VQA-BC: ROBUST VISUAL QUESTION ANSWERING VIA BIDIRECTIONAL CHAINING

Mingrui Lao[†] *Yanming Guo*[⋆] *Wei Chen*[†] *Nan Pu*[†] *Michael S. Lew*[†]

[†] LIACS Medialab, Leiden University
[⋆]College of Systems Engineering, National University of Defense Technology

ABSTRACT

Current VQA models are suffering from the problem of over-dependence on language bias, which severely reduces their robustness in real-world scenarios. In this paper, we analyze VQA models from the view of forward/backward chaining in the inference engine, and propose to enhance their robustness via a novel Bidirectional Chaining (VQA-BC) framework. Specifically, we introduce a backward chaining with hard-negative contrastive learning to reason from the consequence (answers) to generate crucial known facts (question-related visual region features). Furthermore, to alleviate the over-confident problem in answer prediction (forward chaining), we present a novel introspective regularization to connect forward and backward chaining with label smoothing. Extensive experiments verify that VQA-BC not only effectively overcomes language bias on out-of-distribution dataset, but also alleviates the over-correct problem caused by ensemble-based method on in-distribution dataset. Compared with competitive debiasing strategies, our method achieves state-of-the-art performance to reduce language bias on VQA-CP v2 dataset.

Index Terms— Visual question answering, language bias, forward/backward chaining, label smoothing

1. INTRODUCTION

Visual Question Answering (VQA) [1] is an 'AI-complete' task that aims to predict correct answers based on given image-question pairs. With the thriving of attention mechanism [2–5] and multi-modal fusion techniques [6–9], VQA models have shown significant performance on in-distribution datasets [1, 10]. However, many researches [10, 11] pointed out that most models encounter the unwanted shortcut 'language bias', where they are prone to over-reliance on the fallacious correlations between question patterns and frequent answers, thereby neglecting the fine-grained analysis of visual information. This undesirable behavior causes VQA models fail to be robust against label distribution shift, and perform worse on the out-of-distribution datasets [11].

Recently, a variety of de-biasing strategies [12–22] are proposed to alleviate this issue. Among these strategies, ensemble-based methods [16, 17] are more efficient alternative, which train a question-only model to capture data bias, and adjust logits through fusing the outputs of question-only and VQA model. However, we declare that there are still some shortcuts in these methods: 1) they adjust weights for VQA samples mainly based on the captured bias from training distribution, and may ignore whether VQA model understands question-related visual information for each sample. Previous works attempt to employ external explanations such as VQA-HAT [23], VQA-X [24] or QA proposal object set as supervision to strength the visual grounding, which is not accessible and flexible. 2) they are prone to over-correct language bias, and excessively focus on less-biased samples.

In this paper, we analyze VQA models from the concept of forward and backward chaining in the inference engine [25]. Current models follow the forward chaining for multi-modal understanding, where the reasoning starts with the known facts (question and image) to deduce new information (correct answer). Due to the imbalanced label distribution, language bias causes VQA model tend to exploit partial known facts (question patterns), but easily obtain less training error. The opposite inference mode is backward chaining, which starts with goals and works backwards from the consequent to the antecedent to validating if any data supports the consequence. In contrast to forward chaining, the concept of backward chaining is rarely explored in current VQA models.

To enhance the robustness and alleviate the language bias of VQA models, we propose a novel Bidirectional Chaining (VQA-BC) approach, whose core ideology is divided into two parts: backward chaining and introspective regularization. Specifically, backward chaining utilizes the goal (answer) as the drive, and further generates the crucial visual facts selected in forward chaining, without using any extra annotations. In addition, we propose a hard-negative contrastive learning to enlarge the discrepancy between reconstructed visual features from correct and related incorrect answers. Introspective regularization is to leverage the validating results in backward chaining, and then alleviate the over-confident problem in forward chaining through label smoothing.

To demonstrate the effectiveness of our method, we combine VQA-BC with the ensemble-based LMH [17] approaches. Considerable ablations studies show that our approach not only reduces language bias with a remarkable gain of 7% on out-of-distribution VQA-CP v2 dataset, but also alleviates the over-correct problem in LMH with an im-

provement of around 6% on in-distribution VQA v2 dataset. Finally, we compare VQA-BC with competitive debiasing strategies, and our method achieves state-of-the-art performance on VQA-CP v2, without using extra annotations.

