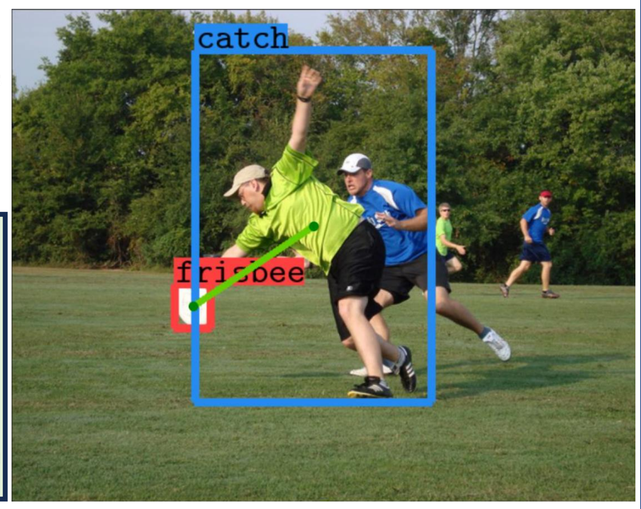
Boosting Zero-Shot HOI Detection with Vision-Language Transfer

Sandipan Sarma, Pradnesh Kalkar, Arijit Sur
Department of CSE, Indian Institute of Technology Guwahati, India

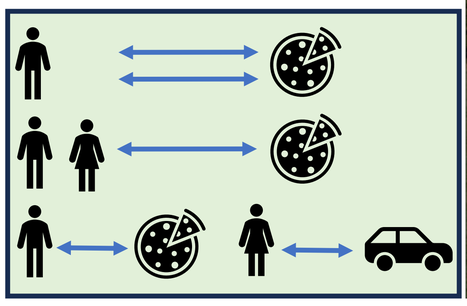
Human-Object Interaction Detection

Objectives

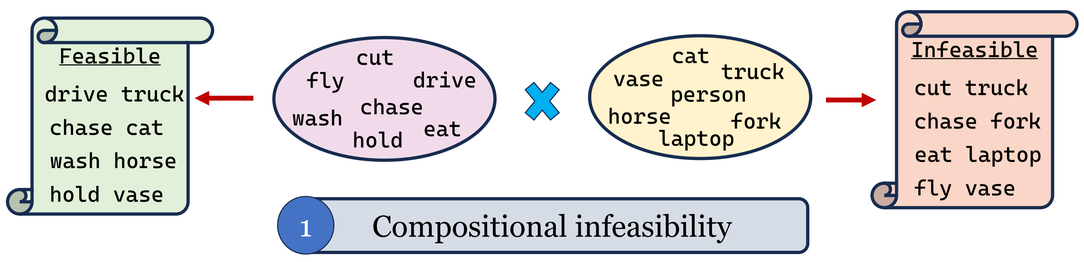
- ❑ Localize interactive human-object pairs
- ❑ Recognize interactive object
- ❑ Recognize interaction class



