Supplementary Material for HOTR: End-to-End Human-Object Interaction Detection with Transformers

Bumsoo Kim[1,2] Junhyun Lee[2] Jaewoo Kang[2] Eun-Sol Kim[1,†] Hyunwoo J. Kim[2,†]

[1]Kakao Brain [2]Korea University

{bumsoo.brain, eunsol.kim}@kakaobrain.com

{meliketoy, ljhyun33, kangj, hyunwoojkim}@korea.ac.kr

Summary

In this supplement, we provide more detailed analysis for HOTR and the experiments that has not been provided in the main paper due to the limited space. This includes i) Detailed explanation of the transformer encoder-decoder architecture of HOTR, ii) More experimental results of HOTR, and iii) Qualitative results for HOTR.

1. Transformer Architecture of HOTR

Figure 1 shows a more detailed illustration of the overall pipeline of HOTR. Our architecture is composed with three main components: 1) a shared encoder with CNN backbone and multi-head self-attention, 2) a parallel decoder each with a self-attention layer and an encoder-decoder attention layer, and finally 3) the recomposition that gets the bounding box for the human box and object box based on the pointers inferred from the interaction decoder to complete final HOI triplets.

2. Experimental Results for τ

Table 1 provides experimental results for the temperature factor τ. We conducted the experiment with a light-weight version of HOTR (with 1 layer decoder with $d = 1024$). We fixed the other experimental settings with the value that appeared best in this setting, $\tau = 0.1$.

τ	$AP^{\#1}_{\text{role}}$
0.05	52.2
0.08	54.3
0.1	**55.2**
0.5	47.2
1.0	44.9

Table 1. Effect of temperature factor τ for HOTR in the V-COCO test set.

3. Qualitative Analysis for HOTR

Here, we present qualitative results for HOTR. HOTR directly predicts the human region, object region and the class of the interaction at once with a set prediction approach. Given that there are k ground-truth interacting pairs in the image (a single interacting pair might have multiple interactions), the top-k <human,verb,object> triplet with the highest HOI score is visualized. The red box and yellow box denotes all the ground-truth human and object regions that are involved in an interaction, respectively. Since the evaluation metric is based on mAP, we visualize the triplets that are correctly detected within the top-k predictions in green. Note that HOTR can detect various types of interactions including one-to-one, many-to-one (multiple persons interacting with a single object), one-to-many (one person interacting with multiple objects), and many-to-many (multiple persons interacting with multiple objects) relationships.

