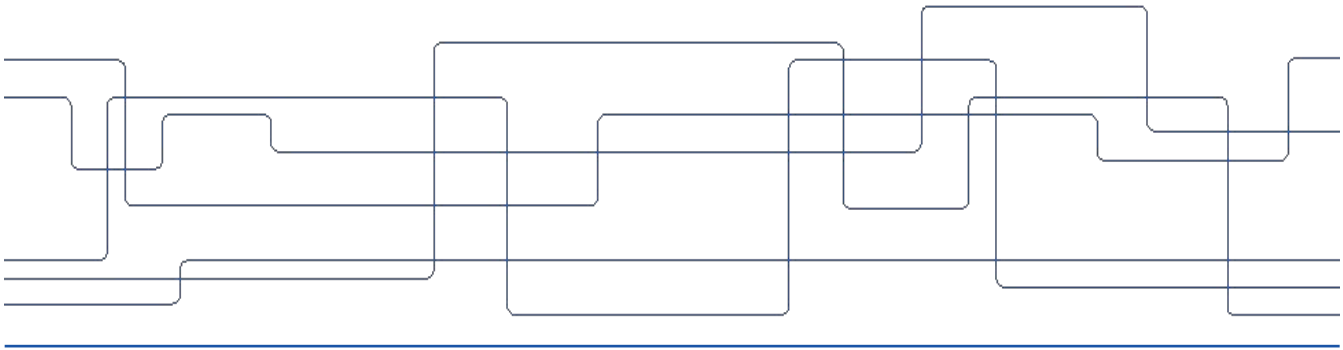
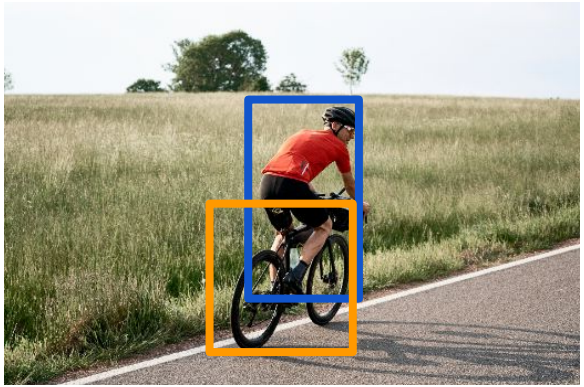
Visual Relationships

Federico Baldassarre
3 December 2019



Visual Phrases / Triplets

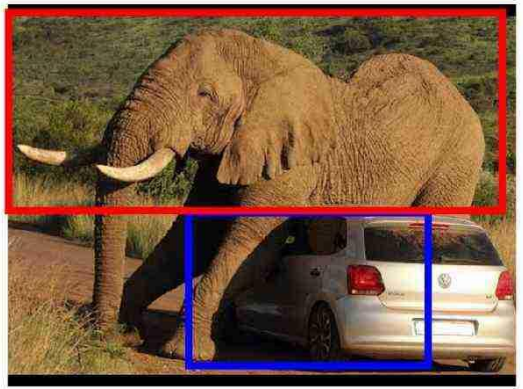


person - ride - bike

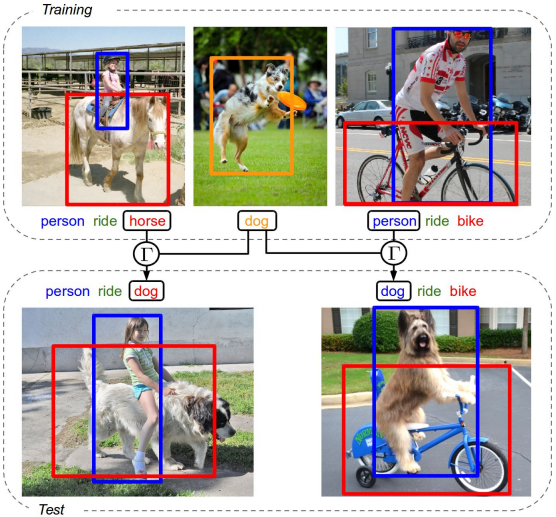
Challenges

- Expensive to annotate
- Unseen objects
- Unseen predicates
- Unseen (s, p, o) triplets

