Improving Closed and Open-Vocabulary Attribute Prediction Using Transformers

Khoi Pham[1], Kushal Kafle[2], Zhe Lin[2], Zhihong Ding[2], Scott Cohen[2], Quan Tran[2], Abhinav Shrivastava[1]

[1]University of Maryland, College Park [2]Adobe Research

ECCV TEL AVIV 2022

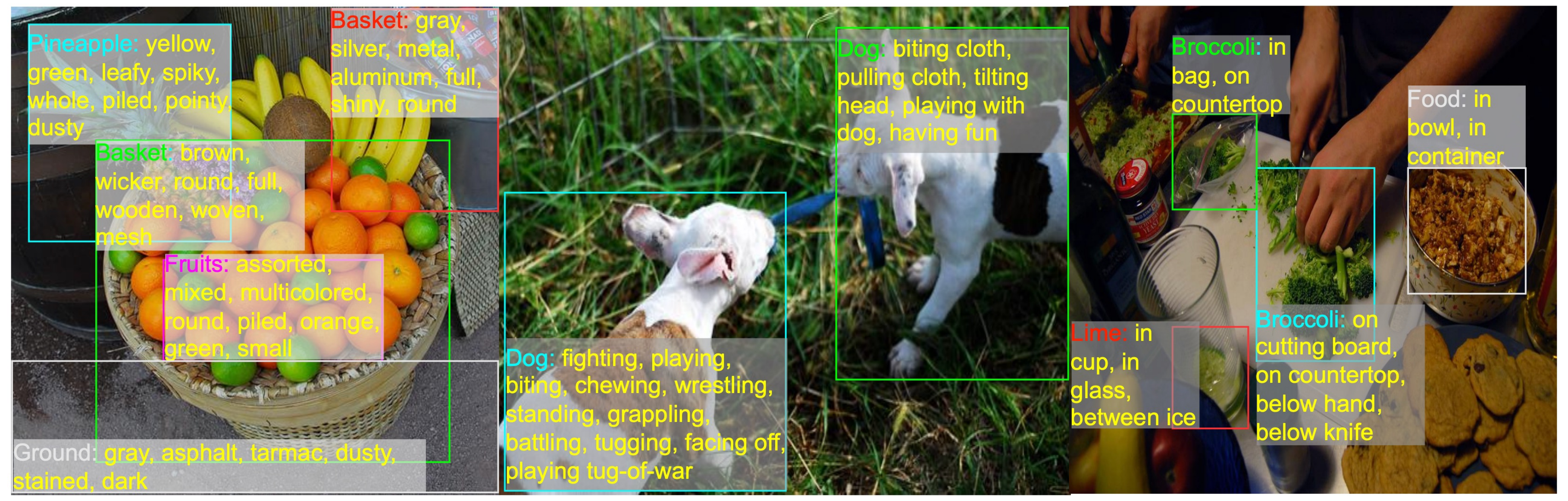
Project page: https://vkhoi.github.io/TAP/

Motivation

Limitations of existing work

- Focus on object physical properties (*adjectives*) and ignore interaction-based properties.
- *Visual relationship detection* study object interactions but require object localization → difficult for large-scale data collection.
- Attributes are abundant in existing image-text datasets but have not been utilized for large-scale attribute learning.

