ISSN 2616-5775 Vol. 4, Issue 1: 75-81, DOI: 10.25236/AJCIS.2021.040112

Object tracking in siamese network with attention mechanism and Mish function

Fangbin Zhang[1, *], Xiaofeng Wang[2]

[1]College of Information Engineering, Shanghai Maritime University, Shanghai, 201306, China
[2]Shanghai Maritime University, 201306, China
**Corresponding author: 610362980@qq.com*

Abstract: *In order to improve the recognition and tracking ability of the fully-convolutional siamese networks for object tracking in complex scenes, this paper proposes an improved object tracking algorithm with channel attention mechanism and Mish activation function. First, the channel attention mechanism is introduced into the model, and different weights are assigned to each channel to improve the network's representation ability. At the same time, the Mish function is used to replace the ReLU activation function in the network. The smooth Mish function can make better information enter the network, thereby obtaining better accuracy and generalization. Finally, the gradient centralization is embedded in the stochastic gradient function, so as to improve the generalization performance of the network and make the training more efficient and stable. The experiment was performed on the OTB50 and VOT2018 data sets, and the improved algorithm achieved better performance than the original algorithm.*

Keywords: *Object tracking, channel attention mechanism, Mish function, gradient centralization*

1. Introduction

Object tracking is a very critical and challenging research field in the field of computer vision. It has application in video surveillance, human-computer interaction, and robotics. It is a very important part of artificial intelligence and has attracted the attention of scholars from all over the world. Therefore, the research on object tracking algorithms is of great significance. At present, there are three mainstream object tracking algorithms, namely correlation filtering, deep learning and a combination of the two.

The method based on correlation filtering regards the process of object tracking as the process of correlation filtering on the search area image. More representatively, Bolme et al [1] proposed the MOSSE (Minimum Output Sum of Squared Error) algorithm, which makes correlation filtering in a true sense for online object tracking. MOSSE filtering is based on an adaptive training method, which only requires one of frame of image to generate stable correlation filtering. Henriques et al. [2] proposed KCF (Kernelized Correlation Filter) and DCF (Dual Correlation Filter) to improve the robustness of the tracker to motion blur and illumination changes in complex environments. The CN (Color Name) proposed by Danelljan et al.[3] uses multichannel color features to enhance the ability to characterize the target object. With the development of deep learning, object tracking algorithms based on CNN (Convolutional Neural Network) have developed rapidly. One way of thinking is the combination of deep learning and related filtering. For example, trackers such as ECO [4] and C-COT [5] combine the powerful feature extraction capabilities of CNN with traditional frameworks and achieve good tracking results. Another idea is to construct an end-to-end deep network object tracking method, such as MDNet [6] through multi-domain learning, and SANet [7] that combines features of convolutional neural networks and recurrent neural networks. Among them, the appearance of the siamese network structure can achieve a balance between accuracy and speed, which has greatly improved compared with the previous methods, and has become a research hotspot in this field. Tao et al [8] proposed that SINT(Siamese Instance Search for Tracking) was the first to use the siamese structure for object tracking. Subsequently, Bertinetto et al. [9] proposed a faster SiamFC(Fully-Convolutional Siamese Networks for Object Tracking). Later, Li et al. [10] used the RPN (Region Proposal Network) for reference and proposed SiamRPN, which solved the problem of SiamFC's poor robustness in the face of target scale changes. In recent years, deep learning-based object tracking algorithms have made great progress, but they still face many challenges, such as occlusion, illumination changes, scale changes, and deformation. Therefore, this paper improves the algorithm based on SiamFC to improve the accuracy of object tracking. The main contributions of this article include:

(1) Introduce the channel attention module to improve the expressive ability of the network;

(2) Replace the ReLU function with the Mish function to make better information penetrate the network and get better accuracy and generalization;

(3) Embed gradient centralization technology into the stochastic gradient function to improve the generalization performance of the network;

2. Related work

2.1 SiamFC tracking framework

SiamFC was proposed by Luca Bertinetto et al. It is the pioneering work of siamese networks and has achieve good results in the field of single object tracking. SiamFC belongs to a discriminative tracking algorithm. SiamFC solves the object tracking problem by using the similarity learning method. Through learning, a function f(z, x) is obtained. This function obtains a score map by comparing the template area z and the search area x. The higher the score, the higher the similarity between the area and the template area, the greater the possibility that the target is in the area. The function is as follows:

$$f(z,x) = g\big(\varphi(z), \varphi(x)\big) \tag{1}$$

Where φ is a feature extraction function and g is a cross-correlation operation function.