2. PRELIMINARY

Paradigm of VQA Model: We denote the VQA dataset as $\mathcal{S} = \{I_i, Q_i, A_i\}_{i=1}^{N}$, with N triplets of images $I_i \in \mathcal{I}$, questions $Q_i \in \mathcal{Q}$, and the ground truth answer distribution $A_i \in \mathcal{A}$. VQA model aims to predict the correct answer distribution A_i based on given image I_i and question Q_i. The common function of VQA is formulated as :

$$P_{vqa}(A_i|I_i, Q_i) = C(f_{vqa}(I_i, Q_i)), \tag{1}$$

where $f_{vqa}(\cdot)$ implies the VQA model to integrate visual and textual features, and $C(\cdot)$ denotes the classifier to project the multi-modal fusion feature into the answer dictionary space. Most works train VQA models with binary-cross-entropy loss to optimize their learning parameters:

$$\mathcal{L}_{\text{vqa}} = \sum_{i}^{N} \sum_{j}^{|\mathcal{A}|} a_{ij}^{*} \log p_{ij} - (1 - a_{ij}^{*}) \log(1 - p_{ij}), \tag{2}$$

where a_{ij}^{*} is the ground truth of the j^{th} answer candidates in the i^{th} VQA sample, and p_{ij} is the corresponding prediction.

Ensemble-Based Method: Ensemble-Based method has become a widely-used benchmark to reduce language bias for out-of-distribution datasets. In this paper, we build VQA-BC on the ensemble-based LMH [17] approach, where an additional question-only branch was built to compute a biased prediction $P\left(a_j \mid Q_i; \phi\right)$, and ϕ refers the parameters of the extra model. Then, LMH apply a Learned-Mixin(LM) function to obtain a ensemble-based distribution $P_e(a_j)$ by fusing the prediction of two models together:

$$P_e(A_i) = LM(P\left(A_i \mid I_i, Q_i; \theta\right), P\left(A_i \mid Q_i; \phi\right)). \tag{3}$$

3. PROPOSED METHOD

The overview framework of our Bidirectional Chaining is illustrated in Fig 1, which involves three crucial concepts:

3.1. Backward Chaining

Backward chaining is to endow VQA model with the capacity of reverse thinking, and validate whether the important visual information identified by forward chaining supports the correct answer. As showed in Fig 1(a), unlike the forward VQA function that predict correct answer based on given image-pair, the backward chaining attempt to reason started from the answer (*tennis*) to forward-based crucial visual information (*tennis racket*). Specifically, given the visual attention map $att \in \mathbb{R}^k$ of k image regions from the forward VQA model, we select the region features with maximum attention weight $\bar{v}_i$ as the most significant visual information based on forward chaining. Then, we remove $\bar{v}_i$ from the raw image inputs I_i to obtain the corrupted visual inputs I_i^c. For textual input, we combine the word embeddings of question and correct answer with concatenation to form the answer-driven textual explanation T_i. Then, we feed aforementioned multi-modal inputs into the weight-shared VQA model to acquire the backward fusion features. Finally, we pass through the feature into a generator to produce a reconstructed visual feature $\bar{v}_i^r$:

$$\bar{v}_i^r = G(f_{vqa}(I_i^c, T_i)), \tag{4}$$

where $G(\cdot)$ is the generator composed by a non-linear fully connected layer, and the VQA function $f_{vqa}(\cdot)$ shares parameters with that in the forward chaining. In training phase, we apply the Mean Squared Error (MSE) to measure the average squared difference between $\bar{v}_i^r$ and $\bar{v}_i$, and the loss function in backward chaining L_{bc} is :

$$L_{bc} = \frac{1}{K} \sum_{k=1}^{K} \left\| \bar{v}_i^r - \bar{v}_i \right\|^2, \tag{5}$$

where K denotes the dimensionality of visual region feature.

3.2. Hard-Negative Contrastive Learning

The core motivation of backward chaining is to reason from the answer, whereas one potential problem is that the VQA model may neglect the answer feature in the textual input. Therefore, we propose a hard-negative contrastive learning to leverage the answer information by discriminating the inpainted region features from different answers.

Given the correct answer distribution A_i, we first identify its most hard-negative answer candidate a_i^h via the comparison between the real-time prediction and the ground truth:

$$a_i^h = \underset{a_{ij} \in \mathcal{A}}{\arg\max} \left(P_{vqa}\left(a_{ij} \mid I_i, Q_i\right) - A_{ij} \right), \tag{6}$$

where A_{ij} refers to the value of the j^{th} answer in A_i. Then, we concatenate the raw question with the selected hard-negative answer to form a fake textual explanation T_i^f, and then generate hard-negative reconstructed region feature $\bar{v}_i^h$.

