best effect in the 10th round, and then the accuracy decreased. In order to obtain the overall optimal effect, the results of the 10th round were taken as an alternative for further analysis. As shown in Fig. 6, it can be seen that the experimental loss decreased rapidly in the first five rounds of training. Then the decreasing amplitude gradually became smaller and stable in the 22nd round. In order to achieve the overall optimal result and avoid the overfitting phenomenon, the experiment was set as 22 epoches and the results of the 22nd round were compared and analyzed.

4.2. Optimal parameters. DCF-VQA was combined with UpDn (Agrawal *et al*., 2018) model and the S-MRL (Yang *et al*., 2016) model to select optimal parameters for the following experiments. This paper analyzes the nonlinear fusion mode and the distribution mode of learnable parameters.

The DCF-VQA method is combined with the UpDn model by using two different nonlinear fusion modes (HM and SUM), and the comparison results are shown in Table 1. The experimental results of fusion using HM and SUM are superior to those of the basic model of UpDn, and the results are better under the condition of the SUM fusion. There is a 4.32% gap in the total index, and a 15.23% gap in the Yes/No index and only the Num index is slightly lower than that of the HM fusion model. The fusion using the SUM method is better than that using the HM method.

As the basis, the S-MRL model is combined with DCF-VQA, and two different nonlinear fusion methods (HM and SUM) are also used for comparison. The comparison results are shown in Table 2. The combined model is superior to the S-MRL model and shows a great improvement. The SUM fusion method has advantages

Table 3. Accuracy of learnable parameters in different distributions (%).

		All	Yes&No	Num	Other
S-MRL		38.36	42.85	12.81	43.20
HM	random	31.27	29.69	**42.87**	28.91
	prior	46.29	61.88	20.03	45.33
	uniform	**50.39**	**75.75**	19.48	**45.59**
SUM	random	27.52	28.00	**37.88**	24.42
	prior	38.06	41.43	14.90	42.64
	uniform	**55.10**	**88.99**	27.60	**44.88**

in the total index, the Yes/No index and the Num index, but has disadvantages in the Other index. By comparing the results of DCF-VQA and the two models, it can be concluded that the SUM fusion method is more suitable for the whole model structure and can produce the best result. Therefore, the SUM fusion method is used to carry out the next experiment.

Since the inputs to the neural model must be valid inputs, the model takes learnable parameters as inputs, blocking V and Q to remove the language bias. In order to find the influence of learning parameters on the model under different distribution conditions, the following experiments were conducted. Based on the S-MRL model and combined with DCF-VQA, the random distribution, prior distribution and uniform distribution methods were used to process the learnable parameters in the model. The results are shown in Table 3. Under the conditions of prior and random distributions, the model results are even worse than for the S-MRL basic model, and only the Num index yields the highest value under the fusion mode of HM and SUM, while the other indexes are lower than the model under the uniform distribution. Under such a distribution, the effective implementation of the natural direct effect (NDE) can be ensured and language bias can be effectively removed. Therefore, the following experiments are conducted with the uniform distribution of learnable parameters.

4.3. Comparison of existing methods. In this section, the proposed DCF-VQA method is compared with the existing ones on the VQAv2 and VQA-CP datasets, and the robustness of the model and the effectiveness of the proposed method are analyzed by comparing the results on the two datasets with more detailed quantitative indicators.

DCF-VQA is combined with the basic models UpDn and S-MRL to provide the required features for DCF-VQA and remove the language bias existing in the original model. The combined model was analyzed and compared with other six models on the common data set VQAv2. The experimental results are shown in Table 4. GVQA (Agrawal *et al.*, 2018) converts images and problems into feature vectors, and uses multi-layer perceptrons to fuse images and problems. Although the multi-layer gating mechanism is used for the multi-modal interaction, it still cannot handle more complex images and problems, so it performs poorly on the VQAv2 datasets.

The SAN (Yang *et al.*, 2016) model adopts a multi-level attention mechanism, which can better represent images and problems, but simple stacked attention cannot make full use of multi-modal information, and there is still the problem of language bias. By fusing the information of images and questions, the UpDn model generates a feature vector containing more contextual information, pays better attention to the key information in the questions, and selects the parts related to the questions from the images for attention, which can better capture the semantic relationship between the answers and questions. Because the model does not balance the relationship between images and texts, the UpDn model can also better capture the semantic relationship between the answers and questions. Thus, there is still the problem of language bias.