The SiamFC model consists of two parts: Feature extraction and similarity measure. The feature extraction stage uses AlexNet as the backbone network to extract features. The second part is a similarity measurement function, the input is the characteristics of the template image and the characteristics of the search area, and the result is a score map.

The loss function used by the SiamFC model is logistic loss function, the formula is as follows:

$$l(y,v) = log(1 + exp(-yv)) \tag{2}$$

$$L(y,v) = \frac{1}{|D|}\sum_{u\in D} l(y[u], v[u]) \tag{3}$$

Where v is the single response value output by the network, y is the label, y∈{-1,1}, D is the generated response graph, u is a value in D, and |D| is the size of the response graph. The label paper in the response diagram is marked as follows:

$$y[u] = \begin{cases} +1 \; if \; k\|u-c\| \le R \\ -1 \quad otherwise \end{cases} \tag{4}$$

Where c is the center of the response graph, u is any point in the response graph, and k is the multiple that the response graph is reduced after passing through the network.

The optimization function is SGD (Stochastic Gradient Descent) for parameter update:

$$arg_{\theta}^{min} \, {}_{(z,x,y)}^{E} L(y, f(z,x;\theta)) \tag{5}$$

2.2 Mish activation function

The activation function is a non-linear point-by-point function, which is responsible for introducing non-linearity into the linear transformation input in the neural network layer. Most of the current neural networks use the ReLU activation function, because it has the advantages of making the network training faster and making the grid sparse. Although it has better performance and stability than other activation functions, ReLU is not without its shortcomings. One of the shortcomings is usually called Dying ReLU, which is that when the gradient value is too large, the weight will be negative after updating, and it will become 0 after ReLU, resulting in no further updates. Later Prajit Ramachandran et al. [11] proposed the Swish activation function, which is a more powerful activation function. Compared with ReLU, Swish's smooth and continuous contours can better carry out information dissemination. Later, Diganta Misra [12] proposed the Mish activation function, which is a novel self-adjusting non-monotonic activation function. Compared with the different tasks of ReLU and Swish in computer vision, Mish tends to match or improve the performance of neural network architecture。Therefore, this paper introduces the Mish function into the model to make the backbone network perform better.

2.3 Attention mechanism

Attention mechanism is a mechanism that focuses on local information. The essence is to locate the information of interest and suppress useless information. Similar to the selective visual attention mechanism of humans, when humans see an image, they will efficiently allocate limited attention resources to focus on the key target area, obtain the detailed information of the desired focus, and suppress other useless information. The attention mechanism is widely used in the field of computer vision. SENet [13] proposed a channel attention mechanism, which automatically obtains the importance of each feature channel through learning, and then enhances useful features according to the importance and suppresses features that are not useful for the current task. SKNet [14] proposed a nonlinear method to aggregate the information of multiple convolution kernels and improved SENet. CBAM [15] proposed a model that combines channel attention mechanism and spatial attention mechanism. Inspired by this, this article introduces the channel attention mechanism into the model.

2.4 Gradient Centralization

Optimization technology is essential for efficient training of deep neural networks. Gradient descent has always been a crucial part of training deep neural networks. It directly affects the convergence speed of the model. Whether the gradient is processed properly or not determines whether the effect of the trained model meets expectations. Due to the characteristics of the neural network itself, as the number of model layers increases, gradient descent becomes more and more difficult. How to alleviate these problems is also a challenge for deep learning. For the study of gradients, many methods have been proposed in recent years. Gradient centralization was proposed by Hongwei Yong et al[16], which can be easily embedded in the existing optimization function and achieve good results. Therefore, in this paper, the gradient centralization technology is embedded in the stochastic gradient descent (SGD) function.

3. Model design

3.1 Model structure

The overall structure of the model is shown in Fig 1. The backbone network is AlexNet, the last three fully connected layers are removed, and the template image and the search area are extracted through the same weight of AlexNet. The difference with SiamFC is that this article replaces the ReLU activation function in the network with the Mish activation function. Embed gradient centralization technology into the optimization function, and replace SGD with SGD_GCC. Then the feature map is sent to the channel attention module, and the channel attention module adjusts the weights of different channels. Finally, the features extracted by the two branches are sent to the cross-correlation layer to obtain a 17x17 score map.

