sensors

MDPI

Article

Adversarial Learning with Bidirectional Attention for Visual Question Answering

Qifeng Li [1,2,3,*], Xinyi Tang [1,3] and Yi Jian [1,3]

[1] Shanghai Institute of Technical Physics of the Chinese Academy of Sciences, Shanghai 200083, China; gq227@mail.sitp.ac.cn (X.T.); jianyi@mail.sitp.ac.cn (Y.J.)
[2] School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing 100049, China
[3] Key Laboratory of Infrared System Detection and Imaging Technology, Chinese Academy of Sciences, Shanghai 200083, China
* Correspondence: liqifeng@ustc.edu

Abstract: In this paper, we provide external image features and use the internal attention mechanism to solve the VQA problem given a dataset of textual questions and related images. Most previous models for VQA use a pair of images and questions as input. In addition, the model adopts a question-oriented attention mechanism to extract the features of the entire image and then perform feature fusion. However, the shortcoming of these models is that they cannot effectively eliminate the irrelevant features of the image. In addition, the problem-oriented attention mechanism lacks in the mining of image features, which will bring in redundant image features. In this paper, we propose a VQA model based on adversarial learning and bidirectional attention. We exploit external image features that are not related to the question to form an adversarial mechanism to boost the accuracy of the model. Target detection is performed on the image—that is, the image-oriented attention mechanism. The bidirectional attention mechanism is conducive to promoting model attention and eliminating interference. Experimental results are evaluated on benchmark datasets, and our model performs better than other models based on attention methods. In addition, the qualitative results show the attention maps on the images and leads to predicting correct answers.

Keywords: bidirectional attention; adversarial learning; visual question answering; attention visualization; feature fusion; feature selection; attention mechanism

Citation: Li, Q.; Tang, X.; Jian, Y. Adversarial Learning with Bidirectional Attention for Visual Question Answering. *Sensors* **2021**, *21*, 7164. https://doi.org/10.3390/s21217164

Academic Editors: Friedhelm Schwenker and Mariofanna Milanova

Received: 30 September 2021
Accepted: 25 October 2021
Published: 28 October 2021

1. Introduction

Computer vision and natural language processing have developed rapidly for a long time. As a result, more and more cross fields of machine vision and natural language processing have emerged. Visual question answering (VQA) is one of the most important research fields. To better solve the problem of visual question answering, it is necessary to understand the image that the method focuses on and the deep visual features contained in the correct area of the image.

The attention mechanism is a human-specific brain signal processing mechanism [1–4]. In a world where we unintentionally or intentionally receive different signals every day, mainly from sight and hearing, top–down attention mechanisms allow us to quickly notice important signals in the environment. Although this kind of attention pays attention to the important signals of the surrounding environment quickly, this kind of unidirectional attention is still insufficient in efficiency. That is, there is no in-depth mining of features from the perspective of images. So, we choose a bidirectional attention method, including bottom–up attention and top–down attention, to make up for this shortcoming [5].

The answer of a good VQA model should be no different from that of humans. The adversarial learning is motivated by Generative Adversarial Nets (GANs). In the process of recognizing objects, humans enhance the cognitive ability of objects through contrast.

Since the adversarial learning was proposed, it has achieved a series of good achievements in the field of artificial intelligence [6–8]. In this article, we propose using deep learning features with two different pictures, which match with related questions to infer the best answer. Through adversarial learning, the model will pay attention to significant parts of the target image that are distinguished from the adversarial image [9].

In this paper, we propose a bidirectional attention mechanism through adversarial learning for improving the understanding facility for the fusion of language and visual representations. The main flow of the model followed is illustrated in Figure 1. Given a question, a related image, and an unrelated image, we use a bidirectional attention network to obtain a target attention embedding, which can find the part of the image that is most relevant to the problem. Meanwhile, by subtracting the target attention embedding and the adversarial attention embedding, we conduct the adversarial learning to filter out useless parts of the image and highlight the most meaningful image part for the question.

