VLMs and LLMs Can Help Detect Human-Object Interactions with Weak Supervision

Mesut Erhan Unal Adriana Kovashka

University of Pittsburgh

School of Computing and Information

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Introduction

In this work, we tackle HOI detection with the weakest supervision setting in the literature, using only image-level interaction labels, with the help of a pretrained vision-language model (VLM) and a large language model (LLM). We first propose pruning non-interacting human/object proposals, exploiting the grounding capability of the vision-language model (VLM). Second, we use a large language model (LLM) to query which interactions are possible given an object category to restrict model's output space. Lastly, an auxiliary weakly-supervised preposition prediction task will make the model explicitly reason over spatial space.

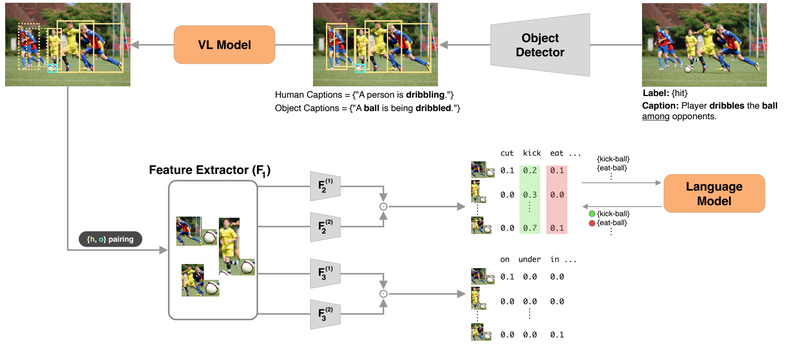

Figure 1. An overview of our approach during training. After retrieving human and object proposals from an object detector, our method first prunes non-interacting human/object proposals with the help of a VLM, calculating an interaction score for each proposal. Next, we pair the remaining human-object proposals and run those pairs through a two-stream feed-forward neural net (F_2) that operates on F_1's output space. Finally, image-level predictions are calculated by summing F_2's output over region pairs. We query an LLM to restrict our model's output space only to meaningful interactions. In order to improve our model's spatial reasoning capability, we formulate a weakly-supervised preposition prediction task wherein supervision comes from prepositions extracted from captions. During inference, we drop the proposal pruning and preposition prediction modules, requiring only an image to detect HOI instances.

Approach

- **Weakly-supervised HOI detection:** Inspired by weakly-supervised object detection (WSOD) literature, we formulate our task as a multiple instance learning (MIL) problem. Similar to (1), we split the classification layer (F_2 in Fig. 1), into a two-stream head (i.e. $F_2^{(1)}$ and $F_2^{(2)}$) where one models the distribution over interaction set given a human-object pair while the other models the distribution over human-object pairs given an interaction class.
- **Pruning non-interacting proposals:** To identify interacting human-object pairs among a large candidate pool, we propose to exploit the implicit grounding capability of a VLM. We manually build a set of human and object captions using verbs and nouns extracted from the original image captions. We then run these newly created captions through a VLM to identify the image regions that are possibly tied to an interaction. Finally, we prune non-interacting human and object proposals based on this grounding information.
- **Suppressing implausible interactions:** Previous work (2; 3) has shown that it can be beneficial to restrict output space only to meaningful interactions, conditioning on some type of lookup table in which plausible interactions are encoded. In this work, we use an LLM as knowledge base, hypothesizing it would learn natural co-occurrences throughout training. For an object class o, we plug "what does a person do with o?" as the question and interaction classes as the answer set to an LLM fine-tuned for multiple choice question answering. We then use the output probability distribution over interaction classes as our lookup table.
- **Weakly-supervised preposition prediction:** Prior work (4) demonstrates that encoding discrete pairwise spatial relations (e.g. inside of, contains) improves performance on tasks that require explicit spatial understanding, such as TextVQA. Inspired by this, we formulate a preposition prediction task in which pairwise features are mapped to discrete spatial labels in weakly-supervised manner, in the unique context of HOI detection. For this task, we employ a two-stream head (F_3 in Fig. 1) similar to our weakly-supervised HOI detection formulation.
- **Extracting interaction labels from captions:** Our learning procedure requires image-level ground-truth interaction labels for supervision. However, one can utilize captions to extract such labels to further relax the level of supervision. We demonstrate that it is possible to learn an HOI detector on image-caption pairs scraped from the web (i.e., Conceptual Captions).

Experiments

We use the well-established HOI detection benchmark datasets, HICO-DET and V-COCO, in our experiments. We use weakly-supervised adaptation of SCG (3) as our baseline since it is one of the best performing fully-supervised two-stage HOI detector with a publicly-available implementation.