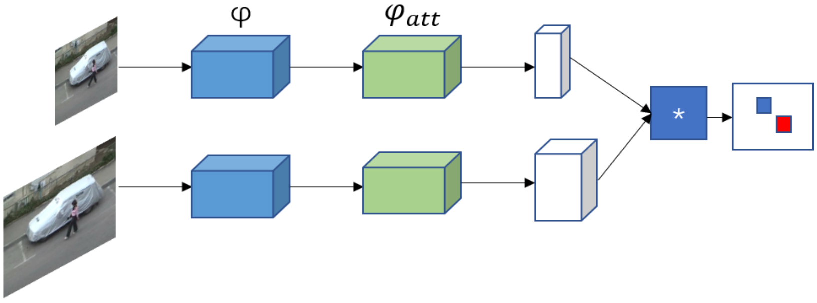

Figure 1: Overall architecture of model

3.2 Mish function

The activation function is a nonlinear point function, and its task is to introduce nonlinearity into the neural network layer. In order to make the network perform better, this article introduces the Mish activation function to replace the ReLU function. Mish is a new self-regular non-monotonic activation function, which is inspired by the self-gating feature of Swish. The mathematical definition of Mish function is as in formula (6), the comparison between Mish function and ReLU function is shown in Fig 2. As can be seen from the figure, Mish has a slight tolerance for negative values, which will produce a better gradient flow. In addition, a smooth activation function allows better information to penetrate the neural network, resulting in better accuracy and generalization.

$$f(x) = xtanh(ln(1 + e^x)) \tag{6}$$

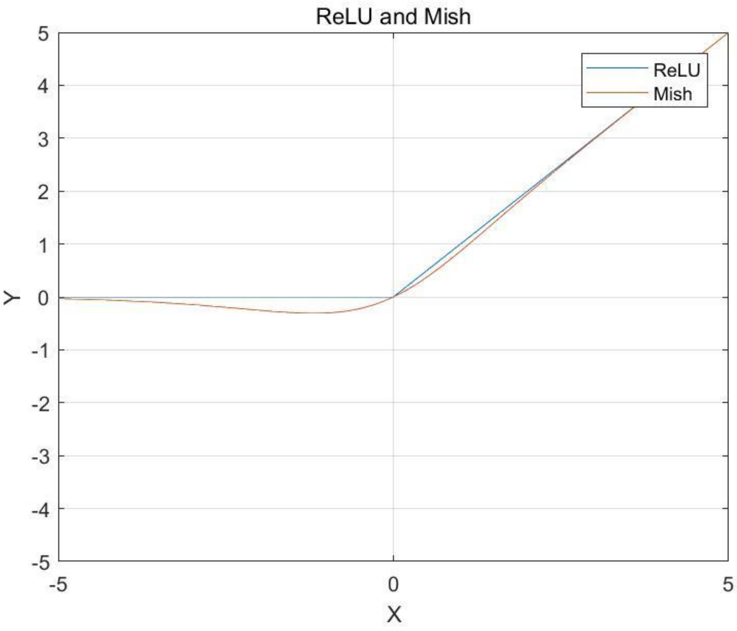

Figure 2: Comparison of Mish function and ReLU function

3.3 Channel attention module

The most important thing in the convolutional neural network is the convolution operator, which enables the network to fuse spatial information and channel information when extracting features of each layer. Based on this spatial composition relationship of features, Hu et al. proposed a new unit called "squeeze-excitation" module. This module first clearly shows the interdependence between channels, and then adaptively calibrates the channels according to the dependence. Related feature responses to improve the quality of features generated by the network. Inspired by this, this article introduces the channel attention mechanism into the model to improve network performance. By using global information, the mechanism biasedly emphasizes useful information features and suppresses features with limited effects. This attention mechanism is divided into three parts: squeeze, excitation, attention.

Squeeze function:

$$z_c = F_{sq}(u_c) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} u_c(i, j) \tag{7}$$

This function makes a global average value, adds up all the eigenvalues in each channel and averages it. This operation gets the global description feature.