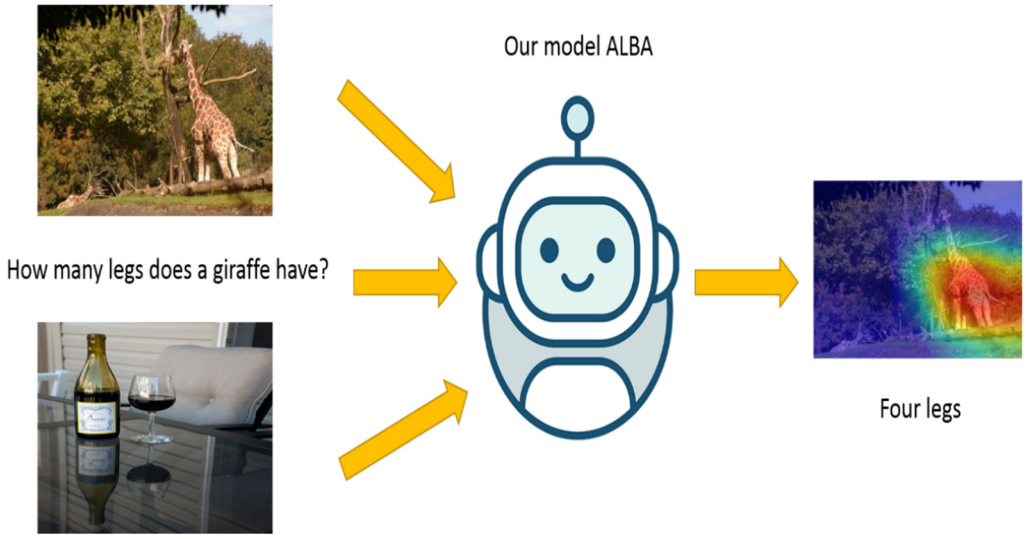

Figure 1. As an example, our model fuses features of the target image and the adversarial image according to the question, predicts the correct answer, and visualizes the attention.

2. Related Work

In this section, we briefly introduce the previous research of VQA, especially the development in the field of adversarial learning and attention mechanism.

2.1. Attention Mechanism

The attention mechanism has proved to be very effective in many tasks, and visual question answering is no exception. Several deep neural networks based on an attention mechanism have been proposed for visual question answering, in which question-oriented attention combined with image regions are commonly used.

The creation and use of attention maps have aroused great interest. Lu et al. [10] proposed a co-attention model for visual question answering that figures out image and question attention. Shih at al. [11] presented a way of learning to answer visual questions by selecting image areas related to text-based queries. Yang et al. [12] created a stacked attention network, which generates multiple attention maps on the image in a sequential manner, in order to increase the reasoning steps and improve the reasoning ability.

Early VQA research started with the attention mechanism concerning the question. Xu et al. [13] proposed a spatial memory network to apply question-guided spatial attention on certain parts of the image, which are related to either a single word or the whole question. Lu et al. produced the co-attention mechanism that creates attention on a certain area of the image and related question words. Shi et al. [14] suggested Question Type

Guided Attention (QTA), which takes advantage of information about the question type to efficiently balance the bottom–up and top–down visual features.

2.2. Glove

In the training or testing phase, the input of each instance is a target image, an adversarial image, and a question. Spaces and punctuation are used to divide the question into individual words. Any number is taken as a word. The maximum number of words in each question is reduced to 15 words, which can keep the semantics of the question and computational efficiency. Extra words are automatically omitted. Probably only 0.1% of the questions exceed 15 words. Each word is encoded by the glove embedding algorithm [15], which is a look-up table with 300-dimensional vectors. The public algorithm we used has been trained on Wikipedia. Questions less than 15 words are padded zero at the end of the question. The size of a resulting vector matrix is 15×300, and then, it is passed through the GRU (Gated Recurrent Unit). After processing fifteen words sequentially, we got the 512-dimensional question vectors.

2.3. Faster-RCNN

The same limitation of most models mentioned above is to use a full image to create global features to represent the visual input. This will feed noise to the model and then negatively affect the final result. In fact, we make use of Faster-RCNN as a segmentation method to divide the image into 36 different regions which are 36 different pooled convolutional feature vectors [16,17]. In this way, we can obtain image features directly related to the question, eliminate noise interference, and facilitate the attention of the model to the most relevant objects.

2.4. Adversarial Learning

The method of adversarial learning is inspired by human perception mechanisms [18]. In the process of answering questions, the more suitable and reasonable answers are given out by people comparing the differences between the information to find contradictions and differences [19–22]. Adversarial learning has made a series of good achievements in the field of artificial intelligence. Chen et al. [23] take adversarial learning as a segmentation network to automatically create realistic composite images. Wu et al. [24] presented an approach of combining Reinforcement Learning and Generative Adversarial Networks whose advantage is to overcome the relative shortage of training data. Our work is motivated by the effectiveness of the adversarial learning, but we cautiously extend it to our application VQA. Through the adversarial network, the model will pay attention to two parts: one is the area that is most relevant to the question in the target image, and the other is the part of the adversarial image that is the most different from the question. By contrast between two images, the mode learned how to find the best answer.

3. Proposed Method

This section is divided into five subheadings, which introduce the internal structure of the model, the experimental methods, and the experimental procedures in detail.

