Chen et al. EURASIP Journal on Wireless Communications and Networking (2019) 2019:125
https://doi.org/10.1186/s13638-019-1444-y

EURASIP Journal on Wireless Communications and Networking

RESEARCH Open Access

Synthetical application of multi-feature map detection and multi-branch convolution

Jin Chen[1,2], Rong Liu[1,2], Ying Tong[1,2*] and Hanling Wu[3]

Abstract

Two methods for improving the detection performance of neural networks are introduced in this paper, multi-feature map detection and multi-branch convolution structure. The former is to analyze the features of each convolution layer in the network separately, because these features have different resolutions and correspond to objects of different sizes. Finally, the comprehensive judgment of the analysis results can give better consideration to the overall situation and improve the accuracy of detection. The multi-branch convolution structure uses convolutions of different sizes on multiple branches to process input in parallel, and these branches are independent of each other. Finally, the feature maps corresponding to different receptive fields from each branch are combined and analyzed comprehensively. In this paper, the application process of the above two methods is described in combination with classical neural networks, such as the single shot multibox detector (SSD) and receptive field block (RFB) net.

Keywords: Multi-branch convolution, Multi-feature map detection, Receptive field block (RFB) net , Single shot multibox detector (SSD)

1 Introduction

Neural network has been widely concerned since it was proposed. Based on the deepening understanding of the delicate working process of the human visual system, experts have developed many network versions whose performance has been continuously improved. After a convolutional neural network was proposed, it has entered a new stage. Improving network accuracy has always been a common concern [1]. The key point is to improve the speed of network operation and the precision of detection as much as possible, but often the two are not balanced. Supposing that we increase the depth and breadth of the network to improve the accuracy immoderately, then too much computation will seriously affect the detection speed of the network and vice versa. In view of this contradiction, scholars have proposed many methods to improve the performance of the network, including the multi-feature map detection and the multi-branch convolution structure.

The main idea of multi-feature map detection is to detect the feature maps obtained from convolution layers of different sizes separately and then judge all the detection results synthetically, so as to get the best output. SSD network is a typical case of multi-feature map detection, which is improved on the basis of VGG16 network. It has been found that it can improve the running speed of the network while ensuring the accuracy and has been widely used [2]. In the third part of this paper, we will introduce multi-feature map detection combined with the SSD Net. The multi-branch convolution structure uses different size convolution layers to detect the input synchronously and then aggregates and analyses the results obtained on each convolution branch. The typical network using this structure is the inception network, which will be introduced in detail in the fourth part of this paper.

RFB Net is a representative network which combines multi-feature map detection with multi-branch convolution structure. It was proposed by experts after being inspired by the relationship between the size and the eccentricity of the receptive field in the human retina. This network is composed of RFB modules with multi-branch convolution structure on the basis of SSD Net. It has been

* Correspondence: tongying2334@163.com
[1]Tianjin Key laboratory of Wireless Mobile Communications and Power Transmission, Tianjin Normal University, Tianjin 300387, China
[2]College of Electronic and Communication Engineering, Tianjin Normal University, Tianjin 300387, China
Full list of author information is available at the end of the article

Springer Open

© The Author(s). 2019 **Open Access** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

proved that it can improve the detection accuracy while maintaining the speed [3]. The fifth part of this paper will take RFB Net as an example to explain the comprehensive application process of multi-feature map detection and multi-branch convolution structure.

2 Base net—VGG16 network

Most of the neural networks using multi-feature graph detection method and multi-branch convolution structure mentioned in this paper take VGG16 as the basic network and then add an auxiliary network. Therefore, in this section, the VVG16 network is introduced briefly at first.

VGG16 network is a very deep convolutional network, which was proposed by *Karen Simonyan* and *Andrew Zisserman*. Initially, they wanted to verify whether the depth of the network was proportional to its accuracy. Continuous testing eventually led to the realization of the VGG16 network, and the answer was affirmative. VGG16 is a very classical network in the development of neural network. Many of the networks used for target recognition are improved on the basis of VGG16 network, such as SDD and the RFB network mentioned in this paper. Most of the basic networks used in the early stage of these networks are based on VGG16.