Visual Relationships



car under elephant



Weakly-Supervised Learning of Visual Relations

Julia Peyre, Josef Sivic,
Ivan Laptev, Cordelia Schmid
ICCV 2017

Detecting unseen visual relations using analogies

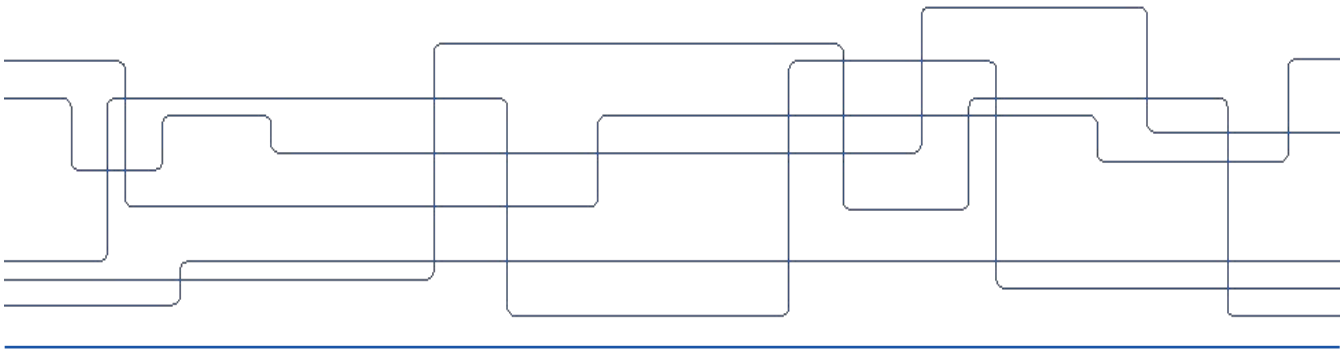
Julia Peyre, Ivan Laptev,
Cordelia Schmid, Josef Sivic
ICCV 2019

KTH ROYAL INSTITUTE
OF TECHNOLOGY

Weakly-Supervised Learning of Visual Relations

Julia Peyre, Josef Sivic, Ivan Laptev, Cordelia Schmid

ICCV 2017

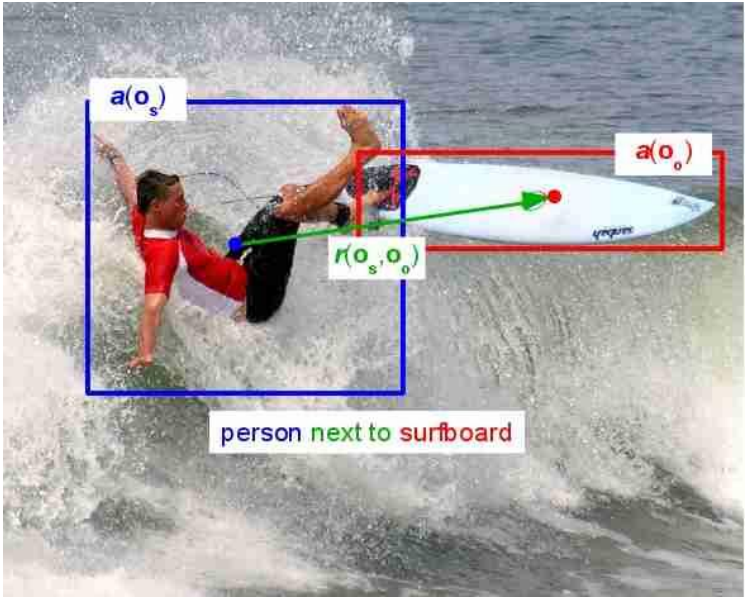


Weakly-Supervised Learning of Visual Relations

Contributions

- Encoding of appearance and spatial configuration
- Weakly-supervised training (image labels only)
- New dataset of Unusual Relations

