2021 20th International Symposium on Distributed Computing and Applications for Business Engineering and Science (DCABES)

An effective lightweight attention mechanism

1st Zhipeng Liu
School of Artificial Intelligence and Computer Science
Jiangnan University
Wuxi, China
952738565@qq.com

2nd Wei Fang
School of Artificial Intelligence and Computer Science
Jiangnan University
Wuxi, China
fangwei@jiangnan.edu.cn

3rd Jun Sun
School of Artificial Intelligence and Computer Science
Jiangnan University
Wuxi, China
junsun@jiangnan.edu.cn

***Abstract*—Aiming at the problem of large parameters and poor portability of attention mechanism modules, an extremely lightweight attention mechanism is designed, which uses a combination of spatial attention and channel attention to enhance the model's attention on important information. From the experimental results, the lightweight attention mechanism can add to the deep convolutional neural network which introduces with negligible parameters and calculations, allows the model to focus more on useful information and improve the accuracy of the object detection task. It surpasses SENet(Squeeze and Excitation Networks) in performance and is easily transplanted to mainstream deep convolutional neural networks such as ResNet.**

Keywords-Convolutional neural network; Object detection; Attention mechanism; Lightweight;

I. INTRODUCTION

In the context of the era of big data and artificial intelligence, the rapid development of deep learning has led to major developments in computer vision tasks. Object detection has also developed vigorously as a core task in the field of computer vision. However, due to the large scale of data, the complexity and variety of data sources, ever-changing environmental backgrounds, and limited computing resources, target detection tasks still face many challenges, and they are always struggling on the road to improve detection accuracy and detection speed.

Object detection [1] is one of the important tasks of computer vision, and it is considered to be the main pioneer in solving the problem of semantic understanding of the surrounding environment. Object detection is the basis for solving complex high-level vision tasks such as segmentation, scene understanding, target tracking, image description, event detection and activity recognition. Object detection has a wide range of applications in many fields of artificial intelligence and information technology, including robot vision, autonomous driving, human-computer interaction, content-based image retrieval, intelligent video surveillance, and augmented reality. In human vision, the attention mechanism is inseparable at all times, as a result of it exists all the time, it is often ignored. For example, when reading books and newspapers, people will pay attention to pictures and text information. At this time, information such as paper color and material will not be the focus. Humans will treat different information differently and pay more attention to valuable information.

Nowadays, deep learning [2] has developed rapidly. Computer vision is mostly based on neural networks [3]. Neural networks can self-learn and extract features from input pictures and other information. For these features, we cannot generalize and perform equivalent processing, because different features play different roles for reasoning. The degree of help is different, so it is particularly important for the neural network to construct the attention mechanism. Realize the neural network to focus on the effective area autonomously, reduce the attention to irrelevant information, and improve the discriminative ability of the network, which is the core of the attention mechanism in computer vision. The neural network can learn from itself, which provides the possibility for parameter learning of the attention structure. At the same time, the attention mechanism can allow humans to better understand the world expressed by the neural network, and provide a visual basis for network structure optimization and parameter adjustment.

In recent years, most of the research on the attention mechanism of computer vision is based on masks to generate attention structure modules. In the training process of neural networks, the required attention areas in the pictures are learned, the weights are updated, and the masks are generated. It is weighted with the original features to form an attention mechanism, which in turn allows the network to pay more attention to the features of discriminative regions. Mask-based attention mechanisms are divided into soft attention mechanisms and hard attention mechanisms. The soft attention mechanism pays more attention to the channel or spatial information, which can be differentiated. This allows the parameters of the soft attention structure to be calculated in the process of backpropagation of the neural network, learn and adjust during the iterative process. After the training is completed , soft attention can be directly generated by the network, so soft attention is also called determinable attention. Hard attention pays more attention to pixels, emphasizing the attention of each pixel, similar to a binary method, its prediction is a dynamic random process, and it is a kind of non-differentiable attention that cannot be self-learned through the network. Realization is usually done through reinforcement learning.