As the name implies, the network structure of VGG16 has 16 layers, including 13 convolution layers and 3 fully connected layers. As shown in Fig. 1, the front end of the whole network is a combination of convolution layer and pooling layer, and the end is three fully connection layers in turn. All the convolution layers adopt a convolution core of 3×3 size instead of a convolution core of gradually increasing size. In this way, the number of parameters can be reduced, so that the computation of the network can be reduced and the detection speed can be improved. Why do experts take this approach? As we all know, convolution network extracts some features of an image by convolution operation of input image and convolution core. The convolution core here is equivalent to a filter, and the image features filtered by different filters are widely divergent. Therefore, the parameter setting of convolution core is very important, and the number of parameters directly determines the calculation of the whole network. For example, under the same conditions (assuming the proportion of channels is C), two connected 3×3 convolution layers need to initialize weights of $2 \times (3 \times 3) \times C = 18C$, while a 5×5 convolution layer needs to initialize weights of $5 \times 5 \times C = 25C$, which increases by 39%. Similarly, it can be calculated that a 7×7 convolution layer needs weights of 81% more than three connected 3×3 convolution layers. Therefore, in VGG16, two cascades of 3×3 convolution layers are used instead of 5×5, because they cannot only achieve the same effect but also reduce the computational complexity. Similarly, the 7×7 convolution layer can be replaced by three 3×3 convolution cascades.

Therefore, not every convolution layer is followed by a pooling layer. As shown in Fig. 1, there are five pooling layers in VGG16 network. Each pooling layer is associated with its front-end convolution heap. For example, the combination of C1, C2, and P1 is equivalent to a 5×5 convolution core connected to a pooling layer, or C5, C6, and C7 combined with P3 achieve the effect of the combination of a 7×7 convolution layer and a pooling layer. From this point of view, VGG16 can also be regarded as a neural network with an increasing convolution core, but its computational complexity is greatly reduced.

All in all, VGG16 is a very classical deep convolution network, which has been widely used in the field of visual representation. The next few networks are improved on the basis of VGG16. If you want to get the training and testing process of VGG16 network, please refer to [4].

3 Multi-feature map detection

Through the validation of experts, the detection accuracy of the network can be improved by expanding the depth

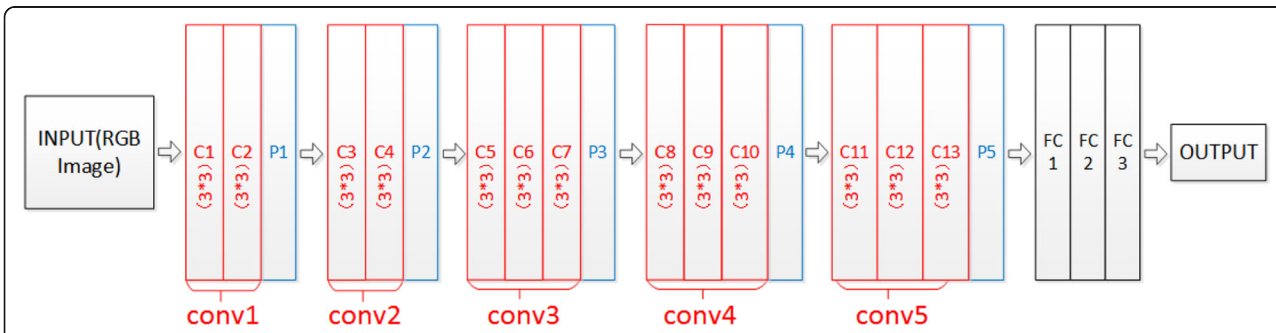

Fig. 1 Network structure of VGG16. Among them, C1–C13 represents convolution layer, P1–P5 represents pooling layer, and FC1–FC3 represents full connection layer. C1 and C2 as a whole can be regarded as Conv1 of 5×5 convolution layer, C5; C6, and C7 as a whole as Conv3 of 7×7 convolution layer; and so on. There are some hidden layers in the network structure, which will not be discussed here. Please refer to [4] for all information

of the network, that is, increasing the number of convolution layers. In traditional deep convolution networks, the feature map obtained from the last convolution layer is analyzed. In some cases, some details will be omitted, which will affect the final detection accuracy. Multi-feature map detection can effectively avoid this problem. Its main idea is to analyze the features of each convolution layer in the network separately, and then to determine the best output of all the results obtained. This method detects the feature maps corresponding to different receptive fields separately, and obtains the results with different resolutions. It can take into account the details when processing the whole image, so that smaller objects can also be detected well, thus improving the detection accuracy of whole network.

SSD network is a typical convolution network using multi-feature map detection, which is improved on the basis of VGG16 network. Neural network framework can be divided into two kinds: two-stage detection and one-stage detection. The former includes region proposal, which reduces the detection speed, while the latter omits this step and processes the whole input image directly, so the network running speed is naturally improved. SSD belongs to the latter, that is, it only uses a single deep convolution network. At present, the detection accuracy of perfect two-stage convolution network is very high, but its speed is not very ideal, while SSD network can significantly improve the detection speed on the basis of maintaining the same accuracy, so it can be used in real-time detection field.