Encoding of appearance and spatial configuration



Visual appearance

- Faster R-CNN features (VGG-16 on ImageNet)
- L2 normalization
- PCA 4096 -> 300
- L2 normalization

$$a(o_s), a(o_s) \in \mathbb{R}^{300}$$

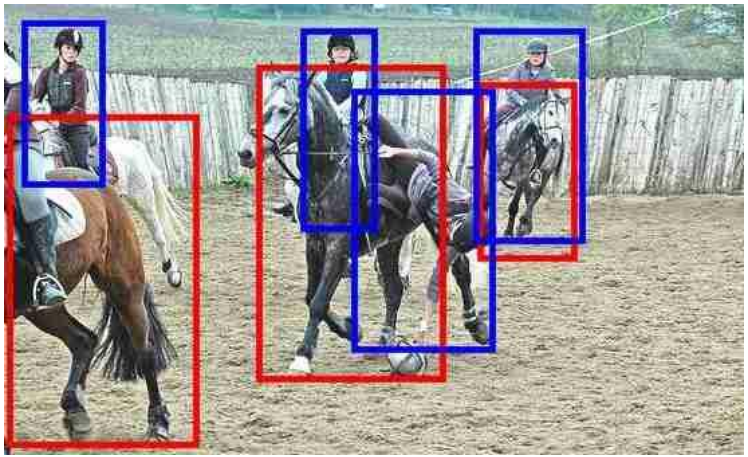
Spatial configuration

- Box sizes, displacement and area as 6D vector
- Gaussian Mixture Model with 400 gaussians

$$r(o_s, o_o) \in \mathbb{R}^{400}$$

Is all the normalization and playing around with GMM and PCA needed?

Weakly-supervised training (image labels only)



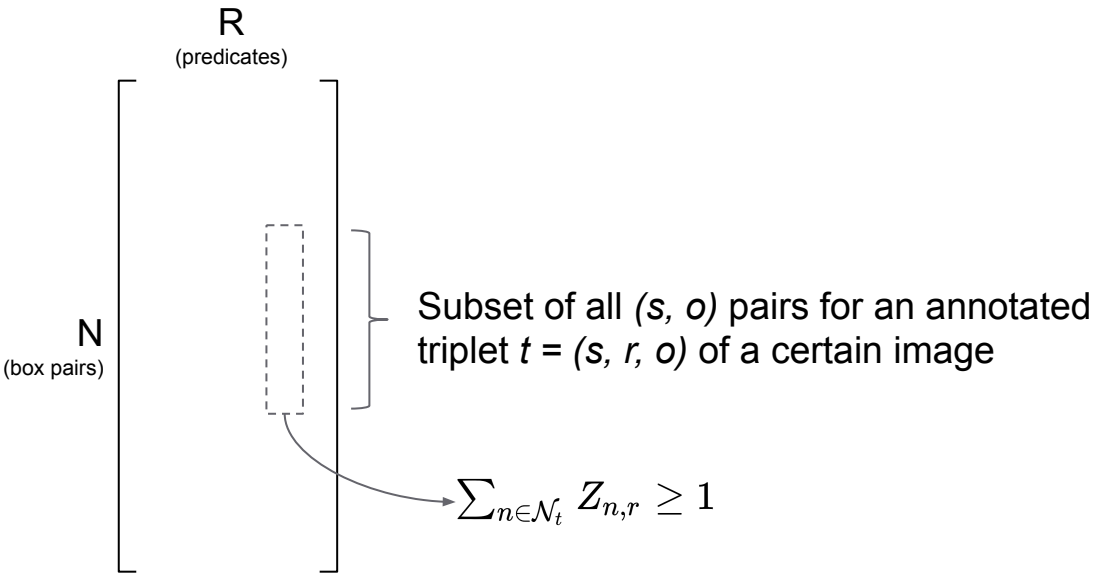
person falling horse

- Only one annotation with
  - s = person
  - o = horse
- 4 person boxes
- 3 horse boxes
- 12 candidate pairs

Weakly-supervised training (image labels only)

$$\min_Z \min_W \frac{1}{N}||Z - XW||^2 + \lambda||W||^2$$

$$s.t. \sum_{n \in \mathcal{N}_t} Z_{n,r} \geq 1 \quad \forall t \in \mathcal{T}$$

