Detecting Unseen Visual Relations Using Analogies

Julia Peyre, Ivan Laptev, Cordelia Schmid, and Josef Sivic

Presentation by Madeline Schiappa







CZECH INSTITUTE OF INFORMATICS ROBOTICS AND CYBERNETICS CTU IN PRAGUE



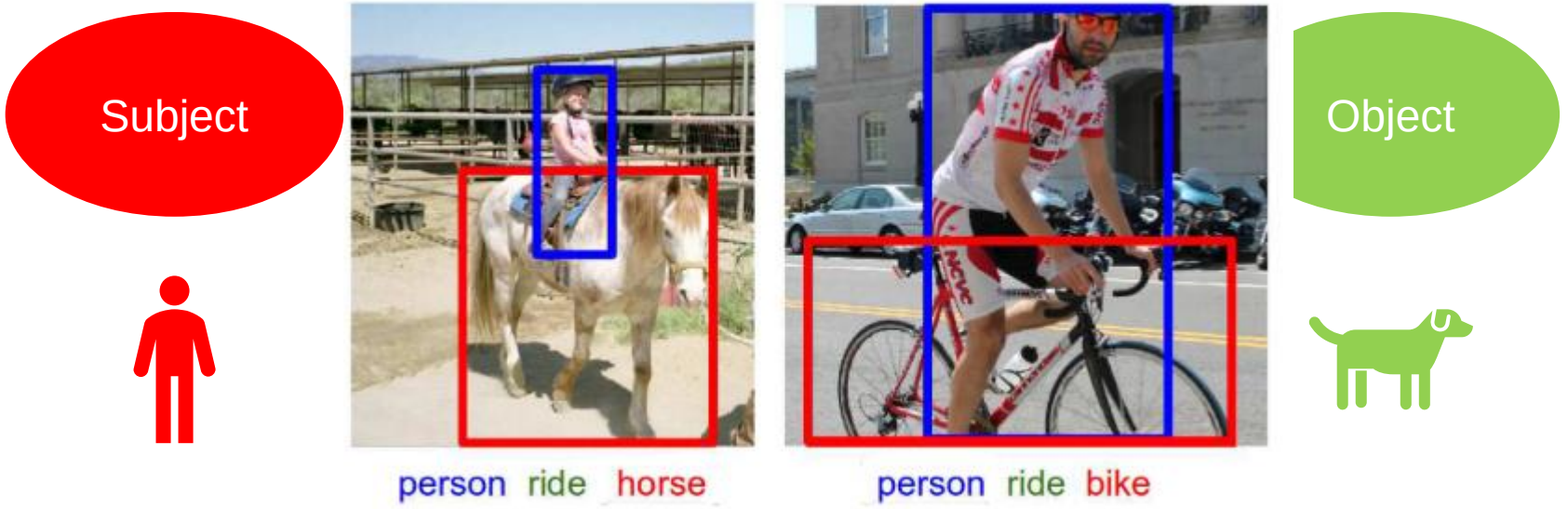


UCF
CENTER FOR RESEARCH IN COMPUTER VISION

Introduction



Visual Relations - Triplets



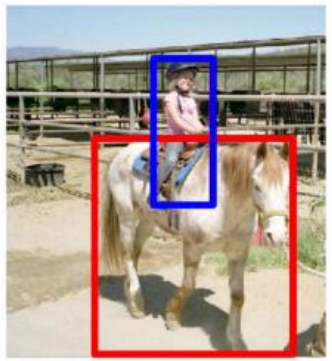
Recognize and **localize** unseen interactions in images

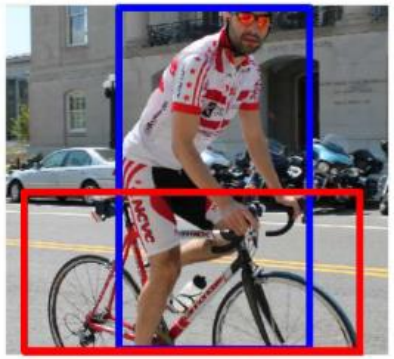
Problem

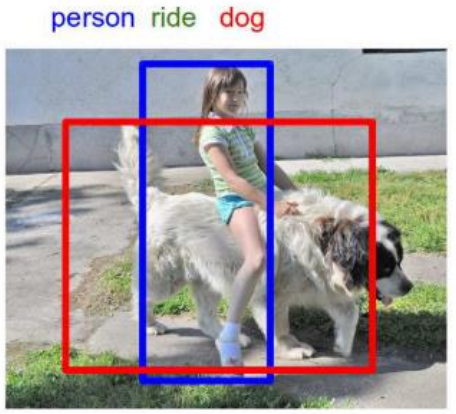
- The large variety of entities and possible relationships between them

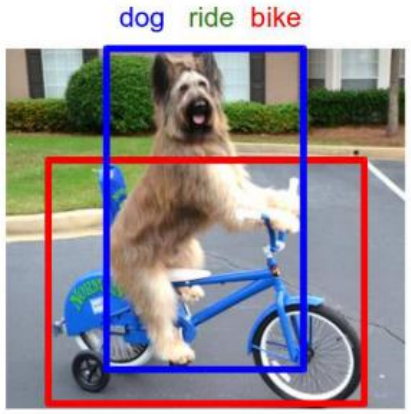
Person watch dog Person pet dog Person watch horse Dog watch horse

- Handle unseen visual triplets
  - Out of vocab unigram and unseen triplets









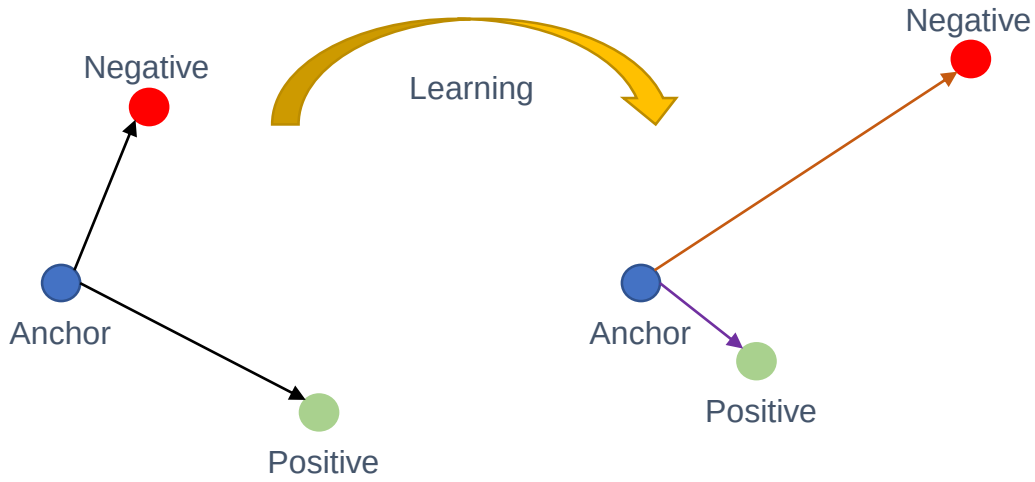
Triplet Loss

- At same time:
  - reduce the distance between similar things, p.
  - increase distance between different things, n.

