electronics

MDPI

Article

Multi-Modality Global Fusion Attention Network for Visual Question Answering

Cheng Yang, Weijia Wu, Yuxing Wang * and Hong Zhou *

Key Laboratory for Biomedical Engineering of Ministry, Zhejiang University, Hangzhou 310027, China; zijingyang@zju.edu.cn (C.Y.); weijia_wu@yeah.net (W.W.)

* Correspondence: wangyuxing@zju.edu.cn (Y.W.); hongzhou_zju@163.com (H.Z.)

Received: 13 October 2020; Accepted: 6 November 2020; Published: 9 November 2020

Abstract: Visual question answering (VQA) requires a high-level understanding of both questions and images, along with visual reasoning to predict the correct answer. Therefore, it is important to design an effective attention model to associate key regions in an image with key words in a question. Up to now, most attention-based approaches only model the relationships between individual regions in an image and words in a question. It is not enough to predict the correct answer for VQA, as human beings always think in terms of global information, not only local information. In this paper, we propose a novel multi-modality global fusion attention network (MGFAN) consisting of stacked global fusion attention (GFA) blocks, which can capture information from global perspectives. Our proposed method computes co-attention and self-attention at the same time, rather than computing them individually. We validate our proposed method on the two most commonly used benchmarks, the VQA-v2 datasets. Experimental results show that the proposed method outperforms the previous state-of-the-art. Our best single model achieves 70.67% accuracy on the test-dev set of VQA-v2.

Keywords: visual question answering; global attention mechanism; deep learning

1. Introduction

With the development of deep learning, researchers have made great progress in many computer vision tasks in the last several years, e.g., classification [1,2] and detection [3,4]. Along with computer vision, natural language processing has made considerable progress as well, e.g., reading comprehension [5,6] and information extraction [7]. Since vision and language are the main forms of human communication, tasks that need to understand language and vision simultaneously have gained increasing attention, including visual captioning [8–10], visual grounding [11], and visual question answering (VQA) [12,13]. VQA is a system that has to answer free-form questions by reasoning about presented images [14]. It has many applications in practice, such as assisting vulnerable (and blind) people to access image information and improving human-machine interaction.

VQA is a challenging task since it requires a high-level understanding of both questions and images, along with visual reasoning to predict the correct answer. In recent years, many methods have been proposed to improve the performance of VQA models. Almost all of these methods are based on an attention mechanism, and they focus on how to adaptively select important features that can help with the correct answer. The basic idea of attention is that certain visual regions in an image and certain words in a question are more informative than others for answering a given question. The early methods only computed the attention of visual regions from the question [15,16]. Recent works have shown that it is also

important to learn the textual attention of question words from the image, and learning co-attention for textual and visual synchronously can lead to more accurate prediction [17]. However, these co-attention networks lose sight of self-attention, which has been proven effective for Transformer [18].

To overcome the weakness of lacking self-attention, two intra- and inter- modality attention models DFAF [19] and MCAN [20] were proposed to model both self-attention in each modality and co-attention between textual and visual modality. DFAF first generates an inter-modal attention flow to pass information between the image and question [19] and then calculates intra-modal attention based on the inter-modal attention flow to capture intra-relations. MCAN designs a self-attention (SA) unit and a guided-attention (GA) unit to respectively model intra-modality and inter-modality interactions and then combines SA and GA to construct the basic layer. Interestingly, both of these models can be cascaded in depth; therefore, they can support more complex visual reasoning.

Despite attention-based models having already attained great achievement, they only model relationships between individual regions in an image and words in a question. Therefore, we propose a novel multi-modality global fusion attention network (MGFAN) with a GFA block. Global fusion attention (GFA) blocks can capture information from global perspectives. Our proposed method computes co-attention and self-attention at the same time, rather than individually. The details of our method will be explained in Section 3.

Our contribution can be summarized in the following:

- We propose a novel MGFAN for VQA that can compute attention considering global information.
- In the MGFAN, we combine self-attention and co-attention into a unified attention framework.
- Our proposed approach outperforms the previous state-of-the-art methods on the most used benchmark for VQA: VQA-v2.

