Advances in Artificial Intelligence, Big Data and Algorithms
G. Grigoras and P. Lorenz (Eds.)

doi:10.3233/FAIA230927

Target Tracking Algorithm Based on Hybrid Attention Unification Framework

Xiaoyu LI[1], Xiao DAI and Dan WANG
Chinese Flight Test Establishment, China

Abstract. In this paper, we propose a compact tracking framework built on top of Transformer. Our core design utilizes both attentional and convolutional operations and proposes a hybrid attentional module for simultaneous feature extraction and fusion of feature information between the target and the search image. This simultaneous modeling scheme allows the extraction of detailed features of the target and the fusion of features between the template and the search area. We use an attention mechanism with deep convolution to enhance the local attention to the target, as a way to ensure that the tracker achieves a balanced global and local attention for visual image target tracking. The experimental results show that the proposed method surpasses most of the mainstream trackers, and the operation speed achieves the real-time requirement with guaranteed accuracy.

Keywords: target tracking, hybrid attentional module, feature extraction, computer vision, fusion of feature information

1. Introduction

Target tracking has been a long-standing problem in computer vision for decades, with the aim of estimating the initial state of an arbitrary target in a video sequence. This technology has found its application in a wide range of areas, including but not limited to human-computer interaction and visual surveillance.

The currently popular trackers [1], [2], [3], [4], [5], [6] usually use a multi-stage pipeline, which consists of three parts: (1) a dedicated component for extracting relevant image features from both the tracking template image and the search image; (2) a feature fusion module to fuse and interact the feature information from the template image and the search image to let the tracker know to tracked in the search image; (3) a prediction head for pinpointing the target and estimating its bounding box. In general, the feature fusion module is the key to the tracking algorithm, as it is responsible for fusing the information between the target and the search image. Some algorithms have achieved competitive results by replacing the attention operation with a simple correlation operation in the feature fusion approach, but these methods still go through the CNN first to extract the feature information of the target, and these feature map information obtained through the CNN is limited because they are generally pre-trained after a given object category, which is difficult to apply to category-ambiguous tracking tasks and may ignore the finer tracking structure information.

To overcome the above problems, we integrate and unify visual feature extraction and target information in a single module, and streamline the three stages to two stages.

[1] Corresponding Author: Xiaoyu LI, Chinese Flight Test Establishment; e-mail: lixiaoyucfte@163.com

We propose a streamlined interaction scheme with self-attention to the target and search images then cross-attention to both, where self-attention focuses on the features of the self and cross-attention is responsible for mixing the information of both, and the improved tracker has stronger robustness.

2. Related Work

Siamese network-based trackers compute tracking as a similarity problem and have received much attention for their good balance between accuracy and speed.SiamFC for visual tracking was first proposed by Bertinetto et al[1], which uses inter-correlation operations to compute similarity between targets and search regions for target localization. Li et al[7] applied it to a Siamese network framework, referred to as SiamRPN, to solve the problem that SiamFC has no scale estimation, and the SiamRPN tracker predicts the target scale by pre-generated anchor point regression to obtain the bounding box with the best results, which solves the scale change problem of the target. However, the simple representation of box usually generates great loss when the object is rotated. Siammask [8] combines target tracking with segmentation to generate prediction boxes by predicting the mask segmentation map of the target, which solves the problem of loss of prediction boxes due to target rotation. To reveal the powerful depth features extracted by deep networks, siamrpn++[9] and SiamDW[10] solved the problem of using only shallow networks in target tracking by replacing deep networks into target tracking, which led to improved performance and overcame the problem of assigning large weights to image centers during feature learning by deep networks, allowing deep networks such as ResNet to be used as backbone networks for target tracking tasks.