$$\sum_{i}^{N}\left[\left\|f(x_i^a)-f(x_i^{\textcircled{p}})\right\|_2^2-\left\|f(x_i^a)-f(x_i^{\textcircled{n}})\right\|_2^2+\alpha\right]$$

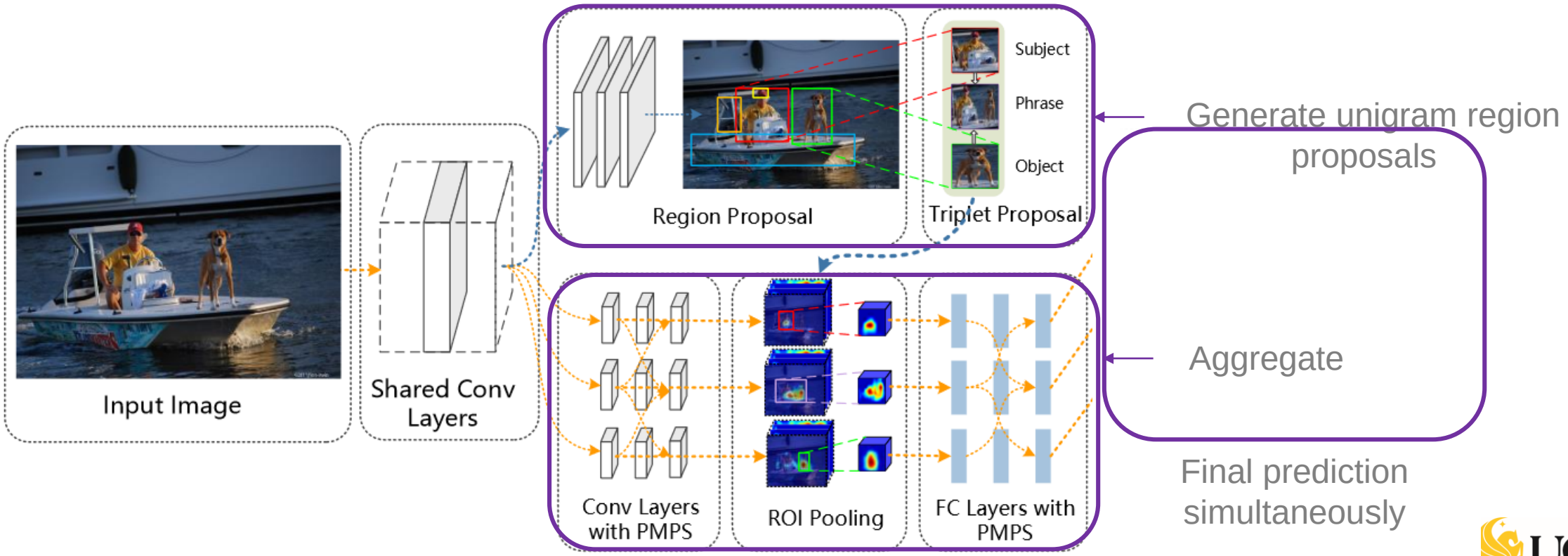
Minimize

Maximize



Compositional Approach

- Subject, predicate and object detected separately then aggregated together.