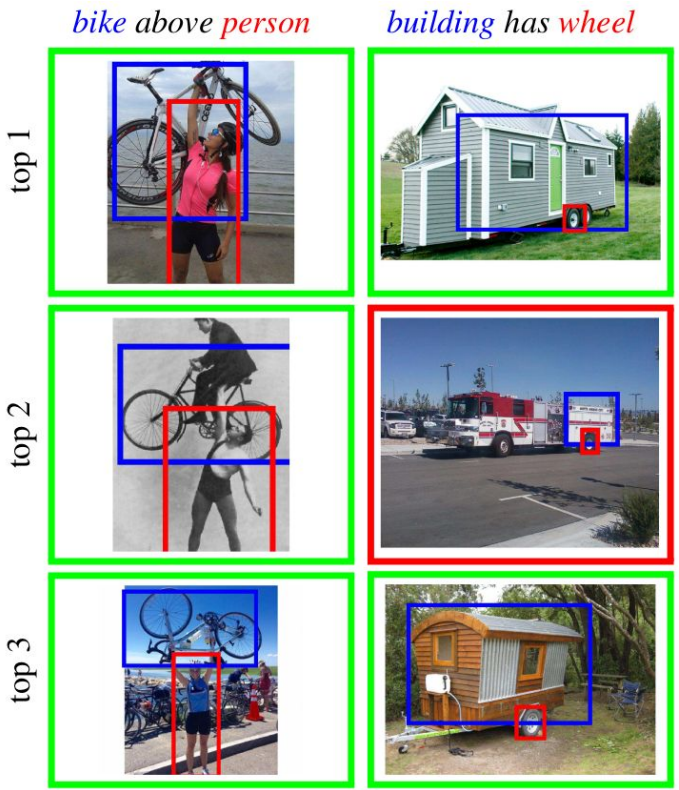


Main Results

		Predicate Det.		Phrase Det.		Relationship Det.	
		All	Unseen	All	Unseen	All	Unseen
	Full sup.						
a.	Visual Phrases [44]	0.9	-	0.04	-	-	-
b.	Visual [32]	7.1	3.5	2.2	1.0	1.6	0.7
c.	Language (likelihood) [32]	18.2	5.1	0.08	0.00	0.08	0.00
d.	Visual + Language [32]	47.9	8.5	16.2	3.4	13.9	3.1
e.	Language (full) [32]	48.4	12.9	15.8	4.6	13.9	4.3
f.	Ours [S]	42.2	22.2	13.8	7.4	12.4	7.0
g.	Ours [A]	46.3	16.1	14.9	5.6	12.9	5.0
h.	Ours [S+A]	50.4	**23.6**	16.7	**7.4**	14.9	**7.1**
i.	Ours [S+A] + Language [32]	**52.6**	21.6	**17.9**	6.8	**15.8**	6.4
	Weak sup.						
j.	Ours [S+A]	46.8	19.0	16.0	6.9	14.1	6.7
k.	Ours [S+A] - Noisy	46.4	17.6	15.1	6.0	13.4	5.6

There are no standard deviations of multiple runs, but it's possible that the weakly-supervised training scheme (j) performs just like a noisy training scheme (k)

Unusual Relations Dataset



	With GT	With candidates		
	-	union	subj	subj/obj
Chance	38.4	8.6	6.6	4.2
Full sup.				
DenseCap [22]	-	6.2	6.8	-
Reproduce [32]	50.6	12.0	10.0	7.2
Ours [S+A]	**62.6**	**14.1**	**12.1**	**9.9**
Weak sup.				
Ours [S+A]	58.5	13.4	11.0	8.7
Ours [S+A] - Noisy	55.0	13.0	10.6	8.5

Table 2: Retrieval on UnRel (mAP) with IoU=0.3

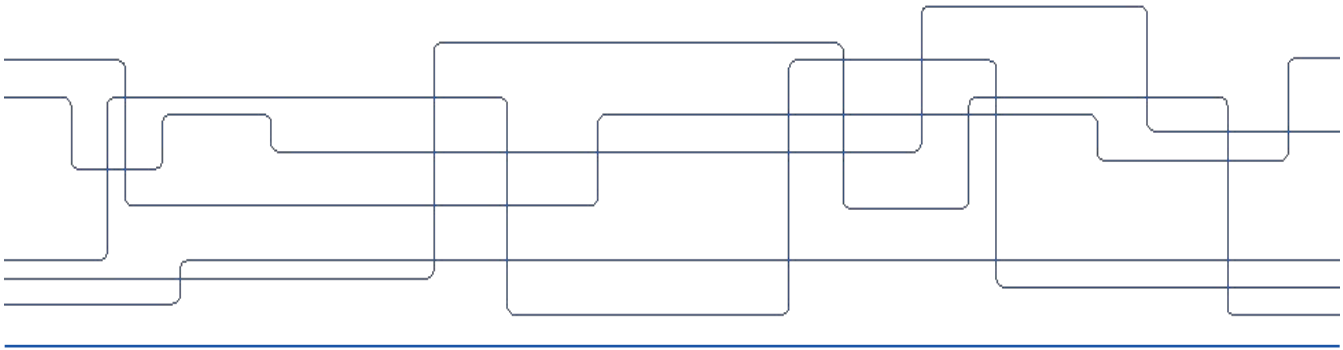
Using candidate boxes gives much lower performances, can this be a problem of the object detection part?