3. Experimental Methods

3.1. Hybrid Attention

For the target tracking problem, a hybrid attention module is proposed whose inputs are a template image and a search image. Unlike the original multiple attention, the hybrid attention module performs a dual hybrid attention operation on the target template and the search region sequences. It performs self-attention on the markers themselves in each sequence to capture information about the target or the search region. At the same time, it performs cross-attention between the markers in both sequences to enable information exchange between the target template and the search region. Formally, given a sequence consisting of a target and a search, we first reshape it into a two-dimensional feature map. To allow the model to better obtain local perception of the image, we perform a separable deep convolutional projection of each feature map. Then, each feature map for target and search is flattened to produce values for query, key, and attention operations. In Figure 1 we do separate self-attention for the target and search images with the aim of having them reinforce their own self-features, and then connect the target and search images and do cross-attention with the aim of having their feature information fully integrated. To reduce the capacity of the model and the inference speed, we choose to do cross-attention

only in the search branch to achieve an asymmetric structure of the cross-attention mechanism.

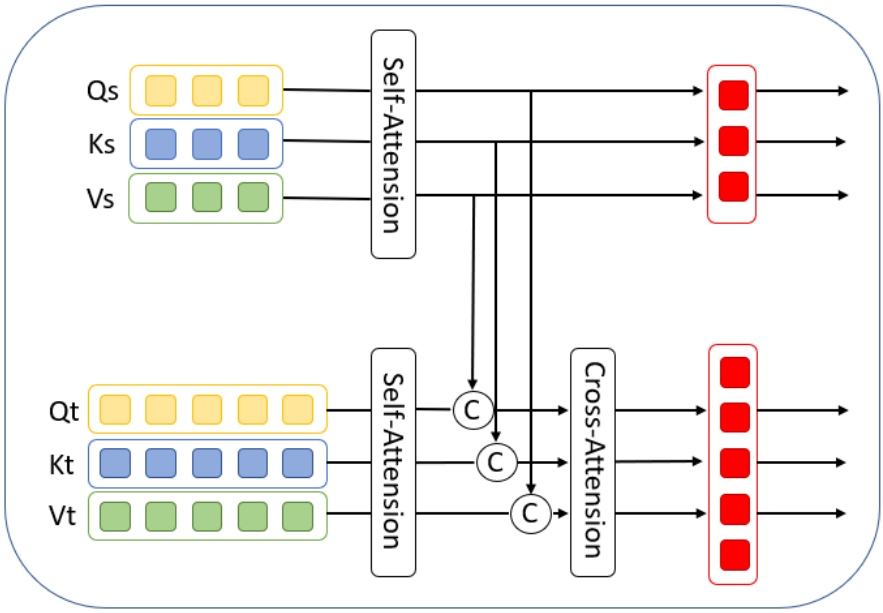

Figure 1 Attentional perception

Attention is expressed as:

$$Self - Attension = MultiHead(Q, K, V) \tag{1}$$

$$Cross - Attension = MultiHead(concat(qs, qt), concat(ks, kt), concat(vs, qv)) \tag{2}$$

where K, V are the inputs of any branch, Self-Attension is the self-attention mechanism, and Cross-Attension is the cross-attention mechanism.

3.2. DWConv

Compared with traditional convolution operations, the main difference of DWConv is that its convolution kernels are separated, i.e., they are divided into two parts for convolution, corresponding to the depth and width-height dimensions, respectively. First, for each channel of the input, a convolution kernel equal to the number of channels is used to convolve the channels, and then the result is superimposed in the depth dimension to obtain the final output. This operation maintains the output quality with a small number of convolutional kernels, and therefore reduces the parameters and computation.

As illustrated in Figure 2, to capture better local information, we re-reshape the feature map after the attention operation into a 2D vector, perform DWConv operation on it, and finally reconvert the 2D vector to 1D to facilitate its subsequent operation.