Method	Sup.	Backbone	Role AP
VSGNet (5)	Full	RN152	57.00
SCG (3)	Full	RN50 FPN	58.02
IDN (6)	Full	RN50	60.30
HOTR (7)	Full	RN50+Transformer	**64.40**
MX-HOI (8)	Weak+	RN101	-
AlignFormer (9)	Weak+	RN50	**14.15**
Baseline (3)	Weak	RN50 FPN	20.05
Ours	Weak	RN50 FPN	**29.59**
Ours-CC	Weak-	RN50 FPN	**17.71**

Method	Sup.	Backbone	mAP
VSGNet (5)	Full	RN152	19.80
SCG (3)	Full	RN50 FPN	21.85
IDN (6)	Full	RN50	23.36
HOTR (7)	Full	RN50+Transformer	**23.46**
MX-HOI (8) †	Weak+	RN101	16.14
AlignFormer (9) †	Weak+	RN50	**19.26**
Baseline (3)	Weak	RN50 FPN	7.05
Ours	Weak	RN50 FPN	**8.38**

Our method improves absolute 9.54% over weakly-supervised variant of SCG and absolute 15.54% over AlignFormer, which uses stronger supervision in the form of image-level <interaction, object> labels (above left). Our method trained on the Conceptual Captions (Ours-CC) also surpasses AlignFormer and performs comparable to weakly-supervised SCG, even though we extract image-level interaction labels from captions. Unsurprisingly, AlignFormer was not affected heavily by increased combinatorial complexity over the <interaction, object> joint space on HICO-DET thanks to its stronger supervision than ours (above right). Also worth noting is that both AlignFormer and MX-HOI use an object detector fine-tuned on HICO-DET (denoted by †) while we do not.

Method	Agent AP (Δ)	Role AP (Δ)
Baseline (3)	32.41	20.05
+Pruning	33.88 (+1.47)	21.80 (+1.75)
+Suppressing	37.04 (+4.63)	28.28 (+8.23)
+Preposition	40.53 (+8.12)	29.59 (+9.54)

Method	mAP (Δ)
Baseline (3)	7.05
+Pruning	7.55 (+0.50)
+Suppressing	7.81 (+0.76)
+Preposition	8.38 (+1.33)

Method	Agent AP (Δ)	Role AP (Δ)
Baseline (3)	17.71	14.33
+Pruning	19.44 (+1.73)	15.95 (+1.62)
+Suppressing	20.00 (+2.29)	18.23 (+3.90)
+Preposition	20.75 (+3.04)	17.71 (+3.38)

To demonstrate the effectiveness of our contributions, we incrementally ablate them over the baseline weakly-supervised SCG on V-COCO (left), HICO-DET (center) and Conceptual Captions (right). While all of our contributions clearly improve the performance over the baseline, results also show that caption-dependent parts of our method (proposal pruning and preposition prediction) are not affected heavily from the caption source.

References

[1] H. Bilen and A. Vedaldi, "Weakly supervised deep detection networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2846–2854, 2016.

[2] T. Gupta, A. Schwing, and D. Hoiem, "No-frills human-object interaction detection: Factorization, layout encodings, and training techniques," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9677–9685, 2019.

[3] F. Z. Zhang, D. Campbell, and S. Gould, "Spatially conditioned graphs for detecting human-object interactions," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 13319–13327, 2021.

[4] Y. Kant, D. Batra, P. Anderson, A. Schwing, D. Parikh, J. Lu, and H. Agrawal, "Spatially aware multimodal transformers for textvqa," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*, pp. 715–732, Springer, 2020.

[5] O. Ulutan, A. Iftekhar, and B. S. Manjunath, "Vsgnet: Spatial attention network for detecting human object interactions using graph convolutions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13617–13626, 2020.

[6] Y.-L. Li, X. Liu, X. Wu, Y. Li, and C. Lu, "Hoi analysis: Integrating and decomposing human-object interaction," *Advances in Neural Information Processing Systems*, vol. 33, pp. 5011–5022, 2020.

[7] B. Kim, J. Lee, J. Kang, E.-S. Kim, and H. J. Kim, "Hotr: End-to-end human-object interaction detection with transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 74–83, 2021.

[8] S. K. Kumaraswamy, M. Shi, and E. Kijak, "Detecting human-object interaction with mixed supervision," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 1228–1237, 2021.

[9] M. Kilickaya and A. W. Smeulders, "Human-object interaction detection without alignment supervision," in *British Machine Vision Conference*, 2021.