Figure 2 shows the basic structure of SSD network, which is a typical network using multi-feature map detection. It can be divided into two parts: basic network and auxiliary network. Among them, the basic network adopts the general structure of VGG16 network, only on the basis of which some changes have been made. The original 13 convolution layers have not changed: Conv1_1–Conv5_3 in Fig. 2 correspond to C1–C13 in Fig. 1, respectively; the main changes occur in the full connection layer. The first two full connection layers are transformed into convolution layers, that is, from FC1 and FC2 in Fig. 1 to Conv6 and Conv7 in Fig. 2, while the last full connection layer (FC3 in Fig. 1) is deleted. Following the basic network is the auxiliary network, which is the most distinctive part of SSD.

As can be seen from Fig. 2, the auxiliary network is a series of convolution layers, and the size of the layers is gradually reduced. In this way, the receptive field corresponding to each node in the feature map after each convolution naturally decreases. Therefore, many feature maps with different proportions can be obtained from the auxiliary network. If these feature maps are detected by convolution separately, the results will be different in resolution [5]. Finally, these results are superimposed and screened to get the best one, which is the main idea of multi-feature map detection.

Among other neural networks such as VGG16, feature detection is usually carried out on the feature map obtained at the last layer of the network, while SSD detects several feature maps generated during the operation of the network separately. As shown in Fig. 2, SSD uses convolution layers of different sizes to detect the feature maps of Conv4_3, Conv7, Conv8_2, Conv9_2, Conv10_2, and Conv11_2, respectively. The final result is obtained by summing up the results obtained at each level and judging them. A series of fixed default boxes are pre-correlated on each feature map. After the detection of the convolution layer (i.e. the top convolution in Fig. 2), the default box

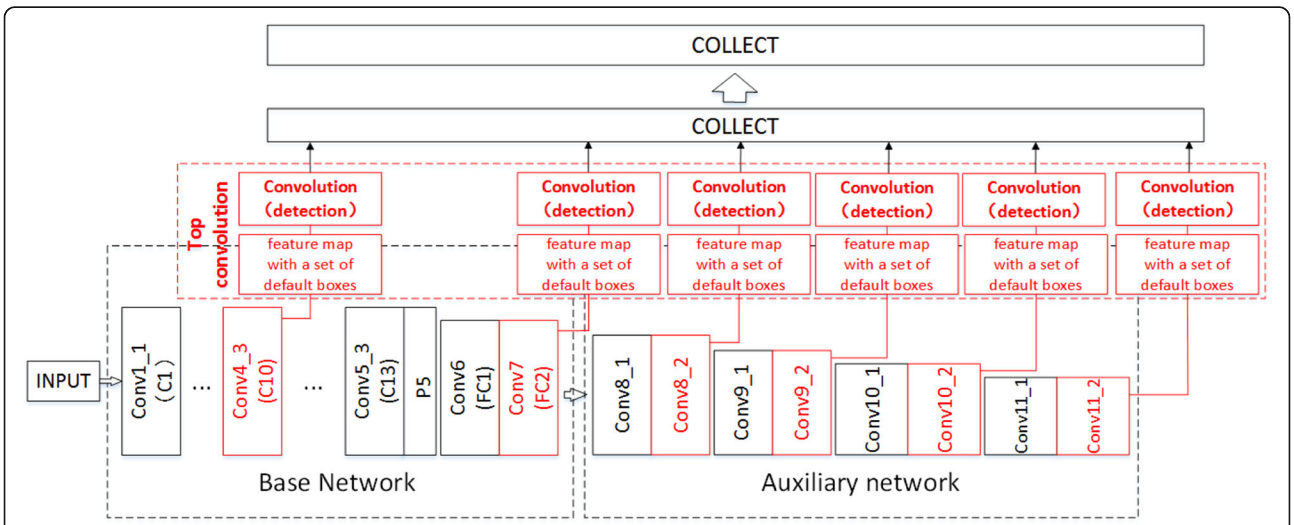

Fig. 2 The network structure of SSD. It adds auxiliary network and top convolution layer to the basic network improved by VGG16 and detects the features of different sizes of feature maps, which improves the accuracy of the network

closest to the shape of the object is selected and adjusted to output as the object's boundary box. At the same time, the probability that the object exists in the boundary box and the offset of the boundary box from the original default box will be the output [6]. Therefore, the output from the top convolution is a series of boundary boxes. After summarizing these boundary boxes, the detection results closest to the object can be selected through non-maximum suppression [7]. It can be seen that there are many top convolution operations in SSD, that is, the method of multi-feature map detection is adopted. This is because the feature maps obtained by different convolution kernels correspond to different receptive fields, so the detection results at the same location in different feature maps are not identical. The bigger the size of feature map, the better the processing of the whole image, but the details cannot be well reflected. Therefore, using different sizes of feature map to detect synchronously, we can take into account the size differences of different objects in a picture, so as to achieve a better detection effect.