In this section, we introduce the architecture of our model in Figure 2. As the Figure 2 shows, it demonstrates the combination of adversarial learning and bidirectional attention mechanism to solve the problem of visual question answering, with bidirectional attentions over regions of the image. In order to maintain transparency, we will list the exact steps and specific hyperparameter values in the model in detail, which result in its best performance.

3.1. Finding Adversarial Images

Although few images in the VQA2.0 dataset are somewhat similar, it is necessary to choose an image that is very different from the target image in order to ensure the effectiveness of the adversarial learning network.

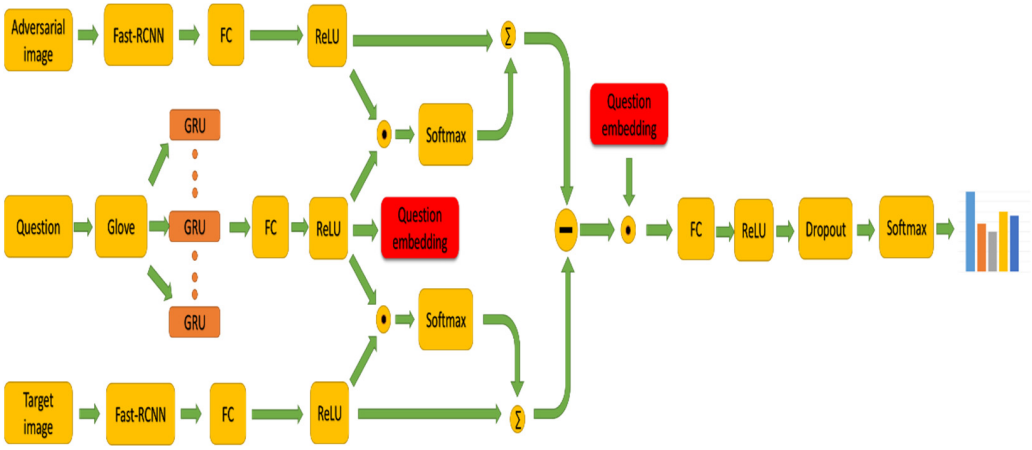

Figure 2. Overview of the proposed model. The target image–question pair and the adversarial image–question pair perform feature extraction and feature fusion operations respectively and then predict the correct answer through the classification network.

In our experiment, we use a histogram to compare the similarity of two images and select adversarial images. The histogram correlation d is given by [25,26]:

$$d(H_1, H_2) = \frac{\sum_I (H_1(I) - \overline{H}_1)(H_2(I) - \overline{H}_2)}{\sqrt{\sum_I (H_1(I) - \overline{H}_1)^2 \sum_I (H_2(I) - \overline{H}_2)^2}} \quad (1)$$

$$\overline{H}_k = \frac{1}{N} \sum_J H_k. \quad (2)$$

N is the total number of histogram bins. H_1 and H_2 represent the histogram of the target picture and the histogram of the confrontation picture respectively. $\overline{H}_k$ denotes the average of the histogram.

First, we select the target image as the reference image. Second, we randomly pick up another image in the dataset. Third, we convert two images into HSV format. Finally, we calculate the correlation of two image histograms. The greater the correlation of the histogram, the more similar the two images. If the correlation between the two pictures is less than 0.2, we select these two images as a set of input.

3.2. Image Feature Selection

As a target detection method based on convolutional neural network, Faster-RCNN first uses a set of basic layers including convolutional layers, ReLU layers, and pooling layers to extract the feature maps of the image [17]. The feature maps are shared for the subsequent RPN layer and fully connected layer.

In the RPN network, in order to achieve the matching of the common feature box in the original image with the candidate box of each target [27], it is necessary to use anchor boxes to set the target block diagram. Each feature vector in the convolution feature map corresponds to a small area in the original image, and the size of the area is determined by the convolution kernel. With each point on the feature map as the center, anchor boxes of different sizes and different proportions can be generated, corresponding to different areas on the original image. As a result of the actual changes in the size of the anchor boxes and the size of the target, usually the corresponding area cannot be completely matched with the target frame. This results in the absence of the target in the RPN built-in candidate area. In order to obtain a more accurate candidate area and complete the area proposal link, it is necessary to optimize the preset area and generate k frames of different proportions and different sizes to adapt to the target with each anchor point as the center.

The pooling layer extracts proposal feature maps after integrating feature maps and proposals information and sends them to the subsequent fully connected layer to determine the target category.

The classification layer uses proposal feature maps to distinguish the category of the proposal and utilizes bounding box regression to obtain the final precise position of the detection frame [28].

3.3. Bidirectional Attention

The target image is passed through a Faster R-CNN to obtain k × 2048-dimensional vector representation, where K represents the number of features on an image. Each feature is a 2048-dimensional vector, which encodes a specific area on the image. The Faster R-CNN can extract K (K = 36) distinctive objects from the image, which greatly provides the information related to the problem.