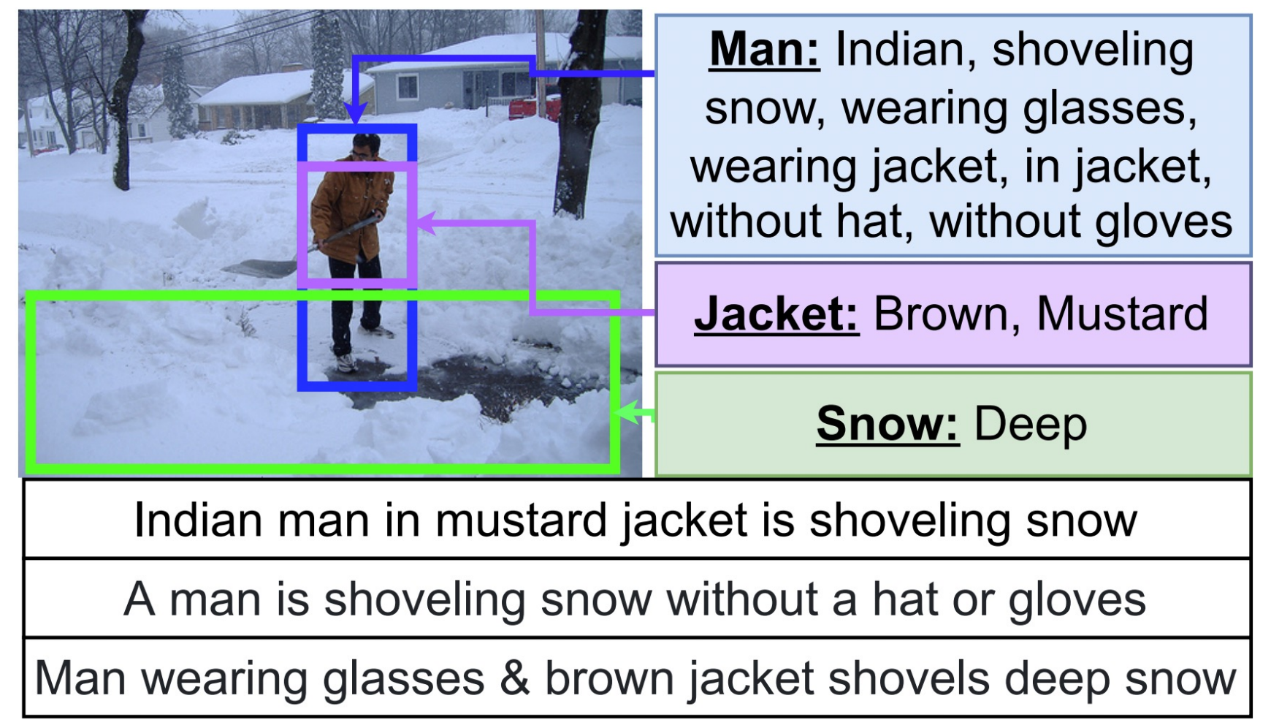
Proposal Large-scale attribute learning from image-text datasets, extendable to open-vocabulary attribute prediction that allows to recognize arbitrary textual attribute phrases.


Pineapple: yellow, green, leafy, spiky, whole, piled, pointy, dusty
Basket: gray, silver, metal, aluminum, full, shiny, round
Basket: brown, wicker, round, full, wooden, woven, mesh
Fruits: assorted, mixed, multicolored, round, piled, orange, green, small
Ground: gray, asphalt, tarmac, dusty, stained, dark
Dog: fighting, playing, biting, chewing, wrestling, standing, grappling, battling, tugging, facing off, playing tug-of-war
Dog: biting cloth, pulling cloth, tilting head, playing with dog, having fun
Broccoli: in bag, on countertop
Food: in bowl, in container
Lime: in cup, in glass, between ice
Broccoli: on cutting board, on countertop, below hand, below knife

Adjective Verb & Verb-Object Preposition-Object

Large-Scale Attribute (LSA) Dataset

Datasets	# images	# instances	# attr annotations	Type of grounding
VG + GQA	108k	6.5M	10.1M	Box
Flickr30K-Entities	32k	285k	503k	Box
MS-COCO + COCO-Attrs	122k	1.2M	2.2M	Ungrounded + Box
Localized Narratives	312k	1.4M	1.7M	Mouse trace
Total	420k	9.5M	14.6M	



Man: Indian, shoveling snow, wearing glasses, wearing jacket, in jacket, without hat, without gloves

