Mitigating Language Biases In Visual Question Answering Through The Forgotten Attention Algorithm

Anonymous ACL submission

Abstract

At present, in the field of Visual Question Answering (VQA), a model's ability to comprehend various modalities is crucial for accurate answer reasoning. However, recent studies have uncovered prevailing language biases in VQA, where reasoning frequently relies on incorrect associations between questions and answers, rather than genuine multi-modal knowledge-based reasoning. Thus, it is of great challenge to reveal the accurate relationship between image and question. The key idea of this work is inspired by the process of answering questions of human beings, where people always gradually reduce the focus area in the image with the aid of question information until the final related area is retained. More specifically, we introduce a novel attention algorithm, named the Forgotten Attention Algorithm (FAA), where this algorithm gradually **"forgets"** some visual contents after several rounds. This deliberate forgetting process concentrates the model's "attention" on the image region that is the most relevant to the question. As a result, it can enhance the integration of image content and thus mitigate language biases. We conducted comprehensive experiments on the VQA-CP v2, VQA v2, and VQA-VS datasets to validate the efficiency and robustness of the algorithm.

1 Introduction

In recent years, Visual Question Answering (VQA) has become one of the prominent tasks in the field of deep learning (Hudson and Manning, 2019a), achieving significant accomplishments in various applications, such as intelligent service systems (Luo et al., 2023; Wang et al., 2022). However, recent research has found that many existing VQA methods tend to rely on false associations between questions and answers, without sufficiently extracting accurate visual information from images to answer questions. For example, when answering questions "What color?", some VQA models are inclined to use the most common answers from training data of that type, like "yellow," rather than extracting genuine color information from images. Additionally, some studies (et al., 2021; Liu et al., 2022) have indicated deficiencies in the existing methods' understanding of images, resulting in answers generated by the model relying on image regions with low relevance to the questions. In other words, specific methods often provide correct answers based on incorrect image regions, which does not genuinely reflect the model's performance in the question-answering task. Consequently, the factors affecting the robustness of VQA models can be summarized into two primary aspects: inherent biases in the language distribution of training and testing datasets, and the improper shortcut biases caused by the inadequate utilization of visual information (Liu et al., 2023).

The state-of-the-art and noteworthy methods primarily revolve around data augmentation techniques and attention-based approaches. Data augmentation methods (Chen et al., 2020) aim to enhance a model's understanding of critical features within the data by expanding the dataset with samples, such as counterfactual instances and additional annotations (Liang et al., 2020; Gokhale et al., 2020), which help eliminate biases and enhance robustness (Agarwal, 2020; Wen et al., 2021) by obtaining more critical sample features and supplementary information. However, it is still of great interest and challenge to remove the language biases in VQA model without resorting to data augmentation (Niu et al., 2021). Regarding attention-based methods(et al., 2017), the majority currently integrate these into pre-trained models for efficient feature fusion (Tan and Bansal, 2019; Yu et al., 2019; Lu et al., 2016; Lu et al., 2022; Anderson, 2018), with limited emphasis on fully utilizing visual information.

Therefore, we believe that effectively utilizing

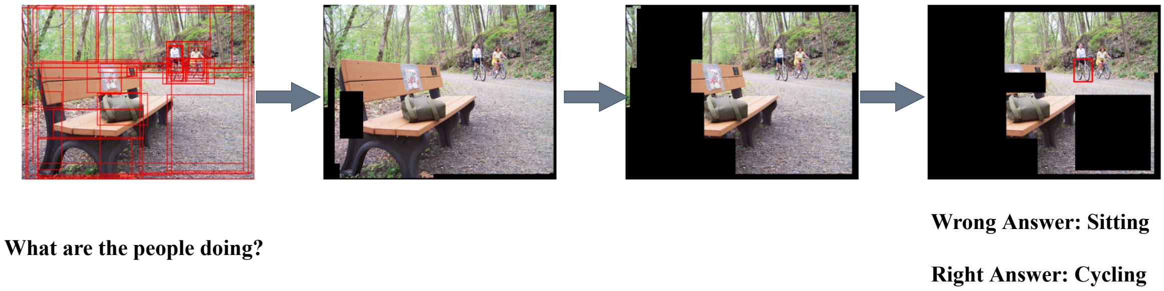

Figure 1: Due to the presence of biases, the influence of the size of prominent objects in the image on model reasoning leads to incorrect answers, while the image regions relevant to the answers often occupy a small portion. FAA achieves this by masking irrelevant regions in the image, allowing the model to focus on image details for inference.