The S-MRL model uses a scene graph to learn multiple relations, so as to make better use of the semantic relationship between the images and the problems. However, the above methods all have the problems of insufficient use of image information and linguistic bias, resulting in only superficial correlation between image information and case. When the model predicts the answer, it relies more on the relationship between the question and the answer. CF-VQA using counterfactual causality can reduce the linguistic biases existing in the model and effectively improve the robustness of the model to interference, but it cannot completely eliminate these linguistic biases.

The DCF-VQA model proposed in this section considers the multi-feature method to capture the relationship between images and problems from the perspective of different features, so as to better remove the bias from the text. Compared with the CF-VQA method, DCF-VQA achieved higher accuracy on the VQAv2 data and significantly improved the Num index. When combined with the UpDn model, the results are similar in the total index, with an increase of 0.17% in the binary question Yes/No, but a decrease in the open question Other and the quantity question Num. When combined with the S-MRL model, there is a 0.35% improvement in the total index, a 0.32% and 0.21% improvement in the binary questions Yes/No and open questions Other, and a 0.95% improvement in the Num index.

In order to test the robustness of the model under interference, six models were compared again on the VQA-CP dataset, and the experimental results are also shown in Table 4. The robustness of the model was observed by calculating the accuracy difference between

Table 4. Accuracy (%) comparison of existing methods on the VQAv2 and VQA-CP datasets.

Model	Base	VQAv2				VQA-CP			
		All	Yes/No	Num	Other	All	Yes/No	Num	Other
GVQA	–	48.24	72.03	31.17	34.65	31.3	57.99	13.68	22.14
SAN	–	52.41	70.69	39.28	47.84	24.96	38.35	11.14	21.74
UpDn	–	63.48	81.18	42.14	55.66	39.74	42.27	11.93	46.05
S-MRL	–	63.10	–	–	–	38.46	42.85	12.81	43.20
LXMERT	–	61.16	78.24	44.71	51.89	46.23	42.84	18.91	**55.51**
CF-VQA(SUM)	UpDn	**63.54**	82.51	43.96	**54.30**	53.55	91.15	13.03	44.97
CF-VQA(SUM)	S-MRL	60.94	81.13	43.86	50.11	55.02	90.32	22.37	45.47
DCF-VQA(SUM)	UpDn	63.53	**82.68**	42.42	54.28	54.40	**91.38**	15.67	45.64
DCF-VQA(SUM)	S-MRL	61.29	81.54	**44.77**	50.32	**55.10**	88.99	**27.60**	44.88

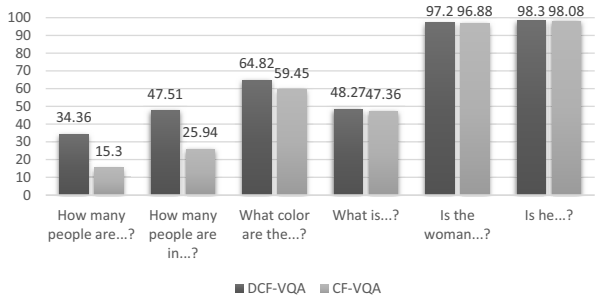

Fig. 7. Comparison of detailed accuracy (%) between DCF-VQA and CF-VQA on the VQA-CP dataset.

the two datasets. GVQA, which is unable to process complex images and problems, also has poor results on the VQA-CP dataset. Compared with the results in the VQAv2 dataset, the total accuracy of the model decreases by 16.94%, the duality index Yes/No decreases by 14.04%, and the quantity index Num decreases by 17.49%. The openness index fell by 12.51 % on the Other part; SAN models using multi-level attention mechanisms experienced a 27.45% decline in the total metrics, 31.71% decline in Yes/No metrics, 28.14% decline in the Num metrics, and 26.1% decline in the Other metrics.

The UpDn model, which integrates the information of images and problems, decreases by 23.74% in the total index, 38.91% in the Yes/No index, 30.21% in the Num index, and 9.16% in the Other index. The S-MRL model, which uses scenario diagrams, has a 24.64% decline in the overall indicator. The LXMERT model (Hashemi *et al.*, 2023) has the best results on Other, but it is pulled apart by DCF-VQA on other metrics. The DCF-VQA model combined with UpDn decreased the total index by 9.13%, increased the Yes/No index by 8.7%, decreased the Num index by 27.75% and decreased the Other index by 8.64%. The DCF-VQA model combined with S-MRL decreased the total index by 6.19%, increased the Yes/No index by 7.54%, decreased the Num index by 17.17% and decreased the Other index by 5.44%.