Our model adopts the attention method corresponding to the human question-guided model, with different features from the model of Anderson et al. [5] that focuses on the attention mechanism of image features.

The output of the Faster R-CNN (v_i) first passes through a Rectified Linear Unit (ReLU) activation layer [29] to avoid gradient explosion and gradient disappearance, and then, it passes through a fully connected layer to get a K × 512- dimensional vector m_i. w_i is a learned parameter vector

$$m_i = w_i f(v_i). \tag{3}$$

Each feature vector mi (i = 1 . . . K) and question embedding q are combined with an element-wise multiplication [30]. The resulting vector g is a fusion of image features and embedding of the question.

$$g = m_i \odot q \tag{4}$$

The fusion vector is normalized over all features with a softmax function.

$$\lambda = \text{softmax}(g) \tag{5}$$

Then, the image features m_i (i = 1 . . . K) from an entire image are weighted by the normalized values and summed to obtain a 512-dimensional vector V_T.

$$V_T = \sum_{i=1}^{K} \lambda_i m_i \tag{6}$$

The same processing steps are taken for the adversarial image to get a 512-dimensional vector V_A. Note that this attention mechanism is two-way attention, as opposed to simple one-way attention and a more complicated attention mechanism.

3.4. Adversarial Attention

Based on the adversarial learning, some features of the representation of the target image V_T not related to the question have been removed by subtracting the image features of the adversarial image V_A. So, we get some image features without redundancy. The representations of the target image V_T and of the adversarial image V_A are processed by element-wise subtraction. W_1 is a learned parameter vector.

$$h = w_1(V_T \ominus V_A) \tag{7}$$

3.5. Classifier

In the train set, each question has 10 answers, and these answers constitute the label of training classification, in which we select the correct answer appearing more than 7 times. Therefore, we got a total of N = 3129 candidate answers. Improving the efficiency of the VQA model is actually improving the accuracy of the multi-label classification task. In the

VQA2.0 training set, each answer is labeled with soft accuracies in the range of 0 to 1. The score of each answer is different based on the judgment of different people.

The fusion of image features and question embedding h is fed into a ReLU activation layer, passes through a fully connected layer and a dropout layer, and then finally passes through a softmax layer [31] to predict a score for each candidate answer.

$$\hat{s}_{ij} = \text{softmax}\left\{f_d[w_s.\, f_s(h)]\right\} \tag{8}$$

where $w_s \in R^{N \times 512}$ is a learned weight matrix, f_s is a ReLU activation layer, and f_d is a dropout layer ($p = 0.5$).

The softmax layer normalizes the answer score to (0,1). Then, we use a loss method similar to binary cross-entropy [32]. This layer can be seen as using logistic regression to predict the correct answer. The final function is

$$L = -\sum_{i}^{M}\sum_{i}^{M} s_{ij}\log\left(\hat{s}_{ij}\right) - \left(1 - s_{ij}\right)\log\left(1 - \hat{s}_{ij}\right) \tag{9}$$

where M and N represent respectively the number of training questions and the number of candidate answers. S is the soft ground-truth scores of truth answers. Soft scores as targets maintain more effective target signals than binary labels.

4. Evaluation

In the section, for analyzing our proposed VQA model, we focus on analyzing how the glove algorithm affects the question embedding and accuracy of the model performance. In addition, we present additional experiments to prove the superiority of our model by comparing with other models. Finally, we analyze the specific impact of the bidirectional and adversarial attention mechanism.

4.1. Dataset

The VQA2.0 dataset [33] is the richest and most extensive dataset, which is the updated version of VQA1.0 dataset. It increases the diversity of answers to each question to minimize the impact of dataset priors. The answers to each question are divided into three categories: yes/no, number, other. Each question includes ten candidate answers. The VQA2.0 dataset was selected as the official dataset of the VQA Challenge, which includes 443,757 train, 214,354 validation, 447,793 test questions.

4.2. Experimental Setting

Our model adopts single network learning instead of ensemble learning. Each of our networks is trained based on the VQA2.0 dataset. The model is trained multiple times on the VQA2.0 training dataset, and the optimal parameters are selected. The highest accuracy rate on the validation test VQA2.0 is selected as the result. We performed each step of the experiment four times, each time using different random seeds. In order to evaluate the performance of the model, the standard VQA metric [34] is used to calculate the accuracy, which minimizes the accidental noise from the annotators of the ground truth answers.

4.3. Ablation Study

In order to obtain the optimal model, we have done relevant comparative experiments through the controlled variable method. By changing the model parameters and model structure, the most suitable architecture and hyperparameter values can be selected. In addition, we evaluate the sensitivity of each part to the final experimental results.