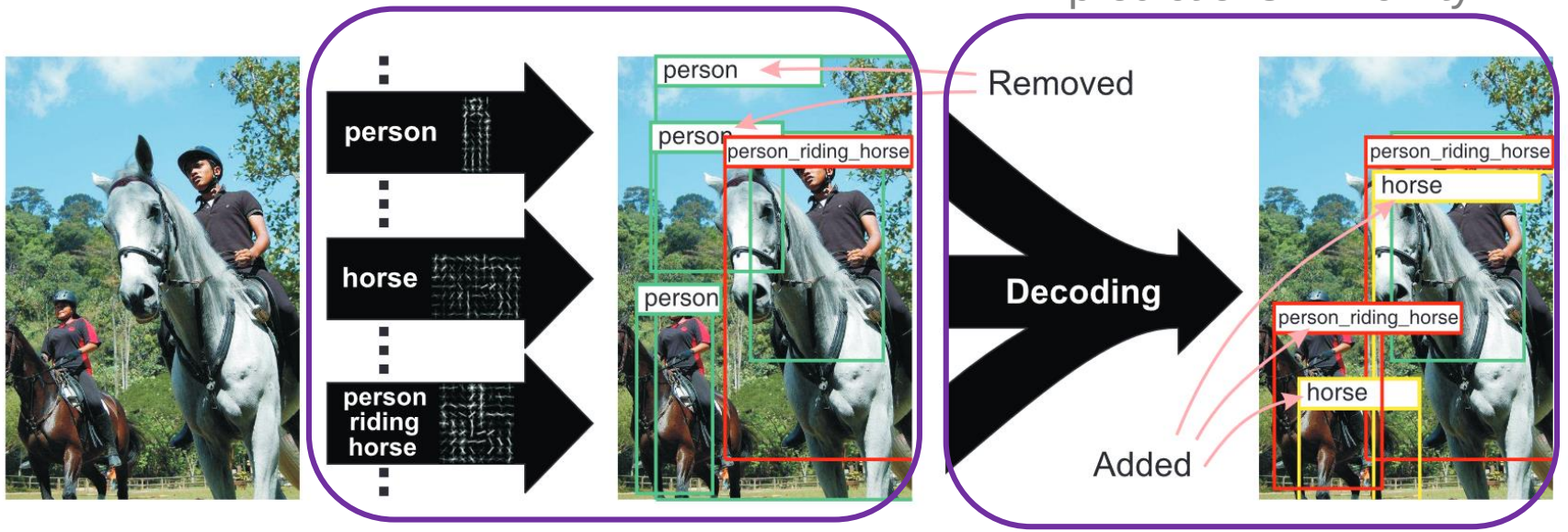
Visual Phrase Models

- Treat triplet as single entity
- Struggle with larger vocabulary



Correct predictions by looking at other predictions in vicinity

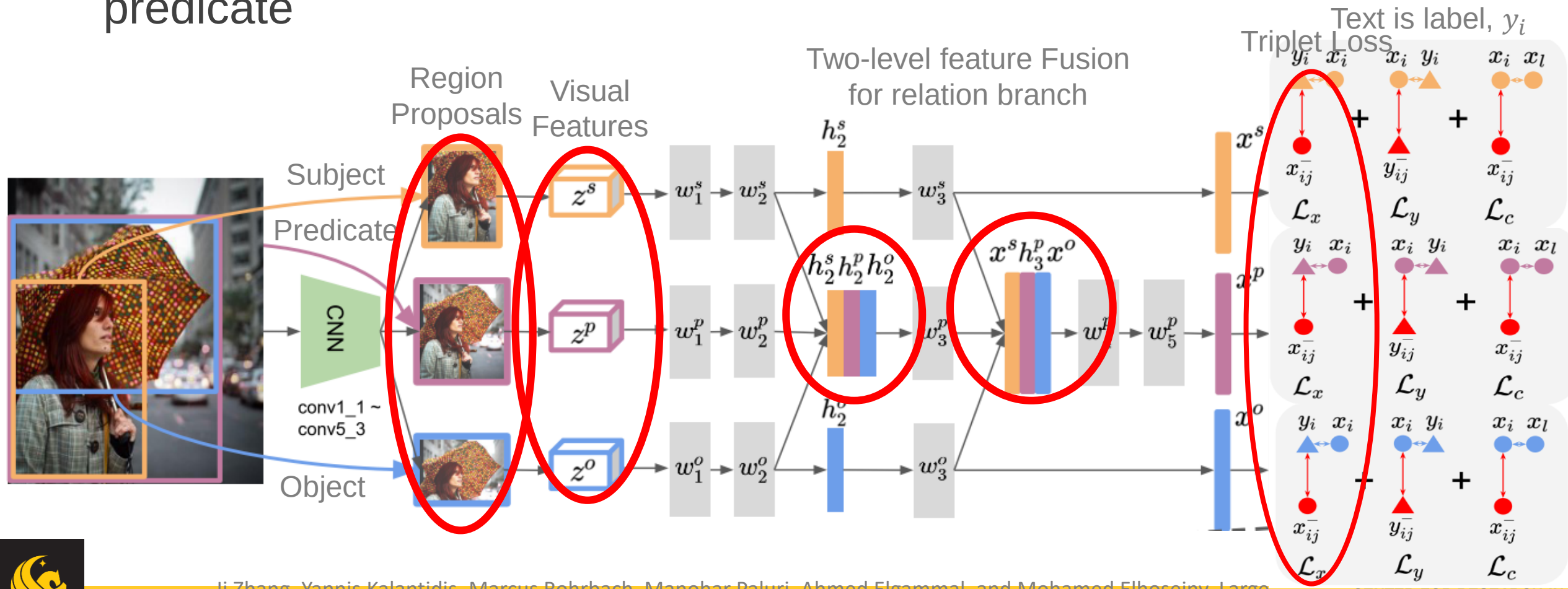
Predict VP region + objects



Mohammad Amin Sadeghi and Ali Farhadi. Recognition using visual phrases. In CVPR, 2011

Visual-Semantic Embeddings

- Learn separate visual-semantic spaces for subject, object, and predicate



Method

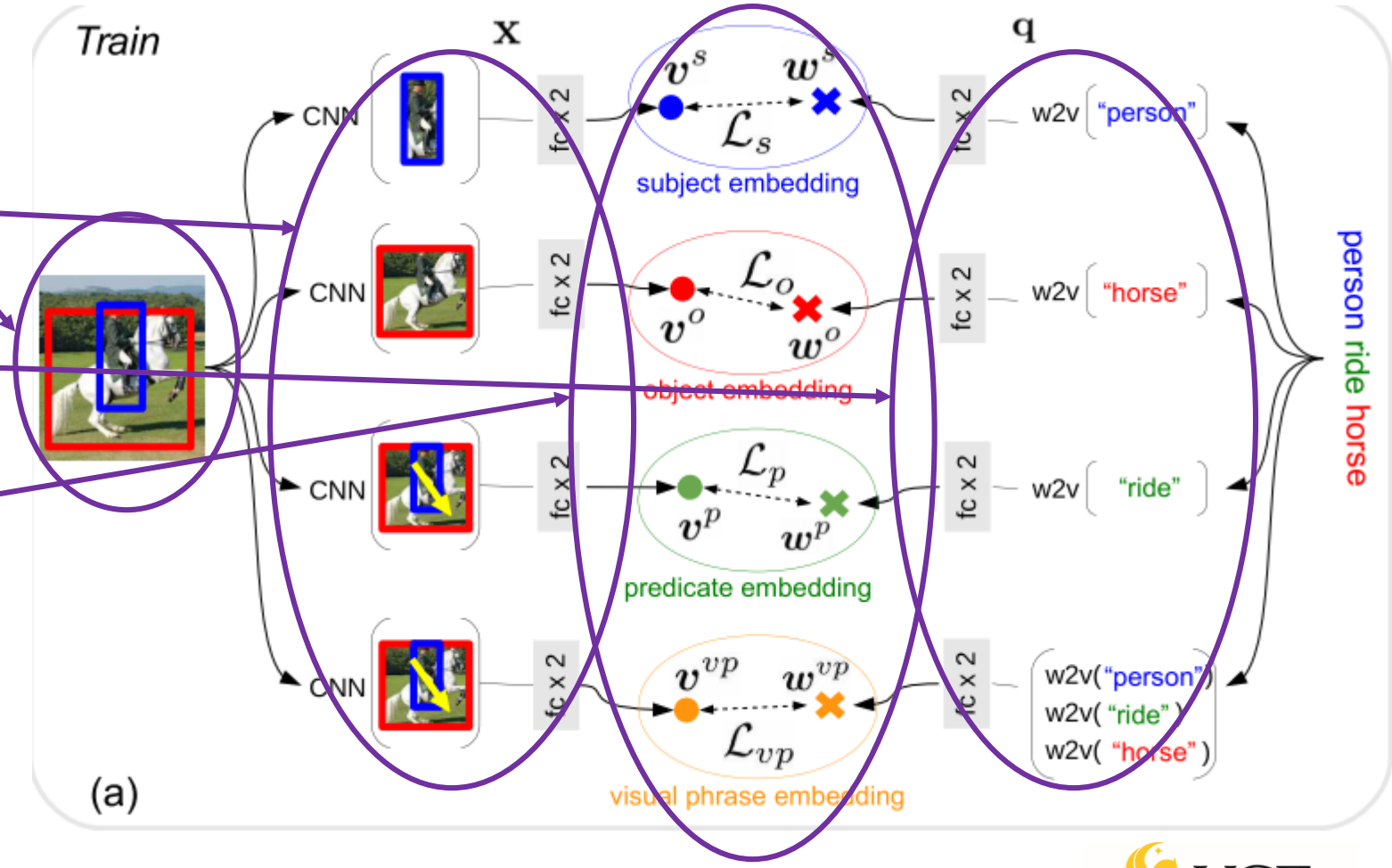
Phase 1

Pair of objects i

Visual Embeddings

Textual Embeddings

Final Embeddings



projection functions

Trained

$$v_i^b = f_v^b(x_i) \qquad f_v^b : \mathbb{R}^{d_v} \rightarrow \mathbb{R}^d$$

$$w_t^b = f_w^b(q_i) \qquad f_w^b : \mathbb{R}^{d_q} \rightarrow \mathbb{R}^d$$

Peyre, J., Sivic, J., Laptev, I., & Schmid, C. (2020). Detecting Unseen Visual Relations Using Analogies. ICCV, 2020.