In our hard-negative contrasitive learning, we define the inpainted region feature based on correct answer $\bar{v}_i^r$ as the anchor sample, the removed raw feature $\bar{v}_i$ as the positive sample, and the hard-negative reconstructed feature $\bar{v}_i^h$ as the negative sample. Then, we apply the cosine similarity to measure the distance between different features. The similarity score between the anchor $\bar{v}_i^r$ and positive sample $\bar{v}_i$ is as follows:

$$\cos(\bar{v}_i^r, \bar{v}_i) = \frac{\bar{v}_i^r \cdot \bar{v}_i}{\|\bar{v}_i^r\| \|\bar{v}_i\|}. \tag{7}$$

Likewise, we can compute the similarity between anchor and negative as $cos(\bar{v}_i^r, \bar{v}_i^h)$. Finally, we propose to employ the contrasitve loss [26] to enlarge the discrepancy between features generated from correct and hard-negative answer:

$$L_{cl} = \mathop{\mathbb{E}}_{\bar{v}_i^r, \bar{v}_i, \bar{v}_i^h} \left[-\log \left(\frac{e^{cos(a, \bar{v}_i)}}{e^{cos(\bar{v}_i^r, \bar{v}_i)} + e^{cos(\bar{v}_i^r, \bar{v}_i^h)}} \right) \right]. \tag{8}$$

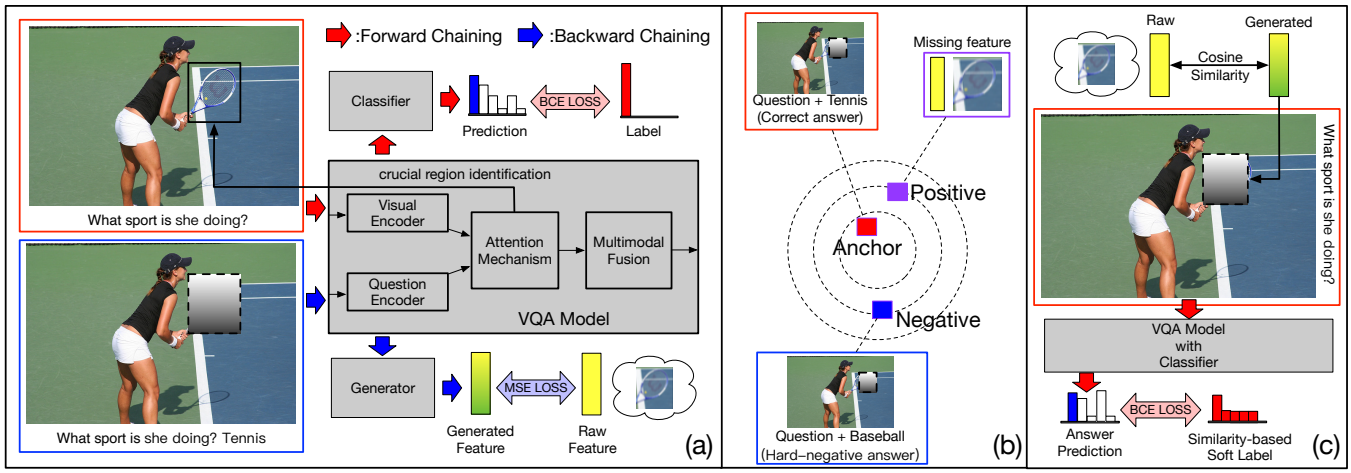

Fig. 1. Overview framework of VQA-BC, which consists of three crucial components: (a) backward chaining with common forward chaining (answer prediction). (b) hard-negative contrastive learning. (c) introspective regularization.

3.3. Introspective Regularization

Based on our proposed backward chaining, one potential challenge is how to improve the consistency of bidirectional chaining, and alleviate the overconfident problem caused by language bias in the forward chaining. In this section, we propose a introspective regularization to tackle this problem.

