

Challenges

Interaction category prediction is limited by

- detection performance (previous two-stage methods)
- predefined interaction locations (union boxes or interaction midpoints of previous one-stage methods)



ect" in yellow, "drive" (b) Interaction midpoints: verb (a) Union boxes: verb ' in purple, matched anchor in red





Aggregated Features



Adaptive Feature Aggregation Adaptive Set-based Ground-truth Matching (c) Our adaptive set prediction method: verb "drive" in purple, "direct" in yellow. Interaction vectors point from human centers to object centers. The features aggregated by queries are visualized at left.



Sunday Full Rare Non-Rare # Full Rare Non-Rare Non-Rare # Full Rare Non-Rare Non-Rare Rare Non-Rare Non-Rare Rare Non-Rare Rare Non-R																													
$ \begin{array}{ c c c c c c c c c c c c c c c c c c c$	Strategy	Full	Rare	Non-Rare	$K \mid$	Full	Rare	Non-Rar	e #Para	ameters	λemb	Full	Rare	Non-Rare	10-10-10 KG KG	eligitati deventation	Finetune	1470		24275 (1.100)	Defau	lt	5-6770-254784	Know Ob	ject				
Embedding 28.65 23.95 30.05 8 28.87 24.25 30.25 16 25.357 M 0.5 28.87 24.25 30.25 15 23.08 29.93 52.537 M 0.5 27.84 21.71 29.67 (more stage Method: (mor	Vector	28.56	24.13	29.88	4	28.21	22.65	29.87	52.4	527 M	0.05	28.31	23.65	29.70	Method	Backbone	Detection	Extra	Time (ms) / FPS	Full	Rare	Non-Rare	Full	Rare	Non-Rare	Method	Backbone	Extra	mAP_{role}
$ \begin{array}{ c c c c c c c c c c c c c c c c c c c$	Embedding	28.65	23.95	30.05	8	28.87	24.25	30.25	52.5	530 M	0.1	28.87	24.25	30.25	Two-stage Method: InteractNet [12]	ResNet-50-FPN	×	×	145 / 6.90	9.94	7.16	10.77	-	-	-	Two-stage Method:			
(a) Matching Strategy: Analysis of different matching strategies, i.e., interaction vector and semantic embeddings.(b) Dimension of Semantic Embeddings: Choice of di- of different settings of loss weight.(c) Weight Coefficient λ_{emb} : The effects 	Combined	28.87	24.25	30.25	10 32	28.36	23.08	30.16	52.	537 M 549 M	0.5	27.84	21./1	29.67	GPNN [34] iCAN [10]	Res-DCN-152 ResNet-50	×	×	- 204 / 4.90	13.11 14.84	9.34 10.45	14.23 16.15	- 16.26	- 11.33	- 17.73	GPNN <i>et al.</i> [26]	Res-DCN-152	×	40.0 44.0
semantic embeddings.Decoder LayersEmbeddingsIA AttentionFullRareNon-Rare#ParametersMon-Rare#Parame	(a) Matching Strategy: Analysis of different (b) Dimension of Semantic Embeddings: Choice of di- matching strategies, <i>i.e.</i> , interaction vector and (c) Weight Coefficient λ_{emb} : The effects of different settings of loss weight.									No-Frills [14] PMFNet [40]	ResNet-152 ResNet-50-FPN	×	P P	494 / 2.02 253 / 3.95	17.18 17.46	12.17 15.65	18.68 18.00	- 20.34	- 17.47	21.20	1CAN [6] DRG [5]	ResNet-50 ResNet-50-FPN	X L	45.3 51.0					
$\frac{1}{14 \text{ Attens}} = \frac{1}{14 \text{ Attens}} $	semantic en	beddings.		I	Decoder Lavers Embeddings IA Attention Euli Dare				Doro	Non Para #Parameters				DRG [9] IP-Net [42]	ResNet-50-FPN Hourglass-104	×	L X	200 / 5.00	19.26 19.56	17.74 12.79	19.71 21.58	23.40 22.05	21.75 15.77	23.89 23.92	VSGNet [30]	ResNet-152	× ×	51.0 51.8	