KTH ROYAL INSTITUTE
OF TECHNOLOGY

Detecting unseen visual relations using *analogies*

Julia Peyre, Ivan Laptev, Cordelia Schmid, Josef Sivic

ICCV 2019

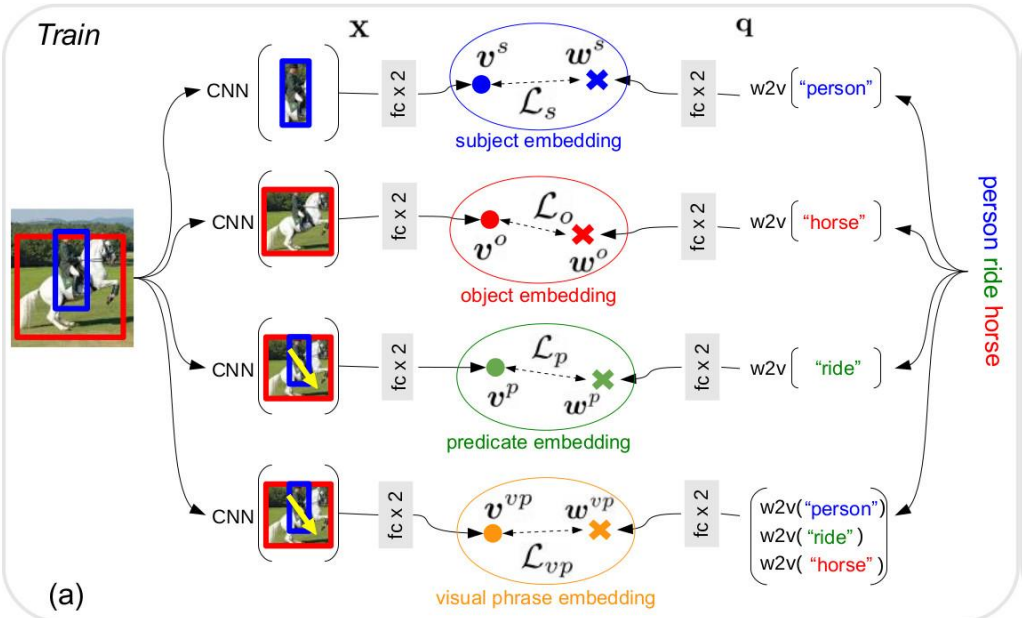


Detecting unseen visual relations using *analogies*

Contributions

- Combine compositional and phrase-based approaches
- Retrieval using *analogy* transformations

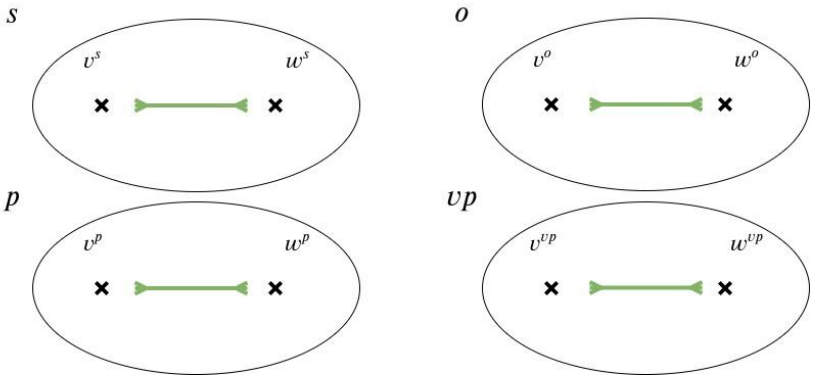
Compositional and phrase-based approaches