Jacket: Brown, Mustard

Snow: Deep

Indian man in mustard jacket is shoveling snow

A man is shoveling snow without a hat or gloves

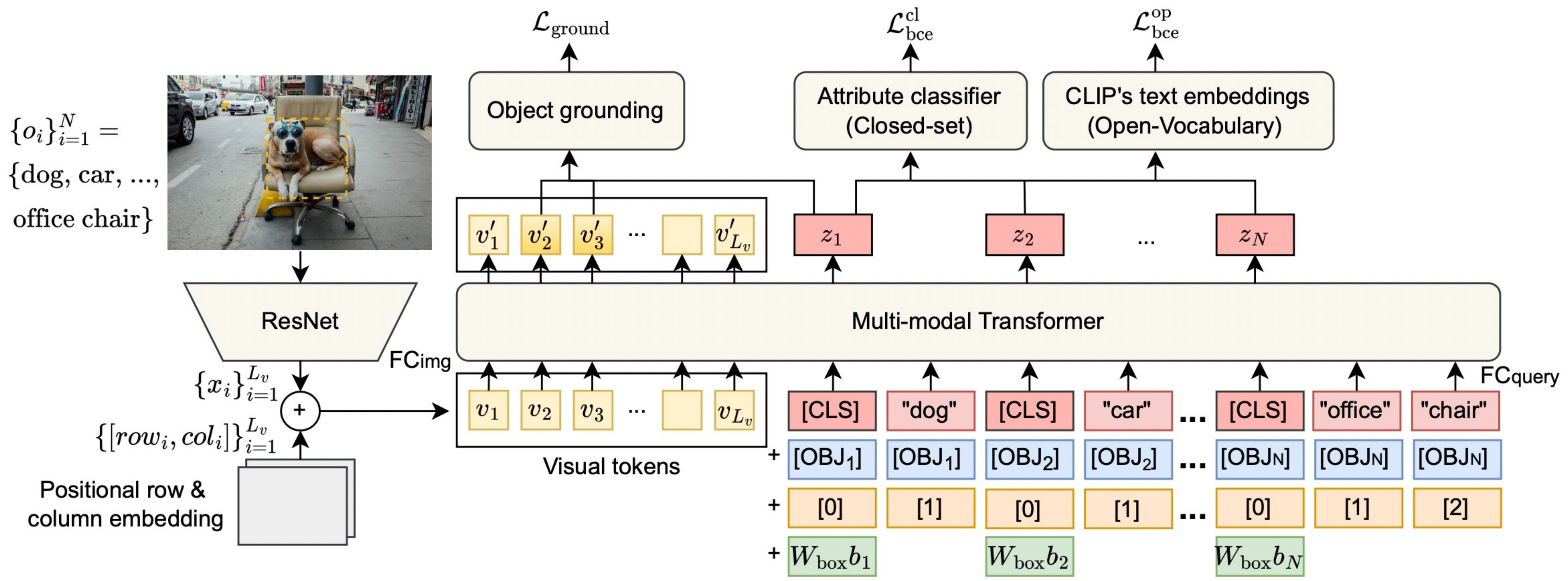
Man wearing glasses & brown jacket shovels deep snow

Flickr30K-Entities captions

Attribute types	# of classes in $\mathcal{C}_s$
Adjective	1251
Verb	950
Interaction	1278
Location	2047
Total	5526

Statistics of attributes

Transformer for Attribute Prediction (TAP)


$\mathcal{L}_{\text{ground}}$
$\mathcal{L}^{\text{cl}}_{\text{bce}}$
$\mathcal{L}^{\text{op}}_{\text{bce}}$
Object grounding
Attribute classifier (Closed-set)
CLIP's text embeddings (Open-Vocabulary)
$\{o_i\}_{i=1}^N =$ {dog, car, ..., office chair}
v'_1 v'_2 v'_3 ... v'_{L_v}
z_1 z_2 ... z_N
ResNet
FCimg
Multi-modal Transformer
FCquery
$\{x_i\}_{i=1}^{L_v}$
$\{[row_i, col_i]\}_{i=1}^{L_v}$
Positional row & column embedding
v_1 v_2 v_3 ... v_{L_v}
Visual tokens
[CLS] "dog" [CLS] "car" ... [CLS] "office" "chair"
+ [OBJ1] [OBJ1] [OBJ2] [OBJ2] ... [OBJN] [OBJN] [OBJN]
+ [0] [1] [0] [1] ... [0] [1] [2]
+ $W_{\text{box}}b_1$ $W_{\text{box}}b_2$ $W_{\text{box}}b_N$