Through the introduction of SSD network, we can understand the main process of multi-feature map detection more concretely, which has a very positive impact on improving the accuracy of the network, proving that it is a very effective detection method. For more specific training and testing process about SSD, please refer to [2].

4 Multi-branch convolution

The main idea of multi-branch convolution structure is to use convolutions of different sizes to process input in parallel. There are many convolution branches in the structure, and the convolution size on each branch is different. When the network runs, each branch will convolute with the input map independently at the same time and finally aggregate the results of each branch. Next, an example is given to illustrate the operation process of the structure.

The inception module in the inception network is a typical representative of multi-branch convolution structure. The inception network composed of the inception module has achieved good results in the 2014 ILSVRC

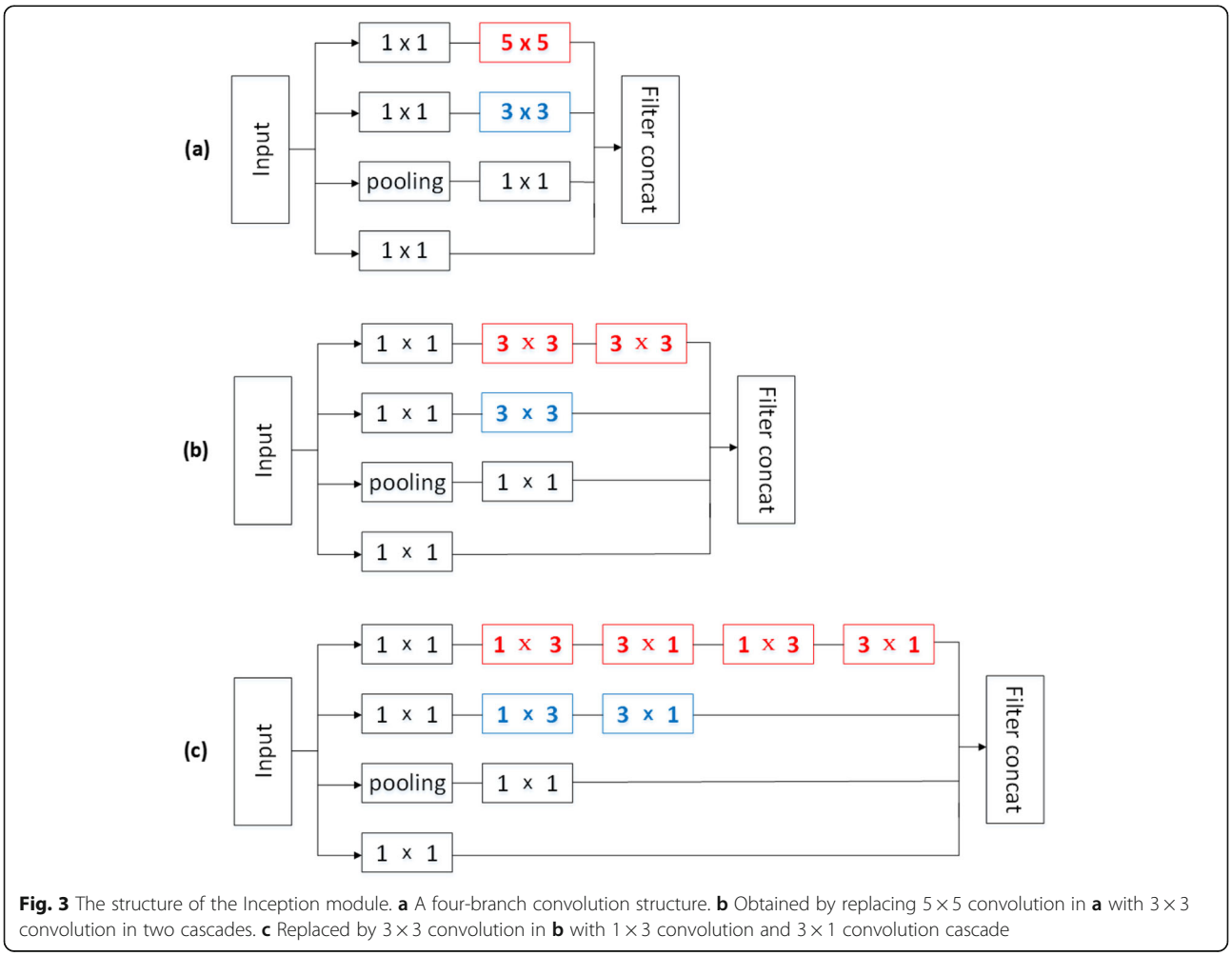

Fig. 3 The structure of the Inception module. **a** A four-branch convolution structure. **b** Obtained by replacing 5 × 5 convolution in **a** with 3 × 3 convolution in two cascades. **c** Replaced by 3 × 3 convolution in **b** with 1 × 3 convolution and 3 × 1 convolution cascade