Table 5. Comparison of the DCF-VQA and S-MRL models on the VQA-CP dataset.

Model	All	Yes&No	Num	Other
UpDn(only q)	18.36	39.45	11.1	9.3
UpDn(only v)	17.56	55.35	0.58	2.42
UpDn(q&v)	48.06	73.6	15.22	43.68
DCF-VQA(q&v)	**54.4**	**91.38**	**15.67**	**45.64**

In the VQA-CP data set, compared with CF-VQA, which eliminates language bias, DCF-VQA combined with the UpDn model improves the total index by 0.85%, the Yes/No index by 0.23%, the Num index by 2.64%, and the Other index by 0.67%. The combined DCF-VQA and S-MRL model increased the total index by 0.08%, decreased the Yes/No index by 1.33%, increased the Num index by 5.23% and decreased the Other index by 0.59%. Compared with CF-VQA, DCF-VQA has a small overall improvement, but it has a large increase in the Num index. Compared with other problems, quantitative problems require more information from images to distinguish the difference between different objects and the relationship between similar objects, which can better test the model's ability to understand and represent images.

Figure 7 lists more detailed indicators in the VQA-CP dataset. On quantitative questions such as: "How many people…?", the model represented by multiple visual features can better distinguish different targets and capture similar targets, and the model accuracy has been improved. However, the overall accuracy is still lower than that of binary questions. For instance, questions like "What color?" and "What type?" are about 50% correct, whereas "yes/no" questions have more than 90% accuracy.

The accuracy rates of a single-mode model and a multi-mode model in the VQA-CP dataset are shown in Table 5. The UpDn model is modified as the basic one, and the accuracy rates of the three models with problem (only Q), picture (only V), problem and picture ($Q\&V$) as input are shown on the VQA-CP dataset. On the VQA-CP dataset, the accuracy of the model using only

single-mode data is poor, and most questions cannot be answered correctly. The model using multi-modal data has a higher accuracy, and the DCF-VQA strategy can significantly upgrade the anti-bias ability of the model, and the accuracy of each indicator has been greatly improved.

4.4. Model example display. This section provides several examples of the DCF-VQA model. As shown in Fig. 8, the model DCF-VQA stripped of semantic bias can use more information from the interaction between the image and the image question, pay better attention to the image area related to the question, and give the correct answer. However, it is difficult for the DCF-VQA model to accurately learn the fine-grained features in an image and efficiently determine the corresponding attention regions when encountering image–text pairs with small required image feature regions. As a result, the model often gives incorrect answers to questions that require attention to detailed aspects of the image, as illustrated by the last two examples in Fig. 9.

Figure 9 shows several examples of the same questions. For questions about what color is the frisbee or what food is in the box, the DCF-VQA model demonstrates its capability to respond accurately without succumbing to linguistic biases, providing correct answers across various images. This performance underscores the model's resilience against linguistic bias and showcases its enhanced proficiency in utilizing visual features for answer prediction.

5. Conclusions

In this study, the utilization of image data in the model was improved based on a multi-feature enhancement approach. Based on this approach, an improved DCF-VQA model was proposed, which achieves a deeper understanding of the relationship between images and text by incorporating DCT-transformed image features. The effect of single-mode bias is mitigated by employing counterfactual causality, which reduces the apparent correlation between the image and the corresponding answer. The experimental results show that feature fusion using multi-feature representations significantly enhances the model's focus on image data, reduces surface connections between images and answers, and improves the accuracy of the model. In the future, we plan to explore the potential of relational networks in recognizing differences and connections between various targets, as well as investigate the association between image data and answers. Integration of image convolutional neural networks to filter out noise and present image data more comprehensively is also being considered.

Acknowledgment

This research is supported by the following projects:

- Special Fund Project for Basic Scientific Research of the Zhongyuan University of Technology (project no. K2021TD05),
- Key Research Projects of Higher Education Institutions in Henan (project no. 23A520022),
- Henan Postgraduate Education Reform and Quality Improvement Project (project no. YJS2022KC19).

References

Abbasnejad, E., Teney, D., Parvaneh, A., Shi, J. and van den Hengel, A. (2020). Counterfactual vision and language learning, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, USA*, pp. 10044–10054.