Phase 1 - Training

- $i \in \{1, \ldots, N\}$: Each (s, o) candidate bounding boxes
- $y_t^i = [0,1]$ if i^{th} pair of boxes can(not) be described as relation $t = (s, p, o)$
- Triplet loss

$$\mathcal{L}_b = \sum_{i=1}^{N} \sum_{t \in V_b} 1_{y_t^i = 1} \log\left(\frac{1}{1 + e^{-\mathbf{w}_t^{b\,T} \mathbf{v}_i^b}}\right) + \sum_{i=1}^{N} \sum_{t \in V_b} 1_{y_t^i = 0} \log\left(\frac{1}{1 + e^{-\mathbf{w}_t^{b\,T} \mathbf{v}_i^b}}\right)$$

For each input type b

$$\mathcal{L}_{joint} = \mathcal{L}_s + \mathcal{L}_o + \mathcal{L}_p + \mathcal{L}_{vp}$$

Peyre, J., Sivic, J., Laptev, I., & Schmid, C. (2020). Detecting Unseen Visual Relations Using Analogies. ICCV, 2020.

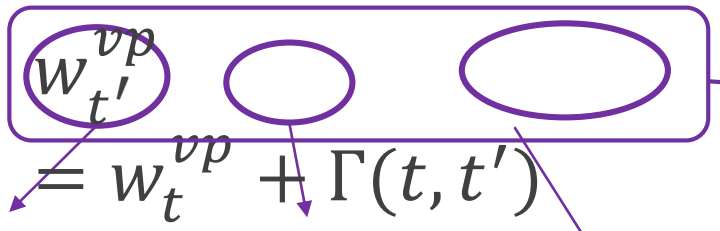
Phase 2

- *Transfer by Analogy* : transfer learned representation to unseen

Choosing Source Triplets:

$G(t,t') = \sum_{b\in(s,o,p)} \alpha_b w_t^{b\,T} w_{t'}^b$

where $G(t,t') \in [0,1]$ because $\sum_b \alpha_b = 1$

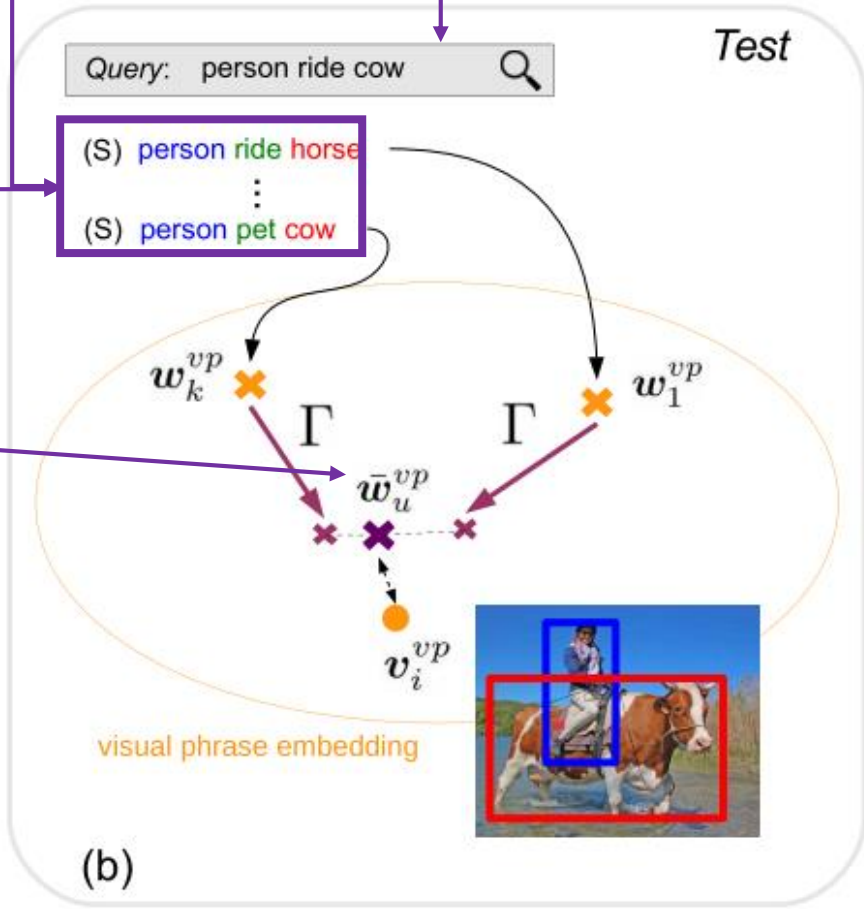


Embedding of target triplet

Embedding of source triplet

Transformation function, $t \rightarrow t'$

Aggregate



$$\Gamma(t,t') = \mathrm{MLP}\begin{bmatrix} \boldsymbol{w}_{s'}^{vp} - \boldsymbol{w}_s^{vp} \\ \boldsymbol{w}_{p'}^{vp} - \boldsymbol{w}_p^{vp} \\ \boldsymbol{w}_{o'}^{vp} - \boldsymbol{w}_o^{vp} \end{bmatrix}$$

$$w_s^{vp} = f_s^{vp}(\boldsymbol{q}_{[s,0,0]})$$
$$w_p^{vp} = f_p^{vp}(\boldsymbol{q}_{[0,p,0]})$$
$$w_o^{vp} = f_o^{vp}(\boldsymbol{q}_{[0,0,o]})$$

Peyre, J., Sivic, J., Laptev, I., & Schmid, C. (2020). Detecting Unseen Visual Relations Using Analogies. ICCV, 2020.