(Large Scale Visual Recognition Challenge), whose most prominent feature is the use of the inception module. The basic structure of the inception module is shown in Fig. 3. It can be seen that it is a multi-branch convolution structure, and each module consists of four branches. The inception module initially proposed is shown in Fig. 3a, which contains four branches: a 1×1 convolution branch, a 1×1 convolution plus 3×3 convolution branch, a 1×1 convolution plus 5×5 convolution branch, and a pooling layer plus 1×1 convolution branch. The input feature map is processed by four branches and the results are jointly entered into the filter combiner to get the final output.

It can be noted that 1×1 convolution is basically used in each branch of the inception module. The purpose of this is to reduce the number of channels in the feature map, that is, to reduce the dimension. In the "Base net—VGG16 network" section, we point out that in order to reduce the computational complexity of VGG16 network, two cascaded 3×3 convolutions can be used instead of one 5×5 convolution and three cascaded 3×3 convolutions can be used instead of one 7×7 convolution. The inception module also uses this method to improve the network, so as to get the structure shown in Fig. 3b, which replaces the 5×5 convolution layer with two cascaded 3×3 convolutions [8]. Figure 3c is improved on the basis of Fig. 3b, in which 1×3 convolution and 3×1 convolution cascade are used to replace 3×3 convolution layer. Assuming that for any $n \times n$ convolution layer, $1 \times n$ convolution and $n \times 1$ convolution cascade are used instead, it can be found that the computational complexity is reduced by $(n \times n - 2 \times n)/n \times n$ times. It can be concluded that when the value of n is greater than 2, the total amount of calculation will be reduced, which is very important for the network. Therefore, in the Inception Module shown in Fig. 3c, all $n \times n$ convolution layers are replaced by $1 \times n$ convolution and $n \times 1$ convolution cascade, thus reducing the computational complexity. For training and testing details of inception network, please refer to [9, 10].

5 The Synthetical application

The main ideas and application examples of multi-feature graph detection and multi-branch convolution structure are explained in the above two parts. It can be seen that these two methods are very helpful to improve network performance. At present, the idea of applying the above two methods to neural networks has been realized, and RFB Net is one of them. Practice has proved that the network performance has been better optimized, that is, it can maintain high-speed detection while improving accuracy.

RFB Net is a new convolution neural network proposed by Liu et al [3]. It is improved by SSD network, which mainly simulates the relationship between the size and eccentricity of receptive fields of cells in human visual system. As mentioned earlier, researchers have found that a series of cells in the visual cortex are associated with different size areas of the retina, called the receptive field of the cells. When the receptive field is stimulated, the activity of the related cells will be activated. Through continuous research, scholars have found that the size of receptive field is proportional to the eccentricity, and tried to embody this rule in the neural network, resulting in RFB Net. After testing, it has been proved that RFB Net can improve the operation accuracy while maintaining the detection speed.

RFB Net integrates multi-branch convolution module and multi-feature feature map detection, that is, it adds multi-branch convolution module similar to the inception network on the basis of retaining multi-feature map detection of SSD network, as shown in Fig. 4. By

