An Improved YOLOX Algorithm based on Structural Re-parameterized CBAM for Wild Animals Detection

Yuxin Lyu
School of Software
Southeast University
Suzhou, China
1017495610@qq.com

Xiaobo Lu*
School of Automation
Southeast University
Nanjing 210096, China
xblu2013@126.com
*Corresponding author

***Abstract*—We use the structural re-parameterization method to optimize the generation process of the attention map. CBAM has long been among the most widely used attention techniques because of its simplicity and good performance. The success of RepVGG shows great application potential of the structural re-parameterization methods, which we notice that could be used to further strenthen the ability of the CBAM. Based on YOLOX, we test our method on a self-built wild animals dataset. The result shows better performance compared to the original CBAM. Notebly, the gain is totally FREE OF CHARGE. Compared to the baseline model, no extra parameters or FLOPs are introduced during the inference stage.**

***Index Terms*—Attention mechanism, YOLOX, Structural Re-parameterization, Object Detection**

I. Introduction

In the process of human visual recognition, people tend to pay more attention to the more distinguishable areas in the picture. This mechanism allows the eyes to focus on the objects in the scene more quickly and accurately. In the field of computer vision, this mechanism is imitated by humans, and many attention mechanism methods have been proposed. Some methods focus on the channel dimension, such as Squeeze-and-Excitation networks[1], and others combine channel attention and spatial attention at the same time, such as BAM[2], CBAM[3], etc. In addition, there are a large number of self-attention methods. These methods have achieved superior results in various fields such as classification, detection, and segmentation.

On the other hand, in the latest research on neural network structure design, researchers began to pay attention to how to combine the advantages of neural network width and depth. Some methods focus on how to enhance the feature extraction ability of convolutional neural networks by enriching the topology of the training phase network under the premise of keeping the calculation amount and parameter amount consistent during inference. Typical methods include ACNet[4], RepVGG[5], etc. This type of method adds parallel branches to the convolution layers during training, increases the width of the network, and fuses those branches during inference stage. Those using this idea ar structural heavy parameter method. Using structural heavy parameter methods, we can improve the expressive power of the network without increasing the cost of inference.

We noticed that there are many convolution operations in the generation process of the attention mechanism. However, there are few studies on introducing the idea of structural re-parameterization to the attention module. This limits the exploration of the performance-effect boundary of the attention mechanism. This paper focuses on the application of structural re-parameterization in the spatial attention mechanism, and proposes an re-parameterization method for CBAM. Experiments prove that it is feasible to apply the idea of structural re-parameterization in the generation process of attention mechanism.

As we all know, there exists a problem in the detection field that the receptive field of the feature map does not match the size of the detection target. FPN[6] solves this problem to a certain extent, especially for single-stage detectors such as YOLOX[7], YOLOv3[8], CenterNet[9], etc. FPN can effectively improve the detection effect and is an essential module for single-stage detection algorithms. However, in the FPN method, the feature fusion of each layer is simply performed. When performing feature fusion, a lot of up-sampling and down-sampling processes are involved, resulting in the loss of semantic information and the introduction of noise. Adding fusion may lead to confusion of features.

Some research on FPN focuses on the method of introducing the attention mechanism into FPN. CE-FPN[10] uses channel attention to enhance information between feature map channels. AC-FPN[11] fuses the features of multiple different receptive fields to increase the receptive field while using contextual information to improve the detection effect. In addition, there are some works, such as YOLOF[12], to explore the feasibility of removing FPN. Based on the above analysis, this paper proposes a method that combines FPN and the attention mechanism to achieve better detection results in the single-stage detector YOLOX. The contributions of this

paper can be divided into the following three points:

- A re-parameterized CBAM module is proposed, which uses two parallel branch structures to optimize the generation process of generating spatial attention maps. The module structure is shown in Fig. 4.
- Applying the re-parameterized CBAM module to the FPN structure improves the detection effect of the YOLOX model.
- We tested the proposed method in a self-built wild animal scene data set, which has a strong reference value for the wild animal recognition task scene.