Basic Model, Int × 6 6× × - 27.52 22.04 29.16 50.94 M H Attn × 6, Int w/o emb × 6 6× × 6× × 6× × 6× × 6× × 6× × 6× × 6× × 6× × 6× × 6× × 6× × 6× × 6× × 6× × 6× × 6× × 6× × 9× 2×.13 M 0.18 M<		_	Single I	Branch	6×		X		25.91	17.88	28.31	41	.44 M		VSGNet [39] PD-Net [47]	ResNet-152 ResNet-152-FPN	×	× L	312 / 3.21	19.80 20.81	16.05 15.90	20.91 22.28	- 24.78	- 18.88	- 26.54	PD-Net [36] FCMNet [22]	ResNet-152-FPN ResNet-50	r L X	52.0 52.6 53.1
$+ Int w/emb \times 6$ $6 \times$ \checkmark $ 27.75$ 22.71 29.25 51.34 M $UnionDet [20]$ $ResNet-50-FPN$ \checkmark $78/12.82$ 17.58 11.72 19.33 19.76 14.68 21.27 $AS-Net*$ $ResNet-50-FPN$ \checkmark 47.37 $+ IA Attn \times 3, Int w/emb \times 3$ $3 \times$ \checkmark $3 \times$ 28.63 23.61 30.13 47.20 M \checkmark $78/12.82$ 17.58 11.72 19.33 19.76 14.68 21.27 $AS-Net*$ $ResNet-50$ \checkmark 35.9 $+ IA Attn \times 6, Int w/emb \times 6$ $6 \times$ \checkmark 3.28 28.63 23.61 30.13 47.20 M Λ $AS-Net*$ $ResNet-50$ \checkmark \checkmark $71/14.08$ 24.40 22.39 25.01 27.41 25.44 28.00 P' , 'L' represent the human pose information and the language $+ IA Attn \times 6, Int w/emb \times 6$ $6 \times$ \checkmark $6 \times$ 28.87 24.25 30.25 52.53 M Λ $AS-Net$ $ResNet-50$ \checkmark \checkmark $71/14.08$ 28.87 24.25 30.25 31.74 27.07 33.14		+ IA	Basic Mod Attn×6, Ii	nt w/o emb×6	6× 6×		x	- 6×	27.52 27.96	22.04 23.01	29.16 29.44	50	.94 M .13 M		DJ-RN [23] One-stage Method:	ResNet-50	X	Р	-	21.34	18.53	22.18	23.69	20.64	24.60	One-stage Method:	DecNet 50 EDN	 	
+ IA Attn \times 3, Int w/ emb \times 33×3×3×3×3×3×3×3×47.20 M+ IA Attn \times 6, Int w/ emb \times 66×4×4×4×4×4×4×4×4×4×1× <td< td=""><td></td><td>+ IA</td><td>+ Int w/ Attn$\times 3$, I</td><td>emb×6 Int w/ emb×6</td><td>$6 \times 6 \times$</td><td></td><td>1</td><td>- 3×</td><td>27.75 28.39</td><td>22.71 24.02</td><td>29.25 29.70</td><td>51 51</td><td>.34 M .94 M</td><td></td><td>UnionDet [20] PPDM-Hourglass [27]</td><td>ResNet-50-FPN Hourglass-104</td><td>1</td><td>×</td><td>78 / 12.82</td><td>17.58</td><td>11.72</td><td>19.33 24.32</td><td>19.76 24.81</td><td>14.68</td><td>21.27</td><td>AS-Net*</td><td>ResNet-50</td><td>×</td><td>53.9</td></td<>		+ IA	+ Int w/ Attn $\times 3$, I	emb×6 Int w/ emb×6	$6 \times 6 \times$		1	- 3×	27.75 28.39	22.71 24.02	29.25 29.70	51 51	.34 M .94 M		UnionDet [20] PPDM-Hourglass [27]	ResNet-50-FPN Hourglass-104	1	×	78 / 12.82	17.58	11.72	19.33 24.32	19.76 24.81	14.68	21.27	AS-Net*	ResNet-50	×	53.9
$\mathbf{t}_{0}\mathbf{t}_{0$		+ IA + IA	Att $n \times 3$, Att $n \times 6$, AttAn \times 6, AttAn	Int w/ $emb \times 3$ Int w/ $emb \times 6$	$3 \times 6 \times$		1	$3 \times 6 \times$	28.63 28.87	23.61 24.25	30.13 30.25	47.52	.20 M .53 M		AS-Net* AS-Net	ResNet-50 ResNet-50	×	×	71 / 14.08 71 / 14.08 71 / 14.08	24.40 28.87	22.39 24.25	25.01 30.25	27.41 31.74	25.44 27.07	28.00 33.14	Table 2. Performance comparison on the V-COCO te 'P', 'L' represent the human pose information and th feature respectively. * denotes freezing the instance d			test set. The the language detection re