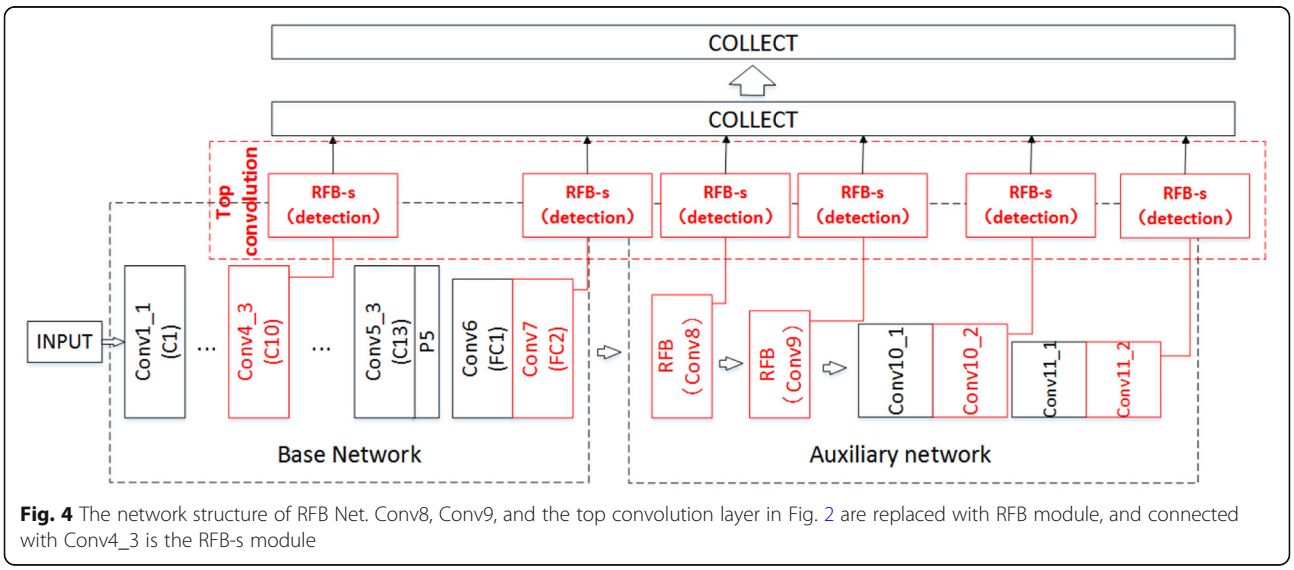

Fig. 4 The network structure of RFB Net. Conv8, Conv9, and the top convolution layer in Fig. 2 are replaced with RFB module, and connected with Conv4_3 is the RFB-s module

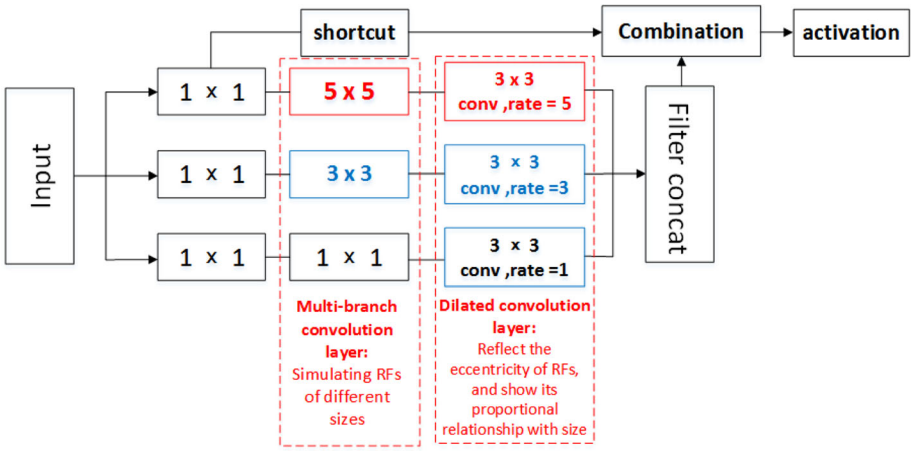

Fig. 5 The structure diagram of the RFB module. Similar to Fig. 3a, the Dilated convolution layer is added on the basis of multi-branch convolution to simulate the proportional relationship between the size and the centrifugal rate of the RFs in the human retina

comparing with Fig. 2, we can find that the RFB module in RFB Net replaces a part of the front-end convolution layer and the top convolution detection layer in SSD network, so the RFB module is the most important part. Figure 5 shows the basic structure of the RFB module. Through observation, it can be found that the structure of the RFB module is very similar to the inception module mentioned in the "Multi-branch convolution" section, which is also a multi-branch convolution structure. However, compared with the inception module, each branch end of the RFB module adds a convolution layer of different rate to become the dilated convolution layer. This is because the size of convolution layer at the end of the branch in inception module is constantly expanding, which only corresponds to the different size of receptive field in human brain and cannot reflect the proportional relationship with eccentricity. Therefore, it is considered that all branches correspond to the same centrifugal rate, that is, all branches are detected in the same center. The relationship between the size of RFs and the centrifugal rate can be achieved by adding different rates of dilated convolution layer in the RFB module. That is, the multi-branch convolution layer with different sizes is used to simulate the RFs of multiple sizes in the human visual system, while the dilated convolution layer immediately following reflects the eccentricity. In addition, shortcut is also added to the RFB module, which is not discussed here. See [11] for details.