Since soft attention can realize self-learning in neural networks and is more widely used, this paper mainly studies the soft attention mechanism and proposes a lightweight attention mechanism that is easy to embed in existing neural networks to improve detection tasks effect.

2473-3636/21/$31.00 ©2021 IEEE
DOI 10.1109/DCABES52998.2021.00042

II. METHODS

From the perspective of spatial attention and channel attention merged, a lightweight attention mechanism is proposed. The overall structure of the attention mechanism is shown in Figure 1. For the input feature map, the mask is obtained by the spatial attention module and then multiplied and weighted by the original feature map. The output result of the spatial attention module and the mask generated by the channel attention module are multiplied and weighted to obtain the final output. The output feature map has the same dimension as the original input feature map. Representatives of spatial attention mechanism are STN [4], SMCA [5], and representative of channel attention mechanisms are SENet [6], FcaNet [7].

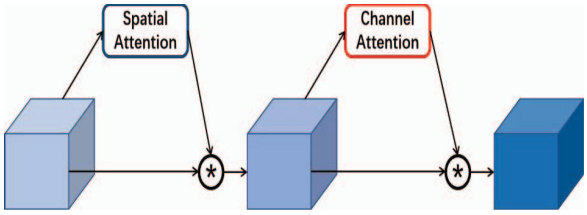

Figure 1. The Architecture of Designed Attention Mechanism.

A. Design of Spatial Attention Module

The purpose of the spatial attention module is to judge the importance of spatial pixels. The structure is shown in Figure 2. For the input feature map, the average value of all channels at each pixel position is calculated as the initial value of the importance for the pixel. Then, the 3×3 convolution and the dilated convolution [8] with the expansion rate of 3 are used in parallel. This operation uses different receptive fields to fully perceive the importance of a pixel compared to the surrounding space pixels, and the use of dilated convolution can increase the receptive field, without introducing additional parameters, the results obtained by the parallel convolution branch are connected to the channels, the feature importance is merged through 1×1 convolution, and then the spatial attention mask and the meta input feature are obtained through the sigmoid function The graph is multiplied and weighted to get the final output.

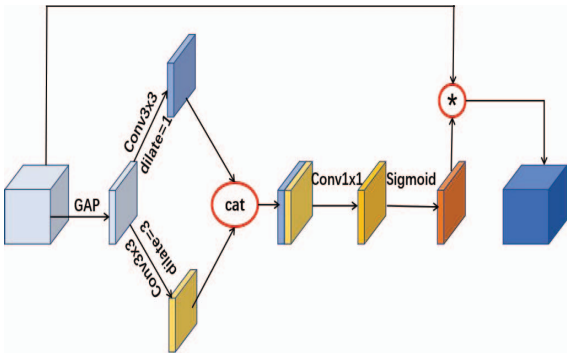

Figure 2. The Architecture of Spatial Attention Module.

B. Design of Channel Attention Module

The purpose of the channel attention module is to judge the importance of channel features. The structure is shown in Figure 3. For the input feature map, global pooling is performed to obtain the incident channel information, and then the channel attention is obtained through a parameterized weighting method. This parameterization method introduces very few parameters, which ensures the lightweight characteristics of the attention mechanism. Finally, the channel attention mask is weighted and multiplied by the original input feature map to obtain the final output.

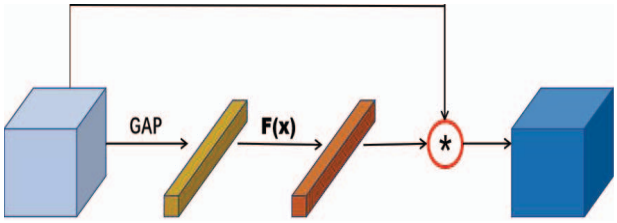

Figure 3. The Architecture of Channel Attention Module.