Phase 2 Example

Intuition:
“person ride horse” – “horse” + “cow” = “person ride cow”

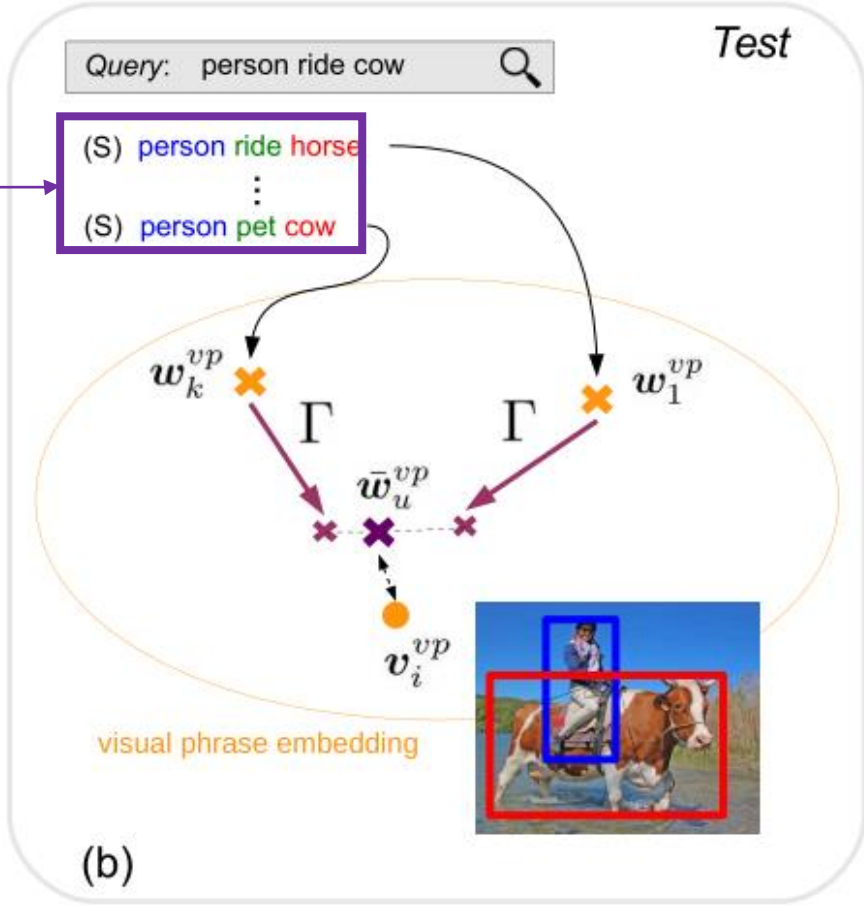
$$w_{t'}^{vp} = w_{t}^{vp} + \Gamma \begin{bmatrix} 0 \\ 0 \\ \boldsymbol{w}_{cow}^{vp} - \boldsymbol{w}_{horse}^{vp} \end{bmatrix}$$

Embedding of target triplet

Embedding of source triplet

Transformation function, $t \rightarrow t'$

Source Triplets



Peyre, J., Sivic, J., Laptev, I., & Schmid, C. (2020). Detecting Unseen Visual Relations Using Analogies. ICCV, 2020.

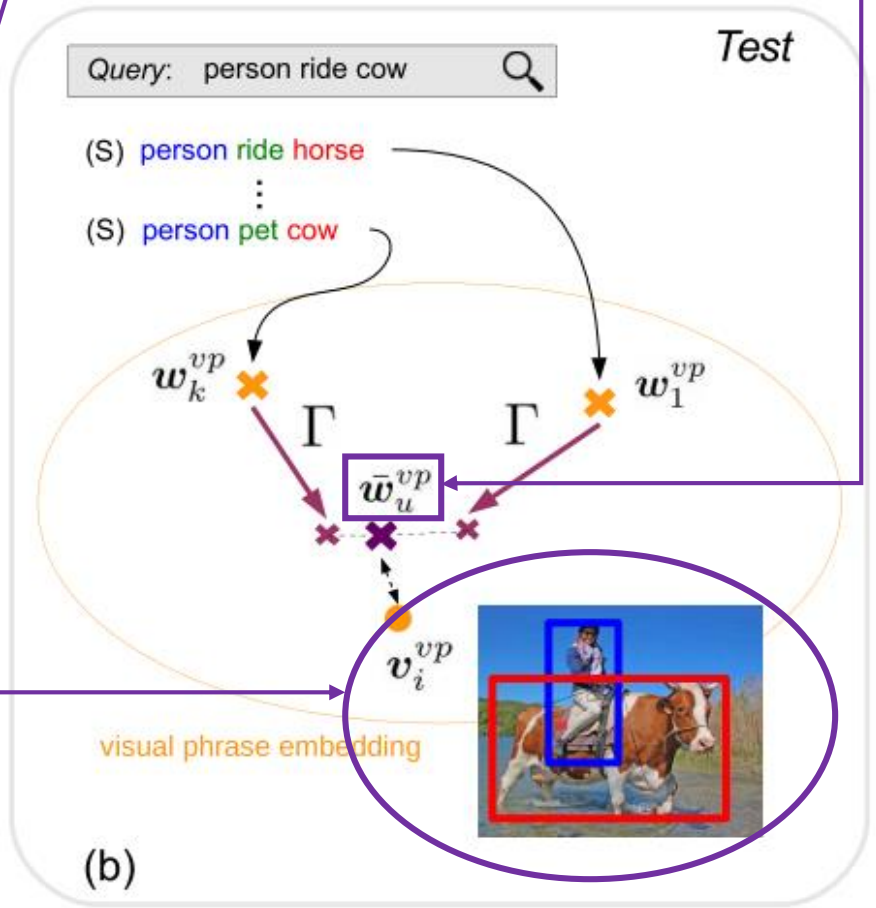
Phase 2 – Training

$$\mathcal{L}_\Gamma = \sum_{i=1}^{N} \sum_{(t,t')\in\mathcal{Q}} 1_{y_{t'}^i=1} \log\left(\frac{1}{1+e^{-(\boldsymbol{w}_t^{vp}+\Gamma(t,t'))^T \boldsymbol{v}_i^{vp}}}\right) + \sum_{i=1}^{N} \sum_{(t,t')\in\mathcal{Q}} 1_{y_{t'}^i=0} \log\left(\frac{1}{1+e^{-(\boldsymbol{w}_t^{vp}+\Gamma(t,t'))^T \boldsymbol{v}_i^{vp}}}\right)$$

- Optimize log-likelihood between pairs
  - Visual features are from target triplet
  - Language features from transformed training triplets

Visual Phrase Embedding of Unseen Triplet

$$\bar{\boldsymbol{w}}_u^{vp} = \sum_{t\in\mathcal{N}_u} G(t,u)\left(\boldsymbol{w}_t^{vp} + \Gamma(t,u)\right)$$