Excitation function:

$$s = F_{ex}(z, W) = \sigma(g(z, W)) = \sigma(W_2 \delta(W_1 z)) \tag{8}$$

δ function is ReLU, and σ is a sigmoid activation function, $W_1 \in R^{\frac{C}{r} \times C}$,$W_2 \in R^{C \times \frac{C}{r}}$.C represents the channel dimension, r represents the number of hidden layer nodes in the fully connected layer. Through training and learning these two weights, a one-dimensional incentive weight is obtained to activate each layer of channels. This operation can learn the nonlinear relationship between each channel.

Scaling function:

$$\widetilde{X_c} = F_{scale}(u_c, s_c) = s_c \cdot u_c \tag{9}$$

This is a process of scaling, and the values on different channels are multiplied by different weights, which can enhance the focus on key channel domains.

The final feature map $\widetilde{X_c}$ will be sent to the cross-correlation layer to calculate the score map.

3.4 SGD_GC

GC (Gradient Centralization) constrains the loss function by introducing new constraints to the weight vector. As shown in the Fig 3, W is the weight, L is the loss function, $\nabla_W L$ is the gradient of weight, and $\varphi_{GC}(\nabla_W L)$is the centralized gradient.This process regularizes the weight space and the output feature space, thereby improving the generalization performance of the model. In addition, the constrained loss function has better Lipschitz properties than the original loss function, making the training process more stable and efficient. The formula of GC is as follows:

$$\varphi_{GC}(\nabla_{W_i}\mathrm{L}) = \nabla_{W_i}L - \mu_{\nabla_{W_i}L} \tag{10}$$

Where $\nabla_{W_i}L$ represents the gradient, i represents the i-th column vector of the gradient matrix, and the formula of $\nabla_{W_i}L$ is as follows:

$$\nabla_{W_i}L = \frac{1}{M}\sum_{j=1}^{M}\nabla_{W_{i,j}}L \tag{11}$$

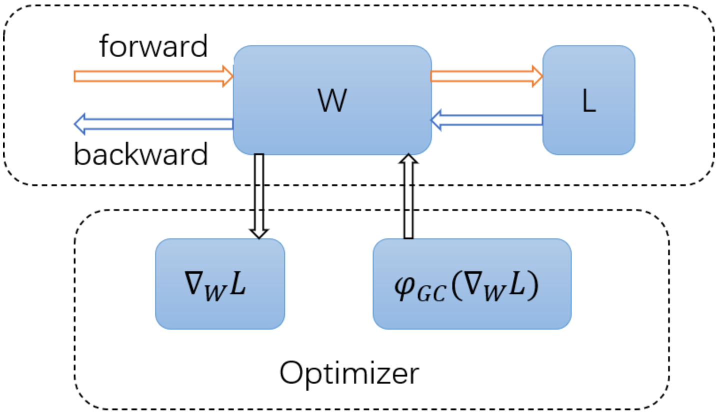

Figure 3: Sketch map for using gradient centralization (GC)

The original text uses Stochastic Gradient Descent (SGD). This article embeds the gradient centralization technology into SGD to form SGD_GC and replace the SGD.

4. Experimental results and analysis

4.1 Training and evalution

The model in this article is trained on the GOT-10K dataset using the GOT-10K Toolkit tool. GOT-10K includes more than 10,000 videos and more than 1.5 million manually labeled target frames. It is composed of five categories: animals, man-made objects, people, natural scenery, and parts. It can be subdivided into 563 target categories, as well as an action category. Divided into 87 actions. The experiment uses the SGD method to optimize, Batch Size is setting to 8, the initial learning rate is 10^{-2}, and the final learning rate is 10^{-4}. Stop after training 50 epochs. In terms of the experimental environment, the system is Ubuntu 16.04, the graphics card is Nvidia GeForce GTX 1080, and the framework is Pytorch. In order to verify the effectiveness of the model in this article, the GOT-10K Toolkit is used to evaluate the two popular data sets OTB50 and VOT2018. OTB50 was proposed by Wu Yi in 2013. It is a benchmark library used in the field of object tracking. It contains 50 artificially labeled video sequences, including fast-moving targets, similar backgrounds and foregrounds, and strong lighting. The VOT2018 is the data set of the VOT Challenge competition. It is another standard data set in the field of single object tracking. It contains 60 artificially labeled color video sequences.