III. EXPERIMENT

A. Experimental Setup

The experiment is based on python 3.6, uses the pytorch framework to build the network, and is carried out in the Linux system environment, uses the GPU model Tesla K80. The experiment is based on the SSD [9] algorithm on the VOC [10] data set. The input image size is 300×300, and the designed attention mechanism is embedded in ResNet18, ResNet34, ResNet50 and ResNet101. By this way, we can obtain DA-ResNet18, DA-ResNet34, DA-ResNet50 and DA-ResNet101, the structure of the DA module embedded in ResNet is shown in Figure 4.

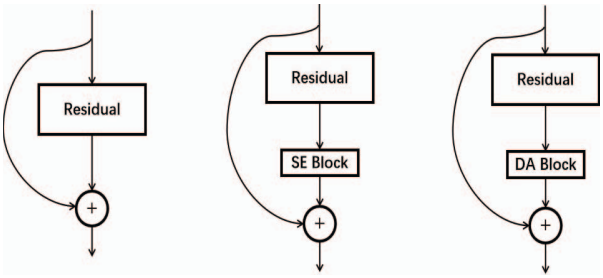

Figure 4. The architecture of DA embedded residual network

B. Evaluation Standard

FLOPs (Floating Point Operations) represent the number of floating point operations, which are usually used as indicators to measure the complexity of the model. For the convolutional layer, the formula of FLOPs is shown as follows. W and H respectively represent the width and height of the input feature vector, C_{in} represents the number

of channels of the input vector, K represents the size of the convolution kernel, and C_{out} represents the number of channels of the output feature vector.

$$FLOPs\_conv = 2 \times H \times W\left(C_{in}K^2 + 1\right)C_{out} \quad (1)$$

For the fully connection layer, the formula of FLOPs is calculated as follows. I represents the dimension of the input and O represents the dimension of the output.

$$FLOPs\_fully = \left(2I - 1\right)O \quad (2)$$

Parameters are also an important indicator to measure the complexity of the model. The formula of calculation is shown as follows. The meaning of the variable is the same as before.

$$Params = C_{out} \times \left(H \times W \times C_{in} + 1\right) \quad (3)$$

Grad-CAM [11] is a gradient-based visualization algorithm that can differentiate different areas in the picture by generating a heat map, so that we can intuitively understand the area of a picture that the neural network cared about, and it is useful for optimizing experiments and understanding the neural network.

Mean Average Precision is used as the evaluation standard of algorithm accuracy, we usually called it mAP. The calculation of mAP is closely related to IoU. The formula of IoU is shown as follows. b_{gt} represents the position information of the real box and b_{pred} represents the position information of the predicted box. By this, we get all boxes' IoU.

$$IoU\left(b_{pred}, b_{gt}\right) = \frac{Area\left(b_{pred} \cap b_{gt}\right)}{Area\left(b_{pred} \cup b_{gt}\right)} \quad (4)$$

The IoU of the predicted box and the real box is calculated, and then the default box is divided into positive or negative examples according to the set IoU threshold. In the image, each category will have real cases (TP, True Positive), false positives (FP, False Positive), true negatives (TN, True Negative), and false negatives (FN, False Negative). The formula of precision is shown as follows.

$$precision = \frac{TP}{TP + FP} \quad (5)$$

Choose 11 different recall rates (0, 0.1, 0.2,..., 0.9, 1.0), calculate the corresponding accuracy rates under different recall rates, and then average the accuracy rates to get the AP, mAP is the average value of all types of AP. The formula of AP is shown as follows.

$$AP_{11point} = \frac{1}{11}\left(\sum x\right)\left(x \in \max precision\right) \quad (6)$$

C. Experimental Result

In order to explore the best combination of the spatial attention module and the channel attention module, the serial and parallel methods of the two method are tested. The experimental results are shown in Table 1. It is not difficult to see from the experimental results that the channel attention is in the front and the spatial attention is in the back, the spatial attention is in the front and the channel attention is in the back, and the parallel combination of the two module can all improve the detection accuracy. The serial mode which the spatial attention is in the front and the channel attention is in the back achieves the highest accuracy. As a result, the combination of the spatial attention part and the channel attention part of the designed lightweight attention module is determined.