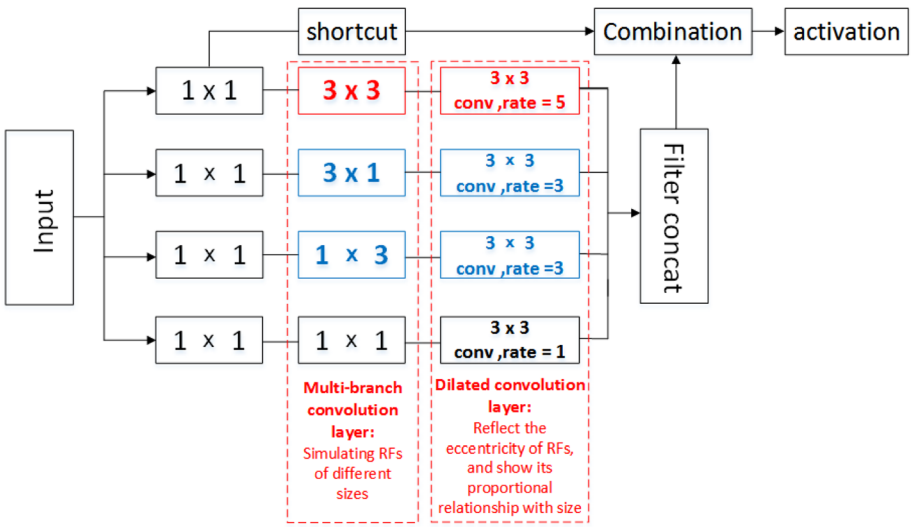

Fig. 6 The structure diagram of RFB-s module. Similar to Fig. 3b and Fig. 3c, two cascaded 3×3 convolutions are used to replace the 5×5 convolution in Fig. 5, and 1×3 convolutions and 3×1 convolution cascades are used to replace the 3×3 convolution layer in Fig. 5. The computational complexity is reduced while the smaller RFs in the shallow retinal map are simulated

As shown in Fig. 6, the basic idea of the improved RFB-s module based on the basic RFB module is still to use small convolution layer cascade instead of large convolution, thus reducing the computational complexity. As mentioned in the section 4, two cascaded 3×3 convolutions are used instead of a 5×5 convolution, and 1×3 convolutions and 3×1 convolution cascades are used instead of 3×3 convolutions, so the smaller RFs in the shallow human retinal mapping are simulated here. As shown in Fig. 4, the RFB-s module is used to detect the feature map obtained by conv4–3.

After introducing the RFB module, we can have a good understanding of RFB Net. By comparing Fig. 4 with Fig. 2, it can be found that the network structure of RFB Net basically follows the main framework of SSD network and still adopts the structure of basic network plus auxiliary network. The main changes occur in the following: (1) RFB Net replaces Conv8 and Conv9 in SSD network with RFB module and (2) all the top convolution layers in SSD are replaced by RFB modules, in which the feature map generated by Conv4_3 is detected by RFB-s modules.

Current testing results have proved that RFB Net can achieve the same accuracy as other detectors while maintaining real-time detection speed, or even better. For specific training and testing process, please refer to [3].

6 Results

Combined with the network models of SSD Net, Inception Network, and FRB Net, we can have a more intuitive understanding of the practical application process of multi-feature map detection and multi-branch convolution structure, and we can also see that it plays a very important role in improving the detection performance of the network. These two methods emphasize the detection of feature maps with different resolutions and the parallel operation of convolutions with different sizes, which is also of great reference significance for the design of subsequent network models.

7 Discussion

This paper mainly introduces the synthetical application process of multi-branch convolution module and the method of multi-feature map detection, which has gone through a long period. First is the formation of the basic network vgg16, which is the basis of many subsequent detection networks, playing a very important role. Subsequently, the main ideas of multi-feature map detection and multi-branch convolution and their application examples, including SSD Net and Inception network, are introduced. Among them, the SSD Net adopts the method of multi-feature map detection, taking into account the different sizes in the whole world, which not only improves the detection accuracy, but also ensures the accuracy. Inception module is a typical multi-branch convolution module, and its Inception network fully embodies the advantages of combining different receptive fields for detection. In recent years, there have been some networks which integrate multi-feature map detection and multi-branch convolution structures, such as the RFB Net. This network simulates the proportional relationship between the size of different receptive fields and the eccentricity in the human visual cortex, improving the operation accuracy while guaranteeing the network speed, and it has been proved and applied.

Abbreviations

ILSVRC: Large Scale Visual Recognition Challenge; RF: Receptive field; RFBNet: Receptive field block net; SSD: Single shot multibox detector; VGG: Visual geometry group

Funding

This work has been partly supported by the National Natural Science Foundation of China (Grant No: 61701344), Tianjin Normal University Application Development Foundation (52XK1601), Tianjin Normal University Doctoral Foundation (52XB1603, 52XB1713), and Tianjin Higher Education Creative Team Funds Program in China.

