Visual Relation Detection

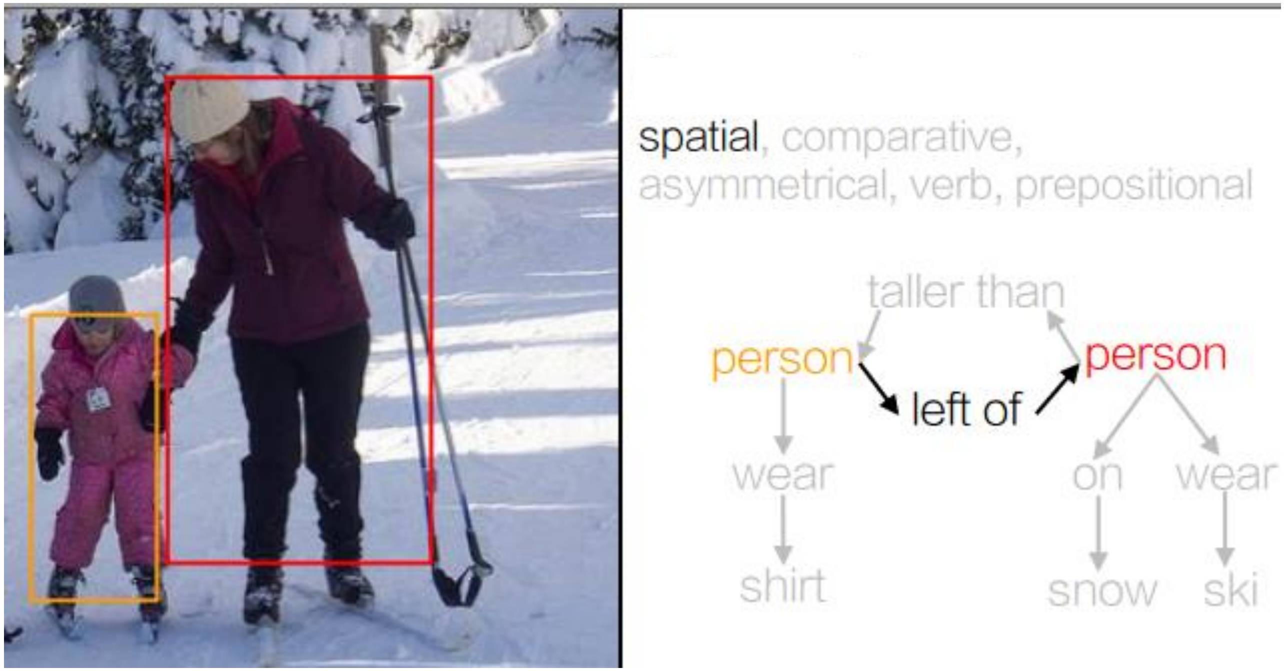
Deep Relation Network

ECCV2016(Oral) Lu C, Krishna R, Bernstein M, et al. Visual relationship detection with language priors[J]. arXiv preprint arXiv:1608.00187, 2016.

CVPR2017(Oral) Dai B, Zhang Y, Lin D. Detecting Visual Relationships with Deep Relational Networks[J]. arXiv preprint arXiv:1704.03114, 2017.