TABLE I. Accuracy results of different combinations of attention modules

Method	mAP
ResNet50	75.9
ResNet50+channel+spatial	76.8
ResNet50+spatial+channel	**77.0**
ResNet50+spatial and channel in parallel	76.6

Compare the parameters and calculations of ResNet18, ResNet34, ResNet50, ResNet101, SE-ResNet18, SE-ResNet34, SE-ResNet50, SE-ResNet101, DA-ResNet18, DA-ResNet34, DA-ResNet50, DA-ResNet101, and use the above-mentioned network as the backbone feature extraction network. SSD is used as the baseline to compare the accuracy. The experimental results are shown in Table 2. From the experimental results, we can see that SE-ResNet50 has increased 2.5150M and 0.0026GFLOPs compared to ResNet50 in terms of parameters and calculations, while DA-ResNet50 has only increased by 0.0004M and 0.0003GFLOPs compared to ResNet50 in terms of parameters and calculations. The amount of our DA attention mechanism's parameters and calculations are negligible. What's more, similar conclusions can be obtained from the comparison results of other items in the table, and our method's accuracy of the detection task is improved more significantly than that of SENet. The results improve that our method is very effective.

TABLE II. DA attention mechanism vs. SE attention mechanism

Backbone	Params	GFLOPs	mAP
ResNet18	11.6895M	1.8214	72.6
SE-ResNet18	11.7766M	1.8215	72.9
DA-ResNet18(ours)	**11.6897M**	**1.8215**	**73.1**
ResNet34	21.7977M	3.6742	74.5
SE-ResNet34	21.9549M	3.6744	74.8
DA-ResNet34(ours)	**21.7980M**	**3.6744**	**75.1**
ResNet50	25.5570M	4.1185	75.9
SE-ResNet50	28.0720M	4.1211	76.5
DA-ResNet50(ours)	**25.5574M**	**4.1188**	**77.0**
ResNet101	44.5492M	7.8444	77.0
SE-ResNet101	49.2923M	7.8492	77.6
DA-ResNet101(ours)	**44.5498M**	**7.8448**	**78.2**

The experiment compared the parameters, calculations and accuracy of the DA module with other lightweight attention mechanisms on the VOC detection task. As shown in Table 3, it can be seen from the experimental results that the amount of parameters and calculations of DA module we

designed is less compared with SGE [12] and ECA [13] attention modules, the accuracy is higher.

TABLE III. DA attention mechanism vs. other attention mechanisms

Method	Params	GFLOPs	mAP
ResNet50	25.557M	4.119	75.9
+SGE	25.557M	4.127	76.3
+ECA	25.559M	4.127	76.8
+DA(ours)	**25.557M**	**4.119**	**77.0**
ResNet101	44.549M	7.844	77.0
+SGE	44.553M	7.858	77.4
+ECA	44.549M	7.858	78.0
+DA(ours)	**44.550M**	**7.845**	**78.2**

Here, ResNet50 and DA-ResNet50 are visualized by using Grad-CAM in layer4_2, as shown in Figure 5. In the same row of pictures, the leftmost is the original picture, the middle is the result of ResNet50 visualization, and the far right is the result of DA-ResNet50 visualization. It is not difficult to see that when the DA module is added, the network pays more attention to effective information, and the area of interest is also more concentration, which proved the effectiveness of the DA attention module.

