

An Interpretable Model for Visual Relationship Detection

Ji Zhang (jz462@rutgers.edu)

Speaker: Kevin Shih (on behalf of Ji Zhang)

Our Model

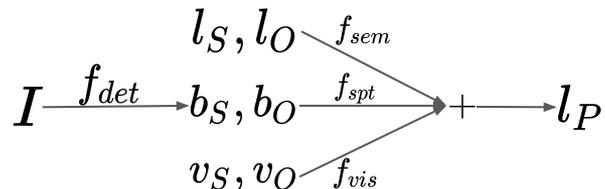
- We want to learn a mapping f from image I to the 3 labels and 2 boxes

$$I \xrightarrow{f} l_S, l_P, l_O, b_S, b_O$$

- We decompose the mapping f into a chain of object detector f_{det} and relationship classifier f_{rel}

$$I \xrightarrow{f_{det}} l_S, l_O, b_S, b_O, v_S, v_O \xrightarrow{f_{rel}} l_P$$

- To predict the predicate, we route the detector information into separate modules for semantic, spatial, and visual information



Why 3 modules?

- Semantic module: linguistic association between S, P, and O can be used as a strong prior
 - Originally from [1]
 - Generally speaking, the types of relationships between two objects are usually limited, e.g., given the subject being person and object being horse, their relationship is highly likely to be “ride”, “walk”, “feed”, but less likely to be “stand on”, “carry”, “wear”, etc.
 - Relationship detection dataset can only contain a limited subset of all possible relationships, making the linguistic association even stronger
 - Implementation:
 - for each image, count the occurrence of l_P given l_S and l_O in the ground truth annotations, and we end up with an empirical distribution of $p(l_P|l_S, l_O)$ for the whole dataset

1. Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. "Neural Motifs: Scene Graph Parsing with Global Context." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

* S is subject, P is predicate, O is object, I is image

Why 3 modules? (cont.)

- Spatial module: particularly useful for spatial relationships (e.g., inside of, under, next to, ...)
 - Visual module is bad at it since convolutional feature maps are too coarse to precisely indicate locations (mentioned in [1])
 - Spatial information is captured by encoding the layout of the three boxes:

$$\langle \Delta(b_S, b_O), \Delta(b_S, b_P), \Delta(b_P, b_O), f(b_S), f(b_O) \rangle$$

where b_S, b_P, b_O are boxes of subject, predicate, object. b_P is the minimum enclosing box of b_S, b_O , and

$$\Delta(b_1, b_2) = \left\langle \frac{x_1 - x_2}{w_2}, \frac{y_1 - y_2}{h_2}, \log \frac{w_1}{w_2}, \log \frac{h_1}{h_2} \right\rangle$$

$$f(b) = \left\langle \frac{x_{min}}{w}, \frac{y_{min}}{h}, \frac{x_{max}}{w}, \frac{y_{max}}{h}, \frac{a_{box}}{a_{img}} \right\rangle$$

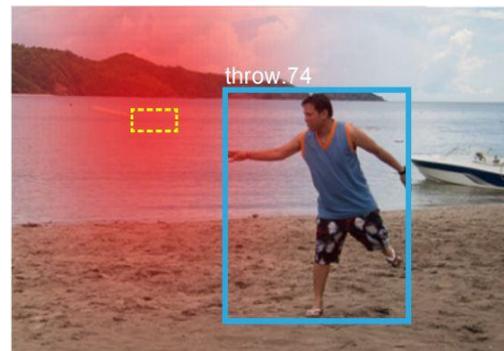
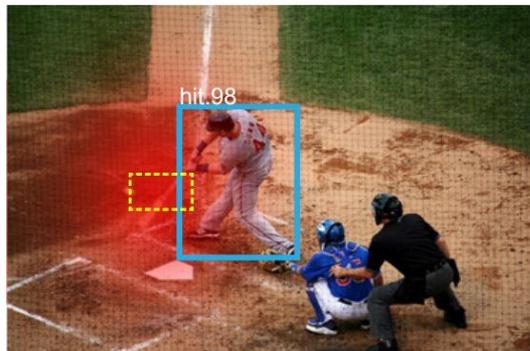
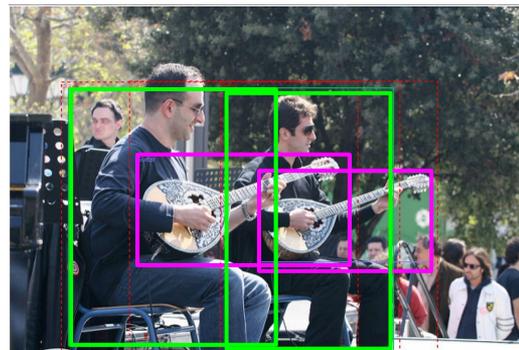
where $b_1 = (x_1, y_1, w_1, h_1)$, $b_2 = (x_2, y_2, w_2, h_2)$, w, h are the width and height of the image, a_{box}, a_{img} are areas of the box and image