image content without data augmentation is an effective approach to mitigating language biases. In Fig. 1, it is evident that prominent objects (i.e., the bench) often dominate the model's attention, causing it to overlook the finer image area that is relevant to the question (i.e., the people). This observation poses a new challenge: how to focus on the right image area that is the most relevant to the question. To address this problem, we are inspired by the process of answering questions of human beings, where people always gradually reduce the focus area in the image with the aid of question information until the final related area is retained.

In this paper, specifically, we introduce a novel attention algorithm, named the Forgotten Attention Algorithm (FAA), where this algorithm iteratively **"forgets"** some visual contents after each round, that is, disregarding irrelevant image information. Through multiple iterations, the model progressively identifies more relevant regions within the image. As shown in Fig. 1, FAA gradually masks less relevant regions, resulting in effectively harnessing related image information. The retained image is then utilized for the final answer reasoning, thus alleviating the influence of salient objects in the image that are not related to the question.

Overall, this paper's contributions are delineated as follows:

1. We introduce a novel forgetfulness attention algorithm (FAA) aimed at mitigating biases in VQA. The FAA achieves robust VQA by focusing on forgetting unimportant information and reinforcing the role of correct visual content in reasoning.

2. On VQA-CP v2, our enhancements in leveraging visual information led to optimal performance. Notably, without additional annotations, our approach attained a 20.78% improvement compared to the UpDn baseline model. Code is available at:https://github.com/EASONGLLL/FAA-VQA.

2 Related work

2.1 Visual Question Answering

The VQA task demands accurate model responses to image-related questions. Since its inception, this field has seen the emergence of various pertinent datasets and multimodal fusion techniques, such as VQA v2(Antol et al., 2015), GQA(Hudson and Manning, 2019b), CLEVR(Johnson et al., 2016), OK-VQA(Marino et al., 2019), and VideoQA(Tu et al., 2013) rooted in video datasets. Presently, methods based on single-stream and dual-stream architectures(Yang et al., 2019; Wang et al., 2019; Izacard and Grave, 2021; Rajpurkar et al., 2018; Chen et al., 2020) achieve high accuracy by extensively pretraining on abundant samples.

2.2 Language Bias

In recent research, researchers have proposed a range of debiasing methods to address language bias concerning existing defined bias issues. These methods include adversarial-based techniques (Ramakrishnan et al., 2018), regularization approaches (Niu et al., 2021; Han et al., 2021; Abbasnejad et al., 2020; Cho et al., 2023; Basu et al., 2023), and data augmentation strategies (Chen et al., 2020; Wen et al., 2021). Our approach focuses on addressing bias issues from the perspective of the visual modality.

2.3 Attention Mechanism

In the context of Visual Question Answering (VQA), attention mechanisms are employed to integrate information from different modalities (et al., 2017), allowing models to focus on the most relevant regions between images and texts. Presently, attention-based methodologies include linear attention (et al., 2016), co-attention (Lu et al., 2016), detection attention (et al., 2017), and relational attention (Wu et al., 2018). Consequently, in our approach, we explore the integration of attention mechanisms into debiasing methods in VQA, strengthening the model's retrieval capabilities between images and questions. Leveraging attention mechanisms enhances the role of visual information, ultimately aiding in debiasing strategies.

3 Method

We now describe the architecture and algorithmic flow of FAA. As shown in Fig. 2, the left side illustrates the primary structure of the UpDn baseline model (Anderson, 2018), responsible for extracting visual-language features. On the right side, there are stacked $Attention\_Layers$ that iteratively mask irrelevant features and make answer predictions.