II. Re-parameterized CBAM

A. Structural re-parameterization

It is known to all that a convolution layer is merely a linear projection without activation functions. Linear operations endow convolution with additivity. Thus given a convolution layer with parallel branches and no nonlinear activations. we can fuse the parallel convolution kernel weights to get an identical convolution layer. The new convolution layer can substitute the original parallel structure. Compared to single feed-forward convolution layer, parallel structure provides more training-time parameters and broader optimization dynamics. Due to this math nature, some new convolutional neural networks are proposed. RepVGG in parallel adds a 1 × 1 convolution layer and a shortcut connection to a convolution layer. Then apply the structure to every convolution layer of an off-the-shelf network architecture. After training, every parallel block is fused separately. The overall structure is depicted in Fig.1. During the inference stage, the re-parameterized network structure is identical to the original counterpart. There is no any additional parameters and FLOPs.

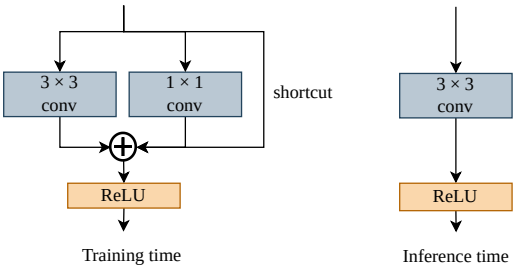

Fig. 1. RepVGG blocks

B. CBAM

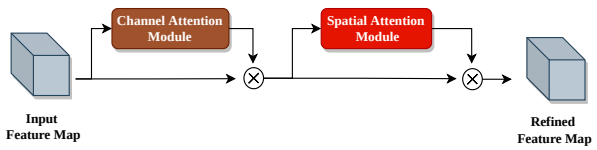

Fig. 2. CBAM structure

In the process of human visual recognition, people tend to pay more attention to more distinguishable areas in pictures. A large number of attention mechanism methods have been proposed, which have achieved superior results in various fields such as classification and detection. Most Existing attention-based methods are plug-and-play and end-to-end trainable modules. Generally speaking, given an intermediate feature map, an attention module generates an attention map along one or more dimensions. An attention map is essentially a mask that is multiplied to the original feature map. The mask enhances the responses of important areas and weakens the remaining. Squeeze-and-Excitation networks focus on the channel dimension. Through the average pooling operation and a nonlinear projection, the SE module generates a channel-wise mask to attend important channels. CBAM further extends the SE modules to the spatial dimension. As shown in Fig.2, CBAM simultaneously utilizes average pooling operations and max pooling operations and sequentially generates channel-refined feature maps and space-refined feature maps.

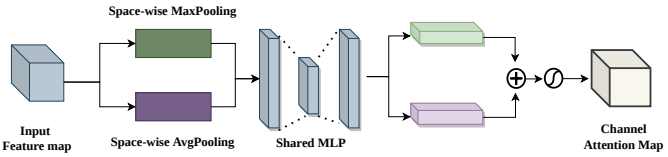

Fig. 3. Channel Attention Module

The generation process of the channel attention mask is shown in Fig.3. Given an input feature map $F \in R^{C\times H\times W}$, we use the max pooling and the average pooling to squeeze its spatial dimensions. Max pooling highlights the distinct information cross different channels, while average pooling gives the spatial statistics. The generated max pooling feature vector and average pooling feature vector then go through a shared MLP with one hidden layer. The hidden layer projects the vectors to a subspace, reducing dimensions to a certain reduction rate. Then an ReLU function is applied. In a nutshell, the channel attention module can be described as:

$$M_c(F) = ReLU(MLP(AvgPool(\mathbf{F}))+MLP(MaxPool(\mathbf{F})))$$

The channel attention module refines the input feature map and generates a intermediate feature map $\mathbf{F_{refined}}$. Through a convolution layer and a sigmoid function, the spatial attention module generates a mask. The mask utilizes the spatial local correlation information. We further strengthen the spatial attention module with novel structural re-parameterization method. The details are described in next subsection.

C. Re-parameterized Spatial Attention Module

The overall structure of our novel re-parameterized Spatial Attention Module is shown in Fig.4. We focus on the generation process of spatial attention maps. Original CBAM method uses a single k × k convolution layer to generate the attention map. Inspired by ACNet, we add a 1 × k convolution layer and a k × 1 convolution layer in parallel. The two asymmetric convolution layers help to strengthen the kernel skeleton of the k × k convolution layer. The structure is similar with ACBlock.