This paper is organized as follows. In Section 2, we cover background material related to the components that we use. Section 3 demonstrates our proposed method. In Section 4, we explain the experiment in detail. We visualize the result of each iteration to explain the behavior of the MGFAN and report the result compared to the state-of-the-art works in Section 5. Finally, in Section 6, we give the conclusion of this paper.

2. Related Work

This section covers previous research and concepts related to the current work.

2.1. Representation Learning for VQA

VQA aims to answer a question in natural language with respect to a given image, so it requires multimodal reasoning over multimodal inputs. With the development of deep learning, learning good representations has been the foundation for advancing natural language processing (NLP) and computer vision (CV). Since [15] proposed using features based on Faster R-CNN, VQA boosted the performance, benefiting from the additional region information. The basic idea is to leverage pre-trained Faster R-CNN to propose some image regions, each with an associated feature vector. We follow this approach in our work.

2.2. Attention Mechanism

The attention mechanism tries to mimic how human vision works, ignores irrelevant information, and focuses on important features. This approach has achieved great success in NLP and CV. Many relational reasoning approaches use an attention mechanism to aggregate contextual information. Attention-based approaches have become mainstream in VQA. We introduce two variants of the attention mechanisms related to our method, self-attention and co-attention.

The self-attention mechanism first transforms features into query, key, and value features. Then, the attention matrix can be calculated by the inner product of the query and key features. After that, we aggregate the original features with the attention matrix to obtain the attended features. A very important article on attention research for non-local neural networks [21] proposed an attention mechanism of non-local information statistics on the basis of capturing the dependence relationship between long-distance features. Transformer [18] is a new model consisting only of self-attention and a feed forward neural network, and the work provided a new concept for designing neural networks. The co-attention-based [22,23] vision and language method models the interactions across the two modalities. MFB [24] is a co-attention mechanism using an end-to-end deep network architecture to jointly learn both the image and question attentions. Reference [25] is similar to our work: it proposes a novel attention mechanism that jointly considers reciprocal relationships between the two levels of visual details.

3. Multi-Modality Global Fusion Attention Network

The structure of the multi-modality global fusion attention network (MGFAN) is depicted in Figure 1. Given an image I and a question Q related to this image, we extract image features with Faster R-CNN [26] pre-trained on the Visual Genome [27] dataset and extract question features with the gated recurrent unit (GRU) [28]. Every question is padded or truncated to a maximum of n words, and each word in the question is initialized with pre-trained GloVe300-Dimensional Word Vectors [29]. The obtained image features R and question features E can be denoted as:

$$R = RCNN(I; \theta_{RCNN}) \tag{1}$$

$$E = GRU(Q; \theta_{GRU}) \tag{2}$$

where $R \in \mathbb{R}^{m \times d_r}$ and $E \in \mathbb{R}^{n \times d_e}$, m means the amount of object regions proposed by Faster R-CNN, and n means the length of a question. We adopt the same setup as [30]: $m = 36$, $n = 14$, $d_r = 2048$, and $d_e = 1024$. This is a common setup used by most of the recent methods for VQA. We also use this setup for the sake of a fair comparison. To reduce the computation and project the features from different modalities into the same vector space, we feed original features R and E into two independent fully-connected layers at first. Then, we obtain mapped features $\bar{R}$ and $\bar{E}$, where $\bar{E} \in \mathbb{R}^{n \times dim}$ and $\bar{R} \in \mathbb{R}^{m \times dim}$. dim represents the common dimension of transformed features from both modalities.

$$\bar{R} = FC_R(R), \ \bar{E} = FC_E(E) \tag{3}$$

In Section 3.1, we will introduce GFA block, the core component of our proposed network. Next, in Section 3.2, we will describe the MGFAN architecture.