3.1 Visual Information Combination

On the left side of Fig. 2, we utilize the UpDn encoding layer to extract features. For a given text, the UpDn leverages a standard GRU to encode each question, generating a question vector. Regarding the provided image, UpDn uses the detected visual features as input. The visual feature set is represented as $F = \{f_1, ..f_i.., f_n\}$, where f_i denotes the feature of the i-th object in the image. In our method, we also incorporate factors such as spatial position. We re-encode all the outputs from Faster-RCNN (Ren et al., 2017) into new visual features. The visual input V is represented as Eq. (1),

$$V = Visual\_Encoder(F, S, Cls, Ari), \quad (1)$$

where $Visual\_Encoder$ represents the visual encoder responsible for re-encoding the four types of features into visual input. These four types of features are represented as visual feature vectors F, spatial features S, classification scores Cls, and attribute information Ari. During the initialization phase, this re-encoded visual data V is introduced as the visual input for the VQA process.

Algorithm 1: Forgetting Attention Algorithm

Input : Representation of Object Detection Outputs:$\mathcal{F}, \mathcal{S}, Cls, Ari$; Text coded representation:Q; Number of layers of attention stack:N; Attention threshold:α.
Output : Predicted answer probability:$\mathcal{A}$.

Initialize:$V \leftarrow [\mathcal{F}, \mathcal{S}, Cls, Ari]$, $k \leftarrow 3$.
Function *FAA(V, Q)*:
  while $n \leq N$ **do**
    $att_v, att_q \leftarrow SelfAttention(V, Q)$
    $V^1, Q^1 \leftarrow att_v \odot V, att_q \odot Q$
    $V^2, Q^2 \leftarrow CrossAttention(V^1, Q^1)$
    $Att \leftarrow V^2 \odot Q^2$
    if $Att \leq \alpha$ **then**
      $V_{mask} \leftarrow 1$;
    else
      $V_{mask} \leftarrow 0$;
    $V^3 \leftarrow V_{mask} \oplus V^2$
    $V, Q \leftarrow V^3, Q^2$
  $\mathcal{A} \leftarrow V^3, Q^2$
return $\mathcal{A}$

3.2 Attention Layers

In the right side of Fig. 2, we have stacked N layers of $Attention\_Layer$ to achieve visual information masking and retrieval. Specifically, the $Attention\_Layer$ module consists of three main components:

1. Initial Impression. After obtaining visual and text features, the next step in our process is to employ the $Self\_Attention$ mechanism. This mechanism helps identify the most critical components within each modality, similar to how humans instinctively react when first encountering an image or text. We establish the model's initial assessment of the pivotal image regions and word vectors within the provided features. As shown in Algorithm 1, it is defined as follows,

$$\begin{aligned} att_v &= Self(V), \\ att_q &= Self(Q), \\ V^1 &= att_v * V, \\ Q^1 &= att_q * Q, \end{aligned} \quad (2)$$

where att_v and att_q represent the initial attention. V^1 and Q^1 represent the features ob-

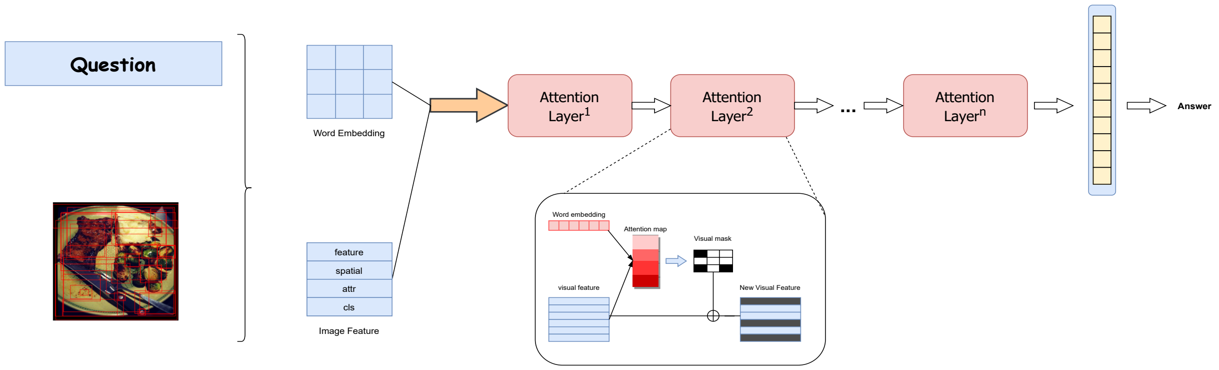

Figure 2: Our proposed FAA follows the architecture of the UpDn baseline model, comprising the feature extraction stage of the UpDn model and the attention layers. The attention layers aim to retrieve information from the encoded question and image features, facilitating a multi-round retrieval process. In each round of retrieval, an image mask matrix is constructed to mask out the information deemed irrelevant by the model during this round, retaining crucial information for subsequent reasoning.