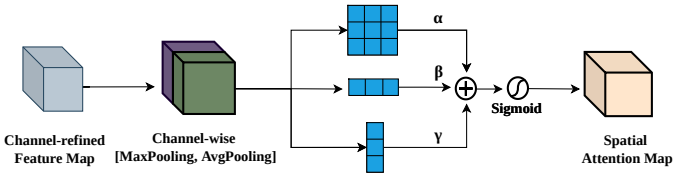

Fig. 4. Re-parameterized Spatial Attention Module

But the generated attention feature maps of the three layers are added with trainable weights rather than direct addition. The re-parameterized spatial attention module can be described as:

$$M_s(\mathbf{F_1}) = \sigma(\alpha \times f^{3*3}(\mathbf{F_1}) + \beta \times f^{1*3}(\mathbf{F_1}) + \gamma \times f^{3*1}(\mathbf{F_1}))$$

In the above equation,σ refers to the sigmoid function. $\mathbf{F_1} = [Avgpool(\mathbf{F_{refined}}); Maxpool(\mathbf{F_{refined}})]$, $\mathbf{F_{refined}}$ refers to the intermediate feature map refined by the channel attention module. α, β, γ are trainable weights. f^{3*3} represents a convolution layer with a 3×3 kernel. Experiments show that the three weights have a great impact on the fusion performance.

Direct addition can be regarded as a manner in which every branch result is multiplied with a fixed weight 1. This manner assumes that each attention map of the three branches is equally important, which is a strong assumption. As shown in Table I, experiments show that the direction addition manner is worse by 1.54% mAP than original CBAM in our Wildlife Dataset. Direct addition without trainable weights gives the worse results in all re-parameterized models. We believe constant weights might affect the grad propagation of the optimization process, leading to poor performance. In comparison, trainable weights give the model more flexibility.

Trainable parameters bring a natural problem, that is, how to choose the initialization method. We conduct experiments to compare two different ways: initializing to 1 and Kaiming initialization. Trainable weights manner with initializing to 1 is better than direct addition manner(fixed weights), but they are all worse than the Kaiming initialization manner. Our novel re-parameterized spatial attention module can be used to substitute the origin SAM. Its implementation is very simple and extensible. The performance gain is totally free.

III. Improve YOLOX with Re-parameterized CBAM

As one of the latest object detection methods, YOLOX aggregates state-of-the-art tricks. For simplicity, we keep only the core mechanisms and architecture of the original version. The overall architecture includes a backbone network DarkNet21 for feature extraction, naive FPN for feature fusion, and a decoupled detection head structure, which includes parallel classification branches and localization branches. We simply insert our powerful re-parameterized CBAM modules in every FPN outflow path, bringing us 3.71% mAP improvement. We use the same label assignment strategy as the origin version. The structure of the YOLOX network is shown in Fig.5.

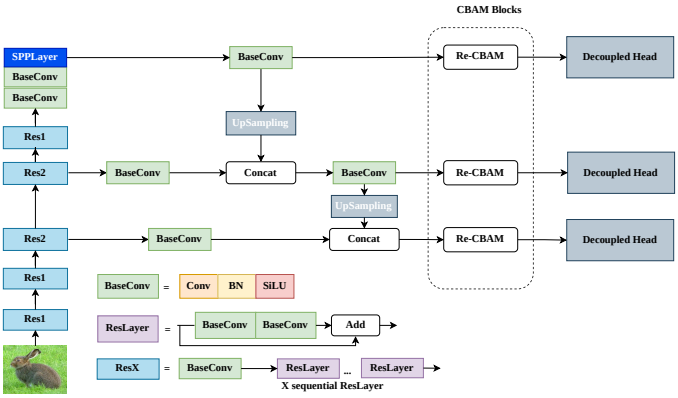

Fig. 5. YOLOX with re-parameterized CBAMs

IV. Experiments and Analysis

This section introduces the evaluation metrics, dataset, specific model structure and hyper-parameters used in detail. All experiments are done on two Nvidia RTX 3060 GPUs.

