Consistency Learning via Decoding Path Augmentation for Transformers in Human Object Interaction Detection

Jihwan Park[1,2] SeungJun Lee[1] Hwan Heo[1] Hyeong Kyu Choi[1] Hyunwoo J. Kim[1,*]
[1]Department of Computer Science and Engineering, Korea University [2]Kakao Brain
{jseven7071, lapal0413, gjghks950, imhgchoi, hyunwoojkim}@korea.ac.kr
{jwan.park}@kakaobrain.com

Summary

In this supplement, we provide the implementation details of cross-path consistency (CPC) learning and additional experimental results. This includes **(1)** the details of CPC loss and hyperparameters, **(2)** results for other HOI transformers that we did not discuss in the main paper, **(3)** limitations of our work, **(4)** negative social impacts, and **(5)** license information.

1. Implementation Details of CPC

Existing HOI transformers [2, 5, 7, 10] have some variants in output logits. We will shortly explain the details of each model's outputs, and specific CPC losses. For simplicity, we discuss the consistency loss between $\mathcal{P}_k$ and $\mathcal{P}_{k'}$ for further explanation, which is defined in the main paper as:

$$\begin{aligned}\mathcal{L}_{\mathcal{P}_k\mathcal{P}_{k'}} = \ & \lambda_h \cdot \mathcal{L}_h\big(\hat{y}^h_k, \hat{y}^h_{k'}\big) \\ & + \lambda_o \cdot \mathcal{L}_o\big(\hat{y}^o_k, \hat{y}^o_{k'}\big) \\ & + \lambda_{act} \cdot \mathcal{L}_{act}\big(\hat{y}^{act}_k, \hat{y}^{act}_{k'}\big)\end{aligned} \tag{1}$$

To avoid clutter, query index $\tilde{\sigma}_{k,n}$ of each path that is matched to the same ground truth label is omitted.

1.1. CPC loss for QPIC

For each HOI triplet element, the outputs of QPIC on a decoding path $\mathcal{P}_k$ are composed of box regression $\hat{b}^h_k$ for human prediction $\hat{y}^h_k$, box regression $\hat{b}^o_k$ and softmax class probabilities $\hat{c}^o_k$ for object prediction $\hat{y}^o_k$, and multi-label class probabilities $\hat{a}_k$ for interaction prediction $\hat{y}^{act}_k$. For (1), we use the following loss:

$$\begin{aligned}\mathcal{L}_{\mathcal{P}_k\mathcal{P}_{k'}} = \ & \lambda_h \cdot \textbf{MSE}\big(\hat{b}^h_k, \hat{b}^h_{k'}\big) \\ & + \lambda_o \cdot \textbf{MSE}\big(\hat{b}^o_k, \hat{b}^o_{k'}\big) + \lambda^{'}_o \cdot \textbf{JSD}\big(\hat{c}^o_k, \hat{c}^o_{k'}\big) \\ & + \lambda_{act} \cdot \textbf{MSE}\big(\hat{a}_k, \hat{a}_{k'}\big)\end{aligned} \tag{2}$$

where **MSE**, **JSD** denotes mean-squared error and Jenson-Shannon divergence. The loss weights, λ_h,λ_o,$\lambda^{'}_o$,λ_{act}, are set to 2.5,2.5,1, and 1, respectively.

1.2. CPC loss for HOTR

The outputs of HOTR on a decoding path $\mathcal{P}_k$ include softmax class probabilities $\hat{c}^h_k$ for human prediction $\hat{y}^h_k$, softmax class probabilities $\hat{c}^o_k$ for object prediction $\hat{y}^o_k$, and multi-label class probabilities $\hat{a}_k$ for interaction prediction $\hat{y}^{act}_k$. CPC loss for HOTR is

$$\begin{aligned}\mathcal{L}_{\mathcal{P}_k\mathcal{P}_{k'}} = \ & \lambda_h \cdot \textbf{JSD}\big(\hat{c}^h_k, \hat{c}^h_{k'}\big) \\ & + \lambda_o \cdot \textbf{JSD}\big(\hat{c}^o_k, \hat{c}^o_{k'}\big) \\ & + \lambda_{act} \cdot \textbf{MSE}\big(\hat{a}_k, \hat{a}_{k'}\big).\end{aligned} \tag{3}$$

The loss weights, λ_h,λ_o,λ_{act} are set to 1,1, and 10, respectively.

Remarks. The loss weights for CPC are hyperparameters. In our experiments, we simply adopted the loss weights of the baseline HOI detectors, *i.e.*, supervision loss weights. For instance, λ_h, λ_o are the same as the loss weights for bounding box regression in the baseline HOI detector. We believe that this is a good starting point for hyperparameter tuning.