Availability of data and materials

Data sharing not applicable to this article as no datasets were generated or analyzed during the current study.

Authors' contributions

JC gives the overall research direction and ideas and helped to draft the manuscript. RL read the relevant literature and books and drafted the article. YT also gives the original ideas and research direction and helped to draft the manuscript. HW participated in its design and coordination. All authors read and approved the final manuscript.

Authors' information

Jin Chen was born in Wuhu, China, in 1976. He received an M.S. degree from Tianjin Normal University and a Ph.D. degree from Tianjin University, in 2005 and 2013, respectively. Since 2005, he has been working at Tianjin Normal University in China. He is an associate professor of Tianjin Key Laboratory of Wireless Mobile Communications and Power Transmission. His research interests include image and acoustic signal acquisition and processing, broadband sensor array signal processing, and artificial intelligence.

Rong Liu was born in ShanXi, China, in 1994. She received a B.S. degree from Tianjin University of Technology and Education in 2016. She is currently working toward an M.S. degree of Tianjin Normal University. Her research interests include software programming and artificial intelligence.

Ying Tong was born in Tianjin, China, in 1982. She received a B.S. and M.S. degree from Tianjin Normal University, and a Ph. D degree from Tianjin University in 2004, 2007, and 2015, respectively. Since 2007, she has been working at Tianjin Normal University in China. She is a lecturer of Tianjin Key Laboratory of Wireless Mobile Communications and Power Transmission. Her research interests include computer vision and digital signal processing.

Hanling Wu was born in Anqing, China, in 1979. He received a B.S. degree from Anhui Normal University and a Ph.D. degree from Institute of Mechanics, Chinese Academy of Sciences, in 2002 and 2009, respectively. Since 2009, he has been working at Beijing Institute of Astronautical Systems Engineering in China. Now his research interests include system design of flight vehicle, acoustic signal acquisition and processing, and artificial intelligence as a senior engineer.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details
[1]Tianjin Key laboratory of Wireless Mobile Communications and Power Transmission, Tianjin Normal University, Tianjin 300387, China. [2]College of Electronic and Communication Engineering, Tianjin Normal University, Tianjin 300387, China. [3]Beijing Institute of Astronautical Systems Engineering, Beijing 100076, China.

Received: 26 January 2019 Accepted: 18 April 2019
Published online: 22 May 2019

References

1. A. Krizhevsky, I. Sutskever, G.E. Hinton, in *NIPS'12 Proceedings of the 25th International Conference on Neural Information Processing Systems, Lake Tahoe, Nevada*. ImageNet classification with deep convolutional neural networks (2012), pp. 1097–1105
2. W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Fu, A.C. Berg, in *ECCV 2016: computer vision – ECCV 2016*. SSD: Single shot multibox detector (2016), pp. 21–37
3. S. Liu, D. Huang, Y. Wang, in *ECCV 2018: Computer Vision – ECCV 2018*. Receptive field block net for accurate and fast object detection (2018), pp. 404–419
4. K. Simonyan, A. Zisserman, in *International Conference on Learning Representations 2015 (ICLR 2015), San Diego, CA*. Very deep convolutional networks for large-scale image recognition (2015)
5. Y. Li, S. Wang, Q. Tian, X. Ding, Feature representation for statistical-learning-based object detection: A review. Pattern Recogn. **48**(11), 3542–3559 (2015)
6. D.K. Guillaumin, V. Ferrari, ImageNet auto-annotation with segmentation propagation. Int. J. Comput. Vis. **110**(3), 328–348 (2014)
7. X. Zeng et al., Crafting GBD-net for object detection. IEEE Trans. Pattern Anal. Mach. Intell. **40**(9), 2109–2123 (2018)
8. K. Zhang, Z. Zhang, Z. Li, Y. Qiao, Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Processing Letters **23**(10), 1499–1503 (2016)
9. C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA*. Rethinking the inception architecture for computer vision (2016)
10. C. Szegedy, S. Ioffe, V. Vanhoucke, A. Alemi, in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA*. Inception-v4, inception-ResNet and the impact of residual connections on learning (2016)
11. K. He, X. Zhang, S. Ren, J. Sun, in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016) , Las Vegas, NV, United States*. Deep residual learning for image recognition (2016), pp. 770–778

Submit your manuscript to a SpringerOpen journal and benefit from:

► Convenient online submission
► Rigorous peer review
► Open access: articles freely available online
► High visibility within the field
► Retaining the copyright to your article

Submit your next manuscript at ► springeropen.com