Journal on Big Data
DOI:10.32604/jbd.2021.017169
Article

WMA: A Multi-Scale Self-Attention Feature Extraction Network Based on Weight Sharing for VQA

Yue Li, Jin Liu* and Shengjie Shang

Shanghai Maritime University, Shanghai, 201306, China
*Corresponding Author: Jin Liu. Email: jinliu@shmtu.edu.cn
Received: 22 January 2021; Accepted: 20 June 2021

Abstract: Visual Question Answering (VQA) has attracted extensive research focus and has become a hot topic in deep learning recently. The development of computer vision and natural language processing technology has contributed to the advancement of this research area. Key solutions to improve the performance of VQA system exist in feature extraction, multimodal fusion, and answer prediction modules. There exists an unsolved issue in the popular VQA image feature extraction module that extracts the fine-grained features from objects of different scale difficultly. In this paper, a novel feature extraction network that combines multi-scale convolution and self-attention branches to solve the above problem is designed. Our approach achieves the state-of-the-art performance of a single model on Pascal VOC 2012, VQA 1.0, and VQA 2.0 datasets.

Keywords: VQA; feature extraction; self-attention; fine-grained

1 Introduction

In recent years, with the development of artificial intelligence technology, methods based on deep learning have achieved great success in computer vision and natural language processing. Traditionally, these two research fields are mutually independent, but multimodal learning for language and vision has gained much attention recently such as Visual Question Answer (VQA) [1], image text retrieval [2–3], and image captioning [4–5]. Among many multimodal machine learning tasks, VQA learns to infer the answer given a real-world image and a question in natural language about the visual content of this image. Thus, VQA is challenging, because it requires a simultaneous understanding of both visual content of images and textual content of questions.

In this paper, we argue that such an effective fine-grained feature extraction module is crucial for VQA. Most existing approaches construct an effective and heavy network to extract discriminative features for image. However, previous methods based on complex DNNs cannot achieve the model of lightweight. The bottom-up attention mechanism [6] is the most straightforward and common solutions to learn discriminative features and has been successively applied in VQA tasks. The given question may strongly relate to only a small part of the image. Therefore, it is intuitive to introduce the bottom-up attention mechanism based on Faster R-CNN [7] into the VQA task to adaptively learn the most relevant image regions for a given question. On the other hand, the features extracted from CNNs with the same structure are the same, but blindly adding different CNNs may make the model more complex and redundant. And single CNNs are insensitive to different scale objects. To tackle above problem effectively, multi-scale deep network is introduced. For instance, Single Shot MultiBox Detector (SSD) [8] and Feature Pyramid Networks (FPN) [9] are the most popular multi-scale models. Motivated by these observations, we propose feature extraction as a combination of self-attention mechanism and multi-scale deep network, which considers two or more advantages. In order to demonstrate the excellence of our designed module, we conduct various ablation experiments and comparative experiments in the following part of this article.

2 Related Work

2.1 Visual Question Answering

Although the field of computer vision and question answering systems based on natural language processing has been developed for nearly half a century, the concept of visual question answering systems was formally proposed in 2015 [1]. The visual question answering system is a multi-modal fusion system that integrates images and text. It usually includes sub-task modules such as image feature extraction, problem semantic analysis, and multi-modal feature fusion.

In recent years, the networks using CNNs for deep and wide feature extraction have also gradually enriched with the development of deep learning technology [10–11]. The VGG network proposed by Simonyan et al. [12] has become one of the most popular networks by relying on the constructed deep convolutional network. The ResNet feature extraction network proposed by He et al. [13] solves the phenomenon of gradient disappearance caused by the increase of the depth of the convolutional network, which lays the foundation for the current feature extraction network. In addition, two-stage detection such as Faster R-CNN designed by Ren et al. [14] extracts target region proposal frames by adding feature-shared RPN networks with little increase in complexity, which greatly improves the accuracy of the network object detection. The one-stage network structure RetinaNet proposed by Lin et al. [15] solves the difficulty of class imbalance in the one-stage structure network through the designed Focal Loss function, which makes the network architecture more lightweight and makes speed and accuracy more good balance.