Task Definition

Task Definition



spatial, comparative, asymmetrical, verb, prepositional

taller than

person

left of

person

wear

shirt

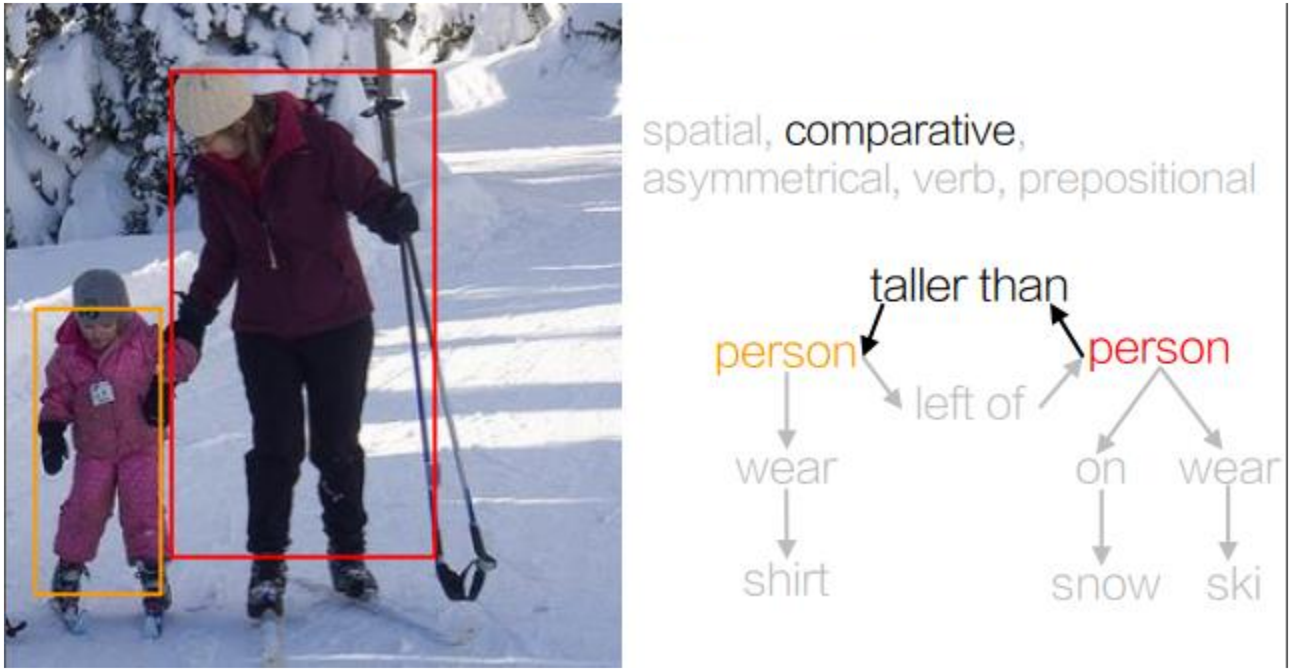
on

wear

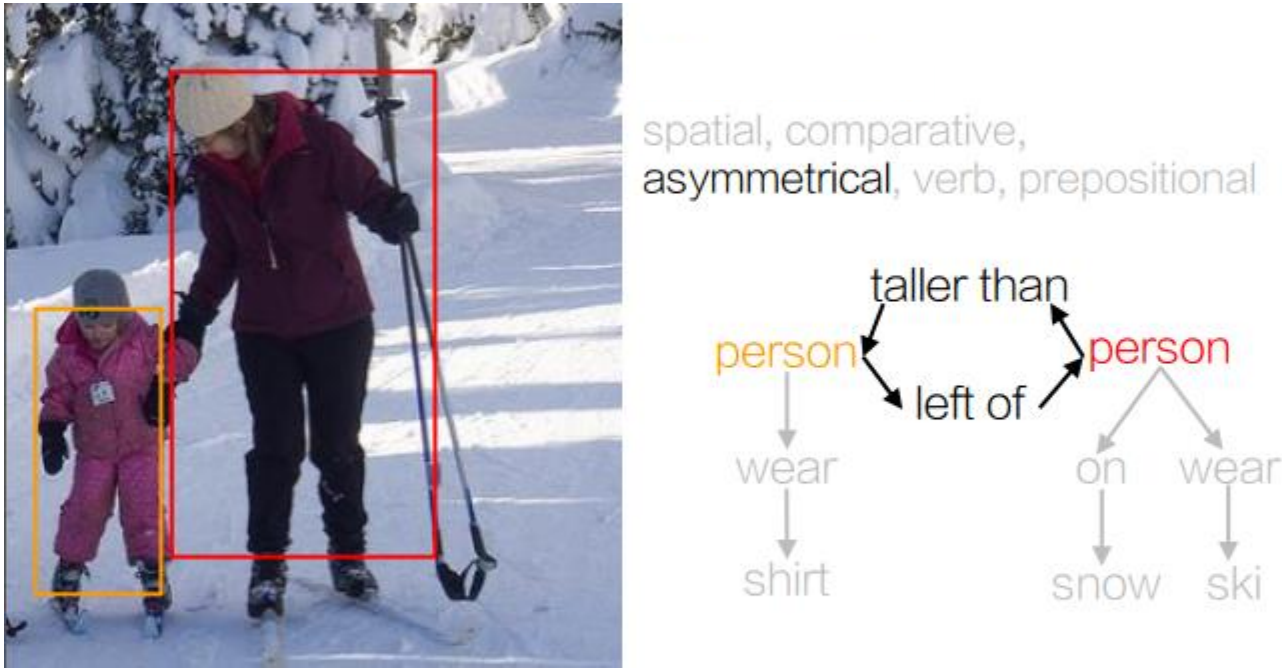
snow

ski

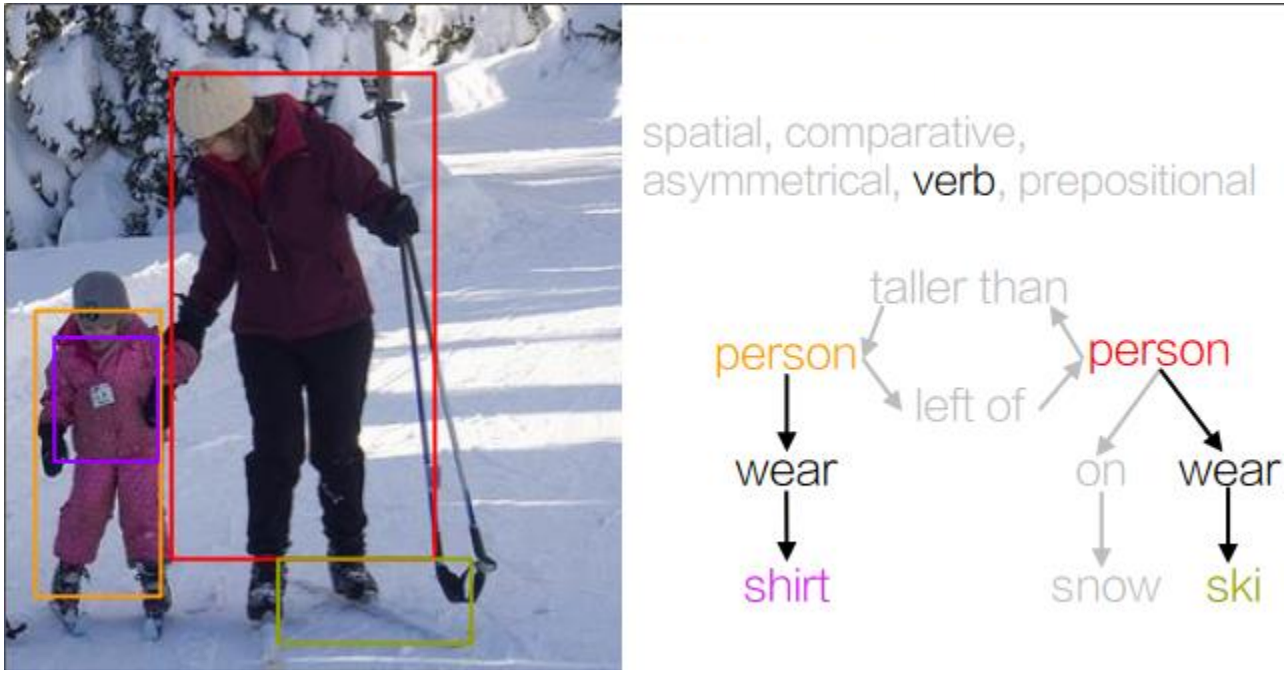
Task Definition



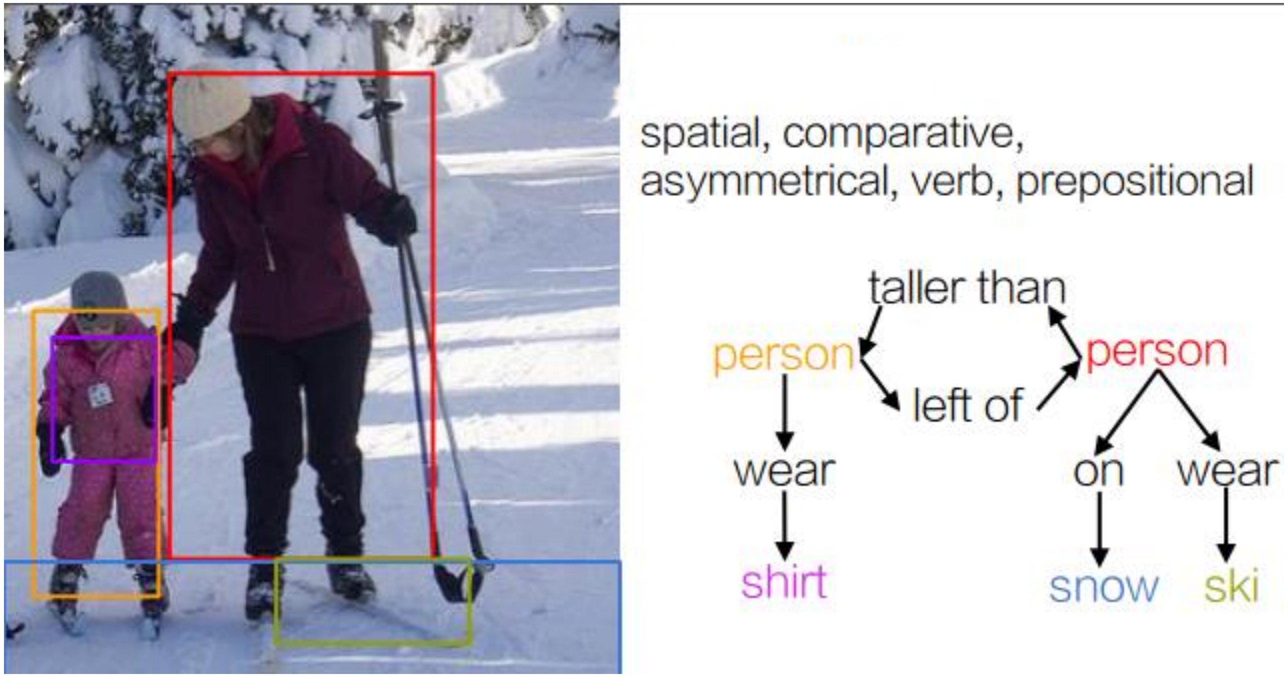
Task Definition



Task Definition



Task Definition



spatial, comparative, asymmetrical, verb, prepositional

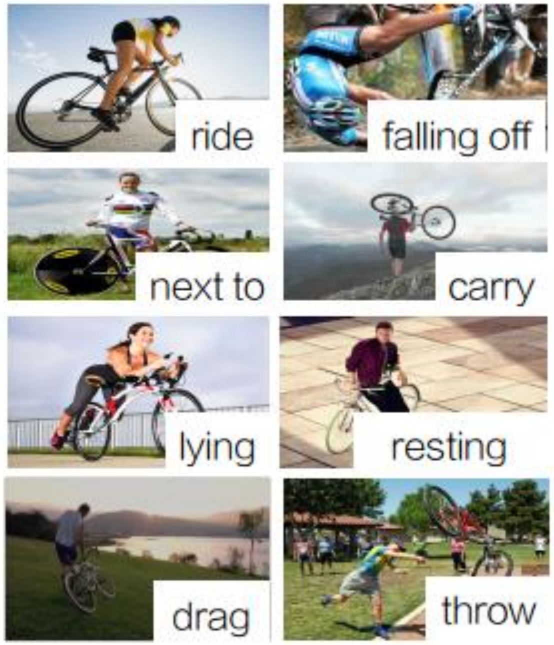
Task Definition

- Prediction recognition
  - Input: img + (lables, BBox) of Subject & Object
  - Output: Triplet (s; r; o), *e.g*. (girl, on, horse)
  - Metric: Recall@50
- Union box detection:
  - Input: img