Image similarity is an important variable that affects the model. Our model uses adversarial images with similarity less than 0.2, and the control group uses adversarial images with similarity greater than 0.2. Compared with the control group, it can be seen from the accuracy rate that almost every item has been improved by almost 1%. This means that the smaller the similarity between two images, the higher the accuracy of the model.

Glove word embedding of dimension 300 followed by a one-layer GRU is adopted in our reference model. Three other word-encoding methods are selected to compare with our reference model. The performance of the first and third methods is slightly reduced by about 1%. The second method is the worst of the three.

Our image features are obtained through a combination of a Faster R-CNN framework and 101-layer ResNet. A fixed threshold is used to limit the number of object detections, and the number of features K perfectly matches the content of the image. The range of parameters K is from 0 to 100. In our experiment, we take K = 36 as the parameter. In this way, we can lower computation and reduce complexity. A 200-layer ResNet as a main option is used in our comparative experiment. The performance of the ResNet-200 features downsampled to 7 × 7 (K = 49) dramatically drops to 69.59%. The ResNet-200 global features (K = 1) are expectedly even worse.

It can be seen from the experimental data on Table 1 that adopting adversarial learning helps improve the accuracy and robustness of the model. Obviously, the model is not only sensitive to the correct answer but can also effectively identify the wrong answer.

Table 1. VQA 2.0 validation score.

	All	**Yes/No**	**Numbers**	**Other**
Reference model	72.12	88.12	53.79	62.38
Image similarity				
Correlation ($p > 0.2$)	71.58	87.1	52.7	61.17
Question embedding				
100-dimensional glove and forward GRU	71.15	87.23	52.69	61.12
200-dimensional glove and forward GRU	70.98	87.01	51.69	62.01
300-dimensional glove and 2-layer forward GRU	71.34	87.54	52.57	61.88
Image features				
ResNet-200 global features (K = 1)	67.12	85.24	51.01	60.98
ResNet-200 features 14 × 14 (K = 196)	68.83	85.84	51.71	61.58
ResNet-200 features downsampled to 7 × 7 (K = 49)	69.59	86.17	51.97	61.93
Attention mechanism				
CNN 7 × 7 features	65.59	84.20	50.75	59.76
ResNet-200 7 × 7 features	66.41	85.68	51.91	60.74
Adversarial learning				
Without adversarial learning	68.78	86.09	52.84	60.62

4.4. Comparison with Existing Methods

In the section, we will introduce the effectiveness and functions of our proposed network. In order to ensure the fairness of the comparison process, the published model we chose does not use additional datasets for training. Table 2 shows the comparison of the results between the proposed method and the other published methods on VQA2.0. Obviously, our model outperforms performs better than the published models in the table by a margin of 0.6% on the test-standard and test-dev dataset. In addition, in three entries, the performance of our model has improved by about 1% compared to second place. This shows that in different categories, the model can understand the complex relationships between question–image pairs. The results indicate that adversarial learning with a bidirectional attention model outperforms the previous published methods on both the test-standard and test-dev datasets using Faster R-CNN and glove.

Table 2. Comparison of the results between the proposed method and the other published methods on VQA2.0 using similar settings.

Model	Test-Dev				Test-Standard
	Overall	Yes/No	Number	Other	Overall
VQA team [35]	57.75	80.5	36.77	43.08	58.16
SMem-VQA [14]	57.99	80.87	37.32	43.12	58.24
SAN [12]	58.7	79.3	36.6	46.1	58.9
FDA [36]	59.24	81.14	36.16	45.77	59.54
HQIC [10]	61.8	79.7	38.7	51.7	62.1
FR-VQA	62.43	77.18	33.52	56.09	-
RAU [37]	63.3	81.9	39	53	63.2
DAN [38]	64.3	83	39.1	53.9	64.2
MLB [39]	65.08	84.14	38.21	54.87	65.07
MFB [30]	65.9	84	39.8	56.2	65.8
DCN [40]	66.89	84.61	42.35	57.31	67.02
Count [41]	68.09	83.14	51.62	58.97	68.41
VisualBERT [42]	71.03	87.39	52.64	61.01	71.19
MCAN+VC [43]	71.5	87.09	53.82	61.97	71.72
ALBA (Ours)	72.12	88.12	53.79	62.38	72.33

4.5. Qualitative Results

In order to understand the internal mechanism of our model, we visualize the attention maps that ALBA generates for predicting answers. Figure 3 indicates the visualization of the related questions and predicted answers from our model ALBA. The color of attention transitions from red to blue, representing the attention weight value from high to low. The image area most relevant to the problem is highlighted in red. The visualization of attention maps indicates that the capability of ALBA is to fuse related and irrelevant image together according to the question to find the correct answers. In addition, we have conducted comprehensive tests on different types of questions. From the results of the four different question types above, it can be seen that the model has good robustness and comprehensive understanding.