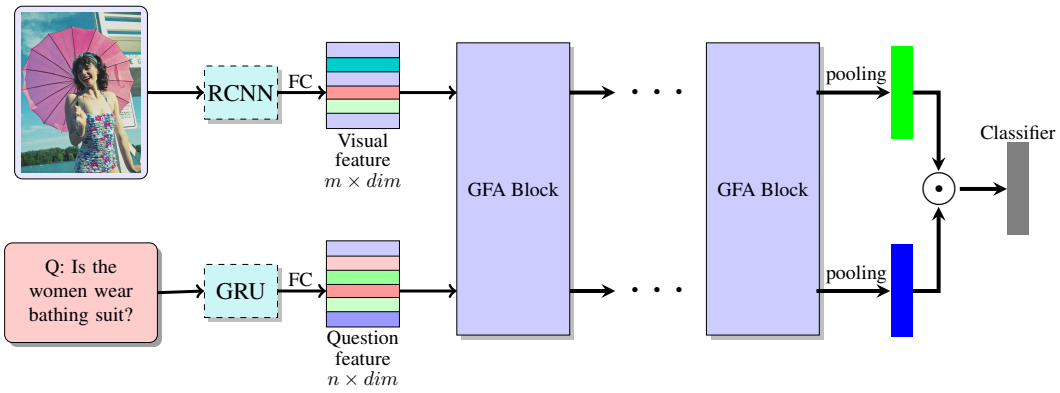

Figure 1. Illustration of the multi-modality global fusion attention network (MGFAN) with the GFA block. The GFA block can be stacked to help the network focus on the most important question words and image regions. GRU, gated recurrent unit.

3.1. Global Fusion Attention Block

Figure 2 illustrates the proposed GFA block. In GFA, we summarize all features into k vectors and then compute the attention such that the attended features conclude about the global information. To retain the discrimination, we element-wisely product the original and attended features. We use residual architecture, as most networks do.

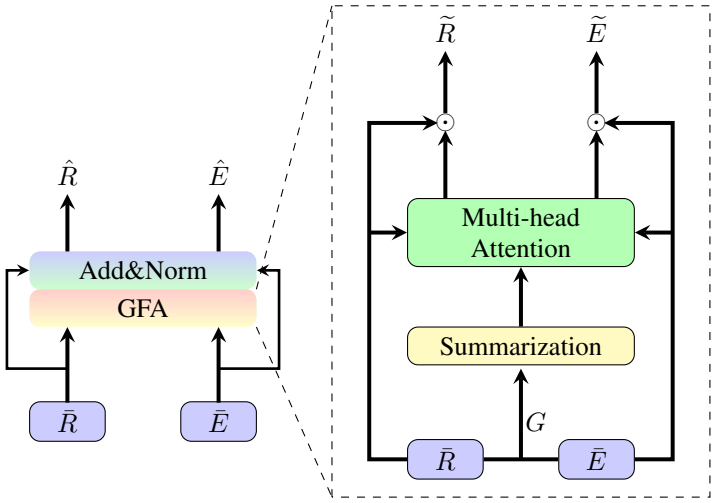

Figure 2. (**a**) Illustration of the global fusion attention (GFA) layer. (**b**)Illustration of the GFA block. See the details in Section 3.1.

Given textual features $\bar{E}$ and visual features $\bar{R}$, we first concatenate them into rows to obtain cross-modal feature matrix G, $G = [\bar{R}; \bar{E}] = [g_1, g_2, \ldots, g_s] \in \mathbb{R}^{s \times dim}$, $s = m + n$. Then, we compute combination weights W_C via:

$$W_C = softmax(W_G G^T + b_G) \tag{4}$$

where $W_G \in \mathbb{R}^{k \times dim}$ and $b_G \in \mathbb{R}^k$ are learnable combination weights. k is the number of fused features, and we try three different k in the experiment to investigate how changes in k affect performance.

The softmax function is applied to all rows of the matrix to generate weights, which sum to one in each row. $W_C \in \mathbb{R}^{k \times s}$, and each row of W_C means a strategy of combining the original features. Now, we can obtain k fused features $\bar{G}$ that can be input features from different global perspectives, $\bar{G} \in \mathbb{R}^{k \times dim}$.

$$\bar{G} = W_C G \tag{5}$$

To calculate the attention, we first transform the visual region and word features into query features, then transform fused features into key and value features, where the transformed features are denoted as $Q_R \in \mathbb{R}^{m \times d}$, $Q_E \in \mathbb{R}^{n \times d}$, K, and $V \in \mathbb{R}^{k \times d}$; we set $k = 6$ as the default. This is a trade-off choice between performance and computing time, and we conduct external experiments to investigate the influence of changing k. d represents the dimension of each feature; we just set $d = 512$ in accordance with practice (it does not really matter).