Object grounding: Train network to softly localize object when grounding supervision is available → can attend to correct image regions when train/test on ungrounded objects

$$\mathcal{L}_{\text{ground}} = \sum_{i=1}^{N} \frac{1}{|O_i^+|} \sum_{j \in O_i^+} -\log\left(\frac{\exp\left(z_i^T v'_j/\tau\right)}{\sum_{k=0}^{L_v-1} \exp\left(z_i^T v'_k/\tau\right)}\right)$$

Classification

$$\mathcal{L}^{\text{cl}}_{\text{bce}}(Y, r) = \sum_{i=1}^{N} \sum_{c=1}^{\mathcal{C}_s} -\mathbb{1}_{[y_{i,c}=1]} p_c \log(\sigma(r_{i,c})) - \mathbb{1}_{[y_{i,c}=0]} n_c \log(1-\sigma(r_{i,c}))$$

Open-vocabulary attribute branch:

- Generate class embedding for attribute *j*

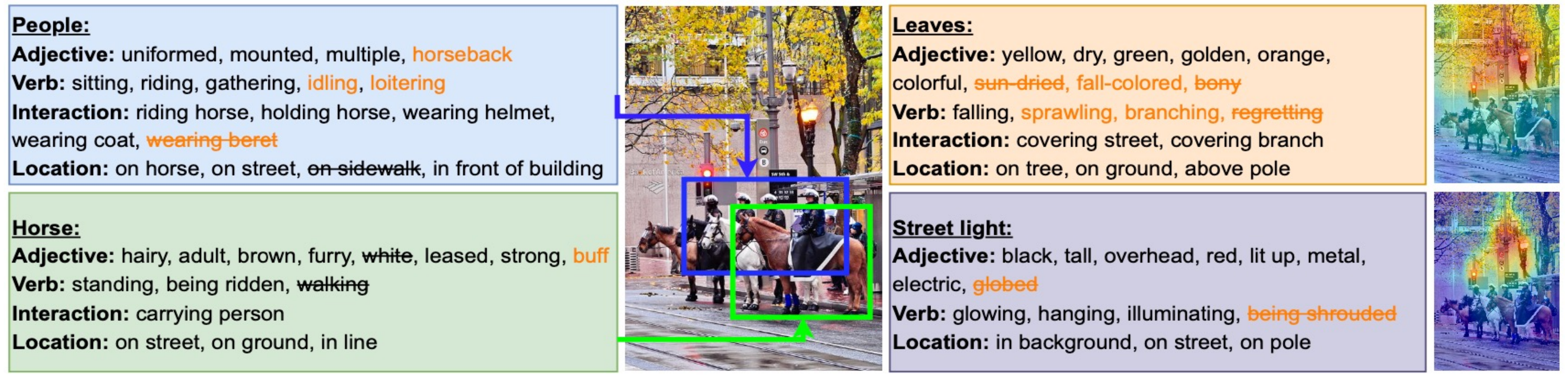
$$q_j = \text{CLIP}(\texttt{`A photo of a <attr> object'})$$