  - Output: Triplet (*s; r; o*)
  - Metric: Recall@50 when IoU thresh=0.5
- **Two boxes detection:**
  - Similar to 2, except treating Subject & Object individually

Visual Relationship Dataset(VGD)

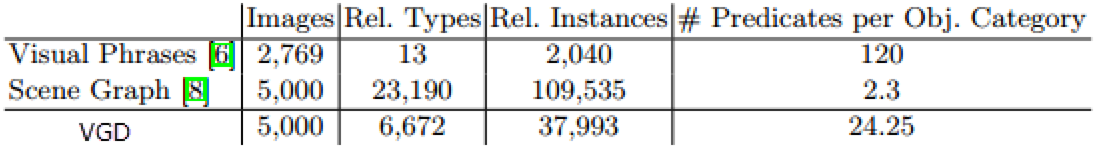
- Observation #1: Number of Visual Phases

	Images	Rel. Types	Rel. Instances	# Predicates per Obj. Category
Visual Phrases [6]	2,769	13	2,040	120
Scene Graph [8]	5,000	23,190	109,535	2.3
VGD	5,000	6,672	37,993	24.25

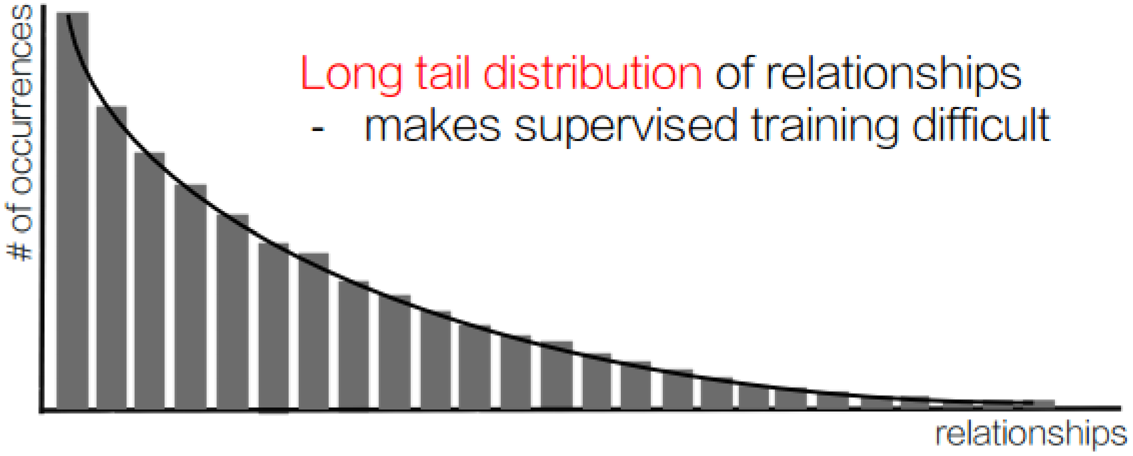


Visual Relationship Dataset(VGD)

- Observation #2: Unbalance Data



	Images	Rel. Types	Rel. Instances	# Predicates per Obj. Category
Visual Phrases [6]	2,769	13	2,040	120
Scene Graph [8]	5,000	23,190	109,535	2.3
VGD	5,000	6,672	37,993	24.25