$$Q_R = FC_r^q(\bar{R}) \quad Q_E = FC_e^q(\bar{E}) \tag{6}$$

$$K = FC^k(\bar{G}) \quad V = FC^v(\bar{G}) \tag{7}$$

Then, we compute the multi-head attention following [18].

$$A(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{8}$$

$$MA(Q, K, V) = [head_1, \ldots, head_i]W^O \tag{9}$$

$$head_i = A(QW_i^Q, KW_i^K, VW_i^V) \tag{10}$$

$$A_R = MA(Q_R, K, V) \tag{11}$$

$$A_E = MA(Q_E, K, V) \tag{12}$$

The attended features are computed from fused features such that each of them contains global information, but lacks details and is not discriminative enough. Therefore, we model the relation [31] between attended features and input features instead of directly outputting attended features. We show the impact of relation modeling in Section 4.3. We use the residual architecture as most networks do. The outputs of the GFA block are:

$$\hat{R} = \bar{R} + \bar{R} \odot A_R \tag{13}$$

$$\hat{E} = \bar{E} + \bar{E} \odot A_E \tag{14}$$

3.2. Network Architecture

In this section, we describe the MGFAN architecture. As Figure 1 illustrates, the extracted multimodal features R and E are first transformed to $\bar{R}$ and $\bar{E}$, then $\bar{R}$ and $\bar{E}$ are fed into L stacked GFA blocks $[GFA^{(1)}, GFA^{(2)}, \ldots, GFA^{(L)}]$. The features are passed in a recursive manner.

$$[R^{(l)}, E^{(l)}] = GFA^{(l)}([R^{(l-1)}, E^{(l-1)}]) \tag{15}$$

where $l \in [1, L]$, $R^{(0)} = \bar{R}$, and $E^{(0)} = \bar{E}$.

Using the attended features $R^{(L)}$ and $E^{(L)}$, we average pool them and element-wise product them. Finally, we train a k-way classifier to predict the answer where k is the number of candidate answers. Since each question has multiple answers, we follow [30] and use the binary cross-entropy loss:

$$\mathcal{L} = \sum_{i=1}^{k} (y_i log(\sigma(p_i))) + (1 - y_i) log(1 - \sigma(p_i)) \tag{16}$$

where p_i is the confidence of each answer and σ is the sigmoid activation function.

4. Experiment

In this section, we conduct experiments to prove the validity of our model on the largest VQA benchmark dataset, VQA-v2.

4.1. Datasets

VQA-v2 is the most commonly used VQA benchmark dataset [32]. It is a dataset containing open-ended questions about images from the MSCOCO dataset [33]. Each image has at least 3 related questions (5.4 questions on average). Each question has 10 ground truth answers and 3 plausible answers. The dataset is split into three subsets: training (80 k images with 444 k questions); val(40 k images with 214 k questions); and testing (80 k images with 448 k questions). The test subset is further split into the test-devand test-stdsets, which are evaluated online with limited attempts [34].

4.2. Experimental Setup

The hyper-parameters we used in the experiment are as follows. We extracted visual features from Faster R-CNN [26], while textual features were encoded by GRU [28]. One image's features were represented as a set of 36 localized regions with dimension 2048. One question's features were presented as a set of 14 words with dimension 1024. Then, we embedded them into 512 dimensions with two independent fully-connected layers. In each GFA block, the latent dimension d was 512, and the number of heads was 8.

For all layers, we clipped the gradients to 0.25. All fully-connected layers had the same dropout rate of 0.1. We set the batch size as 256. We used the Adamax [35] optimizer to train our model. The learning rate was set as 10^{-3} for the first 3 epochs, 2×10^{-3} for the next 7 epochs, and decayed by $1/5$ for the rest of the epochs. While evaluating performance on the val split, we trained the model only on the train split. While evaluating performance on the test-dev or test-standard splits, we trained the model on both the train and val splits and another subset of VQA samples from Visual Genome [27].

4.3. Ablation Studies

We ran extensive ablation studies on the val split of VQA-v2 to investigate why the MGFAN is effective. We checked the influence of the number of fused vectors, the number of GFA stacks, the number of attention heads, with or without modeling relation, and the final feature fusion operator. We set the accuracy of bottom-up [15] as the baseline. The results are shown in Table 1.