Agrawal, A., Batra, D., Parikh, D. and Kembhavi, A. (2018). Don't just assume; look and answer: Overcoming priors for visual question answering, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, USA*, pp. 4971–4980.

AlFawwaz, B.M., AL-Shatnawi, A., Al-Saqqar, F. and Nusir, M. (2022). Multi-resolution discrete cosine transform fusion technique face recognition model, *Data* **7**(6): 80.

Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S. and Zhang, L. (2018). Bottom-up and top-down attention for image captioning and visual question answering, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, USA*, pp. 6077–6086.

Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C.L. and Parikh, D. (2015). VQA: Visual question answering, *Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile*, pp. 2425–2433.

Cadene, R., Dancette, C., Ben younes, H., Cord, M., Parikh, D. (2019). Rubi: Reducing unimodal biases for visual question answering, *Advances in Neural Information Processing Systems* **32**: 3197–3208.

Chen, L., Zheng, Y., Niu, Y., Zhang, H. and Xiao, J. (2023). Counterfactual samples synthesizing and training for robust visual question answering, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **45**(11): 13218–13234.

Clark, C., Yatskar, M. and Zettlemoyer, L. (2019). Don't take the easy way out: Ensemble based methods for avoiding known dataset biases, *arXiv:* 1909.03683.

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

Das, A., Kottur, S., Gupta, K., Singh, A., Yadav, D., Moura, J.M., Parikh, D. and Batra, D. (2017b). Visual dialog, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, USA*, pp. 326–335.

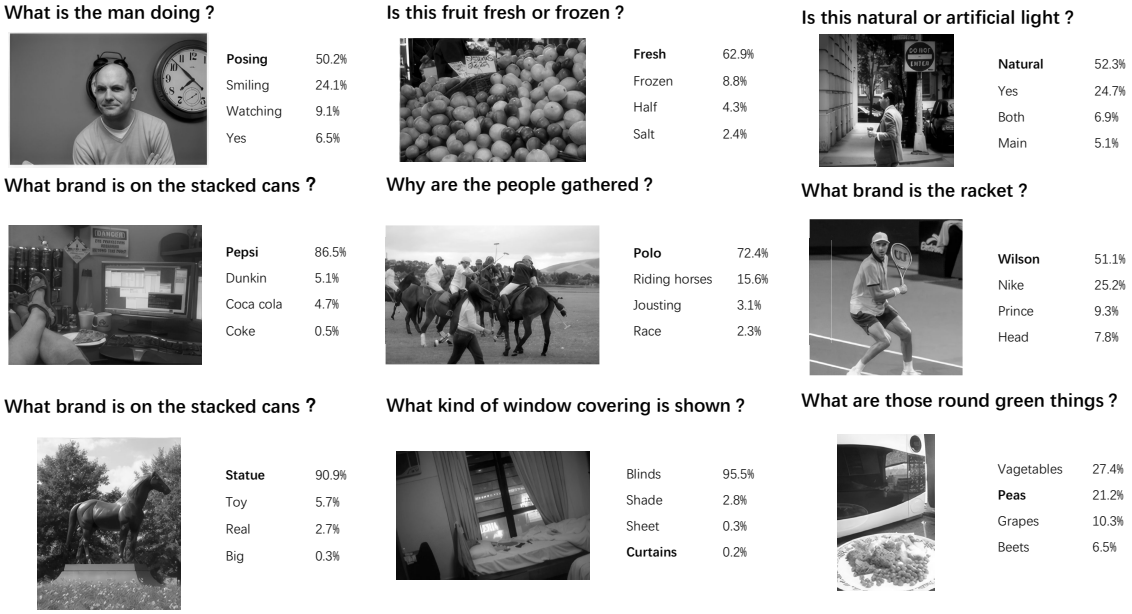

Fig. 8. Examples of DCF-VQA.

Gat, I., Schwartz, I., Schwing, A. and Hazan, T. (2020). Removing bias in multi-modal classifiers: Regularization by maximizing functional entropies, *Advances in Neural Information Processing Systems* **33**: 3197–3208.

Gokhale, T., Banerjee, P., Baral, C. and Yang, Y. (2020). Mutant: A training paradigm for out-of-distribution generalization in visual question answering, *arXiv:* 2009.08566.

Goyal, Y., Khot, T., Summers-Stay, D., Batra, D. and Parikh, D. (2017). Making the V in VQA matter: Elevating the role of image understanding in visual question answering, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, USA*, pp. 6904–6913.