Motivated by these advanced approaches, image feature extraction module can extract enough feature information from the input image of VQA. However, since only a few important regions of the image are required to answer the VQA question, complex object detection models may make the extracted feature information so redundant that resulting in waste of resources. Therefore, we introduced attention mechanism in our approaches.

2.2 Attention Mechanism

In the field of natural language processing, the attention mechanism was first proposed by Bahdanau et al. [16] in 2014. This method allows the model to focus the training on the relevant parts of the input data instead of the irrelevant parts, thereby improving the speed and quality of extracting key information. With the continuous development of deep learning, the field of computer vision also needs to build a neural network with an attention mechanism [16–22]. Through the attention mechanism, the neural network can increase the attention of key regions.

The visual attention mechanism is mostly formed by using masks intensively. The neural network focuses on the most critical parts of the picture by learning the key image features marked by the mask. Jaderberg et al. [17] proposed a network module of a spatial transformer through the attention mechanism in the spatial domain, which transformed the spatial information in the original image into another space and retained the key information. In the channel domain application, the SENet proposed by Hu et al. [18] increased the weight by using the image channel's contribution to key information during the convolution process.

Just recently, Wang et al. [22] innovatively proposed a model combining bottom-up and top-down attention mechanisms for Image Captioning and Visual Question Answering. The model based on top-down attention mechanism is used to learn the weights corresponding to features (generally using LSTMs [23]) to deeply understand the visual images. In other words, bottom-up attention is to extract some important regions of the picture, and each area is represented by a feature vector. Top-down attention is to determine the degree of contribution of features to the text, and then extract the features of the saliency regions that have a large contribution to the description. In addition, Huang et al. [24] proposed a non-local information statistical attention mechanism "self-attention" based on capturing the dependencies between long-distance features. Motivated by these works, we added this attention mechanism to the structure we designed to make the model focus more on the objects in the graph.

3 Method

In this section, we first introduce the residual network module, and then describe in detail the multi-scale convolution network based on weight sharing to which this technology is applied. In addition, we also added the visual attention mechanism to the feature extraction to obtain a new feature extraction network for VQA. Finally, we also proposed a WMA network-based target detection model WMA R-CNN and introduced its construction in detail. For the image feature extraction task of VQA, we define the problem as follows:

$$I = \{i_1, i_2, \ldots, i_m\} \tag{1}$$

where I is a given input image and m is the number of pixels in this image. And feature extraction network outputs the image feature result $\hat{O} = \{\hat{O}_1, \hat{O}_2, \ldots, \hat{O}_N\}$ after convolution pooling according to the input image I, where n is the number of points on the image divided into different subsets by the feature extraction network. And we define the features as follows:

$$r_k = i_j^k j \in \{1,2,\ldots,m\}, k \in \{1,2,\ldots,n\} \tag{2}$$

where r_k is the image feature set extracted by the feature extraction network. Our ultimate target is to enable the model to deal with objects of different sizes while ensuring that the model is lightweight. And it can be sensitive to important objects in the image, so as to avoid wasting computing resources.

3.1 Multi-Scale Convolutional Module

We always hope that the network can obtain the most accurate features with the least resources when use feature extraction networks in object detection tasks. From the perspective of the input image, we hope to reduce the impact of the background of the image during feature extraction and classify objects of different sizes without causing gradient explosions. Not only the input image has a series of factors such as uneven illumination intensity, complex image background and high noise interference in the process of actual image feature extraction, but the feature extraction network cannot satisfy the visual question answering system in agility and efficiency requirements. We propose a new method for image feature extraction. This method cannot only perform deeper feature extraction on the image at different data granularities, but also perform detailed analysis on the image at the microscopic level to exclude various interference factors, so that key information in the image can be efficiently extracted.

