

Transformer Tracking with Controlled Attention for Background Suppression

Janani Kugarajeevan^{1,2,*}, Thanikasalam Kokul³, Amirthalingam Ramanan³, Subha Fernando²

¹Department of Computational Mathematics, Faculty of Technology, University of Moratuwa, Sri Lanka

²Department of Interdisciplinary Studies, Faculty of Technology, University of Jaffna, Sri Lanka

³Department of Computer Science, University of Jaffna, Sri Lanka

*jananitha@univ.jfn.ac.lk

Visual object tracking aims to determine the state of a target object across video frames, starting from its initial position in the first frame. Recently, one-stream Transformer-based trackers have demonstrated impressive performance by jointly processing template and search region tokens through attention mechanisms, allowing effective extraction of target-specific features. However, the high proportion of background in the search region can weaken target representations and reduce the tracker's ability to accurately discriminate target from the background. To address this limitation, we propose a Transformer-based tracking framework with a controlled attention mechanism that suppresses irrelevant attention from background search tokens and enhances target-focused feature learning. In early encoder layers of proposed tracker, attention from search tokens to template tokens is blocked to prevent background interference, while template-to-search attention is allowed for target-specific feature extraction. Following the early encoder layers, a learnable module is introduced to estimate the probability of each search token belonging to the target, using a multilayer perceptron. Based on the module's predictions, search tokens are divided into target and non-target groups, and attention from selected target tokens to the template tokens is enabled to enhance target-focused feature learning. Experiments on GOT-10k dataset demonstrate that our tracker outperforms recent state-of-the-art methods, achieving an average overlap of 73.8%, a success rate of 84.1% at a threshold of 0.5, and a success rate of 70.1% at a threshold of 0.75, while maintaining real-time processing of 39 frames per second.

Keywords: Transformer Tracking, Background Interference, Controlled Attention, Visual Object Tracking