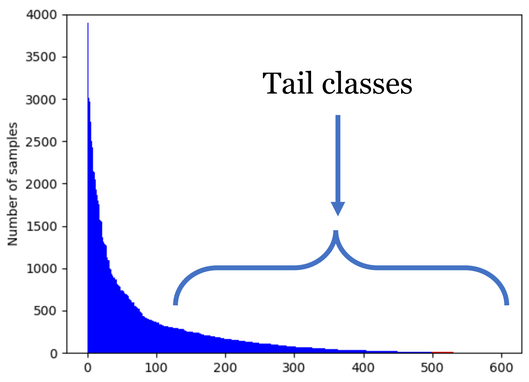
Scenarios to handle



Problems zero-shot learning can mitigate in HOID



1 Compositional infeasibility



ride bicycle

ride elephant

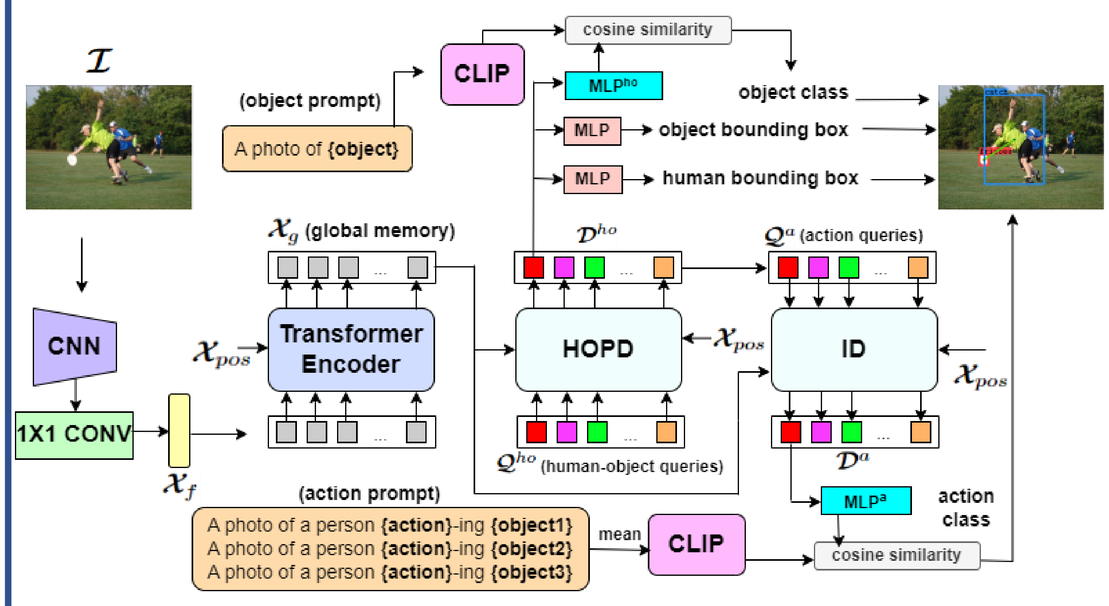
ride boat

ride skateboard

2 Long-tailed distribution

3 Polysemy of verbs

Proposed Framework



- ❑ **Visual** features : Extracted from ResNet-50 backbone
- ❑ **Semantic** features for **object** : CLIP text encoder (objects are well-defined concepts)
- ❑ **Semantic** features for **verbs** : CLIP text encoder (actions are abstract concepts).
- ❑ **Infeasible combinations** like "*feed chair*" are ignored while picking the different objects that can be combined with the action "*feed*" in order to form an interaction
- ❑ **Transformer encoder** captures global information via self-attention
- ❑ **Human-object pair** is anticipated using a pair decoder
- ❑ **Interaction features** are produced by the interaction decoder
- ❑ **Zero-shot** interaction classification is made possible by combing final interaction features with semantic features from CLIP text encoder

Experiments

Dataset: HICO-DET

- ❑ **600 HOI categories** with human and object bounding-box annotations, and interaction labels
- ❑ **Total images** = 47,776
- ❑ **80 object** categories
- ❑ **117 verb** categories

Splits

Setting	Object	Verb	Interaction
UC	All	All	480 S / 120 U
RF-UC	All	All	480 S / 120 U
NF-UC	All	All	480 S / 120 U
UA	All	95 S / 22 U	500 S / 100 U
UO	68 S / 12 U	All	500 S / 100 U