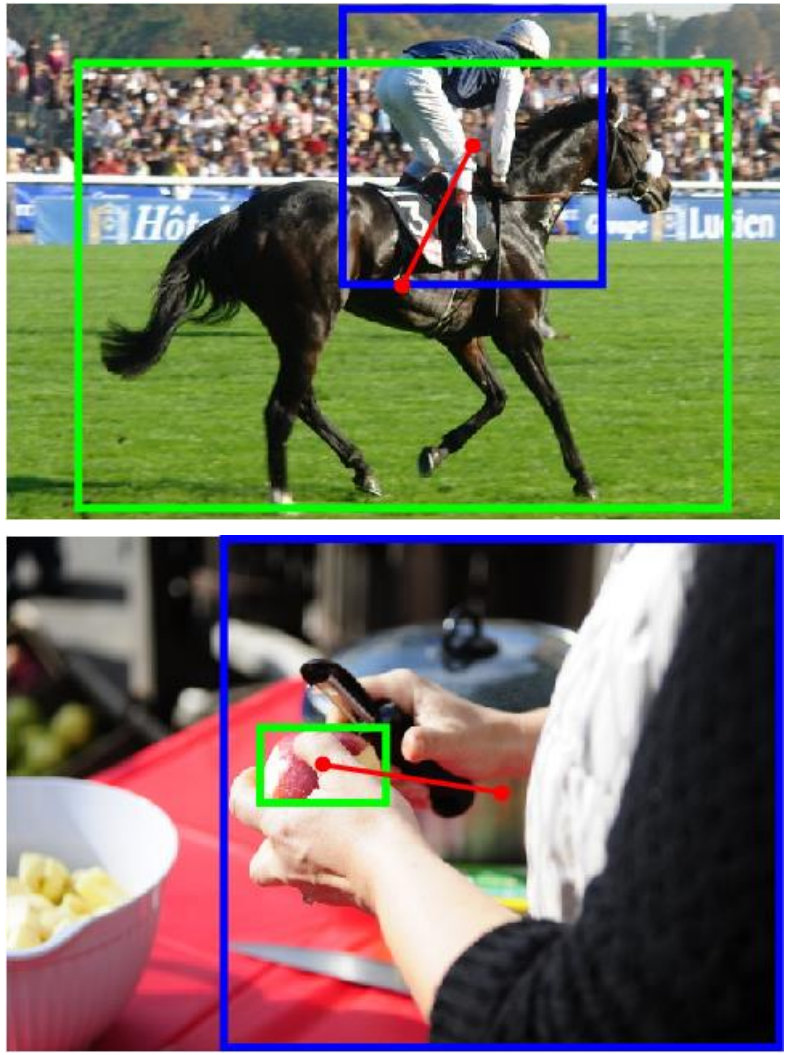


Peyre, J., Sivic, J., Laptev, I., & Schmid, C. (2020). Detecting Unseen Visual Relations Using Analogies. ICCV, 2020.

Experiments

Dataset – HICO-DET

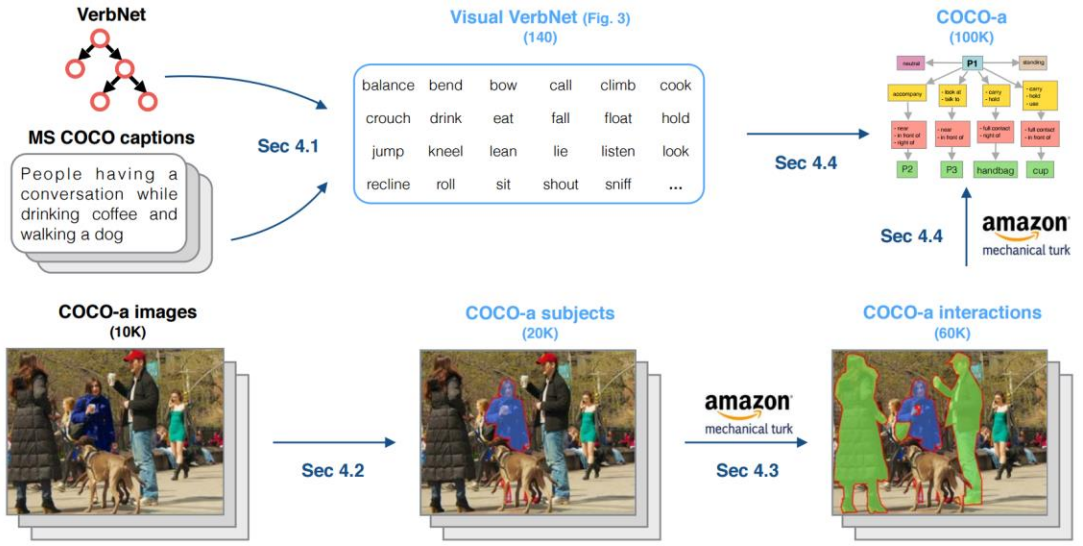
- Human-object interactions
  - 80 categories
- Human and object bounding box annotations
- Positive example of 600 triplets
- 9,360 possible triplets
- Use 25 non-rare triplets as unseen during test time



Yu-Wei Chao, Zhan Wang, Yugeng He, Jiaxuan Wang, and Jia Deng. HICO: A Benchmark for Recognizing Human-Object Interactions in Images. IEEE International Conference on Computer Vision (ICCV), 2015.