From the results of the ablation studies, we can find that relation modeling is important to predict the correct answer. This result proves what we claim in Section 3.1: that the fused features are not discriminative enough. With the number of stacked GFA blocks increasing, the accuracy increases. However, if we stack too many GFA blocks, the model will overfit such that the accuracy decreases. We also performed an ablation study on the influence of the number of fused vectors; too few fused vectors were unable to capture different aspects of the input, leading to decreasing accuracy. Next, we investigated

the influence of multi-head attention. One, 4, and 8 attention heads were experimented with, and eight heads achieved the best performance. Feature multiplication performed marginally better than feature addition and concatenation for the feature fusion method. Finally, we experimented on the influence of the embedding dimension: 512 resulted in better performance than 1024.

Table 1. Ablation studies of our proposed MGFAN on the Visual Question Answering (VQA)-v2 validation dataset. The default setting is represented by the underline.

Component	Setting	Accuracy (%)
Bottom-up	Bottom-up	63.37
Stacked blocks	GFA-3	66.18
	GFA-5	66.34
	GFA-8	**66.49**
	GFA-12	66.26
Relation modeling	Yes	**66.34**
	No	65.41
Fused vectors	4	65.67
	6	66.21
	8	**66.34**
Parallel heads	1 head each 512	65.98
	4 heads each 128	66.22
	8 heads each 64	**66.34**
Final feature fusion	Addition	66.25
	Concatenation	66.29
	Multiplication	**66.34**
Embedding dimension	512	**66.34**
	1024	66.13

5. Results

5.1. Visualization

In Figure 3, we visualize the attention weights of image regions with three GFA blocks to analyze our model. The most selected regions by our attention mechanism are shown as brighter. As shown in Figure 3, iterations through the GFA block tend to gradually focus on the regions that are most relevant to the question. In the first row, the first GFA block pays nearly average attention to several important different regions since our method computes the attention with global fused features and loses some detailed information; we model the relation between attended features and original features. Then, the second GFA block pays most attention to regions related to the question words “they” and “holding”. At last, the third GFA block focuses on the correct region of the answer “surf board” considering the type of question. The second row demonstrates a simpler case. Our model fails to predict the correct answer in the third row, although it focuses on the correct region; this is possible because our model is unable to recognize the chopped banana, and it cannot infer from several foods that it is breakfast.

Figure 3. Visualization of the attention weights of image regions with three GFA blocks.

5.2. Comparison with State-of-The-Art

In this section, we compare our MGFAN with the current state-of-the-art methods on VQA-v2. The results are shown in Table 2. Bottom-up and top-down (BUTP) was the winner of the VQA challenge 2017. This approach first proposed to use features based on Faster R-CNN. All the methods listed in Table 2 use the same bottom-up attention visual features. The test-dev split of VQA-v2 was evaluated online with more attempts than the test-std split and returned more accurate results about different types of questions. Y/N means the answer of the corresponding question is yes or no. Num means the answer of the corresponding question is a number. The best result is highlighted in Table 2, we can see that the MGFAN outperforms the current sate-of-the-art except for Num-type questions.

Table 2. Accuracies (%) of the single model on the test-devand test-stdsplits of VQA-v2. BUTP, bottom-up and top-down.

Model	Test-Dev				Test-Std
	Y/N	Num	Other	All	All
BUTP [15]	81.82	44.21	56.05	65.32	65.67
MUTAN [36]	82.88	44.54	56.50	66.01	66.38
MLB [37]	83.58	44.92	56.34	66.27	66.62
DA-NTN [38]	84.29	47.14	57.90	67.56	67.94
Counter [39]	83.14	51.62	58.97	68.09	68.41
BAN [40]	85.46	50.66	60.50	69.66	n/a
DFAF [19]	86.09	**53.32**	60.49	70.22	70.34
MGFAN (ours)	**86.32**	53.30	**60.77**	**70.45**	**70.67**

5.3. Discussion

Although our method outperforms other methods, it is also important to discuss whether some kind of price has been paid. The core of our method is the GFA block. In the GFA block, we summarize all features into k vectors and then compute the attention map. How to compute the attention map is mentioned in Section 3.1. The matrix representing key features will be $K \in \mathbb{R}^{(m+n)\times dim}$ if we use traditional self-attention and co-attention, but it is $K \in \mathbb{R}^{k\times dim}$. k is smaller than the sum of m and n, where m and n are the counts of the raw visual and text features, respectively. Based on Equation (8), it is obvious that the GFA block saves computing time. Of course, it is worth noting that the MGFAN consists of some stacked GFA blocks. Based on the results of the ablation studies, as the number of GFA blocks increases, the performance increases, as will the computational time and complexity. This will be a trade-off, choosing between computing time or performance. However, in our default setup, where k is 6, the sum of m and n is 50, and the number of stacked GFA blocks is 5, we achieved better performance without increasing the computing time.