tained after fusing the initial attention with the original data.

2. Cross-Modal Retrieval. With the obtained features V^1 and Q^1, we consider using the $Cross\_Attention$ mechanism (Tan and Bansal, 2019) to explore information across modalities. This step is analogous to how humans associate objects with words. We perform cross-modal information retrieval separately in the image and text domains. This is defined as Eq. (3),

$$\begin{aligned} V^2 &= CrossAtt_{v \to q}(Q^1, V^1), \\ Q^2 &= CrossAtt_{q \to v}(V^1, Q^1), \end{aligned} \tag{3}$$

where V^2 and Q^2 represent the feature outputs after conducting cross-modal retrieval for the image and text, respectively. $Cross\_Att$ respectively represents the cross-modal information retrieval layer, with 'image' and 'question' as the primary modalities.

3. Masking Matrix. After cross-modal retrieval, we calculate the masking matrix for V^2 and Q^2. Initially, we employ the Top-Down attention mechanism (Anderson, 2018) to obtain an attention weight matrix Att, which is then compared to a predefined threshold α to determine the masking matrix. As depicted in Algorithm 1, this is defined as Eq. (4),

$$\begin{aligned} Att &= V^2 * Q^2, \\ V_{mask} &= Mask(Att \leq \alpha), \\ V^3 &= V_{mask} \oplus V^2, \end{aligned} \tag{4}$$

where V_{mask} represents the masking matrix, and $Mask()$ denotes the process in which Att is compared to α in Algorithm 1. The value of α is determined by the mean of attention. V^3 represents the features obtained by merging the masking matrix with visual features. $\oplus$ denotes the linear fusion of two types of features.

Specifically, in each $Attention\_Layer$, we establish a masking matrix based on the magnitude of attention weights, which masks regions in the image that contribute less to the answer. Through N such $Attention\_Layers$, we allow the model to progressively identify precise regions with high relevance to the given question.

4 Experiments

4.1 Comparisons with State-of-the-Arts

The experimental results on the VQA-CP v2, VQA v2 and VQA-VS(Si et al., 2022) dataset are displayed in Table 1 and Table 2. Within the table, we list some excellent debiasing endeavors for comparison.

1. We evaluate our approach on three baseline models (UpDn and RUBi), achieving enhancements of approximately 19% and 13% compared to these models.

2. When compared to other attention-based (SCR, AttAlign, HINT) debiasing methods using the same baseline model, our approach delivers performance enhancements in question types requiring more extensive visual in-

Table 1: The results of VQA-CP v2 test set and VQA v2 val set are presented in the following table. Each column illustrates the **Best** performances of each method, excluding data augmentation techniques.