As shown in Fig. 1, we set up convolution kernels with receptive fields of sizes 1 × 1, 3 × 3 and 5 × 5 respectively by analyzing the features of the input image. The network can obtain detailed characteristics of objects of different sizes by setting three parallel branches of convolution kernels of different sizes. In addition, we also split the 3 × 3 and 5 × 5 convolution kernels into 1 × 3 and 3 × 1, 1 × 5 and 5 × 1 convolution kernels. Convolution kernels of size 1 × n and n × 1 have the same receptive field as n × n, and the former will have fewer parameters. Therefore, this architecture will further enhance the agility of our network.

We add dilated convolutions of sizes 1, 2, and 3 to the convolution kernels of 1 × 1, 3 × 3 and 5 × 5. It can control the range of the receptive field of the network through different expansion rates. We assume that the current feature map operation step is s, the dilated convolution of the convolutional layer with an expansion rate of d can increase the receptive field range by $2 \cdot (d-1) \times s$ times. Correspondingly, the dilated convolution of the n layers expansion rate of d can be Increase the receptive field range by $(d-1) \times s \times n$ times. Therefore, the sensitivity of feature extraction network for different sizes of objects is enhanced by expanding convolution, and more comprehensive feature information is captured.

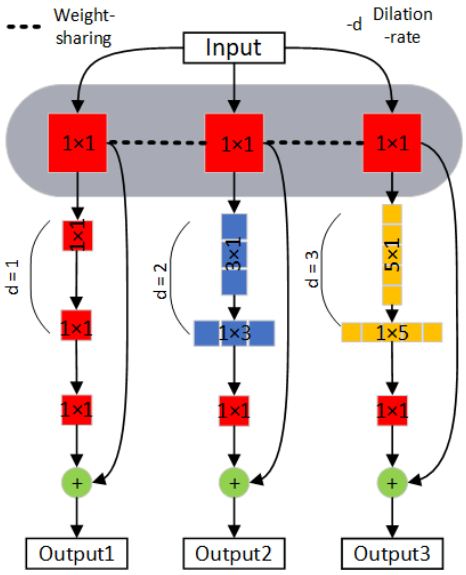

Figure 1: The flowchart of multi-scale convolutional network based on weight sharing

3.2 Multi-Scale Self-Attention Convolutional Network Based on Weight-Sharing (WMA)

Our designed multi-scale lightweight feature extraction network based on weight sharing is an image feature extraction algorithm designed based on convolutional neural networks. It ensures that the number of network parameters is not too high by combining the weight-sharing structure and the multi-scale convolutional layer. However, the reduction in the number of parameters will also make the image features extracted by the network inaccurate. In order to solve this problem, we add the improved visual attention mechanism to extract the key information features in the image, so that the network can focus on the information features of the required objects like the human brain, and then extract more critical feature information. In addition, self-attention directly calculates the relationship between any two pixels in the image to obtain the global geometric features of the image in one step. And it allows the model to learn the dependencies better between global features. The input of the WMA feature extraction network is $I = \{i_1, i_2, \ldots, i_m\}$, and the three outputs after the multi-scale convolution network are $x_{1\times1} \in R^{C\times N}$, $x_{3\times3} \in R^{C\times N}$, $x_{5\times5} \in R^{C\times N}$. Then these outputs are sent to the self-attention branch, and finally the results are accumulated to obtain the output of the WMA network $\hat{O} = \{\hat{O}_1, \hat{O}_2, \ldots, \hat{O}_N\} \in R^{C\times N}$. The $\hat{O}_i$ can be calculated as follows:

$$\hat{O}_i = \sum_{t=1}^{\tilde{C}} \lambda_i O_{t\times t,i} + x_{t\times t,i} \tag{3}$$

where $\tilde{C}$ is the number of branches of the multi-scale convolutional network and $\lambda_i$is a learnable scaler.