The similarity between images is also an important factor that affects the attention mechanism. From the above data and analysis, it can be seen that the smaller the similarity, the more accurately the model can improve. It can be seen from Figure 4 that when the image similarity is greater than 0.2, the model cannot answer the question accurately, and the focus is on the wall that is not related to the question. When the image similarity is less than 0.2, the model can effectively integrate the question features and image features, take advantages of the attention mechanism, and focus on the target.

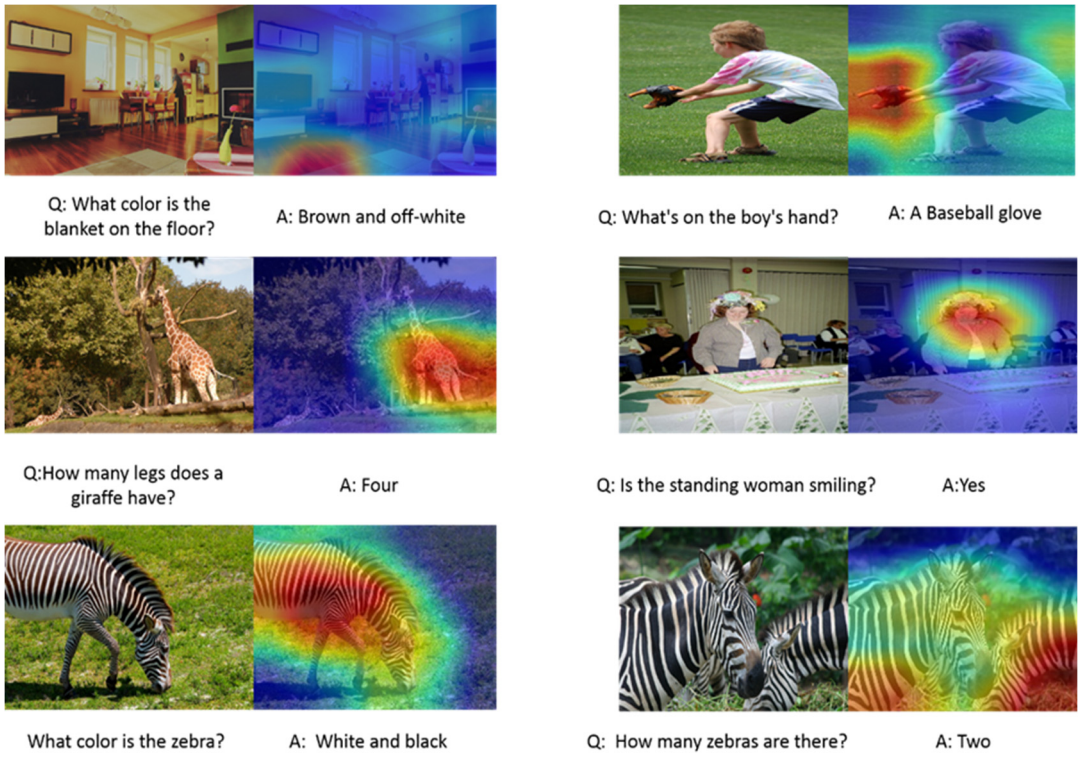

Figure 3. Typical experiment results of related images and questions from the VQA2.0 dataset. The image on the left of each row is related to the question. The image on the right of each row is a visualization of the attention maps related to the question. The text below the image is the question and answer related to the image.

Figure 4. Comparison of the influence of image similarity on attention mechanism.

5. Conclusions

In this paper, we provide three contributions for solving VQA problems. First, we propose to use adversarial learning methods to use images that are not related to the problem to improve the model's understanding and reasoning ability.

Second, the bidirectional attention mechanism is used to extract the features of the effective area of the image and then fuse with the question, which reduces the image

redundancy and the amount of calculation and makes the attention mechanism of the model more effective.

Third, the VQA accuracy of our model is based on Faster-RCNN and glove, and it outperforms other proposed algorithms. Furthermore, our model has strong adaptability and robustness to different types of questions.

Although our model improves the understanding of the problem and thus the accuracy, it does not have the ability to reason. This will be an optional direction for future model improvement. We aim to explore the logic and spatial reasoning capabilities of the model. Although the paper is not a breakthrough in the field, it can give an alternative direction to solve the VQA problem.