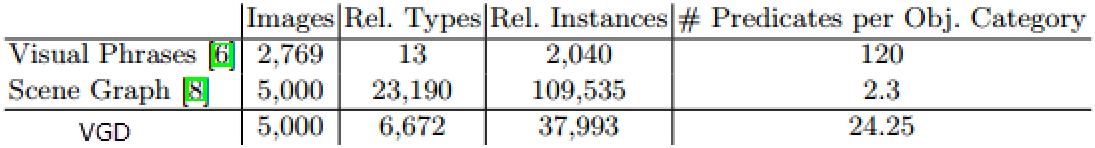


Long tail distribution of relationships

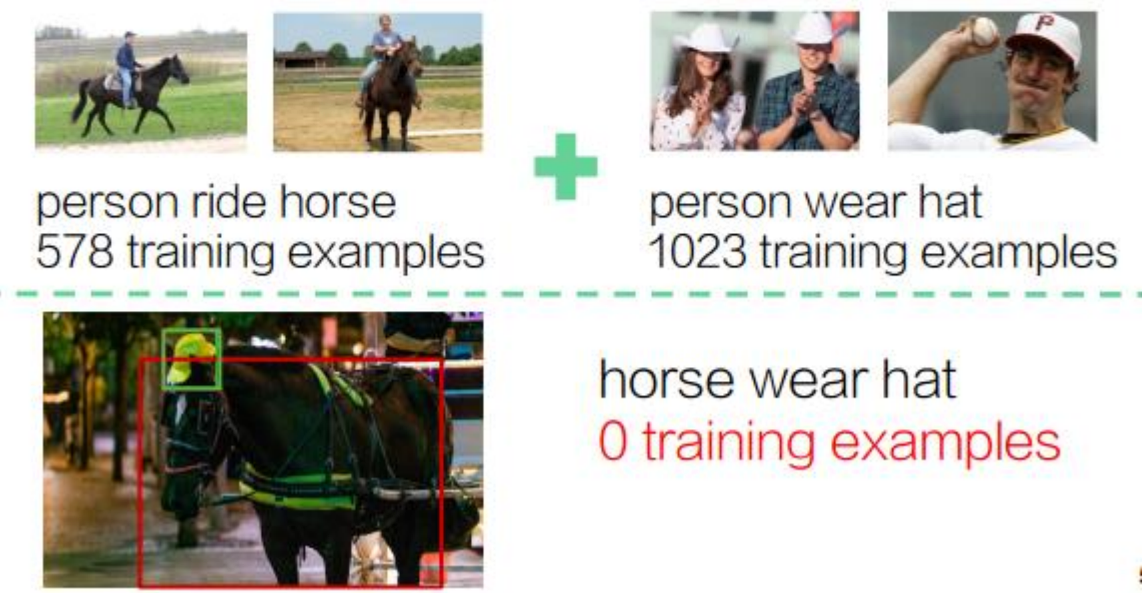
- makes supervised training difficult

Visual Relationship Dataset(VGD)

- Observation #3: Zero Shot Detection



	Images	Rel. Types	Rel. Instances	# Predicates per Obj. Category
Visual Phrases [6]	2,769	13	2,040	120
Scene Graph [8]	5,000	23,190	109,535	2.3
VGD	5,000	6,672	37,993	24.25



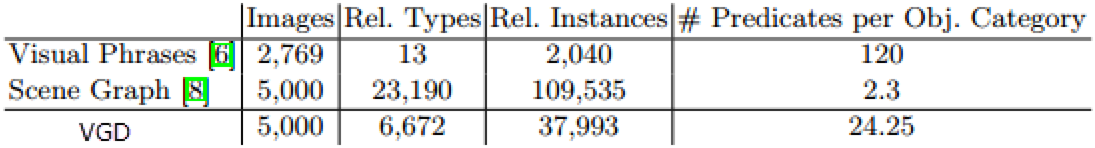
person ride horse
578 training examples

person wear hat
1023 training examples

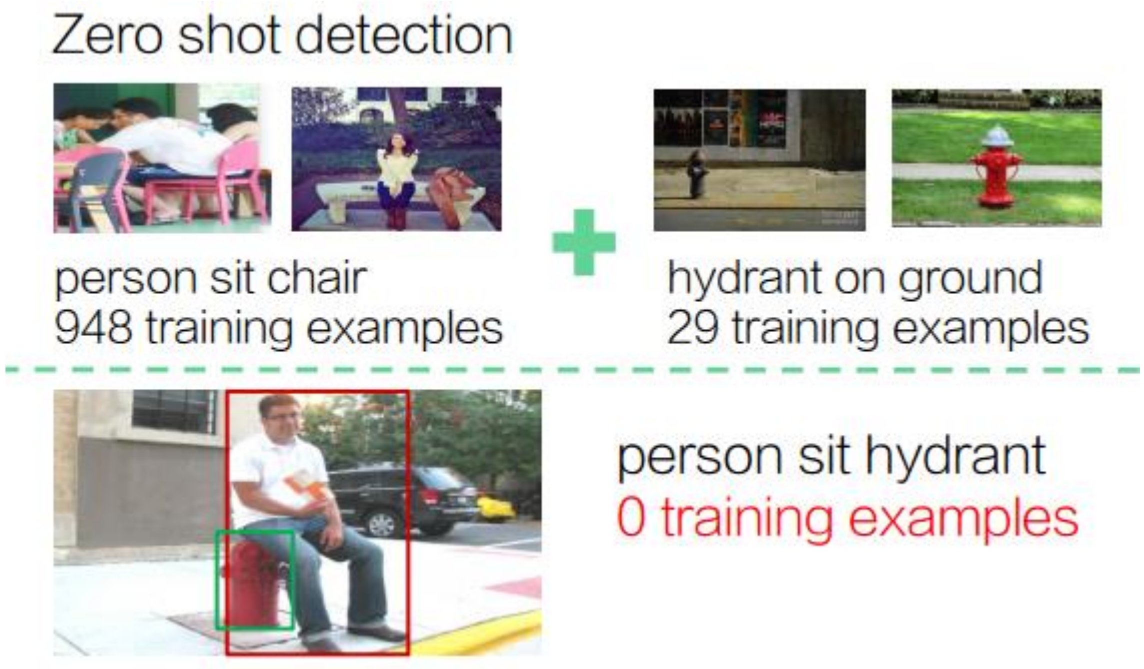
horse wear hat
0 training examples

Visual Relationship Dataset(VGD)

- Observation #3: Zero Shot Detection



	Images	Rel. Types	Rel. Instances	# Predicates per Obj. Category
Visual Phrases [6]	2,769	13	2,040	120
Scene Graph [8]	5,000	23,190	109,535	2.3
VGD	5,000	6,672	37,993	24.25