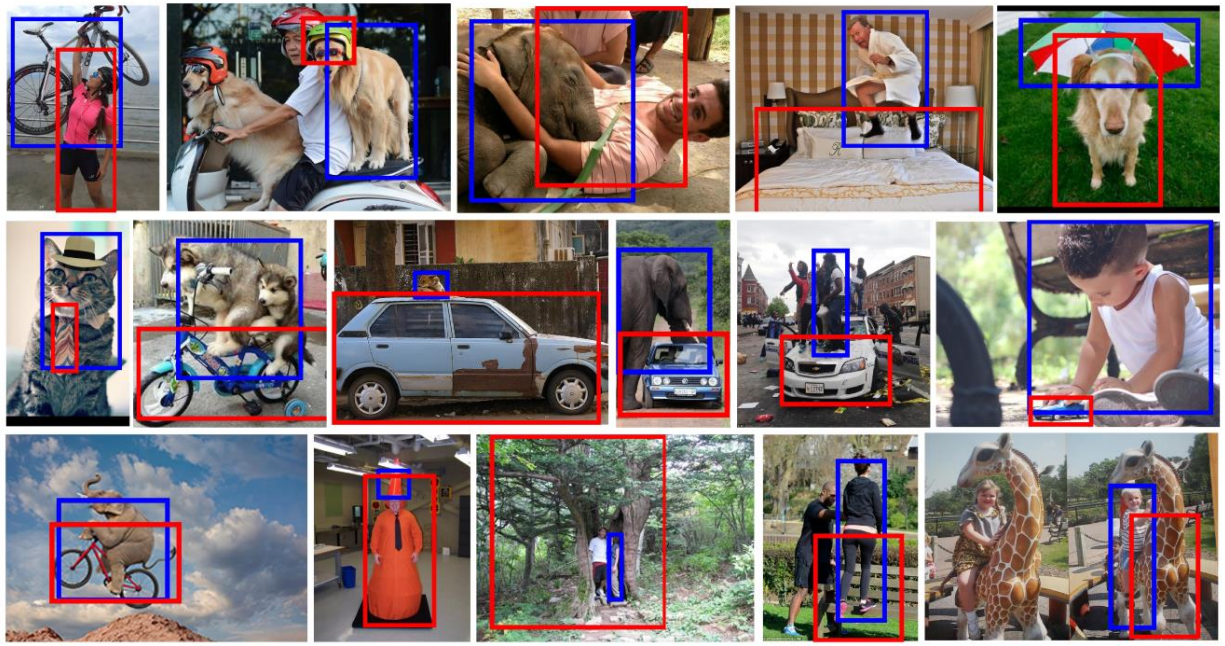
Dataset – COCO-a

- Subset of COCO with human object interactions
  - 80 categories
- Positive examples of 140 triplets
- 1,681 possible triplets
- Used for evaluation on HICO-DET trained model
  - 71% unseen involve out-of-vocabulary *predicate* not seen at training



Dataset – UnRel

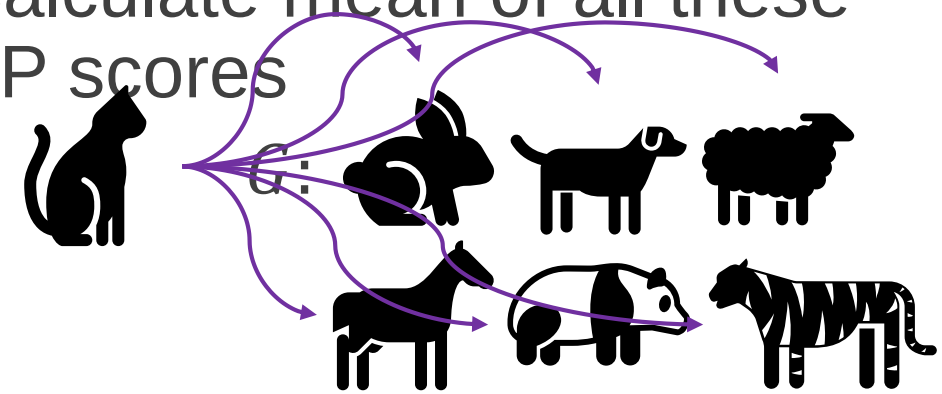
- Triplets but not necessarily involving a human
- Predicates not necessarily actions
- Positive examples of 76 *unusual* triplets
- Used for evaluation only



Peyre, Julia, et al. "Weakly-supervised learning of visual relations." Proceedings of the IEEE International Conference on Computer Vision. 2017.

Mean Average Precision for Image Retrieval

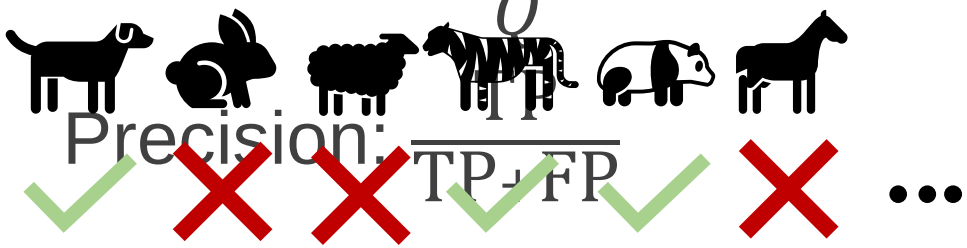
- for a given query, $q \in Q$:
  - calculate average precision, $\text{AveP}(q)$:
- Calculate mean of all these AP scores



- Measure similarity between Q and each G and sort by most similar.

$$\text{MAP} = \frac{\sum_{q=1}^{Q} \text{AveP}(q)}{Q}$$

Precision: $\frac{TP}{TP+FP}$

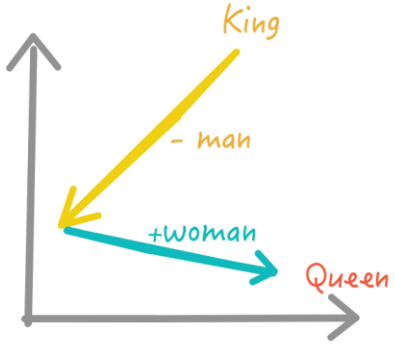


$\frac{TP}{Total}$ $\frac{1}{1}$ $\frac{2}{4}$ $\frac{3}{5}$

$$Overall\ AP = \frac{1}{3}\left(\frac{1}{1} + \frac{2}{4} + \frac{3}{5}\right) = .7$$

Implementation Details

- Evaluation is mAP over set of triplet queries
- Word embeddings from Word2Vec trained on Google news
  - Triplet is aggregated vectors



a) Learns Analogy



b) Similar Words have same angles

- Projection functions are 2 FCs + 1 ReLU
- Pre-extracted candidate pairs
- Visual representation of candidate pair is 1000-dimensional vector
  - Aggregate visual and spatial of objects
  - Subject is only appearance

Peyre, J., Sivic, J., Laptev, I., & Schmid, C. (2020). Detecting Unseen Visual Relations Using Analogies. ICCV, 2020.

Seen Triplets

- mAP results on HICO-DET
- **s+o+p**: the compositional part using all unigram terms
- **s+o+vp**: visual phrase part combined with object scores
- **s+o+p+vp:** visual phrase representation on top of the compositional baseline

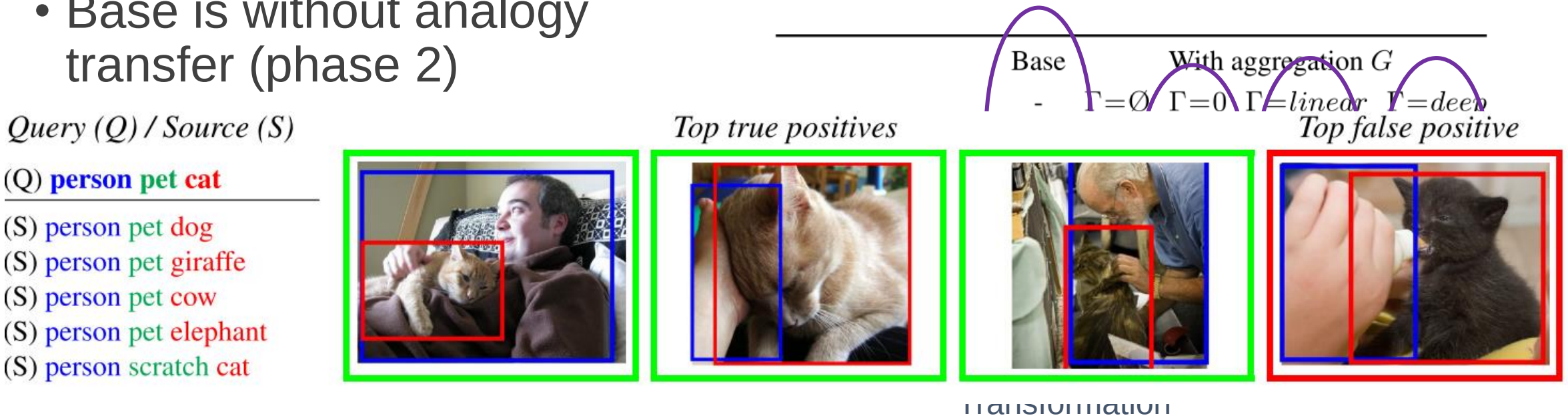
	full	rare	non-rare
Chao [6]	7.8	5.4	8.5
Gupta [13]	9.1	7.0	9.7
Gkioxari [12]	9.9	7.2	10.8
GPNN [34]	13.1	9.3	14.2
iCAN [10]	14.8	10.5	16.1
s+o+p	18.7	13.8	20.1
s+o+vp	17.7	11.6	19.5
s+o+p+vp	**19.4**	**14.6**	**20.9**