Author Contributions: Conceptualization, Q.L. and X.T.; methodology, Q.L. and Y.J.; software, Q.L. and X.T.; validation, Q.L. and Y.J.; formal analysis, Q.L. and X.T.; investigation, Q.L. and Y.J.; resources, Q.L.; data curation, Q.L. and Y.J.; writing—original draft preparation, Q.L.; writing—review and editing, Q.L. and X.T.; visualization, Q.L.; supervision, X.T.; project administration, Q.L.; funding acquisition, X.T. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported in part by the National Pre-Research Foundation of China (Grant No.104040402); in part by the Youth Innovation Promotion Association CAS; in part by the Innovation Project (CX-70) of Shanghai Institute of Technical Physics of The Chinese Academy of Sciences.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Publicly available datasets were analyzed in this study. Available online: https://visualqa.org/download.html (accessed on 26 April 2021).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Guo, W.; Zhang, Y.; Yang, J.; Yuan, X. Re-Attention for Visual Question Answering. *IEEE Trans. Image Process.* **2021**, *30*, 6730–6743. [CrossRef] [PubMed]
2. Zheng, X.; Wang, B.; Du, X.; Lu, X. Mutual Attention Inception Network for Remote Sensing Visual Question Answering. *IEEE Trans. Geosci. Remote. Sens.* **2021**, 1–14. [CrossRef]
3. Rahman, T.; Chou, S.H.; Sigal, L.; Carenini, G. An Improved Attention for Visual Question Answering. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 21–24 June 2021; pp. 1653–1662.
4. Zhang, S.; Chen, M.; Chen, J.; Zou, F.; Li, Y.-F.; Lu, P. Multimodal feature-wise co-attention method for visual question answering. *Inf. Fusion* **2021**, *73*, 1–10. [CrossRef]
5. Teney, D.; Anderson, P.; He, X.; Van Den Hengel, A. Tips and tricks for visual question answering: Learnings from the 2017 challenge. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 4223–4232.
6. Liu, Y.; Zhang, X.; Zhao, Z.; Zhang, B.; Cheng, L.; Li, Z. ALSA: Adversarial Learning of Supervised Attentions for Visual Question Answering. *IEEE Trans. Cybern.* **2021**, 1–14. [CrossRef]
7. Liu, Y.; Zhang, X.; Huang, F.; Cheng, L.; Li, Z. Adversarial Learning With Multi-Modal Attention for Visual Question Answering. *IEEE Trans. Neural Netw. Learn. Syst.* **2021**, *32*, 3894–3908. [CrossRef] [PubMed]
8. Tang, R.; Ma, C.; Zhang, W.E.; Wu, Q.; Yang, X. Semantic equivalent adversarial data augmentation for visual question answering. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2020; pp. 437–453.
9. Ilievski, I.; Feng, J. Generative attention model with adversarial self-learning for visual question answering. In Proceedings of the on Thematic Workshops of ACM Multimedia, Mountain View, CA, USA, 23–27 October 2017; pp. 415–423.
10. Lu, J.; Yang, J.; Batra, D.; Parikh, D. Hierarchical question-image co-attention for visual question answering. *Adv. Neural Inf. Process.Syst.* **2016**, *29*, 289–297.
11. Shih, K.J.; Singh, S.; Hoiem, D. Where to look: Focus regions for visual question answering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016.
12. Yang, Z.; He, X.; Gao, J.; Deng, L.; Smola, A. Stacked attention networks for image question answering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 21–29.
13. Xu, H.; Saenko, K. Ask, Attend and Answer Exploring Question-Guided Spatial Attention for Visual Question Answering. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2016.
14. Shi, Y.; Furlanello, T.; Zha, S.; Anandkumar, A. Question Type Guided Attention in Visual Question Answering. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018.