In Section 5.1, we give a conjecture about the fact that our model fails to predict the correct answer in the third row, although it focuses on the correct region. We should further investigate this to give a definite conclusion. However, due to the imperviousness of deep learning and the huge amount of data, this is a tough task beyond our capacity now. We will do more research in future work.

6. Conclusions and Future Work

In this work, we present a novel multi-modality global fusion attention network (MGFAN) for VQA. We build our proposed model based on the hypotheses that capture information from global perspectives that can provide additional information for deep visual and textual understanding. The MGFAN computes attention weights using global information and unifies self-attention and co-attention into one framework. The GFA block is the core component of the MGFAN; it can be stacked into several layers for better reasoning. We validate our approach on the most commonly used dataset: VQA-v2. The results demonstrate the effectiveness and accuracy of our method. Additionally, we conduct ablation studies to investigate what factors might affect the results.

Our approach provides a new perspective to model the relationship between individual regions in an image and words in a question. The shortcoming of our model is that it sacrifices a certain degree of ability to finely describe local information. This can be improved in our future work. In order to make greater progress in the VQA task, it is important to explore how the neural network makes decisions. For example, more elaborate visualization methods of the computational flow within the neural network are important.

Author Contributions: Conceptualization, C.Y.; investigation, C.Y., W.W.; methodology, C.Y.; resources, C.Y.; formal analysis, C.Y. and W.W.; software, C.Y. and W.W.; writing—original draft, C.Y.; writing—review and editing, W.W., Y.W. and H.Z.; supervision, C.Y. and W.W.; funding acquisition, H.Z. All authors read and agreed to the published version of the manuscript.

Funding: This work is supported by the National Key Research and Development Program of China, through the 2019YFC0118200 project.

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

References

1. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*; MIT Press: Wayne, PA, USA, 2012; pp. 1097–1105.
2. He, K.; Zhang, X.; Ren, S.; Sun, J. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1026–1034.
3. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. SSD: Single shot multibox detector. In Proceedings of the European Conference on Computer Vision, Graz, Austria, 7–13 May 2016; pp. 21–37.
4. Law, H.; Deng, J. Cornernet: Detecting objects as paired keypoints. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; pp. 734–750.
5. Hermann, K.M.; Kocisky, T.; Grefenstette, E.; Espeholt, L.; Kay, W.; Suleyman, M.; Blunsom, P. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems*; MIT Press: Wayne, PA, USA, 2015; pp. 1693–1701.
6. Wang, W.; Yang, N.; Wei, F.; Chang, B.; Zhou, M. Gated self-matching networks for reading comprehension and question answering. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Vancouver, BC, Canada, 30 July–4 August 2017; pp. 189–198.
7. Bhutani, N.; Jagadish, H.; Radev, D. Nested propositions in open information extraction. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Austin, TX, USA, 1–5 November 2016; pp. 55–64.
8. Donahue, J.; Anne Hendricks, L.; Guadarrama, S.; Rohrbach, M.; Venugopalan, S.; Saenko, K.; Darrell, T. Long-term recurrent convolutional networks for visual recognition and description. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Santiago, Chile, 7–13 December 2015; pp. 2625–2634.
9. Ding, S.; Qu, S.; Xi, Y.; Sangaiah, A.K.; Wan, S. Image caption generation with high-level image features. *Pattern Recognit. Lett.* **2019**, *123*, 89–95. [CrossRef]
10. Chen, X.; Zhang, M.; Wang, Z.; Zuo, L.; Li, B.; Yang, Y. Leveraging unpaired out-of-domain data for image captioning. *Pattern Recognit. Lett.* **2018**, *132*, 132–140. [CrossRef]
11. Yu, Z.; Yu, J.; Xiang, C.; Zhao, Z.; Tian, Q.; Tao, D. Rethinking diversified and discriminative proposal generation for visual grounding. *arXiv* **2018**, arXiv:1805.03508.
12. Toor, A.S.; Wechsler, H.; Nappi, M. Biometric surveillance using visual question answering. *Pattern Recognit. Lett.* **2019**, *126*, 111–118. [CrossRef]
13. Zhang, W.; Yu, J.; Hu, H.; Hu, H.; Qin, Z. Multimodal feature fusion by relational reasoning and attention for visual question answering. *Inf. Fusion* **2020**, *55*, 116–126. [CrossRef]
14. Hudson, D.A.; Manning, C.D. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 6700–6709.
15. 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.
16. 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, Las Vegas, NV, USA, 27–30 June 2016; pp. 21–29.