Data set		VQA-CP v2 test				VQA v2 val			
Method	Base	All	Y/N	Num.	Other	All	Y/N	Num.	Other
GVQA	-	31.30	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
UpDn	-	39.96	43.01	12.07	45.82	**63.48**	81.18	42.14	**55.66**
HINT	UpDn	46.73	67.27	10.61	45.88	63.38	**81.18**	42.99	55.56
SCR	UpDn	49.45	72.36	10.93	48.02	62.30	78.80	41.60	54.50
RUBi	UpDn	44.23	67.05	17.48	39.61	-	-	-	-
LMH	UpDn	52.01	72.58	31.12	46.97	56.35	65.06	37.63	54.69
AttAlign	UpDn	39.37	43.02	11.89	45.00	63.24	80.99	42.55	55.22
GGE-DQ-tog	UpDn	57.32	**87.04**	27.75	49.59	59.11	73.27	39.99	54.39
GenB	UpDn	59.15	**88.03**	40.05	49.25	62.74	86.18	43.859	47.03
RMLVQA	UpDn	60.41	**89.98**	45.96	48.74	59.99	76.68	37.54	53.26
FAA(Ours)	UpDn	**60.74**	83.99	**41.45**	**53.85**	62.86	78.65	51.73	54.13
Methods of data augmentation and additional annotation:									
CVL	UpDn	42.12	45.72	12.45	48.34	-	-	-	-
RandImg	UpDn	55.37	83.39	41.60	44.20	-	-	-	-
CSS	UpDn	58.95	84.37	49.42	48.24	59.91	77.25	39.77	55.11
Mutant	UpDn	61.72	88.90	49.68	50.58	62.56	82.07	42.52	53.28
D-VQA	UpDn	61.91	88.93	52.32	50.39	64.96	82.18	44.05	57.54
KDDAug	UpDn	60.24	86.13	55.08	48.08	-	-	-	-
FAA(Ours)	CSS	61.10	83.27	37.82	54.21	-	-	-	-

formation, particularly in "Num." and "Other" question types.

3. We extend the application of FAA to data augmentation methods like CSS, resulting in performance enhancement when combined with CSS.

4. FAA consistently maintains stability and exhibits a certain level of precision and generalization on the VQA v2 dataset.

5. Within the VQA-VS dataset, FAA demonstrates distinct advantages over models employing the same baseline. Additionally, FAA exhibits considerable performance when handling a broader spectrum of bias types.

4.2 Qualitative results

As depicted in Figure 3, the original image, after two rounds of attentional operations, masks out irrelevant areas based on attentional weights in the (1), ultimately identifying the target region relevant to the answer.

In Figure 3, more examples are given to analyze the effect of forgotten attention on changes in image areas. For example, in the example of the (2), the image of the animal is the area where the zebra is located, and there is overlap between some areas that are unrelated to the problem and the zebra, which is covered by the FAA to some extent, but most of the zebra area is still captured by the model. Similarly, in the (3) and (4), the areas of the sign is somewhat obscured, but the model still understands the semantics of the remaining areas of the image and gives the correct answer. In the (5), the final answer area is well preserved due to the size of the relevant image area. In the (6), we give an error example. Although the model correctly answers the relevant questions, the model still locates the wrong image region due to similar semantic information in the image.

4.3 Abalation Experiments

Forgotten Sequence The concept of forgetting attention in this paper is based on the process of human answering relevant questions. The ablation

Table 2: Regarding the experimental outcomes of FAA on the VQA-VS dataset, we have presented the relevant experimental performance reports associated with this dataset. Each column displays the performance results of the corresponding best and second-performing models.

Model	Base	VQA-VS OOD Test Sets									
		Language-based				Visual-based		multi-modality			mean
		QT	KW	KWP	QT+KW	KO	KOP	QT+KO	KW+KO	QT+KW+KO	
UpDn		32.43	45.10	56.06	55.29	33.39	41.31	46.45	54.29	56.92	46.80
+LMH	UpDn	33.36	43.97	54.76	53.23	33.72	41.39	46.15	51.14	54.97	45.85
LXMERT	-	**36.46**	**51.95**	**64.17**	**64.22**	**37.69**	**46.40**	**53.54**	**62.46**	**67.44**	**53.70**
FAA(Ours)	UpDn	32.45	44.6	56.27	54.96	34.75	43.98	44.47	55.69	55.6	46.97

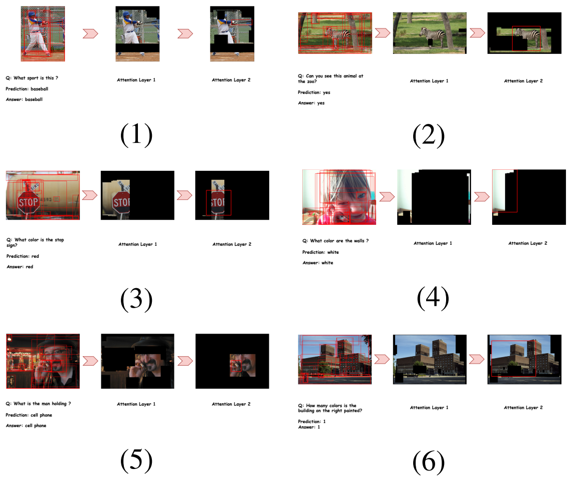

Figure 3: The results of qualitative analysis show the flow of our model when making predictions by masking different image regions so that the model focuses on the effective ones

Table 3: Impact of forgetten order on performance.