The illustration of introspective regularization is in Fig 1(c). Specifically, we add the backward inpainted region feature into the corrupted image feature to form a new image features I_i^b. Then, we combine this image features with the raw question as the backward-chaining driven VQA sample, and re-implement a forward chaining to predict answers. Combined with the Learned-Mixin function (Eq. 3), the predicted result of j^{th} answer candidate for the i^{th} sample a_{ij} is:

$$p_{ij}^{ir} = LM(P\left(A_i \mid I_i^b, Q_i; \theta\right), P\left(A_i \mid Q_i; \phi\right)), \quad (9)$$

Based on the binary-cross-entropy (BCE) loss, our introspective regularization exploits a similarity-based label smoothing to train the backward-chaining driven samples, and its loss function $\mathcal{L}_{ir}$ is defined as:

$$\mathcal{L}_{\text{ir}} = \sum_{i}^{N} \sum_{j}^{|\mathcal{A}|} \epsilon_i A_{ij} \log p_{ij}^{ir} - (1 - \epsilon_i A_{ij}) \log(1 - p_{ij}^{ir}), \quad (10)$$
$$\epsilon_i = \cos(\bar{v}_i^r, \bar{v}_i),$$

where ϵ_i implies the label smoothing factor to soften the ground truth answer distribution A_i, which is computed by the cosine similarity between reconstructed region feature generated in backward chaining and its raw feature. Compared with the standard BCE loss, the loss function equips VQA model with the capacity of introspecting its behaviour in forward chaining from backward chaining.

Training and testing strategies: The training can be divided into three stages. In the first stage, we train the model with forward VQA function with LMH until convergence for well-trained attention mechanism (1st-15th epoch). Then, we integrate backward chaining (with contrastive learning) into forward chaining, and the VQA model is finetuned in the followed 5 epochs. In the final stage, we finetune the whole model for the last 5 epochs with the additional introspective regularization. Especially, the total loss function in the final training stage can be defined as:

$$\mathcal{L}_{task} = \mathcal{L}_{vqa} + \lambda_{bc}\mathcal{L}_{bc} + \lambda_{cl}\mathcal{L}_{cl} + \lambda_{ir}\mathcal{L}_{ir}, \quad (11)$$

where λ_{bc}, λ_{cl} and λ_{ir} are the trade-off factors. For inference, we use the final trained VQA model in forward chaining to predict the correct answer, which is the same as baseline.

4. EXPERIMENTS

4.1. Implementation Details

We apply Faster-RCNN [27] to extract object features with maximum 100 proposals. We set the maximum length of 14 words in a question, where each word is encoded into a word vector with Glove [28]. During training, the learning rate remains unchanged at 2e-3 and the batch-size is set to 512. We utilize a grid search on VQA-CP v2 to set the hyper-parameters, resulting in $\lambda_{bc} = 5$, $\lambda_{cl} = 5$, and $\lambda_{ir} = 3$.

4.2. Compared with the State-of-the-art

For a fair comparison, we group debiasing strategies into non-annotation based approach and annotation based approach. Experiments on VQA-CP v2 dataset demonstrate the effectiveness of alleviating language bias for testing methods. Our VQA-BC achieves state-of-the-art performance at 60.81%, with approximately 8% accuracy boost over the benchmark method LMH. With a deep analysis for the question types on VQA-CP v2, our method remarkably enhances performance by 17% on 'Yes/No', and 15% on 'Number'. These results strongly support that VQA-BC effectively overcomes

Table 1. Comparison with the state-of-the-art on out-of-distribution VQA-CP v2 and in-distribution VQA v2 datasets.

Method	Expl.	VQA-CP v2				VQA v2				Comparison	
		All	Yes/No	Number	Other	All	Yes/No	Number	Other	Gap ↓	Mean↑
UpDn [3]	-	39.89	43.01	12.07	45.82	**63.79**	80.94	42.51	**55.78**	23.90	51.84
HINT [12]	HAT	46.73	67.27	10.61	45.88	63.38	81.18	42.99	55.56	16.65	55.10
SCR [13]	VQA-X	49.45	72.36	10.93	48.02	62.2	78.8	41.6	54.5	12.75	55.83
SCR [13]	HAT	49.17	71.55	10.72	47.49	62.2	78.9	41.4	54.3	13.03	55.69
LMH+CSS [14]	QA	58.21	83.65	40.73	48.14	53.15	61.20	37.65	53.36	5.06	55.68
AdvReg [15]	-	41.17	65.49	15.48	35.48	62.75	79.84	42.35	55.16	21.58	51.96
RUBI [16]	-	45.42	63.03	11.91	44.33	58.19	63.04	41.00	54.43	12.77	51.81
LMH [17]	-	52.87	68.26	35.24	49.65	55.99	62.89	39.00	55.25	3.12	54.43
CF-VQA [18]	-	53.69	**91.25**	12.80	45.23	63.65	**82.63**	**44.01**	54.38	9.96	58.67
GGE-DQ [19]	-	57.32	87.04	27.75	49.59	59.11	73.27	39.99	54.39	1.79	58.22
LMH+Ours	-	**60.81**	86.12	**50.20**	**50.48**	61.74	77.74	39.88	55.37	**0.93**	**61.28**

Table 2. Ablation study for different components in our approaches on VQA-CP and VQA v2 datasets.