4 Experiment

We mainly use the PASCAL VOC 2012 [25] datasets to evaluate the individual object detection capabilities of our designed model, and evaluate the feature extraction performance of our designed model on the VQA system by using the VQA 1.0 and VQA 2.0 datasets [26]. The testing environment was conducted on one single Nvidia GTX 2080Ti graphic card with 16 GB memory,and an Intel(R) Core(TM) i7-7700 3.60 GHz.The ablation experiments of the model are shown in Table 1.

As shown in Table 1 and Fig. 2, the training effects of network models with different configurations are different. Pre-train represents the basic feature extraction network used by each network, multi-branch represents whether the model uses a multi-scale feature extraction network and what form of multi-scale feature extraction network is used. Weight-sharing represents whether weight sharing is adopted mechanism, Attention represents whether the model contains a self-attention mechanism. Baseline is used

as the baseline control network. This network uses VGG16 as the basic feature extraction network without multi-scale feature extraction network and weight sharing mechanism, and its final mAP is 65.7%. In addition, seven other models are respectively comparative experiments set up to detect the Pre-train, Multi-branch, Weight-sharing mechanism and Attention mechanism. When we compare Model 3 and Model 6, it can be seen that the model can obviously learn more image information features when ResNet152 is adopted. Compared with Model 2 and Model 3 or Model 5 and Model 6, we can find that the model using the weight-sharing mechanism can achieve better results. Because this mechanism shares the weight parameters learned between different branches in the multi-scale feature extraction network, which also allows the model to learn more image feature information in the limited parameters. And it also makes the detection effect improved. When comparing Model 1 and Model 2 or Model 4 and Model 5, we find that the model is not sensitive to the feature information of different sizes of objects without applying 5 × 5 branches. On the whole, network with ResNet is better than VGG in feature extraction. Comparing Model 6 and Model 7, we can find that the self-attention mechanism can also improve the detection performance of the model. Because the attention mechanism can make the model pay more attention to the key regions in the image and reduce unnecessary waste of resources to improve the detection effect. The overall detection score mAP on the public Pascal VOC 2012 dataset reached 73.6%, which also confirms the robustness of our proposed image feature extraction method WMA.

Table 1: Ablation study of our WMA R-CNN model on Pascal VOC 2012

Model	Pre-train	Multi-branch	Weight-sharing	Self-Attention	mAP
Baseline	VGG16	-	-	-	65.7
Model 1	Resnet101	{1,3}			68.1
Model 2	Resnet101	{1,3,5}			69.3
Model 3	Resnet101	{1,3,5}	√		71.0
Model 4	Resnet152	{1,3}			70.4
Model 5	Resnet152	{1,3,5}			71.6
Model 6	Resnet152	{1,3,5}	√		72.3
Model 7	**Resnet152**	**{1,3,5}**	√	√	**73.6**

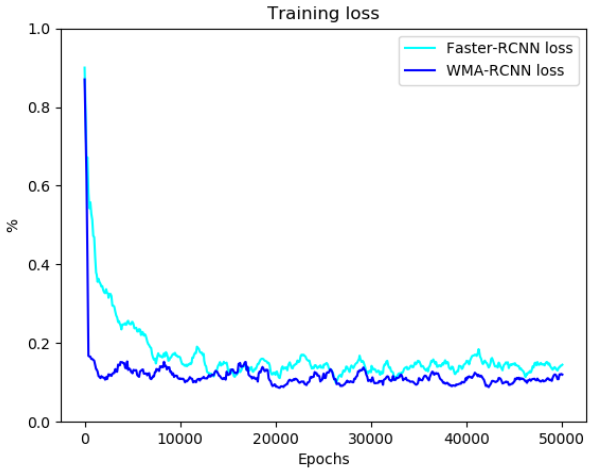

Figure 2: The training loss of our WMA network on Pascal VOC 2012

In order to further measure the performance of our model, we compare the VQA model with WMA technology to the most widely used VQA models in recent years. As shown in Table 2, we performed answer sampling on the VQA 2.0 dataset. Because each image-question pair is annotated with 10 standard answers in the VQA dataset, for each question we only keep the answers that appear more than three