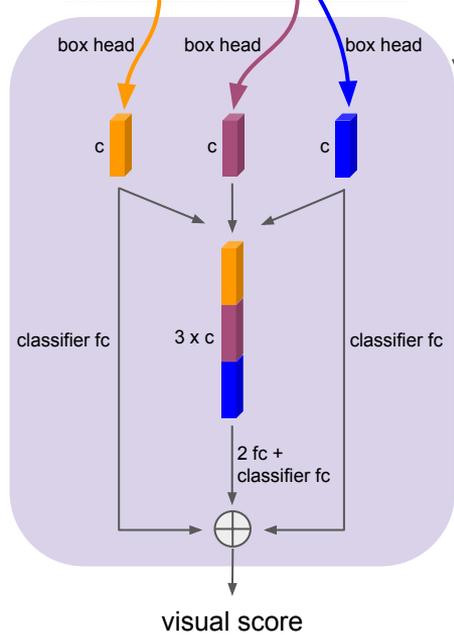
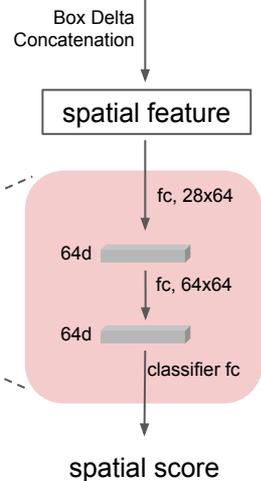
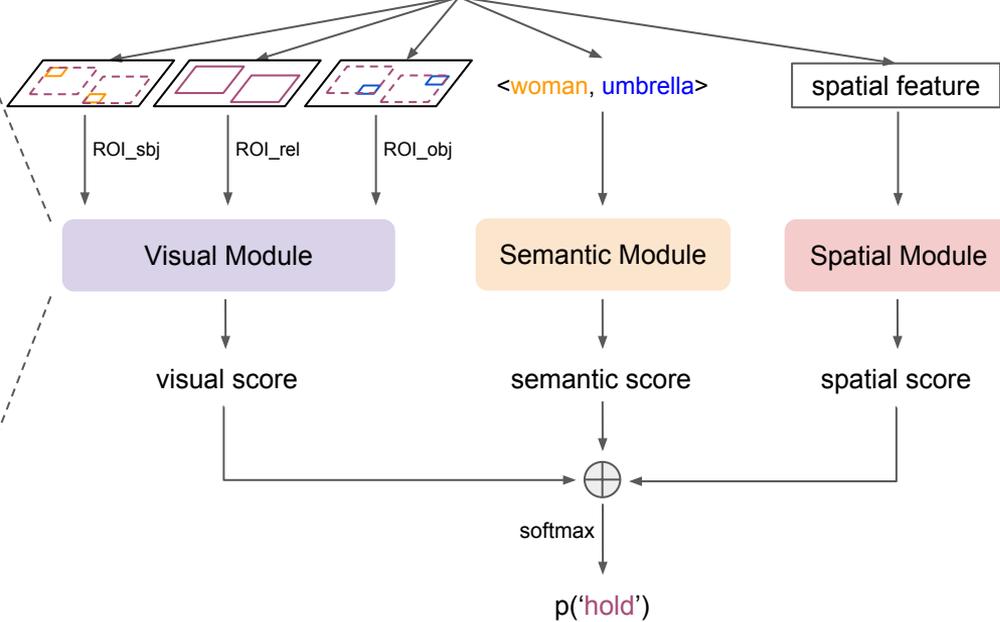
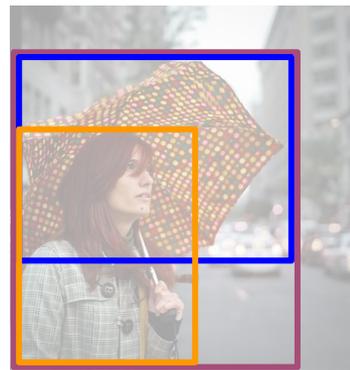
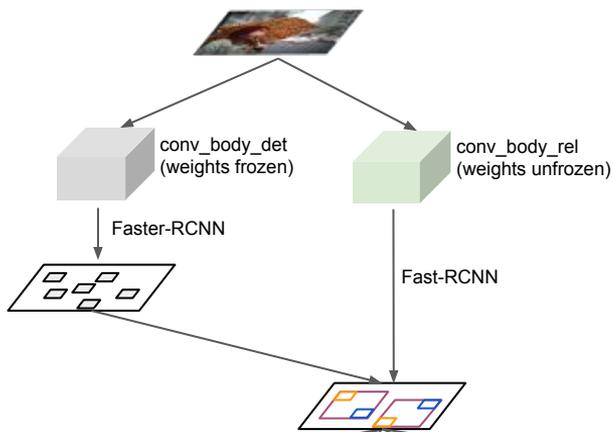
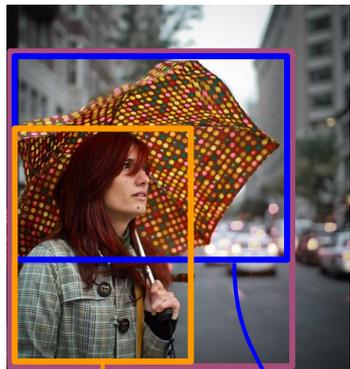
1. Gkioxari, Georgia, Ross Girshick, Piotr Dollár, and Kaiming He. "Detecting and Recognizing Human-Object Interactions." *Conference on Computer Vision and Pattern Recognition*, 2018.