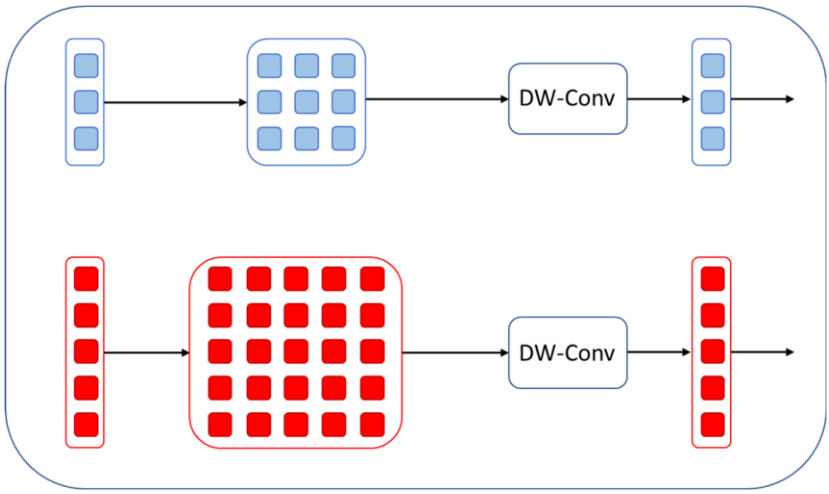

Figure 2 DW-conV local sensing

3.3. Process Structure

The input image first goes through the hybrid attention module to extract its global attention information, and then goes through the DWConv operation to sense the local information, and these two operations are stacked several times to complete the feature fusion operation. Finally, a simple MLP predictor head predicts the position of the target in the image and completes the tracking prediction. The feature map after the hybrid attention and DWConv operations takes into account both global and local feature information, which is less likely to cause semantic loss. Moreover, the asymmetric structure and DWConv we use in the hybrid attention module both greatly reduce the overhead of the model and achieve the speed of real-time while ensuring accuracy.

4. Experimental Analysis

We have experimented our tracker on multiple benchmarks against several popular methods. The details of the experiments are as follows, our tracker was implemented with Python 3.8 and PyTorch 1.12.0. We trained the network on a server with 2 1080 GPUs. For data enhancement, we used level flipping and luminance dithering with search image and template sizes of 320×320 pixels and 128×128 pixels, respectively. The first 500 epochs were used for backbone and head tuning, and another 40 epochs were used for fraction prediction head tuning.

4.1. Introduction to the Dataset

GOT-10k is an extensive target tracking dataset available for training and testing of visual tracking algorithms. The dataset consists of over 10,000 image sequences covering videos of many different scenes and object classes. Each video contains a target object to be tracked, and each video sequence contains hundreds of frames on average. All videos are high-resolution videos captured by desktop-class computers and manually annotated to ensure their accuracy.

LaSOT (Large-Scale Single Object Tracking) is a large-scale single object tracking dataset. The dataset contains about 14,000 video sequences covering more than 1400 different target classes. The length of each video sequence varies between a few frames to thousands of frames, and the resolution is 720p or 1080p. all video sequences are manually annotated to provide the true position of the target tracking.

4.2. Model Performance Comparison

Our method is compared with the mainstream trackers on the GOT-10k dataset and LaSOT dataset, and the experimental results are shown in Table 1. The model of the proposed method can achieve 68.5% AO on the GOT-10k dataset. The experimental data show that the proposed method surpasses most of the mainstream trackers and improves the accuracy of target tracking

Table 1 Comparison of experimental results

	GOT-10K			LaSOT		
	AO	$SR_{0.5}$	$SR_{0.75}$	AUC	P_{Norm}	P
Our	**69.5**	**79.1**	**65.0**	**68.2**	**78.1**	**70.9**
Stark	68.8	78.1	64.1	67.1	77.0	70.3
KeepTracker	-	-	-	67.1	77.2	70.2
SAOT	64.0	75.9	-	61.6	70.8	-
AutoMatch	65.2	76.6	54.3	58.2	-	59.9
TrDiMP	67.1	77.7	58.3	63.9	-	61.4
SiamR-cnn	64.9	73.8	59.7	64.8	72.2	58.9
TREG	66.8	77.8	57.2	64.0	74.1	59.8
SiamAttn	65.2	75.5	55.6	56.0	64.8	-
DTT	63.4	74.9	51.4	60.1	-	-

4.3. Visualization of Tracking Results

We also visualized our tracker with other mainstream trackers on the dataset at the same time, as shown in Figure 3, with our tracker results in red. In the first figure, the other trackers received interference from similar objects outside the target, and some failed to track the target, and our tracker achieved localization of the target; in the second figure, only we localized the bicycle, and the other In the third image, our tracker precisely frames the ship, and the other trackers contain more or less redundant information about the background. In general, our tracker has more robust and accurate results.