4.2 Results and analysis

The experimental model was first evaluated on the OTB50 data set, and the results are shown in Table 1:

Table 1: Evaluation results under the OTB50 benchmark

Tracker	success_score	precision_score	success_rate	speed_fps
SiamFC	0.581	0.783	0.728	143
Ours	0.602	0.796	0.750	76

Compared with the original model, the model in this paper has an increase of 2.1% in success_score, 1.3% in precision_score, and 2.2% in success_rate, and it is still real-time in speed, which proves the effectiveness of this method.

The results of the evaluation carried out on VOT2018 are shown in Table 2:

Table 2: Evaluation results under the VOT2018 benchmark

Tracker	accuracy	robustness	speed_fps
SiamFC	0.489	34	71
Ours	0.511	35	48

In accuracy, it has increased by 2.2%, and the robustness has also been improved, which shows that the object tracking method in this article can better track the target.

5. Conclusion

In order to improve SiamFC's tracking ability in complex scenarios, this article makes the following improvements to SiamFC: The first is to introduce the channel attention mechanism into the network to adjust the weight of each channel to improve the network's feature extraction ability; the second is to replace the ReLU activation function with the Mish activation function to allow better information to enter the network and improve accuracy and generalization ability. The third is to embed gradient centralization into the stochastic gradient descent function to improve the generalization of the network. Compared with the original algorithm, the method in this paper has been improved on the OTB50 and VOT2018 datasets, which verifies the effectiveness of the algorithm. In the future object tracking algorithm research, we will continue to conduct in-depth research on the siamese network architecture.

References

[1] BOLME D S, BEVERIDE J R, DRAPER B A, et al. Visual object tracking using adaptive correlation filters [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2010: 2544 - 2550.

[2] HENRIQUES J F, CASEIRO R, MARTINS P, et al. Highspeed tracking with kernelized correlation filters [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014, 37(3) : 583 - 596.

[3] DANELLJAN M, KHAN F S, FELSBERG M, et al. Adaptive color attributes for real-time visual tracking [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2014: 1090 - 1097.

[4] DANELLJAN M, BHAT G, KHAN F S, et al. ECO: Efficient convolution operators for tracking [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 6638 - 6646.

[5] DANELLJAN M, HAGER G, SHAHBAZ KHAN F, et al. Convolutional features for correlation filter based visual tracking [C]//Proceedings of the IEEE International Conference on Computer Vision Workshops. 2015: 58-66.

[6] NAM H, HAN B. Learning multi-domain convolutional neural networks for visual tracking [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 4293 - 4302.

[7] FAN H, LING H B. SANet: Structure-aware network for visual tracking [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. 2017: 2217 - 2224.

[8] TAO R, GAVVES E, SMEULDERS A W M. Siamese instance search for tracking [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 1420 - 1429.

[9] BERTINETTO L, VALMADRE J, HENRIQUES J F, et al. Fully convolutional siamese networks for object tracking [C]//Proceedings of the European Conference on Computer Vision. 2016: 850 - 865.

[10] LI B, YAN J, WU W, et al. High performance visual tracking with siamese region proposal network [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 8971-8980.

[11] PRAJIT RAMACHANDRAN, BARRET ZOPH, QUOC V LE. Searching for activation functions

[C]//arXiv preprint arXiv: 1710.05941, 2017.

[12] DIGANTA MISRA, LANDSKAPE. A Self Regularized Non-Monotonic Activation Function [C]// arXiv preprint arXiv: 1908.08681v3, 2020.

[13] HU J, SHEN L, SUN G. Squeeze-and-excitation networks [C]//IEEE Conference on Computer Vision and Pattern Recognition, 2018: 7132-7141.

[14] LI X, WANG W, HU X, et al. Selective Kernel Networks [C]//IEEE Conference on Computer Vision and Pattern Recognition, 2019: 510-519.

[15] WOO S, PARK J, LEE J Y, et al. Cbam: Convolutional block attention module [C]//European Conference on Computer Vision, 2018: 3-19.

[16] YONG H, HUANG J, et al. Gradient Centralization: A New Optimization Technique for Deep Neural Networks [C]// arXiv preprint arXiv: 2004.01461v2,2020.