Order	Base	VQA-CP v2 test			
		All	Y/N	Num.	Other
FAA	UpDn	60.74	83.99	41.45	53.85
Reverse	UpDn	37.86	78.00	15.34	23.00
Linear	UpDn	27.13	71.99	6.18	9.37

experiment considers the order of forgetting in the algorithm to verify the validity of the concept. The table shows the effect of three different sequences of attention on the performance of the model:

1. In the method of this article, we follow the normal attention process, pay attention to the image areas that are most relevant to the problem, and forget the results obtained from the irrelevant areas.

2. In contrast to the normal attention flow, the model first notices the areas of the image that are less relevant to the problem, assigns higher weights to them, and finally forgets the areas of the image with lower weights.

3. The feature areas in the image are arranged in the original linear order, and the model forgets the relevant features in the same order.

Table 4: Performance corresponding to different attention thresholds

Thresholds	Base	VQA-CP v2 test			
		All	Y/N	Num.	Other
0.5	UpDn	56.84	83.29	43.87	46.54
0.6	UpDn	60.74	83.99	41.45	53.85
0.7	UpDn	59.75	82.85	35.21	54.38

By comparing the model performance of different forgetting sequences, we were able to observe that the FAA achieved excellent performance by forgetting irrelevant regions, while the other forgetting sequences resulted in decreased performance. This suggests that the human attentional forgetting mechanism on which the FAA is based works.

Attention Threshold Selection In this method, we set thresholds to allow the model to filter image regions to determine which are forgotten and which are retained. Different threshold sizes make the model achieve different performance when forgetting the image region. If the threshold is too high, the model will forget most of the image region, resulting in the model being unable to obtain useful information from the image, while if the threshold is too low, the model will retain useless information, thus degrading the algorithm to a common attention algorithm. Therefore, this section sets up a comparative analysis of different threshold sizes to determine the appropriate parameter as

Table 5: The impact of different feature combinations on VQA modeling.

	Visual	Spatial	CLS	Attribute	All
1	✓				56.34
2	✓	✓			55.30
3	✓	✓	✓		54.53
4	✓	✓	✓	✓	59.02

Table 6: The performance under different number of attention layers.

Layers	All	Yes/No	Num.	Other	Time/Epoch
1	56.95	83.77	42.99	46.72	1048s
2	60.74	83.99	41.45	53.85	1303s
3	57.67	82.66	46.44	47.66	1380s
4	58.56	80.81	50.68	47.26	1532s
5	56.21	84.48	35.66	47.03	1540s

Table 7: Ablation experiments involving FAA and the CSS method

Method	All	Yes/No	Num.	Other
Q-CSS	56.19	80.83	40.33	47.63
CSS	58.17	84.57	46.99	47.40
FAA+Q-CSS	58.31	80.83	48.90	49.10
FAA+CSS	60.09	88.55	53.16	47.09

Table 8: Experiment on the evaluation metric CGD using the FAA method on the VQA-CP v2 dataset. **Best** results are displayed in each column.

Method	CGR	CGW	CGD
UpDn	44.27	40.63	3.91
RUBi	39.60	33.33	6.27
CSS	**46.70**	37.89	8.87
GGE-DQ-iter	44.35	27.91	16.44
GGE-DQ-tog	42.74	**27.47**	15.27
FAA(Ours)	45.09	27.54	**17.56**

the forgetting threshold. As shown in the table 4, the corresponding experimental results of the three threshold sizes selected by us are reported.