Guo, Y., Nie, L., Cheng, Z., Ji, F., Zhang, J. and Del Bimbo, A. (2021). ADAVQA: Overcoming language priors with adapted margin cosine loss, *arXiv:* 2105.01993.

Hashemi, M., Mahmoudi, G., Kodeiri, S., Sheikhi, H. and Eetemadi, S. (2023). LXMERT model compression for visual question answering, *arXiv:* 2310.15325.

Kafle, K. and Kanan, C. (2017a). An analysis of visual question answering algorithms, *Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy*, pp. 1965–1973.

Kafle, K. and Kanan, C. (2017b). Visual question answering: Datasets, algorithms, and future challenges, *Computer Vision and Image Understanding* **163**: 3–20.

Kingma, D.P. and Ba, J. (2014). Adam: A method for stochastic optimization, *arXiv:* 1412.6980.

Kv, G. and Mittal, A. (2020). Reducing language biases in visual question answering with visually-grounded question encoder, *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK*, pp. 18–34.

Li, L., Gan, Z., Cheng, Y. and Liu, J. (2019). Relation-aware graph attention network for visual question answering, *Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea*, pp. 10313–10322.

Liang, Z., Jiang, W., Hu, H. and Zhu, J. (2020). Learning to contrast the counterfactual samples for robust visual question answering, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 3285–3292, (online).

Metwaly, M.K., Elkalashy, N.I. and Zaky, M.S. (2017). Discrete sine and cosine transforms for signal processing spectral overlap saliencies of induction machine, *IEEE Transactions on Industrial Electronics* **65**(1): 189–199.

Niu, Y., Tang, K., Zhang, H., Lu, Z., Hua, X.-S. and Wen, J.-R. (2021). Counterfactual VQA: A cause–effect look at language bias, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, USA*, pp. 12700–12710.

Park, D.H., Hendricks, L.A., Akata, Z., Rohrbach, A., Schiele, B., Darrell, T. and Rohrbach, M. (2018). Multimodal explanations: Justifying decisions and pointing to the evidence, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, USA*, pp. 8779–8788.

Ramakrishnan, S., Agrawal, A. and Lee, S. (2018). Overcoming language priors in visual question answering with adversarial regularization, *Proceedings of the 32nd International Conference on Neural Information Processing Systems, NIPS'18, Montreal, Canada*, pp. 1548–1558.

Selvaraju, R.R., Lee, S., Shen, Y., Jin, H., Ghosh, S., Heck, L., Batra, D. and Parikh, D. (2019). Taking a hint: Leveraging explanations to make vision and language

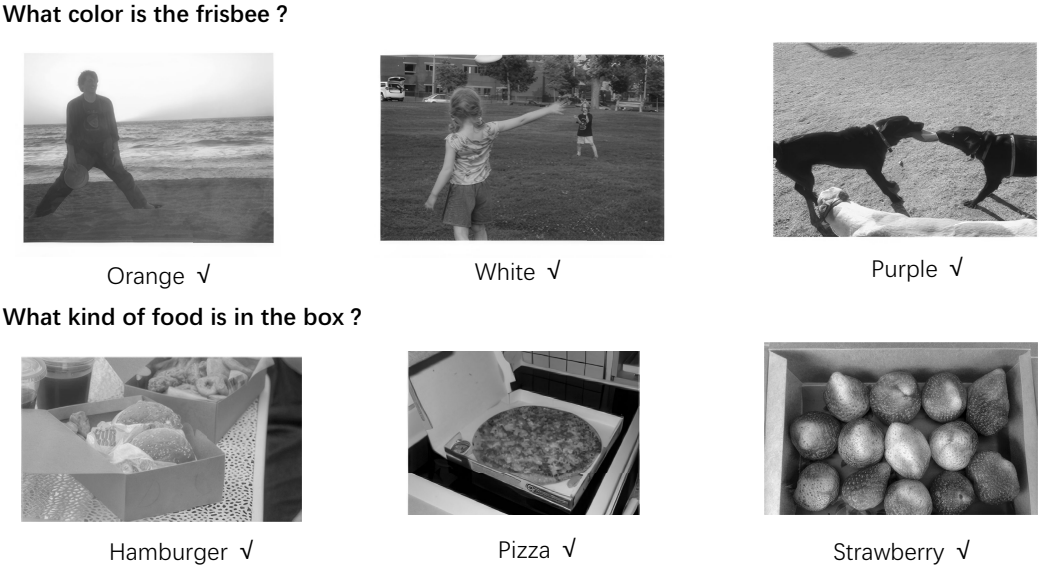

Fig. 9. Samples of DCF-VQA under the same question.

models more grounded, *Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea*, pp. 2591–2600.