1.3. Weight scheduler for total CPC loss

As we mentioned in the main paper, we use a weight scheduler $w(t)$ for the total CPC loss. Weight coefficient increases for the first few epochs (t_{max}) along the sigmoid-shaped function $e^{-0.5(1-x)^2}$, and then the maximum value λ is maintained. The ramp-up function for weight coefficient can be written as

$$w(t) = \lambda \cdot e^{-0.5(1-\min(1,t/t_{\text{max}}))^2}. \tag{4}$$

Table 1, and 2 show λ, t_{max} used for HOTR and QPIC.

2. Comparison with other HOI transformers

We present the experiment results on additional HOI transformers that we did not discuss in the main paper to

[*]corresponding author.

Dataset	epoch	λ	$t_{\max}$
V-COCO	90	1	30
HICO-DET	50	0.2	20

Table 1. **Weight scheduler settings for HOTR**

Dataset	epoch	λ	$t_{\max}$
V-COCO	90	0.2	30
HICO-DET	90	0.2	30

Table 2. **Weight scheduler settings for QPIC**

show the effectiveness of our method in Table 3, 4. Our experiments show that our CPC learning consistently improves the performance of other transformer-based HOI detectors (*e.g.*, HoiT, and AS-Net) on V-COCO and HICO-DET.

	V-COCO		
Method	epoch	λ	AP_{role1}
HoiT	150	-	48.75*
HoiT + ours	150	0.2	**49.34**

Table 3. **Comparison of our training strategy with HoiT on V-COCO**. * signifies our results reproduced with the official implementation codes of [11].

			HICO-DET		
Method	epoch	λ	Full	Rare	Non-Rare
AS-Net	90	-	28.87	24.25	30.25
AS-Net + ours	90	0.5	**29.13**	**25.18**	**30.31**

Table 4. **Comparison of our training strategy with AS-Net on HICO-DET**.

3. Limitations

Apart from no additional computation at inference, the training complexity of our method scales linearly as the number of paths is augmented. In addition, our method mainly targets the HOI detection task; applications on simpler yet more general tasks as image classification or object detection were not covered in our work. Regarding our decoding-path augmentation method, further discussions on how our methods should be applied to tasks with non-separable output (*e.g.* image classification) are required. This may be one interesting future direction of our work.

4. Negative Social Impact

Human-Object Interaction detection is a task that predicts human behavior and localize it. Although we used the popular and public benchmark datasets, the datasets may not sufficiently contain misrepresented population. Potentially the bias of the datasets may lead to the bias of our model predictions. Also, HOI detection can be abused to illegally monitor individual behaviors.

5. Licence

We implemented our framework in PyTorch. Our implementation is based on HOTR [6], QPIC [8], HOITR [11], and DETR [1] licensed under Apache-2.0 License and AS-Net [3] licensed under MIT License. Datasets used in our experiment are V-COCO [4] and HICO-DET [9] released under MIT License.

References

[1] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020. 2

[2] Mingfei Chen, Yue Liao, Si Liu, Zhiyuan Chen, Fei Wang, and Chen Qian. Reformulating hoi detection as adaptive set prediction. In *CVPR*, 2021. 1

[3] Mingfei Chen, Yue Liao, Si Liu, Zhiyuan Chen, Fei Wang, and Chen Qian. Reformulating hoi detection as adaptive set prediction. `https://github.com/yoyomimi/AS-Net`, 2021. 2

[4] Jitendra Gupta, Saurabh Malik. Visual semantic role labeling. In *CVPR*, 2015. 2

[5] Bumsoo Kim et al. Hotr: End-to-end human-object interaction detection with transformers. In *CVPR*, 2021. 1

[6] Bumsoo Kim, Junhyun Lee, Jaewoo Kang, Eun-Sol Kim, and Hyunwoo J. Kim. Hotr: End-to-end human-object interaction detection with transformers. `https://github.com/kakaobrain/HOTR`, 2021. 2

[7] Masato Tamura et al. Qpic: Query-based pairwise human-object interaction detection with image-wide contextual information. In *CVPR*, 2021. 1

[8] Masato Tamura, Hiroki Ohashi, and Tomoaki Yoshinaga. QPIC: Query-based pairwise human-object interaction detection with image-wide contextual information, 2021. 2

[9] Bingjie Xu, Yongkang Wong, Junnan Li, Qi Zhao, and Mohan S Kankanhalli. Learning to detect human-object interactions with knowledge. In *CVPR*, 2019. 2

[10] Cheng Zou, Bohan Wang, Yue Hu, Junqi Liu, Qian Wu, Yu Zhao, Boxun Li, Chenguang Zhang, Chi Zhang, Yichen Wei, et al. End-to-end human object interaction detection with hoi transformer. In *CVPR*, 2021. 1

[11] Cheng Zou, Bohan Wang, Yue Hu, Junqi Liu, Qian Wu, Yu Zhao, Boxun Li, Chenguang Zhang, Chi Zhang, Yichen Wei, and Jian Sun. End-to-end human object interaction detection with hoi transformer. `https://github.com/bbepoch/HoiTransformer`, 2021. 2