Online Resource 1

June 6, 2020

1 Overall Architecture

The paper introduces a modular neural network architecture namely LangTrack that integrates various independent modules to supervise the tracking. The goal of this work is not to redesign the visual trackers. Instead, we would like to explore the user-intention guided tracking method by bringing in language prior. Our method is motivated by the good performance of fast online tracking model SiamMask [2] trained offline and success of visual referring expression. We aim at combining the advantages of visual representation and language embedding to refine the discriminative capacity for object tracking task. Furthermore, we consider this study not just about the cross-modal learning task by combining of CV and NLP, but knowledge-prior guided CV task, as NLP is one expression form of knowledge. This study tries to validate the effectiveness of methodology that human knowledge has its incomparable value to assist specific CV tasks. In reality, it is important for combination of the knowledge human gained and algorithm machine learned to realize applications like video or image analysis, robotics and automatic driving in real life. Besides, we also learn by our study that integration of multimodal information can greatly improve the robustness and effectiveness of the model, which could be adapted in more complex scenarios.

Four modules have been utilized in our network, visual siamese tracking module (VTM), language guide module(LGM), temporal supervision module(TSM) and discriminative integrated module(DIM). Fig.1 illustrates our overall architecture. In terms of initial input, we obtain the visual template from ground truth in the first frame of video, which is feed into visual tracking module with subsequent frames for visual feature extraction. Meanwhile, the template is an input for temporal supervison module to predict the target in the next search frame. In addition, a query describing the target and the search frames are put into language guide module to generate the global context attention for each image.

Visual siamese tracking module aims at extracting visual feature of target template with ResNet-50 and calculating the match scores of search frames of video with Depth-wise cross correlation, and then the module proposes the candidate anchor boxes by positive and negative binary classification and regression branch. Language guide module treats natural language embedding as high-level semantic information to generate the target-driven attention for robust tracking. The word embedding and hidden state are cancatenated as language feature r_t, and feed into dynamic filters.

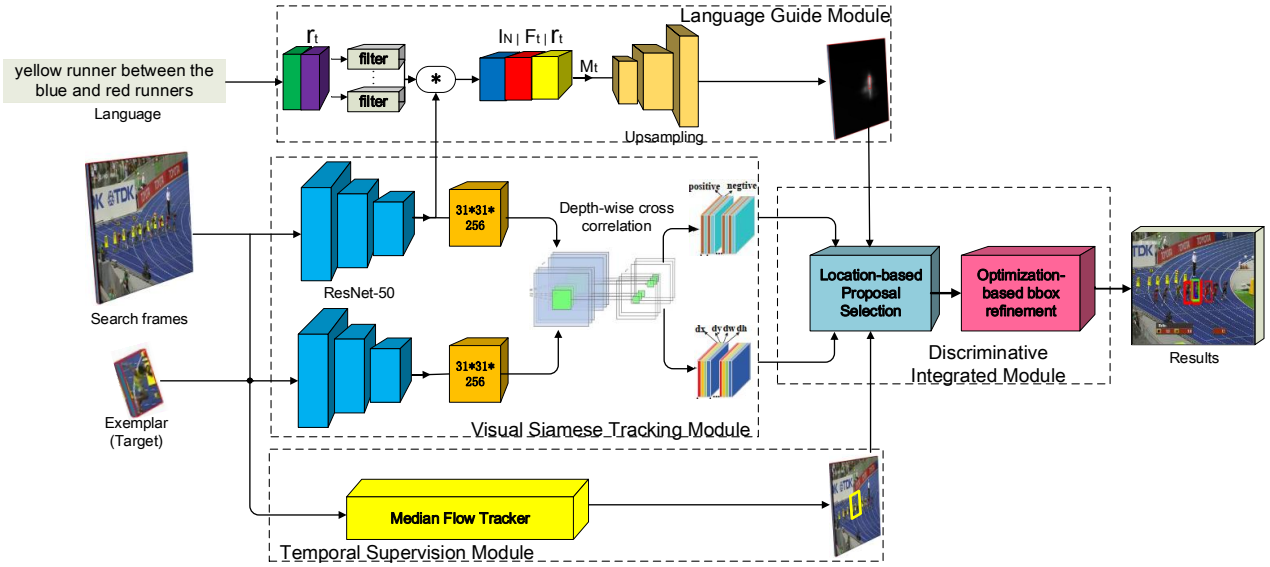

Figure 1: Architecture overview of our LangTrack model, involving four modules: Visual Siamese Tracking Module(VSTM),Language Guide Module(LGM),Temporal Supervision Module(TSM),Discriminative Integrated Module(DIM).

Each output of filters convolves with the visual feature I_N to produce F_t, which combine with visual feature I_N and language feature r_t to generate multimodal response map M_t. An upsampling function is utilized to predict the attention.

Temporal supervision module use optical flow feature with Median flow tracker to predict the target location of current frame based on the previous one, to make use of inter-frame information.

Figure 2: Red bboxes are 50 proposals by VSM, yellow one is from TSM, blue one is from LGM and green one is the final result.

Finally, all the outputs from three branches above are regarded as the inputs of discriminative integrated module to generate the best bounding box of target. Discriminative integration module

synthesizes results of three branches, and introduce a method inspired by IoUnet [1] to single out the best bounding box from candidate proposals. The location-based proposal selection aims to re-rank the proposals based on both classification and location confidence while optimization-based bbox refinement use the neigbouring proposals to refine the bounding box. The result of each module has been demonstrated as Fig.2. The language query of the Fig is *yellow runner between the blue and read runners*.

References

[1] Borui Jiang, Ruixuan Luo, Jiayuan Mao, Tete Xiao, and Yuning Jiang. Acquisition of localization confidence for accurate object detection. 2018.

[2] Qiang Wang, Li Zhang, Luca Bertinetto, Weiming Hu, and Philip H. S. Torr. Fast online object tracking and segmentation: A unifying approach. 2018.