Zero shot detection

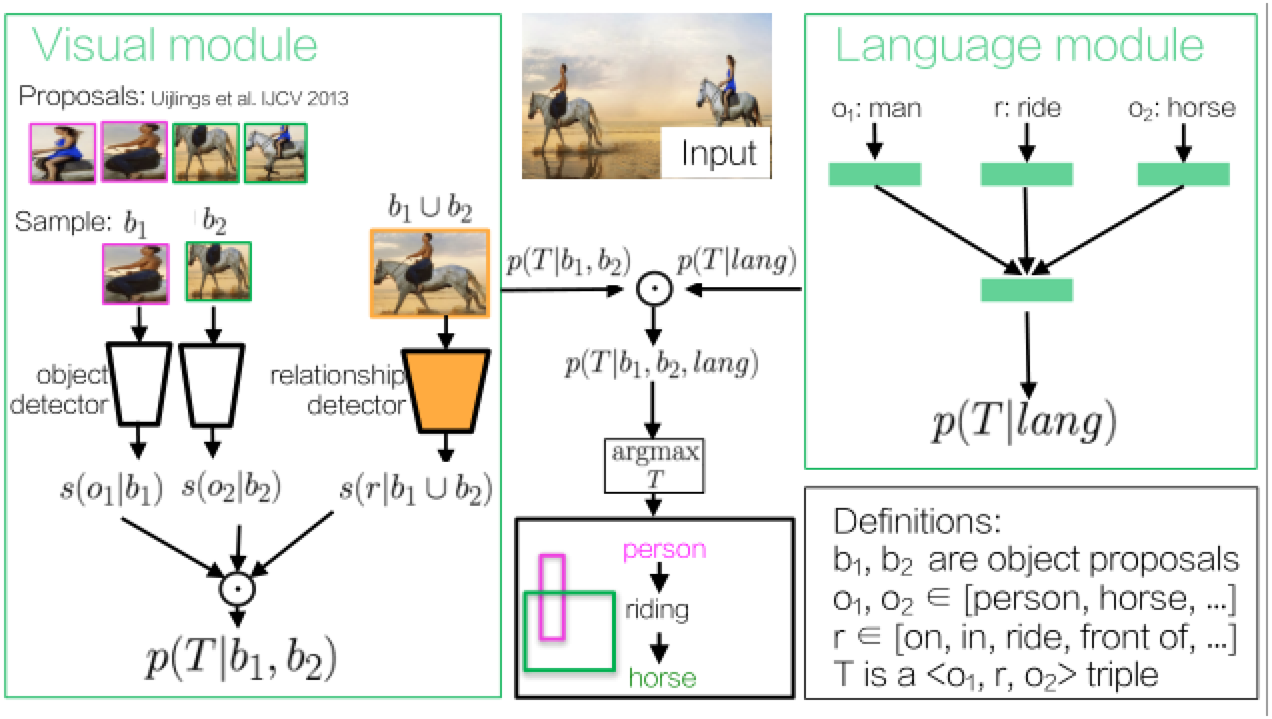
person sit chair
948 training examples

hydrant on ground
29 training examples

person sit hydrant
0 training examples

Related Work

Combine Language Model

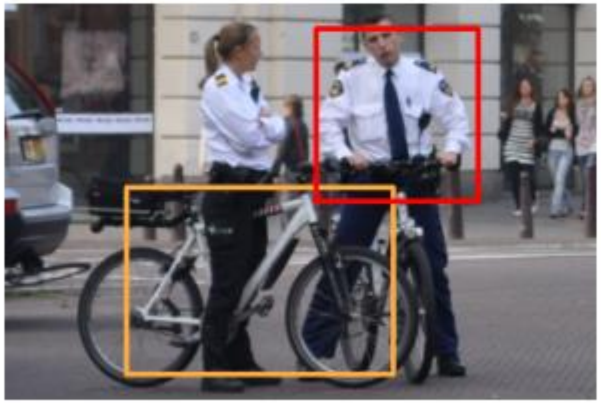


ECCV2016(Oral) Lu C, Krishna R, Bernstein M, et al. Visual relationship detection with language priors[J]. arXiv preprint arXiv:1608.00187, 2016.

Related Work

Combine Language Model

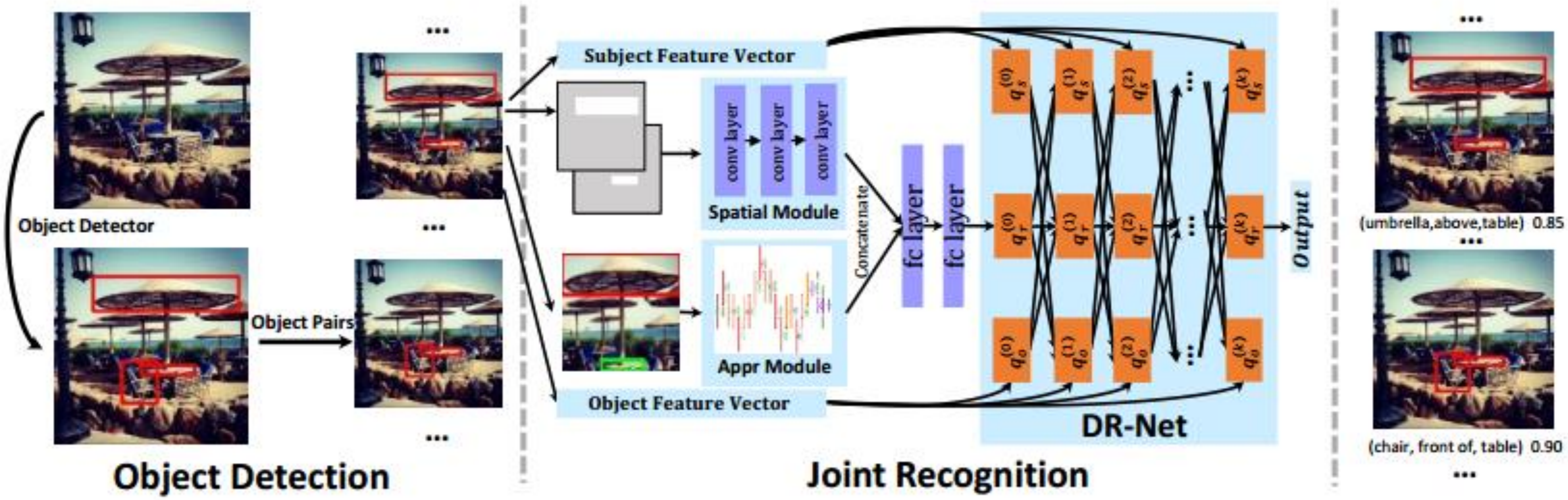
Weakness:



person ride bicycle ☹

DRNet

Pipeline:


Object Detector
Object Pairs
Object Detection
Subject Feature Vector
conv layer
conv layer
conv layer
Spatial Module
Appr Module
Object Feature Vector
Concatenate
fc layer
fc layer
$q_s^{(0)}$ $q_s^{(1)}$ $q_s^{(2)}$ $q_s^{(K)}$
$q_r^{(0)}$ $q_r^{(1)}$ $q_r^{(2)}$ $q_r^{(K)}$
$q_o^{(0)}$ $q_o^{(1)}$ $q_o^{(2)}$ $q_o^{(K)}$
DR-Net
Output
Joint Recognition
(umbrella,above,table) 0.85
(chair, front of, table) 0.90

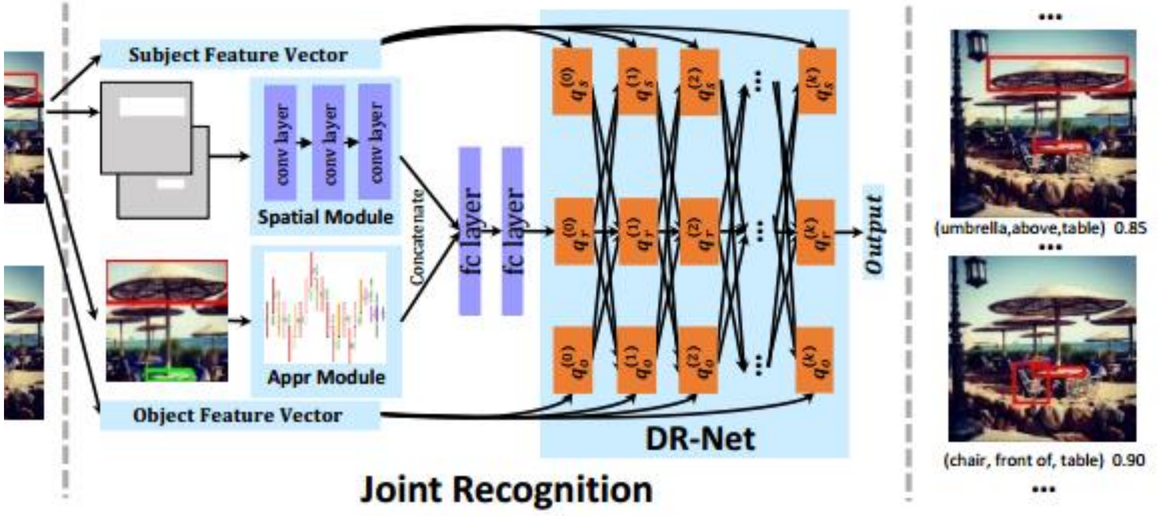
CVPR2017(Oral) Dai B, Zhang Y, Lin D. Detecting Visual Relationships with Deep Relational Networks[J]. arXiv preprint arXiv:1704.03114, 2017.

