Transformer-Based Visual Object Tracking

Subjects: Computer Science, Artificial Intelligence
Contributor: Shuai Wang, Genwen Fang, Lei Liu, Jun Wang, Kongfen Zhu, Silas N. Melo

With the rise of general models, transformers have been adopted in visual object tracking algorithms as feature fusion networks. In these trackers, self-attention is used for global feature enhancement. Cross-attention is applied to fuse the features of the template and the search regions to capture the global information of the object. However, studies have found that the feature information fused by cross-attention does not pay enough attention to the object region. In order to enhance cross-attention for the object region, an enhanced cross-attention (ECA) module is proposed for global feature enhancement.

Keywords: visual object tracking; transformer; global feature enhancement; fast self-attention; enhanced cross-attention

1. Introduction

Single-object tracking is pivotal in computer vision and is tasked with estimating an object's position and motion trajectory across frames. Overcoming challenges like size variations, occlusions, and deformations are persistent issues.

In the field of computer vision, advancements have been made in CNN-based tracking algorithms, such as the Siamese networks, particularly in addressing challenges like size variations, occlusions, and deformations. SiamFC is a classic example that employs cross-correlation operations to compute a score map, determining the target's bounding box by convolving the feature maps of the template and search regions. However, this approach, due to significant differences between the score map and the feature map, overlooks the semantic information of the target, hindering the subsequent classification and regression operations. To solve this problem, many transformer-based trackers are used for feature fusion. MixFormer proposes a mixed attention module (MAM), which applies the attention mechanism to simultaneously extract features and exchange information. OSTrack integrates the target information more widely into the search area to better capture their correlation. In Fast re-OBJ, the authors efficiently leverage the intermediate outputs of the instance segmentation backbone (ISB) for triplet-based training, avoiding redundant feature extraction during both ISB and the embedding generation module (EGM).

With the rise of Vision Transformer, self-attention techniques show promise in visual tasks. However, their quadratic computational complexity has prompted research into mitigating this challenge. Sparse global attention patterns and smaller attention windows can reduce costs, but at the risk of overlooking information or sacrificing long-term dependency modeling. In Fast re-OBJ, the authors efficiently leverage the intermediate outputs of the instance segmentation backbone (ISB) for triplet-based training, avoiding redundant feature extraction during both ISB and the embedding generation module (EGM).

2. Transformer-Based Visual Object Tracking

In recent years, convolutional networks have been widely used in computer vision tasks, including object detection, segmentation, and tracking. SiamFC was the first to apply Siamese networks for visual object tracking, which predicts the object states by calculating the distance between the current frame and the template. SiamFC++ is an improved version of SiamFC that introduces a new network architecture and training method, enabling object tracking in different scenarios. SiamMask uses Siamese networks for both object tracking and segmentation simultaneously. SA-Sia utilizes feature mapping to extract spatial information leading to more accurate and robust object tracking. SAT employs a deep Siamese network to assess the similarity between objects in each frame, offering a novel solution for evaluating and addressing the association challenges in consecutive frames. However, the local matching strategy, which is based on cross-correlation, could lead to sub-optimal results, especially when the object is occluded or partially visible. Moreover, semantic information about the object may be lost during the correlation operation, resulting in imprecise object
The transformer is a popular neural network architecture that is widely used in natural language processing (NLP) tasks. It consists of attention-based encoders and decoders. Self-attention, which is the main module of a transformer, can compute representations of input sequences. It allows each position in the input sequence to focus on the other positions and calculate the weighted averages of their values. DERT and ViT are early methods for introducing transformer models into the field of computer vision. In visual object tracking fields, transformer-based tracking methods achieve significant improvements compared to Siamese network-based trackers. The correlation operation in the Siamese network is replaced by a self-attention module from the transformer in TransT to fuse the information between the template and search region. Stark applies an encoder–decoder structure in tracking, where the encoder models the global spatiotemporal feature dependencies between the object and search region, while the decoder learns embedded queries to predict the spatial location of the object. SwinTrack introduced a fully attention-based transformer algorithm for feature extraction and feature fusion, enabling the complete interaction between an object and a search region in the tracking process. The encoder–decoder framework based on transformers is widely applied to sequence prediction tasks, as self-attention models the interaction between elements in the sequence. Based on these works, researchers find that traditional transformer methods struggle to distinguish between the object to be tracked and similar interfering objects. To enhance the object information within the fused features, the ECA module is proposed to boost the attention on tracking the object. The ECA module can calculate the average attention level for each position in the fused feature sequence and weigh the positions with higher attention levels. This enhances the feature information in the object region and improves matching accuracy.

In recent years, a series of new attention mechanisms, such as fast attention, have garnered research interest. Sparse attention reduces computational costs by constraining attention weights to consider only relationships within a neighborhood. Shaw et al. proposed an attention mechanism to embed relative positional information into attention calculations, allowing the model to better handle dependencies in long sequences. This mechanism has shown excellent performance in tasks involving long sequences, such as text generation. Orthogonal random features are proposed to achieve fast attention operations. It can decompose the attention matrix into the product of random nonlinear functions between the original query and key, avoiding the explicit construction of a quadratic-sized attention matrix.

References


Retrieved from https://encyclopedia.pub/entry/history/show/118652