Model	Component			VQA-CP-v2	VQA-v2
	$\mathcal{L}_{bc}$	$\mathcal{L}_{cl}$	$\mathcal{L}_{ir}$	All	All
LMH				52.87	55.99
A	✓			54.14	57.69
B			✓	53.47	58.24
C	✓	✓		54.22	58.17
D	✓		✓	58.00	61.20
Ours	✓	✓	✓	60.81	61.74

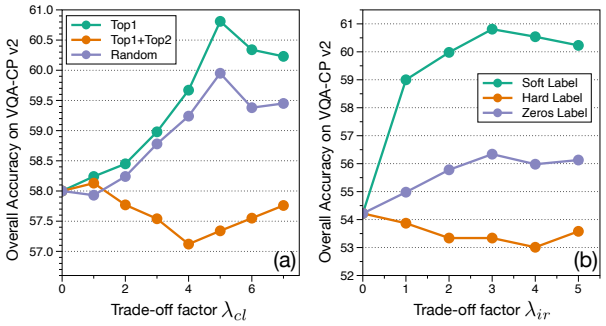

Fig. 2. Different settings for $\mathcal{L}_{cl}$ and $\mathcal{L}_{ir}$ on VQA-CP v2.

language bias on out-of-distribution dataset. Compared with de-biasing strategies focusing on important visual regions (e.g SCR and CSS), our VQA-BC identifies the crucial image region only based on the attention map in its forward chaining, without using any external annotations.

The results in VQA v2 validate the robustness of de-biasing strategies on in-distribution dataset. Although the overall accuracy of VQA-BC is lower than raw UpDn model, it still effectively alleviates the over-correct problem caused by baseline LMH, with a gain of 5.7% on overall accuracy, and 15% on 'Yes/No' question type. From the comprehensive results on both datasets, our method can effectively narrow the gap to 0.93%, and achieve the highest mean score at 61.28% over two datasets among all compared approaches. All these results further demonstrate that, the bidirectional reasoning implemented in our method can not only reduce training bias, but also improve the model robustness.

4.3. Ablation Study

In Tab.2, we conduct extensive ablation studies to analyze the effectiveness of backward chaining ($\mathcal{L}_{bc}$), hard-negative contrastive learning ($\mathcal{L}_{cl}$) and introspective regularization ($\mathcal{L}_{ir}$). Based on LMH, backward chaining (model A) slightly enhances the overall accuracy on both datasets, and plays a fundamental role on enhancing performance driven by contrastive learning (model C) and introspective regularization (model D). In contrast to Model D, adding constrastive learning (Ours) improves performance with an accuracy boost of 2.8% by discriminating inpainted features from different answers, where selecting the Top1 hard-negative answer for contrastive learning obtains best results in Fig. 2(a). Finally, on the basis of Model A and C, introspective regularization significantly boosts accuracies by softening the label in answer prediction. It verifies that the regularization can effectively alleviate the over-confident problem in answer prediction caused by language bias. The results in Fig 2(b) show that the soft label based on backward chaining are remarkably superior to hard and zeros labels.

5. CONCLUSION

In this paper, we proposed to improve the robustness of VQA model via a novel Bidirectional Chaining (VQA-BC). In the framework of VQA-BC, we presented a backward chaining with hard-negative contrastive learning to reason from consequence (answer) to known facts (visual regions). Furthermore, we introduced a novel introspective regularization by label smoothing to overcome the over-confident problem in answer prediction. Extensive experiments showed the effectiveness of our method. In the future, we plan to exploit VQA-BC to enhance model robustness for other multi-modal tasks.

References

[1] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, L. Zitnick, and D. Parikh, "Vqa: Visual question answering," in *ICCV*, 2015.

[2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *NIPS*, 2017.

[3] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," in *CVPR*, 2018.