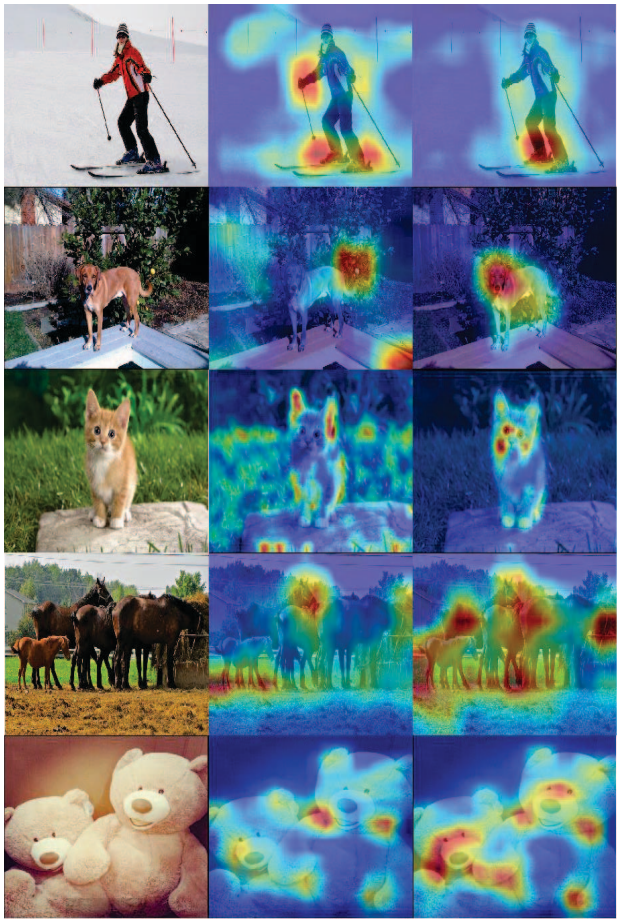

Figure 5. Grad Cam Visualization Comparison

IV. CONCLUSION

Aiming at the problem of large parameters and poor portability of attention mechanism modules, an extremely lightweight attention mechanism is designed, which uses a combination of spatial attention and channel attention to enhance the model's attention to important information. The attention mechanism we proposed is extremely lightweight, which introduces very few parameters and calculations, and improves the detection accuracy significantly. It is easy to be embedded in deep convolutional neural networks such as ResNet, and has strong portability. Experiments have proved that our attention mechanism is very efficient.

The attention mechanism designed in this paper has fewer parameters and calculations, but the improvement effect of the model needs to be further improved. How to design an attention mechanism with fewer parameters, strong portability, less calculation, and obvious improvement effect is the direction of our future work.

ACKNOWLEDGMENT

We wish to thank every member of the team for their efforts. Thank you for your help in my study. I will continue to be enthusiastic in the field of object detection and conduct in-depth research.

REFERENCES

[1] Wu X, Sahoo D, Hoi S. Recent Advances in Deep Learning for Object Detection[J]. Neurocomputing, 2020, 396.

[2] Zhang Q, Yang L T, Chen Z, et al. A survey on deep learning for big data[J]. Information Fusion, 2018, 42:146-157.

[3] Ketkar N. Convolutional Neural Networks[J]. Springer International Publishing, 2017.

[4] Jaderberg M, Simonyan K, Zisserman A, et al. Spatial transformer networks[J]. arXiv preprint arXiv:1506.02025, 2015.

[5] Gao P, Zheng M, Wang X, et al. Fast Convergence of DETR with Spatially Modulated Co-Attention[J]. arXiv preprint arXiv:2101.07448, 2021.

[6] Hu J, Shen L, Sun G.Squeeze-and-excitation networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 7132-7141.

[7] Qin Z, Zhang P, Wu F, et al. FcaNet: Frequency Channel Attention Networks[J]. arXiv preprint arXiv:2012.11879, 2020.

[8] Liu S, Huang D. Receptive field block net for accurate and fast object detection[C]//Proceedings of the European Conference on Computer Vision (ECCV). 2018: 385-400.

[9] Liu W, Anguelov D, Erhan D, et al. Ssd: Single shot multibox detector[C]//European conference on computer vision. Springer, Cham, 2016: 21-37.

[10] Vicente S, Carreira J, Agapito L, et al. Reconstructing pascal voc[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2014: 41-48.

[11] Selvaraju R R , Cogswell M , Das A , et al. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization[C]// IEEE International Conference on Computer Vision. IEEE, 2017.

[12] Li X , Hu X , J Yang. Spatial Group-wise Enhance: Improving Semantic Feature Learning in Convolutional Networks[J]. 2019.

[13] Wang Q , Wu B , Zhu P , et al. ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks[J]. 2019.