times to enhance the validity of the dataset. In addition, we also used the Visual Genome dataset as an extended dataset to train our model. This dataset is three times the size of the model training set. The addition of this dataset greatly enhances the richness of our dataset. Table 2 shows that the bilinear models MCB and MLB have strong advantages compared with other simple VQA models. The accuracy of the VQA model increased by 1.09% and 1.12% respectively when we replaced the feature extraction modules in MCB and MLB with the WMA network. This also shows that the WMA network we designed has the ability to extract fine-grained features, and can focus on key regions in the image. Therefore, it can lay a good foundation for the subsequent feature fusion step and improve the overall efficiency of the VQA model.

Table 2: Comparison of performance of VQA model using our WMA module with the state-of-the-art methods on VQA 2.0. All of these results are obtained by using single model

Model	Test-dev			
	All	Y/N	No.	Other
Ask Your Neur [27]	58.39	78.39	36.45	46.28
SAN [28]	58.7	79.3	36.6	46.1
D-NMN [29]	59.4	81.1	38.6	45.5
MRN [30]	61.68	82.28	38.82	49.25
HieCoAtt [31]	61.8	79.7	38.7	51.7
MCB	66.7	83.4	39.8	58.5
MLB [32]	66.77	84.57	39.21	57.81
WMA with MCB	67.79	83.94	40.67	59.98
WMA with MLB	67.89	85.15	39.85	58.32

Figure 3: VQA examples from VQA 1.0 dataset

Typical examples are shown in Fig. 3. In each example, the left is the original picture, the right is the heat map generated by the intermediate steps of the WMA model, and the bottom is the question-and-answer situation. The VQA model using Faster R-CNN does not correctly identify the number of birds standing on the wall in Fig. 3.

5 Conclusion

In this paper, we design a multi-scale self-attention object detection model based on weight-sharing for VQA (WMA), which can extract fine-grained features of different sizes objects and focus on key regions in the image. So we can improve the overall accuracy of the VQA system. In addition, our designed WMA can also be applied to other object detection tasks as a general feature extraction network. Our experimental results on Pascal VOC 2012, VQA 1.0 and VQA 2.0 datasets demonstrate the effectiveness and robustness of WMA module.

Funding Statement: This work is supported by the National Natural Science Foundation of China (61872231, 61701297).

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] S. Antol, A. Agrawal, J. Lu, M. Mitchell and D. Parikh, "VQA: Visual question answering," *International Journal of Computer Vision*, vol. 123, no. 1, pp. 4–31, 2017.

[2] F. Wu, Z. Yu, Y. Yang, S. Tang, Y. Zhang *et al.,* "Sparse multimodal hashing," *IEEE Transactions on Multimedia*, vol. 16, no. 2, pp. 427–439, 2014.

[3] Z. Yu, F. Wu, Y. Yang, Q. Tian, J. Luo *et al.,* "Discriminative coupled dictionary hashing for fast cross-media retrieval," in *Proc. of the 37th Int. ACM SIGIR Conf. on Research & Development in Information Retrieval*, pp. 395–404, 2014.

[4] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan *et al.,* "Long-term recurrent convolutional networks for visual recognition and description," in *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 2625–2634, 2015.

[5] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville *et al.,* "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. of the 32nd Int. Conf. on Machine Learning*, vol. 14, pp. 2048–2057, 2015.

[6] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson *et al., "*Bottom-up and top-down attention for image captioning and visual question answering," arXiv:1707.07998, 2018.

[7] S. Ren, K. He, R. Girshick and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2017.

[8] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed *et al.,* "SSD: Single shot multibox detector," *European Conference on Computer Vision*, pp. 21–37, 2016.