GMM and PCA from the previous paper have been replaced by MLPs

Training - positive samples

Visual pair *i*



Language triplet *t*

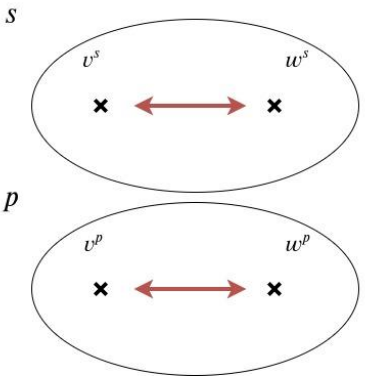
`person - ride - bike`

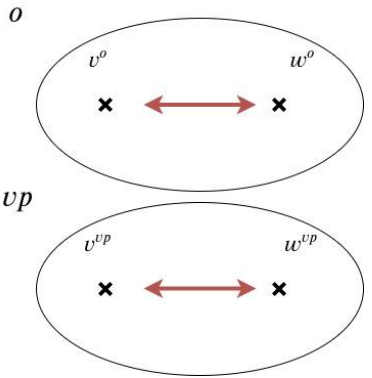
$$\mathcal{L}_b = -\log\left(\sigma\langle v_i^b, w_t^b\rangle\right)$$

In all four spaces {s, p, o, vp}, the visual and language embeddings are pushed closer together

Training - negative samples

Visual pair *i*





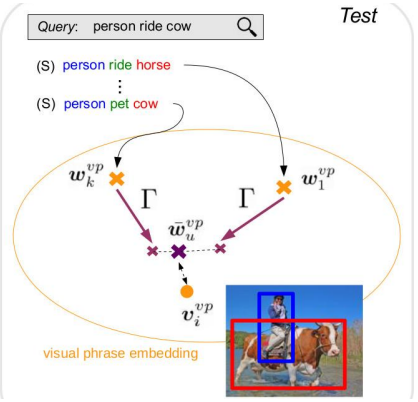
Language triplet *t*

`person - wash - bike`

$$\mathcal{L}_b = -\log\left(1 - \sigma\langle v_i^b, w_t^b \rangle\right)$$

In all four spaces {s, p, o, vp}, the visual and language embeddings are pushed apart (also person - person ?)

Transfer by *analogy*



Retrieval without analogy

- Embed an unseen language query as *(s, p, o, vp)*
- Retrieve box pairs whose embeddings of *(s, p, o, vp)* are close to the query

Retrieval with analogy

- Find seen triplets that are similar enough to the unseen query (similarity is based on *s, p, o*)
- Build a *vp* embedding by aggregating the *vp* embeddings of seen triplets translated by an analogy transformation
- Retrieve box pairs whose embedding *(s, p, o, vp)* is close to the newly built query

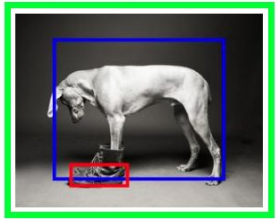
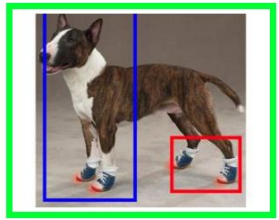
Transfer by *analogy*

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

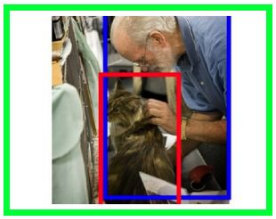
Query (Q) / Source (S)

(Q) **person pet cat**

(S) person pet dog
(S) person pet giraffe
(S) person pet cow
(S) person pet elephant
(S) person scratch cat

Top true positives





Top false positive

Main Results

	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**

Seen triplets

	Base	With aggregation G			
	-	$\Gamma=\emptyset$	$\Gamma=0$	$\Gamma=linear$	$\Gamma=deep$
s+o+p	23.2	-	-	-	-
s+o+vp+transfer	24.1	9.6	24.8	27.6	**28.6**
s+o+p+vp+transfer	23.6	12.5	24.5	25.4	**25.7**
supervised	33.7	-	-	-	-

Unseen triplets

Future directions?

- Contextual query embeddings (BERT ?)
- Different predicate visual representation
- Entity-relationship reasoning

KTH ROYAL INSTITUTE
OF TECHNOLOGY

Thanks