- Attribute prediction score of object *i* against attribute *j*

$$s_{i,j} = (z_i^{\text{T}} q_j/\tau)/(\|z_i\|\|q_j\|)$$

- Trained with BCE loss: $\mathcal{L}^{\text{op}}_{\text{bce}}(Y, s)$

Qualitative Examples

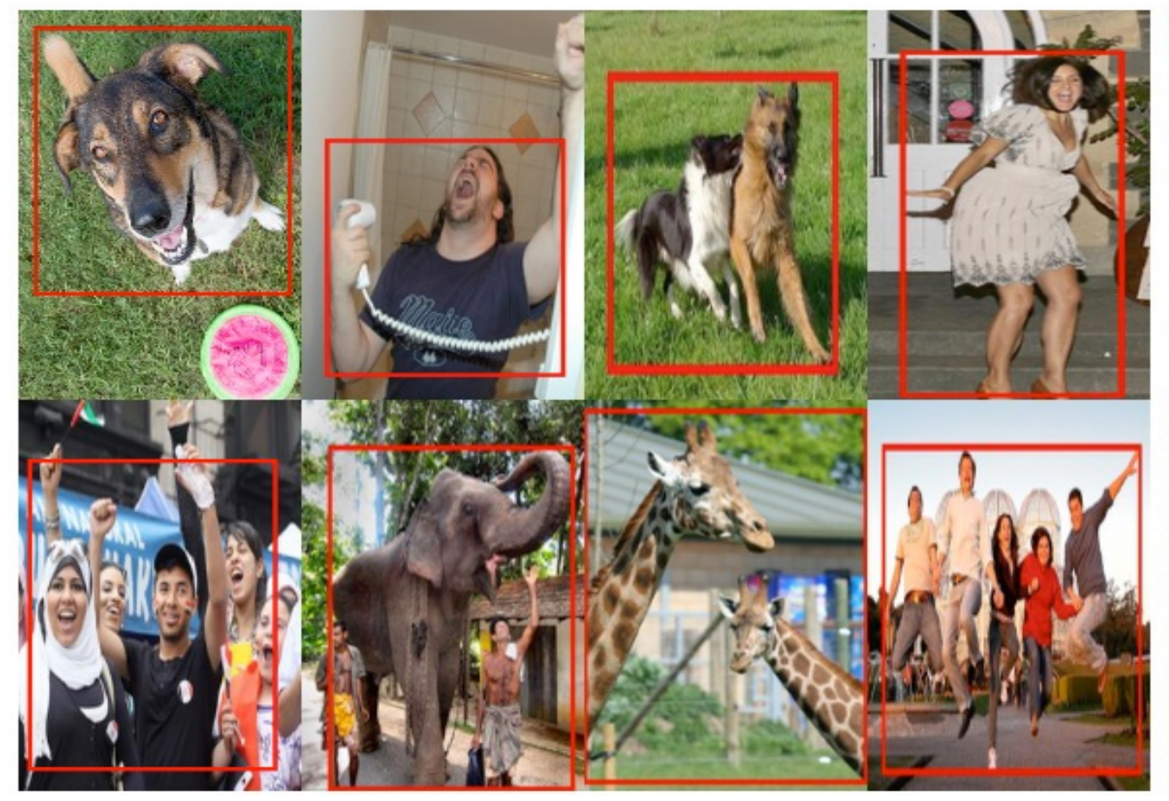


People:
Adjective: uniformed, mounted, multiple, horseback
Verb: sitting, riding, gathering, idling, loitering
Interaction: riding horse, holding horse, wearing helmet, wearing coat, ~~wearing beret~~
Location: on horse, on street, ~~on sidewalk~~, in front of building

