Online Resource 2

June 1, 2020

1 IoU-based bounding box refinement

We find the highest classification score in RPN [3] cannot always reflect the best bounding box when tracking, Traditional NMS(Non-Maximum Suppression) selects the bounding box with maximum classification confidence and eliminates the neighboring boxes. Soft-NMS [1] replace the elimination of boxes by the decrement of confidence to improve the accuracy. Recently, a number of learning-based methods have been proposed as alternatives to the parameter-free NMS and Soft-NMS. [2] finds the misalignment between classification confidence and localization accuracy, and propose the IoUnet to replace classification confidence with the predicted IoU as the ranking keyword in NMS, helping eliminate the suppression failure caused by the misleading classification confidences.

Figure 1: The pink bbox is of the top classification score,which represent the distractors while the green one is the ground truth.

We find that the top classification score has been used in the siamese region proposals networks to localize the final bounding box during tracking, which does not represent the best one. According to Fig.1, the highest score of classification confidence is not the representing the target we are supposed to follow. The most of CNN object detectors depend on the bounding box regression and

non-maximum suppression to localize target, and Intersection over Union (IoU) is the most popular evaluation metric in the benchmarks.

In this paper,we propose to a method to combine the classification and location confidence to figure out the best proposal. In addition, we fine tune the proposal based on neighboring proposals to optimize the final bounding box. When we obtain the three kinds of predicted results from visual, lingual, and inter-frame features, it is important to integrate all of them to achieve better result compared with a single one. Inspired by IoU-guided NMS [2], we work out an approach to integrate the results from three branches by following two step. We first select the best proposal of visual siamese tracking module under the supervision of language-guided and optical flow branches, then an optimization algorithm has employed to refine the chosen bounding box for achieving higher accuracy based on other neighboring high ranking proposals.

The method is illustrated as Algorithm 1 as below.

Algorithm 1 IoU-based Bounding Box Localization and optimization-based refinement.

Input: $B = [b_1, ..., b_n]$,S_{cls},B_{lang},B_{opt},N,M;
N is the number of top bboxes selected from B based on classification. M is the number of neighbouring bboxes used to optimize the best bbox.

Output: p_{final},which denotes the final bbox.

Given $b \in B$,$s_{cls} \in S_{cls}$,B_{lang},B_{opt},N,M;
N>M,$w_{temp} \leftarrow 0$,$s_{temp} \leftarrow 0$;
B_N :top N bboxes based on classification score S_{cls};
$B_N \leftarrow [b_1, ..., b_n]$;
$b_n \leftarrow [cx, cy, w, h]$;
while each $b \in B_N$ **do** $s_n \leftarrow$ IoU(b,B_{lang})+IoU(b,B_{opt})/2
    $s_{max} \leftarrow$ argmax(S_N)
    $p_{best} \leftarrow$ B(s_{max})
    B_M :top M bboxes based on location score S_{loc}
    $B_M \leftarrow [b_1, ..., b_M]$
    for i=2 $\rightarrow$ M **do**
        $w_i \leftarrow s_i \times b_i$
        $w_{temp} \leftarrow w_i$+w_{temp}
        $s_{temp} \leftarrow s_i$+s_{temp}
    end for
end while
$p_{ad} \leftarrow w_{temp}/s_{temp}$;
$p_{final} \leftarrow (p_{best}$+$p_{ad})/2$;
return p_{final};

References

[1] Navaneeth Bodla, Bharat Singh, Rama Chellappa, and Larry S. Davis. Soft-nms – improving object detection with one line of code. 2017.

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

[3] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: towards real-time object detection with region proposal networks. In *International Conference on Neural Information Processing Systems*, 2015.