15. Pennington, J.; Socher, R.; Manning, C.D. Glove: Global vectors for word representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014; pp. 1532–1543.
16. Fang, B.; Sun, F.; Liu, H.; Tan, C.; Guo, D. A glove-based system for object recognition via visual-tactile fusion. *Sci. China Inf. Sci.* **2019**, *62*, 50203. [CrossRef]
17. Sun, X.; Wu, P.; Hoi, S.C. Face detection using deep learning: An improved faster RCNN approach. *Neurocomputing* **2018**, *299*, 42–50. [CrossRef]
18. Sabokrou, M.; Khalooei, M.; Fathy, M.; Adeli, E. Adversarially Learned One-Class Classifier for Novelty Detection. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 3379–3388.
19. Liang, J.; Cao, Y.; Zhang, C.; Chang, S.; Bai, K.; Xu, Z. Additive adversarial learning for unbiased authentication. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2019; pp. 11428–11437.
20. Jaiswal, A.; Wu, Y.; AbdAlmageed, W.; Masi, I.; Natarajan, P. Aird: Adversarial learning framework for image repurposing detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2019; pp. 11330–11339.
21. Agresti, G.; Schaefer, H.; Sartor, P.; Zanuttigh, P. Unsupervised domain adaptation for tof data denoising with adversarial learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2019; pp. 5584–5593.
22. Ma, J.; Liang, P.; Yu, W.; Chen, C.; Guo, X.; Wu, J.; Jiang, J. Infrared and visible image fusion via detail preserving adversarial learning. *Inf. Fusion* **2020**, *54*, 85–98. [CrossRef]
23. Chen, B.-C.; Kae, A. Toward Realistic Image Compositing with Adversarial Learning. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 16–18 June 2019; pp. 8415–8424.
24. Wu, Q.; Wang, P.; Shen, C.; Reid, I.; Hengel, A.V.D. Are You Talking to Me? Reasoned Visual Dialog Generation Through Adversarial Learning. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6106–6115.
25. Zhu, L.Q.; Zhang, Z. Auto-classification of insect images based on color histogram and GLCM. In *Proceedings 2010 Seventh International Conference on Fuzzy Systems and Knowledge Discovery, Yantai, China, 10–12 August 2010*; IEEE: Piscataway, NJ, USA, 2010; Volume 6, pp. 2589–2593.
26. Huang, J.; Kumar, S.R.; Mitra, M.; Zhu, W.J.; Zabih, R. Image indexing using color correlograms. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Juan, Puerto Rico, 17–19 June 1997; IEEE: Piscataway, NJ, USA, 1997; pp. 762–768.
27. Chen, D.; Hua, G.; Wen, F.; Sun, J. Supervised transformer network for efficient face detection. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2016; pp. 122–138.
28. Jiang, H.; Learned-Miller, E. Face detection with the faster R-CNN. In Proceedings of the 2017 12th IEEE International Conference on Automatic Face Gesture Recognition, FG 2017, Washington, DC, USA, 30 May–3 June 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 650–657.
29. Hara, K.; Saito, D.; Shouno, H. Analysis of function of rectified linear unit used in deep learning. In Proceedings of the 2015 International Joint Conference on Neural Networks, IJCNN, Killarney, Ireland, 12–16 July 2015; IEEE: Piscataway, NJ, USA, 2015; pp. 1–8.
30. Yu, Z.; Yu, J.; Fan, J.; Tao, D. Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 1821–1830.
31. Yuan, B. Efficient hardware architecture of softmax layer in deep neural network. In Proceedings of the 2016 29th IEEE International System-on-Chip Conference, SOCC, Seattle, WA, USA, 6–9 September 2016; IEEE: Piscataway, NJ, USA, 2016; pp. 323–326.
32. Ramos, D.; Franco-Pedroso, J.; Lozano-Diez, A.; Gonzalez-Rodriguez, J. Deconstructing cross-entropy for probabilistic binary classifiers. *Entropy* **2018**, *20*, 208. [CrossRef] [PubMed]
33. Goyal, Y.; Khot, T.; Summers-Stay, D.; Batra, D.; Parikh, D. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 6904–6913.
34. Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; Zhang, L. 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, Salt Lake City, UT, USA, 18–22 June 2018; pp. 6077–6086.
35. Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Zitnick, C.L.; Parikh, D. Vqa: Visual question answering. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 2425–2433.
36. Ilievski, I.; Yan, S.; Feng, J. A focused dynamic attention model for visual question answering. *arXiv* **2016**, arXiv:1604.01485.
37. Noh, H.; Han, B. Training recurrent answering units with joint loss minimization for vqa. *arXiv* **2016**, preprint. arXiv:1606.03647.
38. Nam, H.; Ha, J.W.; Kim, J. Dual attention networks for multimodal reasoning and matching. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 299–307.

39. Kim, J.H.; On, K.W.; Lim, W.; Kim, J.; Ha, J.W.; Zhang, B.T. Hadamard product for low-rank bilinear pooling. *arXiv* **2016**, preprint. arXiv:1610.04325.
40. Nguyen, D.K.; Okatani, T. Improved fusion of visual and language representations by dense symmetric co-attention for visual question answering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 6087–6096.
41. Zhang, Y.; Hare, J.; Prügel-Bennett, A. Learning to count objects in natural images for visual question answering. *arXiv* **2018**, preprint. arXiv:1802.05766.
42. Li, L.H.; Yatskar, M.; Yin, D.; Hsieh, C.J.; Chang, K.W. Visualbert: A simple and performant baseline for vision and language. *arXiv* **2019**, preprint. arXiv:1908.03557.
43. Wang, T.; Huang, J.; Zhang, H.; Sun, Q. Visual commonsense r-cnn. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 10760–10770.