Considering the three different choices of attention threshold, in the original attention scheme of UpDn, the contribution degree of different image regions to the answer is realized by the assigned weight, whose value is between 0 and 1. Therefore, we choose the sizes of 0.5, 0.6 and 0.7 for experimental comparison. The final experimental results show that when the threshold size is 0.6, the experimental effect reaches the optimal performance, and the forgetting ability of the model reaches the equilibium.

Visual Information. As shown in Table 5, In our method, the output after Faster RCNN detection is combined as the visual input of the model in this paper. Specifically, we fuse each of the four combinations as new visual feature inputs, confirming the benefit of diverse visual information in model comprehension.

Layers Of Attention. As shown in Table 6, We set up different levels of attention in the method to perform validation experiments. Specifically, when humans answer questions by focusing on different areas in the image, they may go through multiple target shifts to determine the final area, while in the simulation of a computer, this operation can be achieved by setting the number of layers of attention. In this experiment, we set up a total of five different layers, as you can see the model performs the best with three attention layers. This configuration significantly improves performance during inference compared to fewer layers. However, increasing the number of layers yields a slightly worse overall accuracy as well as increasing inference time by nearly 200 seconds. Thus, we settle on three attention layers. Regarding accuracy degradation, we believe that this phenomenon arises due to the fact that the model recognizes incorrect visual information and masks relevant regions, thus hindering accurate answer retrieval.

Q-CSS. In our approach, we opted for a single-word replacement strategy, combined with FAA. The experimental results in Table 7 encompass a partial replication of the Q-CSS strategy from the CSS method and the QV-CSS strategy, incorporating FAA into both Q-CSS and CSS. Notably, our approach exhibits approximately 2% improvement in accuracy over Q-CSS and CSS.

4.4 Analysis of Other Metrics

In our approach, we aim to increase the role of visual content in reasoning. To assess its effectiveness, we use additional metrics. In Table 8, we compare our results with other methods using the CGD metric. For a more detailed understanding of CGD, please refer to (Han et al., 2021; Shrestha et al., 2020). Compared to GGE(Han et al., 2021), our approach performs better in terms of CGD, indicating improved utilization of visual information for answer prediction.

5 Conclusion

In this paper, we introduce a novel attention mechanism, the Forget Attention Algorithm (FAA), aimed at mitigating language bias. We regard language bias as a lack of model comprehension of visual content. We artificially mask the image content in our method using a "forgetting" strategy, enabling the model to mimic human attention flow in each iteration for multi-step reasoning. We experiment with our method on datasets VQA v2, VQA-CP v2, and VQA-VS to validate its effectiveness.

6 Limitations

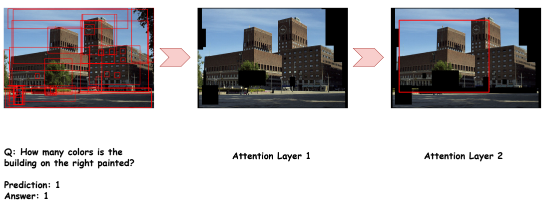

Figure 4: The answer is correct but relies on incorrect visual content.

Firstly, as shown in the Fig. 4, the limitations of the forgetting attention are reflected in the reliance on model knowledge during prediction. When the contents of the images are similar and the model fails to notice detailed information, it focuses on incorrect areas. Secondly, if the model focuses on the wrong areas in the initial rounds, subsequent corrections cannot be effectively made. The model will continue to search for answers within these incorrect regions.

References

Ehsan Abbasnejad, Damien Teney, Amin Parvaneh, Javen Shi, and Antonvanden Hengel. 2020. Counterfactual vision and language learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Vedika et al. Agarwal. 2020. Towards causal vqa: Revealing and reducing spurious correlations by invariant and covariant semantic editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9690–9698.

Peter et al. Anderson. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. *international conference on computer vision*.

Abhipsa Basu, Sravanti Addepalli, and R.Venkatesh Babu. 2023. Rmlvqa: A margin loss approach for visual question answering with language biases. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Universal image-text representation learning. *european conference on computer vision*.

Long Chen et al. 2020. Counterfactual samples synthesizing for robust visual question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10800–10809.