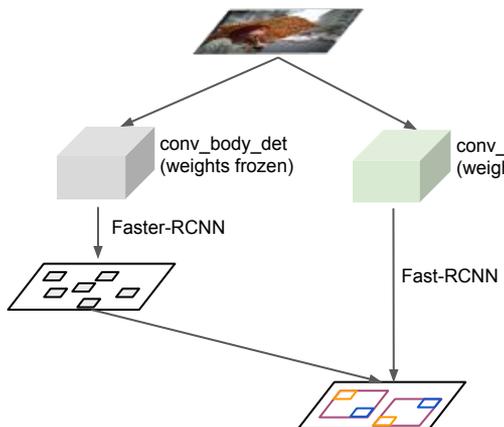
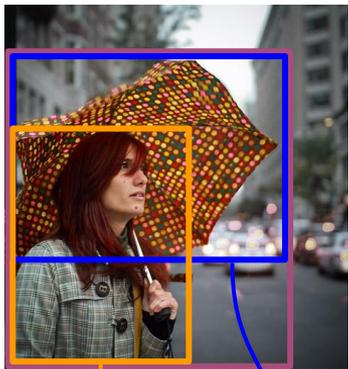
* S is subject, P is predicate, O is object, I is image

Why 3 modules? (cont.)

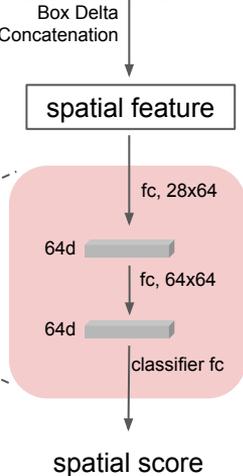
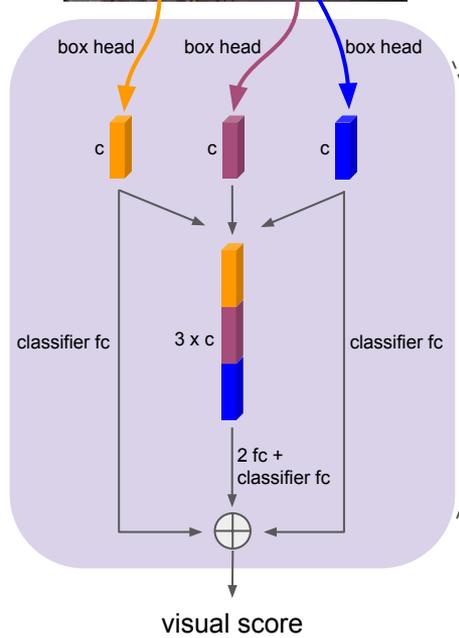
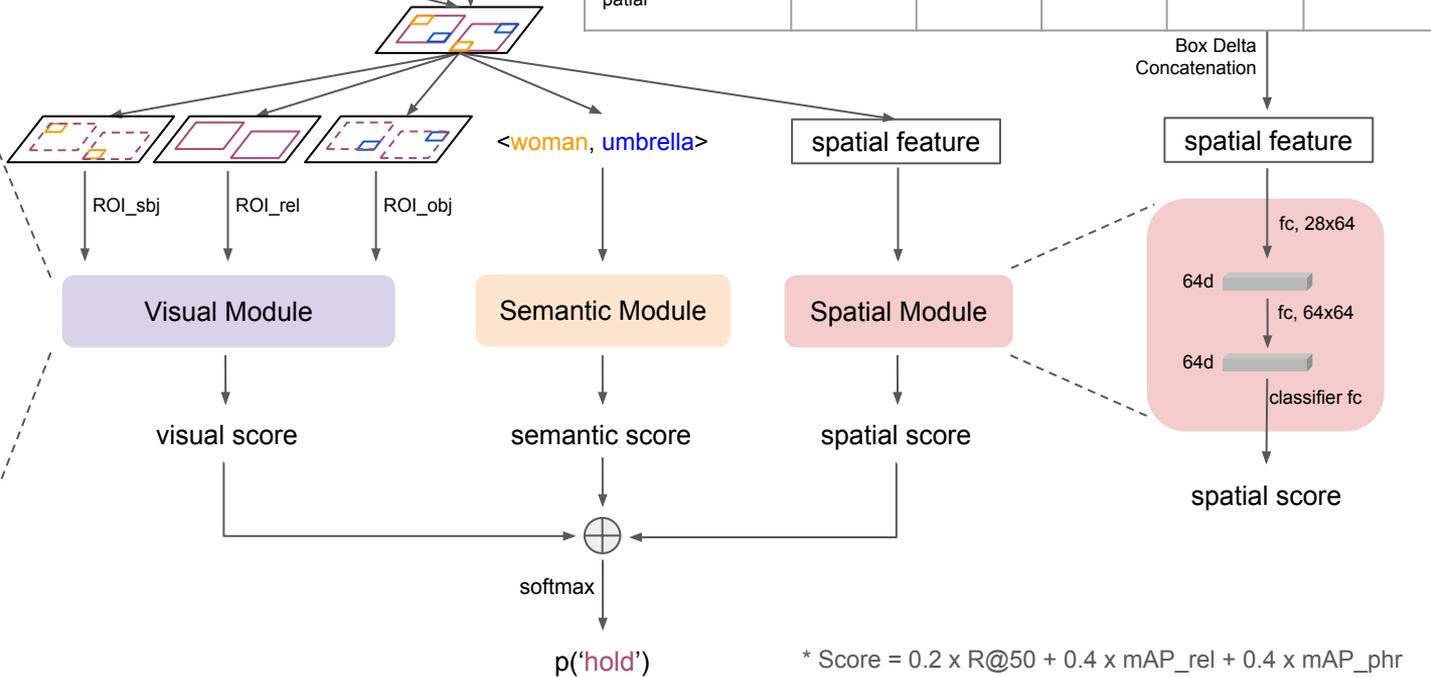
- Visual module: account for all the other types of relationships
 - Visual features disambiguate relationship reference, e.g., which man is playing which guitar
 - Visual features are strong cues for human interactions (i.e., hold, play, wear)
 - We use the output of the last fc of backbone as visual feature







	R@50	mAP_rel	mAP_phr	Score on val set	Score on Kaggle
Semantic Only	72.98	26.54	32.77	38.32	22.21
Semantic+Visual	74.46	34.16	39.59	44.39	-
Semantic+Visual+Spatial	74.40	34.96	40.70	45.14	31.43



* Score = 0.2 x R@50 + 0.4 x mAP_rel + 0.4 x mAP_phr

Per Class AP

	at	on	holds	plays	interacts_with	wears	inside_of	under	hits	mAP_rel
Semantic	28.62	24.52	37.04	27.33	38.37	3.16	16.34	25	38.45	26.53
Semantic+Visual	32.16	35.92	40.47	34.61	42.06	8.29	41.64	25	47.31	34.16
Semantic+Visual+Spatial	32.05	36.06	40.61	32.50	42.39	7.49	43.09	20	60.43	34.96

- “Under” did best with just the semantic module
- The spatial module has clear impact on “inside_of” because it is a spatial relationship. It also improves “hits” much, because “hits” has a clear spatial pattern in this dataset.