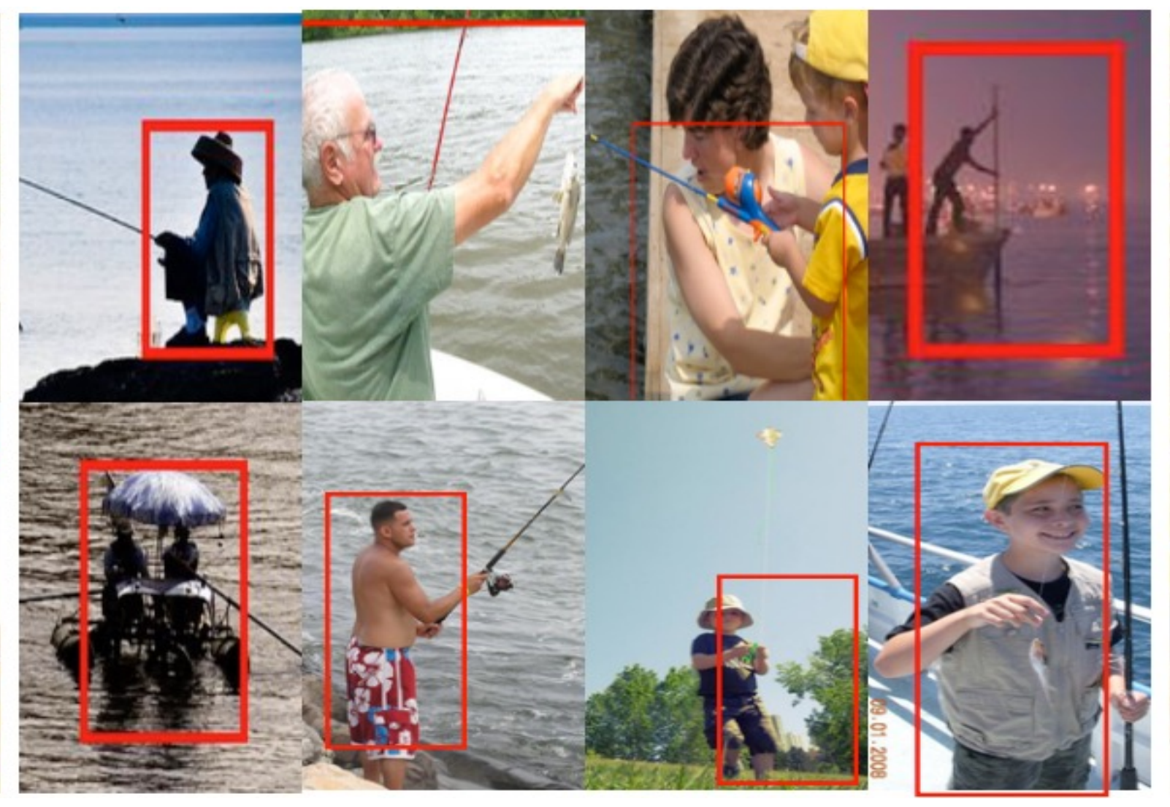
Horse:
Adjective: hairy, adult, brown, furry, ~~white~~, leased, strong, buff
Verb: standing, being ridden, ~~walking~~
Interaction: carrying person
Location: on street, on ground, in line

Leaves:
Adjective: yellow, dry, green, golden, orange, colorful, ~~sun-dried~~, fall-colored, ~~bony~~
Verb: falling, sprawling, branching, ~~regretting~~
Interaction: covering street, covering branch
Location: on tree, on ground, above pole

Street light:
Adjective: black, tall, overhead, red, lit up, metal, electric, ~~globed~~
Verb: glowing, hanging, illuminating, ~~being shrouded~~
Location: in background, on street, on pole



Excited



Fishing

Salmon-colored

Experiments

Methods	LSA pretrained	VAW supervised	mAP	mR@15	mA	F1@15
RN50-Baseline		✓	63.0	52.1	68.6	63.9
ML-GCN		✓	63.0	52.8	69.5	64.1
Sarafianos et al.		✓	64.6	51.1	68.3	64.6
SCoNE		✓	68.3	58.3	71.5	70.3
TAP [**Ours**]		✓	65.4	54.2	67.2	66.4
RN50-Context	✓	✓	67.3	54.1	69.3	66.1
TAP [**Ours**]	✓		67.2	53.8	65.5	61.5
TAP [**Ours**]	✓	✓	**73.4**	**63.3**	**73.5**	**71.1**

- New state-of-the-art result on VAW after pretraining on LSA. Without pretraining, TAP outperforms the baselines and is only lower than SCoNE due to being data hungry.

Methods	AP_{seen}	$\text{AP}_{\text{unseen}}$	$\text{AP}_{\text{overall}}$
CLIP (attribute prompt)	2.53	3.37	2.64
CLIP (object-attribute prompt)	0.97	1.56	1.04
CLIP (combined prompt)	2.81	3.67	2.92
OpenTAP	14.34	7.62	13.59

Methods	Bbox	Pose	CLIP text	mAP
PastaNet	✓	✓		46.3
HAKE	✓	✓		47.1
DEFR-RN50			✓	49.7
OpenTAP			✓	**51.7**

- (Left) On LSA, OpenTAP outperforms CLIP (using custom designed prompts for attribute prediction) → can recognize large # of attributes, even those unseen in the open-world.
- (Right) On HICO, finetuned OpenTAP achieves SOTA human-object interaction classification → can recognize well interaction classes.