CNN

Transformer

Encoder

Backbone

Full Ra	re Non-Rare	K	Full	Rare	Non-Rar	e #Pa	arameters	λomb	Full	Rare	Non-Rare		elatore describe	Finetune	1470			Defaul	t	14770-24780	Know Ob	oject						
8 56 24	13 29.88	4	28.21	22.65	29.87	5	2 527 M	0.05	28.31	23.65	29.70	Method	Backbone	Detection	Extra	Time (ms) / FPS	Full	Rare	Non-Rare	Full	Rare	Non-Rare	Method	Backbone	Extra	mAP _{role}		
8.65 23.	.95 30.05	8	28.87	24.25	30.25	52	2.530 M	0.05	28.87	24.25	30.25	Two-stage Method: InteractNet [12]	ResNet-50-FPN	x	x	145/690	9 94	716	10.77			-	Two-stage Method:					
8.87 24.	.25 30.25	16 32	28.36	23.08	29.93 30.16	52	2.537 M	0.5	0.5 27.84 21.71			GPNN [34]	Res-DCN-152	×	×	-	13.11	9.34	14.23		-	-	InteractNet [8] GPNN et al. [26]	ResNet-50-FPN Res-DCN-152	X X	40.0 44 0		
trategy: A	alvsis of different	(b) Dimension of Semantic Embeddings: Choice of di-						(c) Weight Coefficient λ_{emb} : The effects of different settings of loss weight.			: The effects	iCAN [10] No-Frills [14]	ResNet-50 ResNet-152	×	× P	204 / 4.90 494 / 2.02	14.84 17.18	10.45 12.17	16.15 18.68	16.26 -	- 11.33	17.73	iCAN [6] DRG [5] IP-Net [33]	ResNet-50	X	45.3		
gies, <i>i.e.</i> , interaction vector and mension of semantic embeddings.						B ² 1					eight.	PMFNet [40]	ResNet-50-FPN	×	P	253 / 3.95	17.46	15.65	18.00	20.34	17.47	21.20		ResNet-50-FPN Hourglass-104	L X	51.0 51.0		
dings.		Decoder	Lavers	mbeddings	A Attention	Full	Rare	Non-Ra	e #Par	ameters		IP-Net [42]	ResNet-50-FPN Hourglass-104	×	×	20075.00	19.26 19.56	17.74 12.79	21.58	23.40 22.05	21.75 15.77	23.89	VSGNet [30] PMFNet [31]	ResNet-152 ResNet-50-FPN	X P	51.8 52.0		
Sin	igle Branch	62	×	X	-	25.91	17.88	28.31	41.	.44 M		VSGNet [39]	ResNet-152	×	×	312/3.21	19.80	16.05	20.91	-	-	-	PD-Net [36]	ResNet-152-FPN	L	52.6		
Basic	Model, $Int \times 6$	62	×	X	-	27.52	22.04	29.16	50.	.94 M		DJ-RN [23]	ResNet-50	x	P L P	-	20.81	18.53	22.18	23.69	20.64	24.60	FCMNet [22] One-stage Method:	ResNet-50	X	53.1		
+ IA Allah) + In	nt w/ emb×6	62	×	~		27.96	22.71	29.44	51.	.34 M		One-stage Method:	ResNet-50-FPN	1	×	78 / 12 82	17 58	11 72	10 33	19.76	14 68	21.27	UnionDet [13]	ResNet-50-FPN	X	47.5		
+ IA Attn	$\times 3$, Int w/ emb $\times 6$	62	×	1	3×	28.39	24.02	29.70	51.	.94 M		PPDM-Hourglass [27]	Hourglass-104	1	x	71 / 14.08	21.94	13.97	24.32	24.81	17.09	27.12	AS-Net*	ResNet-50		53.9		
+ IA Attn + IA Attn	\times 3, Int w/ emb \times 3 \times 6, Int w/ emb \times 6	62	×	1	3× 6×	28.63 28.87	23.61 24.25	30.13 30.25	52.	.53 M		AS-Net* AS-Net	ResNet-50 ResNet-50	×	× ×	71 / 14.08 71 / 14.08	24.40 28.87	22.39 24.25	25.01 30.25	27.41 31.74	25.44 27.07	28.00 33.14	'P', 'L' represent the human pose information and the lan feature, respectively, * denotes freezing the instance detection					

(d) Component Analysis: Results of the variants with various components, *i.e.*, interaction branches (Int), instance-aware attention module (IA Attn) and semantic embeddings (emb).

Table 4. Ablation studies of our proposed model on the HICO-DET test set.



Reformulating HOI Detection as Adaptive Set Prediction

Mingfei Chen^{1,3*} Yue Liao^{2*} Si Liu^{2†} Zhiyuan Chen³ Fei Wang³ Chen Qian³ ¹ Huazhong University of Science and Technology ² Institute of Artificial Intelligence, Beihang University ³ SenseTime Research



Solutions

- Applying co-attention to adaptively aggregate interaction-
- relevant features using learnable queries
- Adaptively matching the most suitable ground-truth
- considering both action categories and location distances

Instance-aware Attention

- co-attention manner Semantic embedding

Table 1. Performance comparison on the HICO-DET test set. The 'P', 'L' represent human pose information and the language feature, respectively. * denotes freezing the instance detection related parameters pretrained on the MS-COCO dataset. Our one-stage model with a high inference speed of 71 ms / 14.08 FPS outperforms all previous work by a large margin.

• Involve instructive instance features to interaction branch in

Contributions

- Reformulate HOI detection as set prediction problem, adaptively concentrate on the most suitable features to improve the predicting accuracy
- Propose a novel one-stage transformer-based HOI detection framework (AS-Net)
- Design instance-aware attention module to introduce instance information into the interaction branch
- Maintaining the high efficiency and without any extra features, our method gains 31% relative improvements on HICO-DET, especially 73% on rare HOI categories

• Point to the specific instance more accurately for better matching • Bridging instance branch and interaction branch implicitly

lated parameters pretrained on the MS-COCO dataset.

Attention Visualization



(d) Visual attention of instance-aware attention in (+ IA Attn \times 6, Int w/ emb \times 6).