Shrestha, R., Kafle, K. and Kanan, C. (2020). A negative case analysis of visual grounding methods for VQA, *arXiv:* 2004.05704.

Si, Q., Lin, Z., Zheng, M., Fu, P. and Wang, W. (2021). Check it again: Progressive visual question answering via visual entailment, *arXiv:* 2106.04605.

Surówka, G. and Ogorzałek, M. (2022). Segmentation of the melanoma lesion and its border, *International Journal of Applied Mathematics and Computer Science* **32**(4): 683–699, DOI: 10.34768/amcs-2022-0047.

Teney, D., Abbasnedjad, E. and van den Hengel, A. (2020). Learning what makes a difference from counterfactual examples and gradient supervision, *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK*, pp. 580–599.

Teney, D. and van den Hengel, A. (2019). Actively seeking and learning from live data, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern, Recognition, Long Beach, USA*.

Wu, J. and Mooney, R. (2019). Self-critical reasoning for robust visual question answering, *33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada*, pp. 8604–8610.

Yang, C., Feng, S., Li, D., Shen, H., Wang, G. and Jiang, B. (2021). Learning content and context with language bias for visual question answering, *2021 IEEE International Conference on Multimedia and Expo (ICME), Shenzhen, China*, pp. 1–6.

Yang, L., Xie, T., Liu, M., Zhang, M., Qi, S. and Yang, J. (2023a). Infrared small-target detection under a complex background based on a local gradient contrast method, *International Journal of Applied Mathematics and Computer Science* **33**(1): 33-43, DOI: 10.34768/amcs-2023-0003.

Yang, P., Wang, Q., Chen, H. and Wu, Z. (2023b). Position-aware spatio-temporal graph convolutional networks for skeleton-based action recognition, *IET Computer Vision* **17**(7): 844–854.

Yang, Z., He, X., Gao, J., Deng, L. and Smola, A. (2016). Stacked attention networks for image question answering, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, USA*, pp. 21–29.

Zellers, R., Bisk, Y., Farhadi, A. and Choi, Y. (2019). From recognition to cognition: Visual commonsense reasoning, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, USA*, pp. 6720–6731.

Zhang, L., Liu, S., Liu, D., Zeng, P., Li, X., Song, J. and Gao, L. (2020). Rich visual knowledge-based augmentation network for visual question answering, *IEEE Transactions on Neural Networks and Learning Systems* **32**(10): 4362–4373.

Zhu, X., Mao, Z., Liu, C., Zhang, P., Wang, B. and Zhang, Y. (2020). Overcoming language priors with self-supervised learning for visual question answering, *arXiv:* 2012.11528.

Guan Yang received a BS degree in probability theory and mathematical statistics from the Northwest University, Xian, China, in 1997. He then received his MS in applied mathematics and his PhD in mathematics from Sun Yat-sen University, Guangzhou, China, in 2005 and 2011, respectively. He is currently an associate professor with the School of Computer Science at the Zhongyuan University of Technology. His research interests include computer vision, machine learning and medical image processing.

Cheng Ji is a post-graduate student at the Zhongyuan University of Technology. He holds a BS degree in software engineering from the Changsha University of Science & Technology. His current research interests include computer vision, image processing, and visual question answering.

Xiaoming Liu holds a PhD degree from the Beijing Institute of Technology (BIT), majoring in computer science, software and theory. Now he is a lecturer and master tutor in the Zhongyuan University of Technology. His main research interests include natural language, Chinese information processing and machine learning.

Ziming Zhang is a post-graduate student at the Zhongyuan University of Technology. He holds a BS degree in software engineering from the Zhengzhou University of Light Industry. His current research interests include computer vision, multimodal representation learning, and visual question answering.

Chen Wang holds a BS degree in computer science and technology from Shangqiu University. He is currently pursuing a Master's degree in electronic information at the School of Computer Science, Zhongyuan University of Technology. His present research focuses on machine learning, computer vision, and image processing.

Received: 10 January 2024
Revised: 15 April 2024
Accepted: 20 May 2024