[9] T. Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan *et al.,* "Feature pyramid networks for object detection," *Conference on Computer Vision and Pattern Recognition*, pp. 936–944, 2016.

[10] A. Wulam, Y. Wang, D. Zhang, J. Sang and A. Yang, "A recommendation system based on fusing boosting model and DNN model," *Computers, Materials & Continua*, vol. 58, no. 2, pp. 1003–1013, 2019.

[11] B. Gu, W. Xiong and Z. Bai, "Human action recognition based on supervised class-specific dictionary learning with deep convolutional neural network features," *Computers, Materials & Continua*, vol. 62, no. 3, pp. 243–262, 2020.

[12] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *Computer Science*, 2014.

[13] K. He, X. Y. Zhang, S. Q. Ren and J. Sun, "Deep residual learning for image recognition," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

[14] S. Ren, K. He, R. Girshick and S. Jian, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 39, no. 6, pp. 137–1149, 2017.

[15] T. Y. Lin, P. Goyal, R. Girshick, K. M. He and P. Dollár, "Focal loss for dense object detection," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, pp. 2999–3007, 2017.

[16] D. Bahdanau, K. Cho and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *Computer Science*, 2014.

[17] M. Jaderberg, K. Simonyan, A. Zisserman, K. Kavukcuoglu, "Spatial transformer networks," *Advances in Neural Information Processing Systems*, pp. 2017–2025, 2015.

[18] J. Hu, L. Shen and G. Sun, "Squeeze-and-excitation networks," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 7132–7141, 2018.

[19] L. Itti, C. Koch, W. Way and L. Angeles, "Computational modelling of visual attention," *Nature Reviews Neuroscience*, vol. 2, no. 3, pp. 194, 2001.

[20] Q. S. Zhang, Y. N. Wu and S. C. Zhu, "Interpretable convolutional neural networks," arXiv:1710.00935, 2017.

[21] V. Mnih, N. Heess, A. Graves and K. Kavukcuoglu, "Recurrent models of visual attention," *Advances in Neural Information Processing Systems*, 2014.

[22] F. Wang, M. Q. Jiang, C. Qian, S. Yang, C. Li *et al.,* "Residual attention network for image classification," arXiv:1704.06904, 2017.

[23] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[24] G. Huang, Z. Liu, V. D. M. Laurens and K. Q. Weinberger, "Densely connected convolutional networks," *IEEE Computer Society*, 2016.

[25] T. Y. Lin, M. Maire, S. Belongie, J. Hays and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *European Conf. on Computer Vision*, pp. 740–755, 2014.

[26] Y. Goyal, T. Khot, D. S. Stay, D. Batra and D. Parikh, "Making the V in VQA matter: Elevating the role of image understanding in visual question answering," *IEEE Computer Society,* 2016.

[27] M. Malinowski, M. Rohrbach and M. Fritz, "Ask your neurons: A deep learning approach to visual question answering," *International Journal of Computer Vision,* vol. 125, no. 1, pp. 110–135, 2017.

[28] Z. Yang, X. He, J. Gao, L. Deng and A. Smola, "Stacked attention networks for image question answering," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition,* pp. 21–29, 2016.

[29] J. Andreas, M. Rohrbach, T. Darrell and D. Klein, "Learning to compose neural networks for question answering," in *Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2016.

[30] J. H. Kim, S. W. Lee, D. Kwak, M. O. Heo, J. Kim *et al.,* "Multimodal residual learning for visual QA," *Advances in Neural Information Processing Systems*, pp. 361–369, 2016.

[31] J. Lu, J. Yang, D. Batra and D. Parikh, "Hierarchical question-image co-attention for visual question answering," *Advances in Neural Information Processing Systems*, pp. 289–297, 2016.

[32] J. H. Kim, K. W. On, J. Lim, J. W. Ha and B. T. Zhang, "Hadamard product for low-rank bilinear pooling," arXiv:1610.04325, 2016.