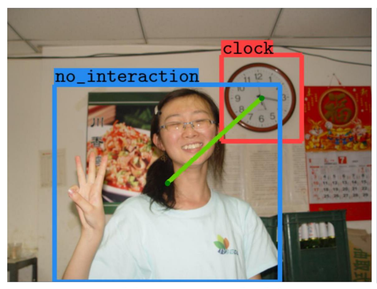
Zero-shot results

Method	Type	Full	Seen	Unseen
VCL baseline [6]	RF-UC	15.56	18.63	3.30
VCL-COCO [6]	RF-UC	16.58	18.84	7.55
VCL-HICODET [6]	RF-UC	21.43	24.28	10.06
ATL [3]	RF-UC	21.57	24.67	9.18
FCL baseline [5]	RF-UC	21.13	24.18	8.94
FCL [5]	RF-UC	22.01	24.23	13.16
THID [8]	RF-UC	22.96	24.32	15.53
Ours	RF-UC	**24.51**	**26.90**	**16.50**
VCL baseline [6]	NF-UC	11.23	12.77	5.06
VCL-COCO [6]	NF-UC	12.76	13.67	9.13
VCL-HICODET [6]	NF-UC	18.06	18.52	16.22
ATL [3]	NF-UC	18.67	18.78	18.25
FCL baseline [5]	NF-UC	18.07	19.22	13.47
FCL [5]	NF-UC	19.37	19.55	18.66
Ours	NF-UC	**20.64**	**20.93**	**19.47**
SHOI [1]	UC	6.26	-	5.62
FG [4]	UC	12.45	12.74	11.31
ConsNet [2]	UC	14.48	14.74	13.46
ZSHOI-AG [3]	UC	11.03	-	9.80
Ours	UC	**23.79**	**24.90**	**19.36**
ConsNet [6]	UA	14.35	14.72	12.50
Ours	UA	**25.78**	**26.67**	**21.36**
FG [4]	UO	13.84	14.36	11.22
ConsNet [2]	UO	14.48	14.67	13.51
ATL baseline [3]	UO	19.33	20.63	12.84
ATL-HICODET [3]	UO	19.36	20.96	11.35
ATL-COCO [3]	UO	20.47	21.54	15.11
FCL baseline [5]	UO	19.45	20.77	12.86
FCL [5]	UO	19.87	20.74	**15.54**
Ours	UO	**21.70**	**23.26**	13.94

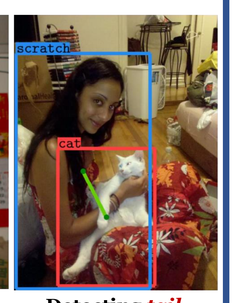
Ablation on text embeddings

Text-Embedding	Type	Full	Seen	Unseen
GAT	RF-UC	21.57	23.56	14.92
CLIP	RF-UC	**24.51**	**26.90**	**16.50**
GAT	NF-UC	18.25	18.37	17.76
CLIP	NF-UC	**20.64**	**20.93**	**19.47**
GAT	UC	21.78	23.25	15.91
CLIP	UC	**23.79**	**24.90**	**19.36**
GAT	UA	18.06	19.77	9.74
CLIP	UA	**25.78**	**26.67**	**21.36**
GAT	UO	19.71	21.21	12.16
CLIP	UO	**21.70**	**23.26**	**13.94**

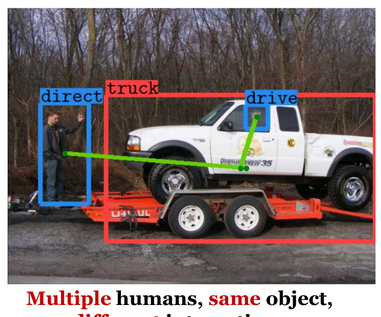
Qualitative results



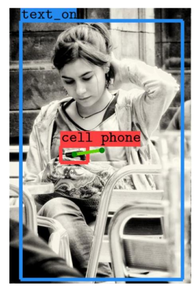
Detected human and object, but **no interaction** detected



Detecting ***tail*** interactions



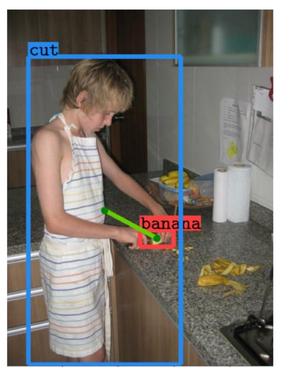
Multiple humans, **same** object, **different** interactions



Interaction with **small** object

Comparative analysis

Failure case of FCL



Our improved results

Takeaways

- ❑ Transformer's **attention** mechanism helps to utilize contextually important cues
- ❑ Joint visual-and-text modeling using **CLIP** helps in generalizing to unseen HOIs
- ❑ The query vectors in our DETR-based framework are vital in projecting an idea about "*what*" visual information about the human-object pairs to look for, with each vector element suggesting "*where*" to look for these pairs within the image. Since the final task is to detect human-object pairs, **unified query vectors** for human-object pairs are important
- ❑ **Improved prompting** to obtain semantic representations of HOI classes from action-object combinations aids in zero-shot HOI detection
- ❑ Despite the unavailability of certain actions and objects (such as in **UA and UO settings**), our method is better at detecting unseen interactions in such challenging settings