A. Evaluation Metrics

To evaluate the model's performance on wildlife detection, we used mAP metrics. mAP is the most commonly used evaluation metrics in target detection algorithms:

- AP (Average Precision) is the area enclosed by the precision-recall curve, which is used to measure the quality of detection. Following the traditions, we choose AP50 to evaluate our model. A box proposal is regarded as positive only if its IOU $>$ 0.5 with a ground truth.
- mAP (mean Average Precision) is the average value of APs with different IOU thresholds, which measures the average quality from AP50 to AP95.

Fig. 6. Wildlife Dataset examples

We use the wildlife dataset constructed by ourselves as the benchmark dataset of the experiment. The training set includes 3680 pictures, and the test set has 525 pictures. The pictures spread in a total of 10 categories, including jungle cats, Asian elephants, tigers, bison, wild horses, gibbons, snow rabbits, bustards, parrots, and black bears. Some image examples are shown in Fig.6.

In all experiments, in order to control variables and ensure that the module comparison effect is not affected by other factors such as randomness, we fixed the random number

TABLE I
AP50 AND MAP OF THE RE-PARAMETERIZED CBAM WITH DIFFERENT CONFIGURATIONS ON WILDLIFE DATASET

Model	CBAM	re-parameterization	trainable weights	initialize to 1	Kaiming Initialization	AP50	mAP
YOLOX						66.69	39.38
YOLOX	✓					70.1	41.7
YOLOX	✓	✓		✓		68.79	40.16
YOLOX	✓	✓	✓	✓		69.50	41.1
YOLOX	✓	✓	✓		✓	**71.22**	**43.09**

seed and kept the hyper-parameters consistent. During training and testing, the size of the input image is all resized to 416, and basic data enhancement such as HSV, random flip, and MixUP[13] is used. We train 100 epochs from scratch. The warm-up learning rate is used in the first 5 epochs, and the batch size is 10 during training.

V. CONCLUSION

We apply the popular structural re-parameterization techniques to strengthen the ability of CBAM. Trainable weights is proven to be essential to the process of fusing the spatial attention convolution kernels. Use Kaiming Initialization helps to make the module optimize better. The novel new method shows great performance on the YOLOX detection structure. Our module is tested on a wild scene dataset. We believe that the algorithm is valuable to the wildlife animal detection task.

REFERENCES

[1] Jie Hu, Li Shen, and Gang Sun. "Squeeze-and-Excitation Networks". In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Computer Society. 2018, pp. 7132–7141.

[2] Jongchan Park et al. "Bam: Bottleneck attention module". In: *arXiv preprint arXiv:1807.06514* (2018).

[3] Sanghyun Woo et al. "Cbam: Convolutional block attention module". In: *Proceedings of the European conference on computer vision (ECCV)*. 2018, pp. 3–19.

[4] Xiaohan Ding et al. "Acnet: Strengthening the kernel skeletons for powerful cnn via asymmetric convolution blocks". In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2019, pp. 1911–1920.

[5] Xiaohan Ding et al. "Repvgg: Making vgg-style convnets great again". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 13733–13742.

[6] Tsung-Yi Lin et al. "Feature pyramid networks for object detection". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 2117–2125.

[7] Zheng Ge et al. "Yolox: Exceeding yolo series in 2021". In: *arXiv preprint arXiv:2107.08430* (2021).

[8] Joseph Redmon and Ali Farhadi. "Yolov3: An incremental improvement". In: *arXiv preprint arXiv:1804.02767* (2018).

[9] Kaiwen Duan et al. "Centernet: Keypoint triplets for object detection". In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2019, pp. 6569–6578.

[10] Yihao Luo et al. "CE-FPN: enhancing channel information for object detection". In: *Multimedia Tools and Applications* (2022), pp. 1–20.

[11] Junxu Cao et al. "Attention-guided context feature pyramid network for object detection". In: *arXiv preprint arXiv:2005.11475* (2020).

[12] Qiang Chen et al. "You only look one-level feature". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2021, pp. 13039–13048.

[13] Hongyi Zhang et al. "mixup: Beyond empirical risk minimization". In: *arXiv preprint arXiv:1710.09412* (2017).