Table 1: Retrieval results on HICO-DET dataset (mAP).

Peyre, J., Sivic, J., Laptev, I., & Schmid, C. (2020). Detecting Unseen Visual Relations Using Analogies. ICCV, 2020.

Transfer by Analogy Unseen Triplets

- 25 zero- shot triplets of HICO-DET
- Base is without analogy transfer (phase 2)



Peyre, J., Sivic, J., Laptev, I., & Schmid, C. (2020). Detecting Unseen Visual Relations Using Analogies. ICCV, 2020.

Transfer by Analogy Unseen, Unusual Triplets

- UnRel dataset
- Used $\Gamma = deep$

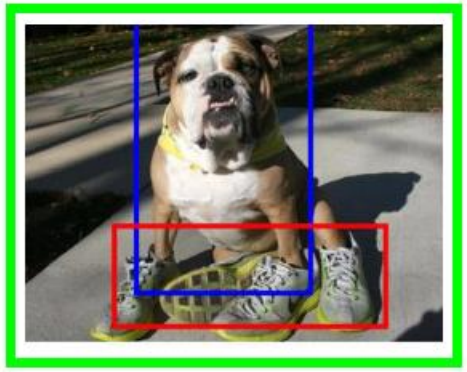
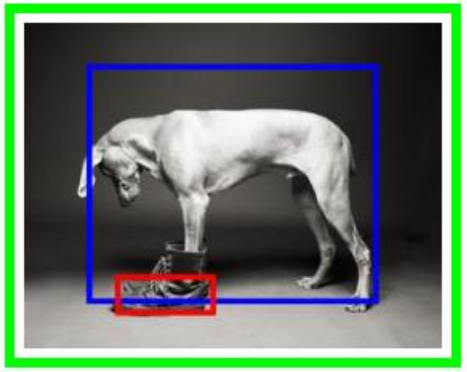
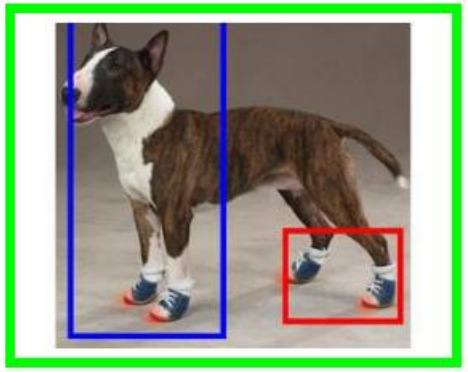
	With GT	With candidates		
	-	union	subj	subj/obj

Query (Q) / Source (S)

(Q) **dog wear shoes**

(S) person wear shoes
(S) person wear shoe
(S) person wear skis
(S) person wear pants
(S) person wear jeans

Top true positives

Top false positive



Table 3: Retrieval on UnRel (mAP) with IoU=0.3.

Peyre, J., Sivic, J., Laptev, I., & Schmid, C. (2020). Detecting Unseen Visual Relations Using Analogies. ICCV, 2020.

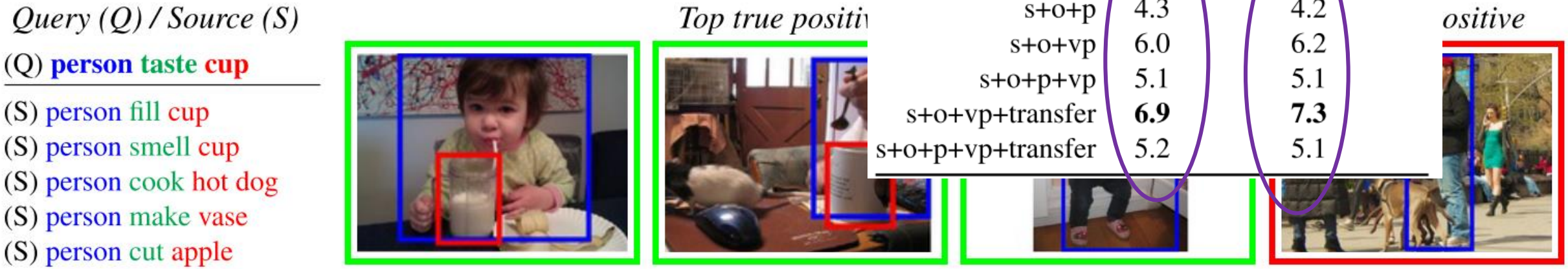
Transfer by Analogy Unseen (out-of-vocabulary) Triplets

- Trained on HICO-DET, evaluated on Coco-a Daaset

All Unseen

Unseen Out-of-vocabulary only

	all	out of vocabulary
s+o+p	4.3	4.2
s+o+vp	6.0	6.2
s+o+p+vp	5.1	5.1
s+o+vp+transfer	**6.9**	**7.3**
s+o+p+vp+transfer	5.2	5.1



Peyre, J., Sivic, J., Laptev, I., & Schmid, C. (2020). Detecting Unseen Visual Relations Using Analogies. ICCV, 2020.

Conclusion

Conclusion

- Authors propose combined method to detect visual relations:
  - Compositional representations
  - Visual phrase representations
- Method involves:
  - Visual and text embeddings
  - Triplet Loss
  - Transfer learning via transformation function

- Their method outperforms existing models in:
  - Seen triplets
  - Unseen triplets
  - Out-of-Vocabulary unseen triplets

Peyre, J., Sivic, J., Laptev, I., & Schmid, C. (2020). Detecting Unseen Visual Relations Using Analogies. ICCV, 2020.

Thank you