†corresponding authors

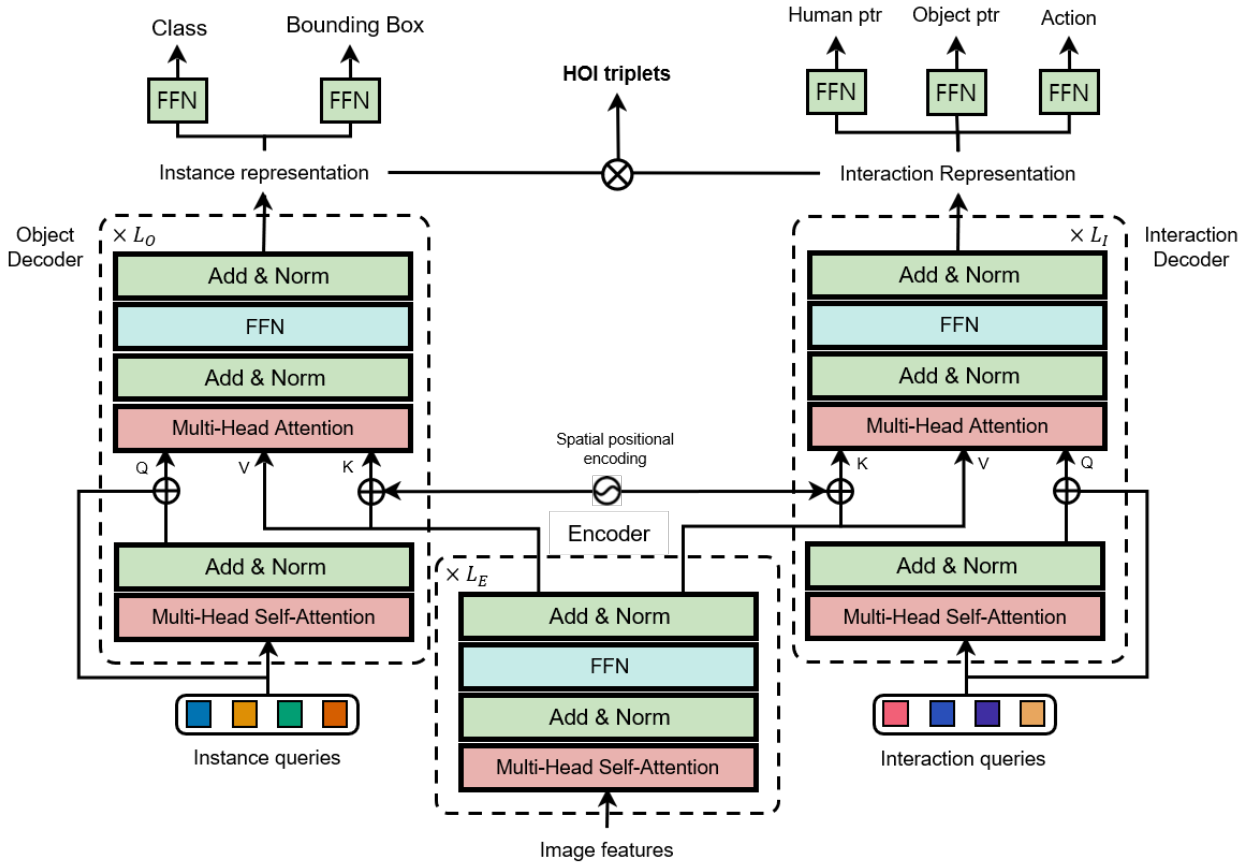

Figure 1. More detailed illustration of the overall pipeline of HOTR. The Instance Decoder and Interaction Decoder runs in parallel, and share the Encoder. The interaction representations predicted by the Interaction Decoder learns to associate the proper ⟨human,object⟩ representations. The final output will be a fixed set of predictions for HOI triplets. The positional encoding is identical to [**?**].

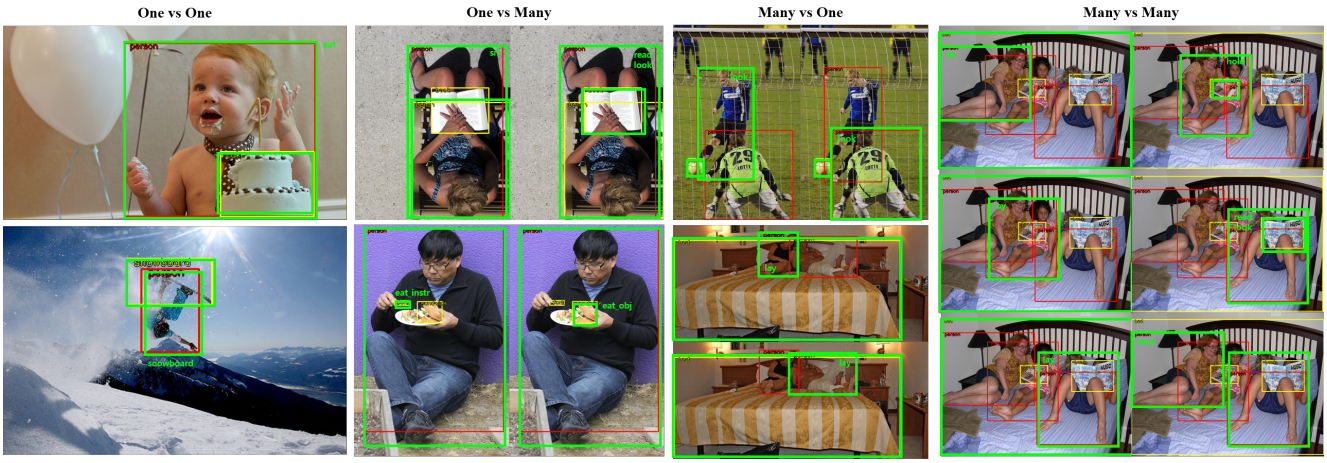

Figure 2. Qualitative Analysis for HOTR. The red box denotes the ground truth human box, yellow box denotes the ground truth object box, and green box denotes the top-k prediction of HOTR that matches the ground truth interaction.