[4] J. Kim, J. Jun, and B. Zhang, "Bilinear attention networks," in *NIPS*, 2018.

[5] Z. Yu, J. Yu, Y. Cui, D. Tao, and Q. Tian, "Deep modular co-attention networks for visual question answering," in *CVPR*, 2019.

[6] J. Kim, K. Woon On, W. Lim, J. Kim, J. Ha, and B. Zhang, "Hadamard product for low-rank bilinear pooling," *CoRR*, vol. abs/1610.04325, 2016.

[7] H. Ben-younes, R. Cadene, M. Cord, and N. Thome, "Mutan: Multimodal tucker fusion for visual question answering," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017.

[8] L. Shi, S. Geng, S. Kai, C. Hori, S. Liu, P. Gao, and S. Su, "Multi-layer content interaction through quaternion product for visual question answering," in *ICASSP*, 2020.

[9] M. Lao, Y. Guo, N. Pu, W. Chen, Y. Liu, and M. S. Lew, "Multi-stage hybrid embedding fusion network for visual question answering," *Neurocomputing*, 2021.

[10] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh, "Making the v in vqa matter: Elevating the role of image understanding in visual question answering," in *CVPR*, 2017.

[11] A. Agrawal, D.Batra, D. Parikh, and A. Kembhavi, "Don't just assume; look and answer: Overcoming priors for visual question answering," in *CVPR*, 2018.

[12] R. Selvaraju, S. Lee, Y. Shen, H. Jin, S. Ghosh, L. Heck, D. Batra, and D. Parikh, "Taking a hint: Leveraging explanations to make vision and language models more grounded," in *ICCV*, 2019.

[13] J. Wu and R. Mooney, "Self-critical reasoning for robust visual question answering," in *NIPS*, 2019.

[14] L. Chen, X. Yan, J. Xiao, H. Zhang, S. Pu, and Y. Zhuang., "Counterfactual Samples Synthesizing for Robust Visual Question Answering," in *CVPR*, 2020.

[15] S. Ramakrishnan, A.Agrawal, and S.Lee, "Overcoming language priors in visual question answering with adversarial regularization," in *NIPS*, 2018.

[16] R. Cadene, C. Dancette, H. younes, M. Cord, and D. Parikh, "Rubi: Reducing unimodal biases for visual question answering," in *NIPS*, 2019.

[17] C. Clark, M. Yatskar, and L. Zettlemoyer, "Don't Take the Easy Way Out: Ensemble Based Methods for Avoiding Known Dataset Biases," in *EMNLP*, 2019.

[18] Y. Niu, K. Tang, H. Zhang, Z. Lu, X. Hua, and J. Wen, "Counterfactual VQA: A Cause-Effect Look at Language Bias," in *CVPR*, 2021.

[19] X. Han, X. Wang, C. Su, Q. Huang, and Q. Tian, "Greedy Gradient Ensemble for Robust Visual Question Answering," in *ICCV*, 2021.

[20] K. Gouthaman and M. Anurag, "Reducing language biases in visual question answering with visually-grounded question encoder," in *ECCV*, 2020.

[21] X. Zhu, Z. Mao, C. Liu, P. Zhang, and Y. Zhang, "Overcoming language priors with self-supervised learning for visual question answering," in *IJCAI*, 2020.

[22] Z. Cheng F. Ji J. Zhang A. D. Bimbo Y. Guo, L. Nie, "Adavqa: Overcoming language priors with adapted margin cosine loss," in *IJCAI*, 2021.

[23] A. Das, H. Agrawal, L. Zitnick, D. Parikh, and D. Batra., "Human Attention in Visual Question Answering: Do Humans and Deep Networks Look at the Same Regions?," *Computer Vision and Image Understanding*, 2017.

[24] D. H. Park, L. A. Hendricks, Z. Akata, A. Rohrbach, B. Schiele, T. Darrell, and M. Rohrbach., "Multimodal Explanations: Justifying Decisions and Pointing to the Evidence.," in *CVPR*, 2018.

[25] A. Al-Ajlan, "The Comparison between Forward and Backward Chaining,," *International Journal of Machine Learning and Computing*, 2015.

[26] F. Wang and H. Liu, "Understanding the behaviour of contrastive loss," in *CVPR*, 2021.

[27] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *NIPS*, 2015.

[28] J. Pennington, R. Socher, and C.D. Manning, "Glove: Global vectors for word representation," in *EMNLP*, 2014.