Pretrain

- **Object detection**
  - Output: BBox + Appearance feature
- **Pair filtering**
  - low-cost neural network
  - Filter out meaningless pair

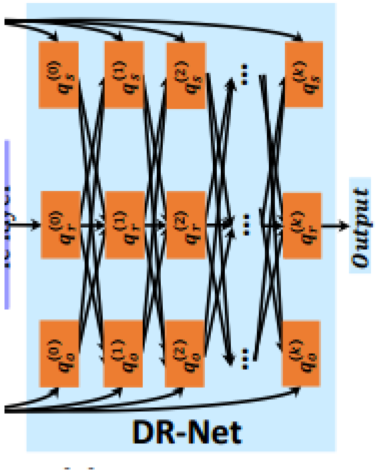
Joint Recognition

- Appearance
- Spatial Configuration
- Statistical Relation


Subject Feature Vector
conv layer
conv layer
conv layer
Spatial Module
Appr Module
Object Feature Vector
Concatenate
fc layer
fc layer
$q_s^{(0)}$ $q_s^{(1)}$ $q_s^{(2)}$ ... $q_s^{(k)}$
$q_r^{(0)}$ $q_r^{(1)}$ $q_r^{(2)}$ ... $q_r^{(k)}$
$q_o^{(0)}$ $q_o^{(1)}$ $q_o^{(2)}$ ... $q_o^{(k)}$
Output
DR-Net
Joint Recognition
...
(umbrella,above,table) 0.85
...
(chair, front of, table) 0.90
...

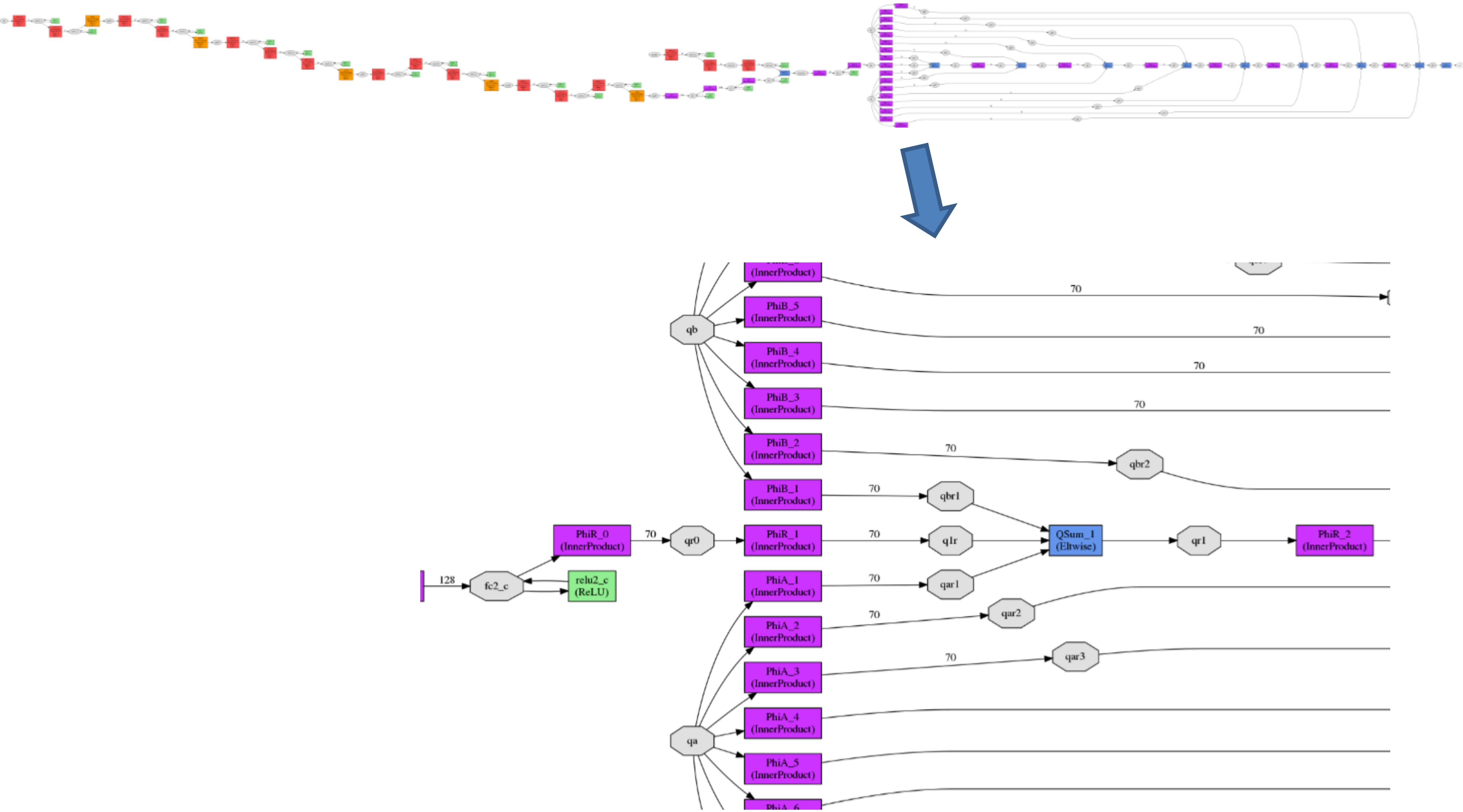
Statistical Relation (DRNet)

- Equivalents between Discriminative Model & Generative Model!



$$
\begin{aligned}
\mathbf{q}_s' &= \sigma\left(\mathbf{W}_a\mathbf{x}_s + \mathbf{W}_{sr}\mathbf{q}_r + \mathbf{W}_{so}\mathbf{q}_o\right),\\
\mathbf{q}_r' &= \sigma\left(\mathbf{W}_r\mathbf{x}_r + \mathbf{W}_{rs}\mathbf{q}_s + \mathbf{W}_{ro}\mathbf{q}_o\right),\\
\mathbf{q}_o' &= \sigma\left(\mathbf{W}_a\mathbf{x}_o + \mathbf{W}_{os}\mathbf{q}_s + \mathbf{W}_{or}\mathbf{q}_r\right).
\end{aligned}
$$