Figure 3 Contrast tracking visualization effect

4.4. Heat Map Visualizations

We also visualize the heat map of the features after we mix attention and DWConv, as shown in Figure 4. The larger red values indicate that the model pays more attention to the ground release of the image. The results show that our proposed method can effectively focus the model's attention on the target to be tracked and thus can localize the target more accurately.

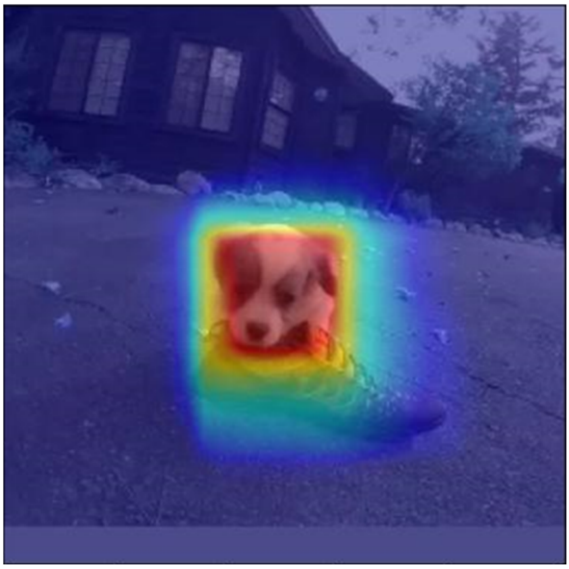

Figure 4 Heat map visualization

5. Conclusion

In this paper, we propose an improved hybrid attention-based target tracking algorithm, and test and experiment on several datasets. The experimental results show that our proposed algorithm has high tracking accuracy and robustness, and can well solve the target tracking problem in complex scenes. We propose a feature fusion method based on hybrid attention and deep convolution, which saves memory expenditure while ensuring accuracy. We have conducted comparative experiments on several datasets, and the results show that our model outperforms other mainstream trackers and achieves real-time tracking.

References

[1] Bertinetto L,Valmadre J, Henriques J F,Vedaldi A, and Philip H S Torr 2016 Fully-convolutional siamese networks for object tracking. In ECCVW.

[2] Bhat G, Danelljan M, Luc Van Gool, and Radu Timofte 2019 Learning discriminative model prediction for tracking. In Proceedings of the IEEE/CVF International Conference on Computer Vision,ICCV, pages 6182–6191

[3] Chen X, Yan B, Zhu J W, Wang D,Yang X and Lu H C 2021 Transformer tracking. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)
[4] Chen Z, Zhong B, Guorong Li, Zhang S P, and Ji R R 2020 Siamese box adaptive network for visual tracking. In IEEE Conference on Computer Vision and Pattern Recognition, CVPR
[5] Cui Y T, Jiang C, Wang L M, and Wu G S 2022 Fully convolutional online tracking. Computer Vision and Image Understanding, 224:103547
[6] Danelljan M,Bhat G,Shahbaz F K and Felsberg M 2018 ATOM: accurate tracking by overlap maximization. In IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2019.
[7] Bo Li, Junjie Yan, Wei Wu, Zheng Zhu, and Xiaolin Hu. High performance visual tracking with siamese region proposal network. In CVPR,
[8] Wang Q, Zhang L, Bertinetto L,Hu M and Torr M 2019 Fast online object tracking and segmentation: A unifying approach. In: CVPR. pp. 1328–1338
[9] Li B,Wu W,Wang Q,Zhang F Y,Xing J L,and Yan J J 2019 SiamRPN++: Evolution of siamese visual tracking with very deep networks. In CVPR
[10] Zhang Z and Peng H 2019 Deeper and wider siamese networks for real-time visual tracking. In: CVPR. pp. 4591–4600