17. Yu, Z.; Yu, J.; Xiang, C.; Fan, J.; Tao, D. Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering. *IEEE Trans. Neural Netw. Learn. Syst.* **2018**, *29*, 5947–5959. [CrossRef] [PubMed]
18. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems*; MIT Press: Wayne, PA, USA, 2017; pp. 5998–6008.
19. Gao, P.; Jiang, Z.; You, H.; Lu, P.; Hoi, S.C.; Wang, X.; Li, H. Dynamic Fusion With Intra-and Inter-Modality Attention Flow for Visual Question Answering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 6639–6648.
20. Yu, Z.; Yu, J.; Cui, Y.; Tao, D.; Tian, Q. Deep Modular Co-Attention Networks for Visual Question Answering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 6281–6290.
21. Wang, X.; Girshick, R.; Gupta, A.; He, K. Non-local neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7794–7803. [CrossRef]
22. Lu, J.; Yang, J.; Batra, D.; Parikh, D. Hierarchical question-image co-attention for visual question answering. In *Advances in Neural Information Processing Systems*; MIT Press: Wayne, PA, USA, 2016; pp. 289–297.
23. Xiong, C.; Zhong, V.; Socher, R. Dynamic coattention networks for question answering. *arXiv* **2016**, arXiv:1611.01604.
24. 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 (ICCV), Venice, Italy, 22–29 October 2017.
25. Farazi, M.R.; Khan, S.H. Reciprocal attention fusion for visual question answering. *arXiv* **2018**, arXiv:1805.04247.
26. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*; MIT Press: Wayne, PA, USA, 2015, pp. 91–99.
27. Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; Kravitz, J.; Chen, S.; Kalantidis, Y.; Li, L.J.; Shamma, D.A.; et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *Int. J. Comput. Vis.* **2017**, *123*, 32–73.
28. Cho, K.; Van Merriënboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; Bengio, Y. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv* **2014**, arXiv:1406.1078.
29. Pennington, J.; Socher, R.; Manning, C. 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.
30. 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.
31. Santoro, A.; Raposo, D.; Barrett, D.G.; Malinowski, M.; Pascanu, R.; Battaglia, P.; Lillicrap, T. A simple neural network module for relational reasoning. In *Advances in Neural Information Processing Systems*; MIT Press: Wayne, PA, USA, 2017; pp. 4967–4976.
32. 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.
33. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In Proceedings of the European Conference on Computer Vision (ECCV), Zurich, Switzerland, 6–12 September 2014; pp. 740–755.
34. Yu, Z.; Cui, Y.; Yu, J.; Tao, D.; Tian, Q. Multimodal unified attention networks for Vision-and-Language interactions. *arXiv* **2019**, arXiv:1908.04107.
35. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
36. Ben-Younes, H.; Cadene, R.; Cord, M.; Thome, N. Mutan: Multimodal tucker fusion for visual question answering. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2612–2620.
37. 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**, arXiv:1610.04325.

38. Bai, Y.; Fu, J.; Zhao, T.; Mei, T. Deep attention neural tensor network for visual question answering. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 20–35.
39. Zhang, Y.; Hare, J.; Prügel-Bennett, A. Learning to count objects in natural images for visual question answering. *arXiv* **2018**, arXiv:1802.05766.
40. Kim, J.H.; Jun, J.; Zhang, B.T. Bilinear attention networks. In *Advances in Neural Information Processing Systems*; MIT Press: Wayne, PA, USA, 2018; pp. 1564–1574.