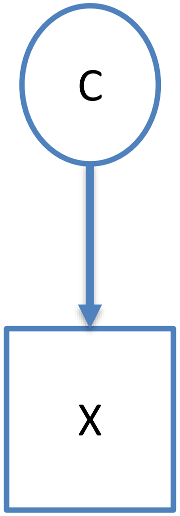
Statistical Relation (DRNet)



Toy Example

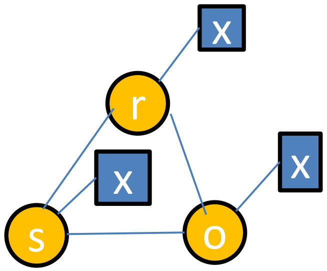
- Bayesian Net
- C is label, $x \sim \mathcal{N}(\mu, \sigma)$ is raw feature/observation
- We prove $p(C_1|x) = \sigma(w^T x + w_0)$

$$
\begin{aligned}
p(C_1|x) &= \frac{p(x|C_1)p(C_1)}{p(x|C_1)p(C_1)+p(x|C_2)p(C_2)} \\
&= \frac{1}{1+\frac{exp[(x-\mu_2)^2/2\sigma^2]p(C_2)}{exp[(x-\mu_1)^2/2\sigma^2]p(C_1)}} \\
&= \frac{1}{1+\frac{p_2}{p_1}exp\left[\frac{\mu_1-\mu_2}{2\sigma^2}x+\frac{\mu_2^2-\mu_1^2}{2\sigma^2}\right]} \\
&= \frac{1}{1+exp(w^Tx+w_0)}
\end{aligned}
$$



DRNet

Representation

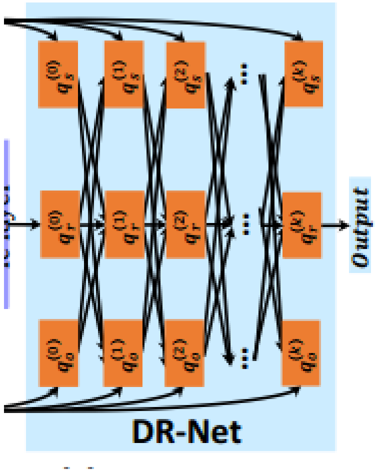


Inference

$$p(r, s, o|\mathbf{x}_r, \mathbf{x}_s, \mathbf{x}_o) = \frac{1}{Z} \exp\left(\Phi(r, s, o|\mathbf{x}_r, \mathbf{x}_s, \mathbf{x}_o; \mathbf{W})\right).$$

$$\Phi = \psi_a(s|\mathbf{x}_s; \mathbf{W}_a) + \psi_a(o|\mathbf{x}_o; \mathbf{W}_a) + \psi_r(r|\mathbf{x}_r; \mathbf{W}_r) + \varphi_{rs}(r, s|\mathbf{W}_{rs}) + \varphi_{ro}(r, o|\mathbf{W}_{ro}) + \varphi_{so}(s, o|\mathbf{W}_{so}).$$

$$p(r|s, o, \mathbf{x}_r; \mathbf{W}) \propto \exp\left(\psi_r(r|\mathbf{x}_r; \mathbf{W}_r) + \varphi_{rs}(r, s|\mathbf{W}_{rs}) + \varphi_{ro}(r, o|\mathbf{W}_{ro})\right).$$



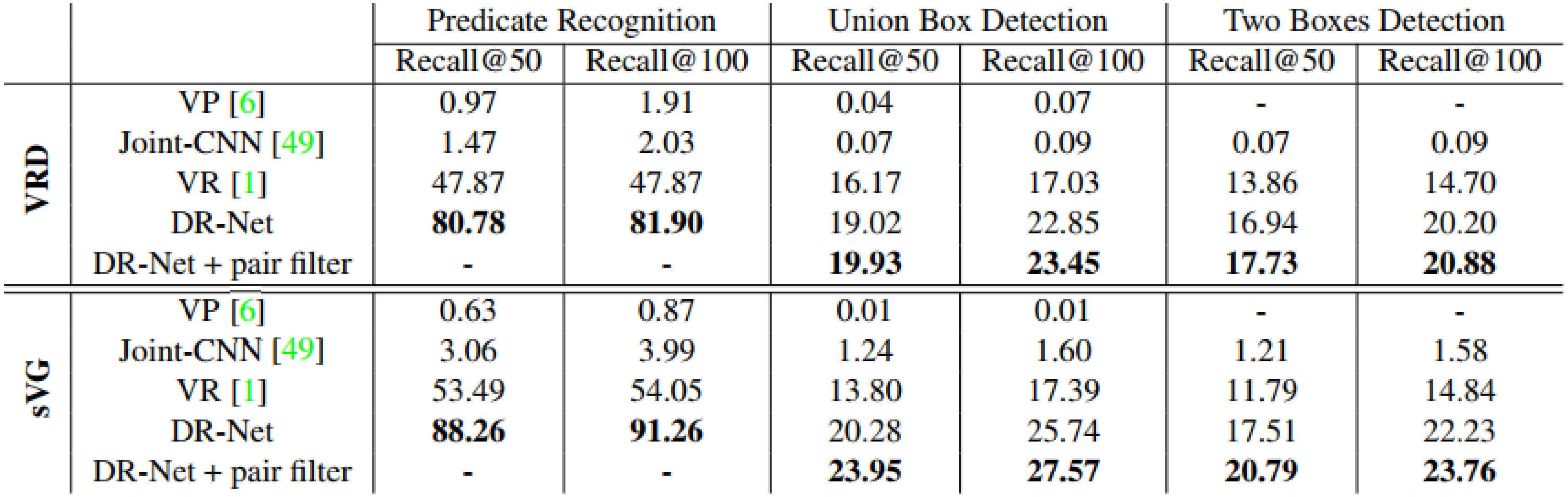
Unroll into a Network

$$\mathbf{q}_r = \boldsymbol{\sigma}\left(\mathbf{W}_r\mathbf{x}_r + \mathbf{W}_{rs}\mathbf{q}_s + \mathbf{W}_{ro}\mathbf{q}_o\right).$$

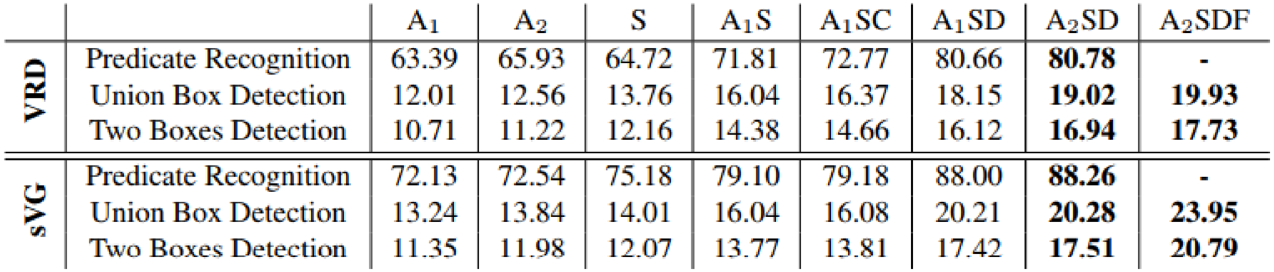
$$\mathbf{q}'_s = \boldsymbol{\sigma}\left(\mathbf{W}_a\mathbf{x}_s + \mathbf{W}_{sr}\mathbf{q}_r + \mathbf{W}_{so}\mathbf{q}_o\right),$$
$$\mathbf{q}'_r = \boldsymbol{\sigma}\left(\mathbf{W}_r\mathbf{x}_r + \mathbf{W}_{rs}\mathbf{q}_s + \mathbf{W}_{ro}\mathbf{q}_o\right),$$
$$\mathbf{q}'_o = \boldsymbol{\sigma}\left(\mathbf{W}_a\mathbf{x}_o + \mathbf{W}_{os}\mathbf{q}_s + \mathbf{W}_{or}\mathbf{q}_r\right).$$