JaeWon Cho, Dong-Jin Kim, Hyeonggon Ryu, and InSo Kweon. 2023. Generative bias for robust visual question answering. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Peter Anderson et al. 2017. Bottom-up and top-down attention for image captioning and visual question answering. *computer vision and pattern recognition*.

Xinzhe Han et al. 2021. Greedy gradient ensemble for robust visual question answering. *arXiv: Computer Vision and Pattern Recognition*.

Zichao Yang et al. 2016. Stacked attention networks for image question answering. *computer vision and pattern recognition*.

Tejas Gokhale et al. 2020. Mutant: A training paradigm for out-of-distribution generalization in visual question answering. *arXiv preprint arXiv:2009.08566*.

Xinzhe Han et al. 2021. Greedy gradient ensemble for robust visual question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1584–1593.

Drew Hudson and Christopher D Manning. 2019a. Learning by abstraction: The neural state machine. *Advances in Neural Information Processing Systems*, 32.

Drew A. Hudson and Christopher D. Manning. 2019b. Gqa: A new dataset for real-world visual reasoning and compositional question answering. *computer vision and pattern recognition*.

Gautier Izacard and Edouard Grave. 2021. Leveraging passage retrieval with generative models for open domain question answering. *conference of the european chapter of the association for computational linguistics*.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. 2016. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. *computer vision and pattern recognition*.

Zujie Liang et al. 2020. Learning to contrast the counterfactual samples for robust visual question answering. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 3285–3292.

Jin Liu et al. 2023. Be flexible! learn to debias by sampling and prompting for robust visual question answering. *Information Processing & Management*, page 103296.

Yibing Liu et al. 2022. Answer questions with right image regions: A visual attention regularization approach. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 18(4):1–18.

Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2016. Hierarchical question-image co-attention for visual question answering. *Neural Information Processing Systems,Neural Information Processing Systems*.

Pan Lu et al. 2022. Co-attending free-form regions and detections with multi-modal multiplicative feature embedding for visual question answering. *Proceedings of the AAAI Conference on Artificial Intelligence*.

Haonan Luo et al. 2023. Depth and video segmentation based visual attention for embodied question answering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6):6807–6819.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. *computer vision and pattern recognition*.

Yulei Niu et al. 2021. Counterfactual vqa: A cause-effect look at language bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12700–12710.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. *meeting of the association for computational linguistics*.

Sainandan Ramakrishnan, Aishwarya Agrawal, and Stefan Lee. 2018. Overcoming language priors in visual question answering with adversarial regularization. *Advances in Neural Information Processing Systems*, 31.

Shaoqing Ren et al. 2017. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, page 1137–1149.

Robik Shrestha et al. 2020. A negative case analysis of visual grounding methods for vqa. *arXiv preprint arXiv:2004.05704*.

Qingyi Si et al. 2022. Language prior is not the only shortcut: A benchmark for shortcut learning in vqa. *arXiv preprint arXiv:2210.04692*.

Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*.

Kewei Tu, Meng Meng, Mun Wai Lee, Tae Eun Choe, and Song-Chun Zhu. 2013. Joint video and text parsing for understanding events and answering queries. *IEEE MultiMedia*.

Zhiguo Wang, Patrick Ng, Xiaofei Ma, Ramesh Nallapati, and Bing Xiang. 2019. Multi-passage bert: A globally normalized bert model for open-domain question answering. *empirical methods in natural language processing*.

Jingyao Wang et al. 2022. Machine learning-based human-robot interaction in its. *Information Processing & Management*, 59(1):102750.

Zhiquan Wen et al. 2021. Debiased visual question answering from feature and sample perspectives. *Advances in Neural Information Processing Systems*, 34:3784–3796.

Chenfei Wu, Jinlai Liu, Xiaojie Wang, and Xuan Dong. 2018. Object-difference attention. In *Proceedings of the 26th ACM international conference on Multimedia*.

Wei Yang, Yuqing Xie, Aileen Lin, Xingyu Li, Luchen Tan, Kun Xiong, Ming Li, and Jimmy Lin. 2019. End-to-end open-domain question answering with bertserini. *north american chapter of the association for computational linguistics*.

Zhou Yu et al. 2019. Deep modular co-attention networks for visual question answering. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

A Example Appendix

This is an appendix.