A. HOI-NMS & HOI-SoftNMS

NMS [6] and its variants [1] can obtain the better performance when mean-Average-Precision (mAP) is used as an evaluation metric and are therefore employed in state-of-the-art object detectors [2,4,5]. However, these approaches cannot be applied directly to the HOI detection since there are a pair of bounding boxes of the human and the object in one HOI prediction. Motivated by the previous [1,6], we present HOI-NMS and HOI-SoftNMS methods to further improve the performance by reducing the number of duplicate HOI predictions in pair-wise level. The pseudo code and the comparison are illustrated as the following.

Algorithm 1 HOI-NMS & HOI-SoftNMS

Input: \( H = \{ h_i = \{ b^H_i, b^O_i, c^O_i, c^I_i \} \}, S = \{ s^I_i \}^N \)

\( H \) is the list of original HOI predictions;

\( b^H_i, b^O_i \) are the bounding boxes of the human and the object of the \( i \)-th HOI prediction;

\( c^O_i, c^I_i \) are the categories of the object and the interaction of the \( i \)-th HOI prediction;

\( s^I_i \) is the interaction score of the \( i \)-th HOI prediction.

1. while \( H \neq \emptyset \) do
2. \( m \leftarrow \arg \max S \)
3. \( D \leftarrow D \cup h_m; H \leftarrow H - h_m \)
4. for \( h_i \in H \) do
5. \( \text{Calculate } \text{IoU}_{\text{hoi}}(h_i, h_m) \) based on Formula 2;
6. if NMS and \( \text{IoU}_{\text{hoi}} > t_{\text{iou}} \) then
7. \( s_i = 0 \)
8. else if NMS and \( \text{IoU}_{\text{hoi}} > t_{\text{iou}} \) then
9. \( s_i = s_i \cdot e^{-\frac{\text{IoU}_{\text{hoi}}^2}{\text{IoU}_{\text{hoi}}^2 - t_{\text{iou}}}} \)
10. else
11. \( s_i = s_i \)
12. end if
13. end for
14. end while

Output: \( D, S \)

Based on \( \text{iou} \) 1 used in the object detectors, we calculate \( \text{IoU}_{\text{hoi}} \) in line 5 to indicate the extent of overlap between two HOI predictions where the Formula 2 shows the calculating process.

\[
\text{iou}(b_i, b_m) = \frac{b_i \cap b_m}{b_i \cup b_m}
\]

\[
\text{IoU}_{\text{hoi}}(h_i, h_m) = \begin{cases} 
0, & \text{If } (c^O_i \neq c^O_m \text{ or } c^I_i \neq c^I_m); \\
\min(\text{iou}(b^H_i, b^H_m), \text{iou}(b^O_i, b^O_m)), & \text{Else};
\end{cases}
\]

Table 1. Comparison against different NMS strategies with their own best performance. Results (Row1/4 vs. Row2/3/5/6) indicate that the performance could be further improved when NMS strategy is employed. Compared with the original CATN with fastText [3] in line 4, HOI-SoftNMS improves mAP-full from 31.62 to 32.40, which obtains the best performance.

Table 2. The effect of different settings of \( t_{\text{iou}} \) on the HICO-DET dataset. We conduct experiments based on the CATN with fastText [3] and evaluate them on the mAP-full metric.

B. Complexity & Effectiveness

Our innovative method can be regarded as a ‘Plug-and-Play’ module that could be readily implemented to effectively promote the performance of transformer-based HOI detection models. Though the complexity of our method may slightly increase compared to baseline models, it is controllable in practical scenes. As shown in Table 3, the inference time only increases 2.4ms per image when a lightweight detector (e.g. Yolov5m) and the parallel architecture are adopted.

Table 3. Our method could obtain significant improvement with the controllable increase of complexity. The inference time only increases 2.4ms per image on one RTX3090 when a lightweight detector, e.g. Yolov5m, is adopted in parallel.
Figure 1. Visualization of a sample from HICO-DET. The human bounding boxes, object bounding boxes, object classes, and verb classes are drawn with blue boxes, yellow boxes, yellow characters, and red characters respectively.

Figure 2. Visualization of decoder attention maps of QPIC (Baseline) v.s. CATN (Ours). As shown in Figure 2(b), our method can lead to more accurate attention so that the aggregated features are more meaningful and more beneficial to interaction classification.

C.1. The effectiveness of CLAM.

To analyze the effectiveness of the proposed category-level attention module (CLAM), we also visualize the attention map of the category of ‘backpack’ in Figure 1(d). The relative regions of both the ‘backpack’ and the interacting human are highlighted in the image, which demonstrates that our CLAM can automatically aggregate the features with rich information of the corresponding category.

C.2. The effectiveness of $Q_{CA}$.

We have conducted experiments on what initialization (all-zeros, random values, and category information) is useful and experiments on where (Object Query, CLAM, and Verb-Classifier) to leverage such information as shown in Tab.5 and Fig.4 in the original manuscript respectively. This research shows that using category info to initialize the Object Query can effectively promote the performance of HOI detection. To clearly explain how the category-aware information contributes to the HOI detection, we also visualize attention maps in transformer-decoder to show that this information and initialization would lead to more accurate attention, as shown in Figure 2.

References