Experiments

Performance



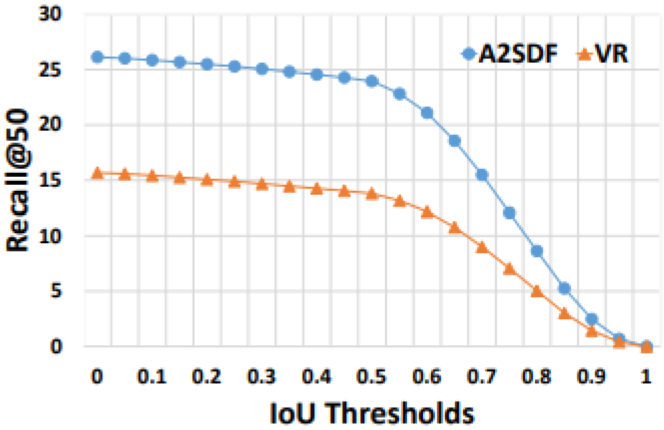
		Predicate Recognition		Union Box Detection		Two Boxes Detection	
		Recall@50	Recall@100	Recall@50	Recall@100	Recall@50	Recall@100
VRD	VP [6]	0.97	1.91	0.04	0.07	-	-
	Joint-CNN [49]	1.47	2.03	0.07	0.09	0.07	0.09
	VR [1]	47.87	47.87	16.17	17.03	13.86	14.70
	DR-Net	**80.78**	**81.90**	19.02	22.85	16.94	20.20
	DR-Net + pair filter	-	-	**19.93**	**23.45**	**17.73**	**20.88**
sVG	VP [6]	0.63	0.87	0.01	0.01	-	-
	Joint-CNN [49]	3.06	3.99	1.24	1.60	1.21	1.58
	VR [1]	53.49	54.05	13.80	17.39	11.79	14.84
	DR-Net	**88.26**	**91.26**	20.28	25.74	17.51	22.23
	DR-Net + pair filter	-	-	**23.95**	**27.57**	**20.79**	**23.76**

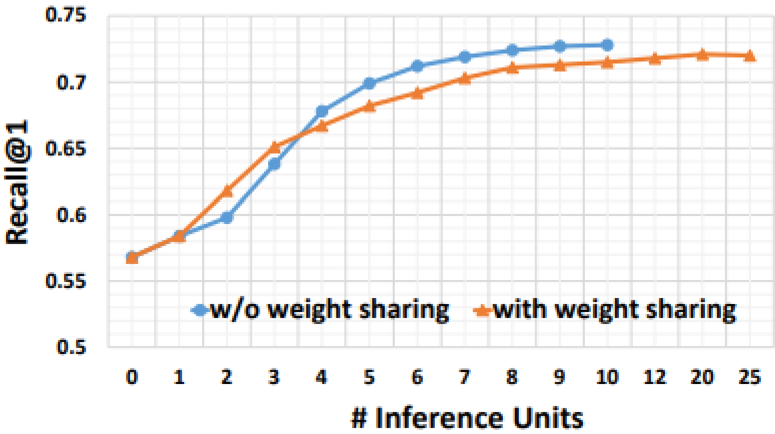


		A_1	A_2	S	A_1S	A_1SC	A_1SD	A_2SD	A_2SDF
VRD	Predicate Recognition	63.39	65.93	64.72	71.81	72.77	80.66	**80.78**	-
	Union Box Detection	12.01	12.56	13.76	16.04	16.37	18.15	**19.02**	**19.93**
	Two Boxes Detection	10.71	11.22	12.16	14.38	14.66	16.12	**16.94**	**17.73**
sVG	Predicate Recognition	72.13	72.54	75.18	79.10	79.18	88.00	**88.26**	-
	Union Box Detection	13.24	13.84	14.01	16.04	16.08	20.21	**20.28**	**23.95**
	Two Boxes Detection	11.35	11.98	12.07	13.77	13.81	17.42	**17.51**	**20.79**

Experiments

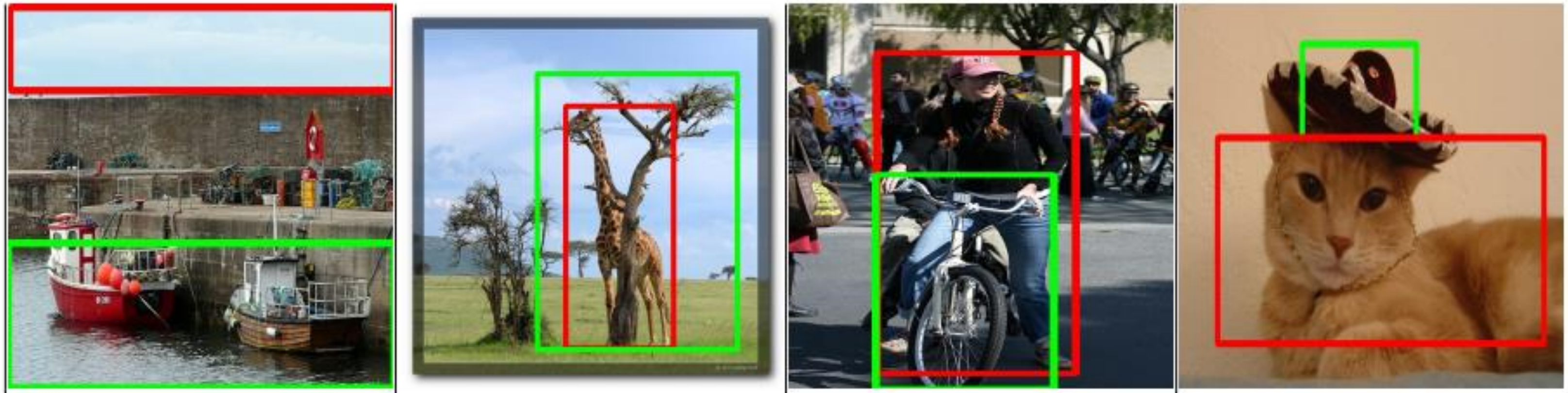
Hyper Param





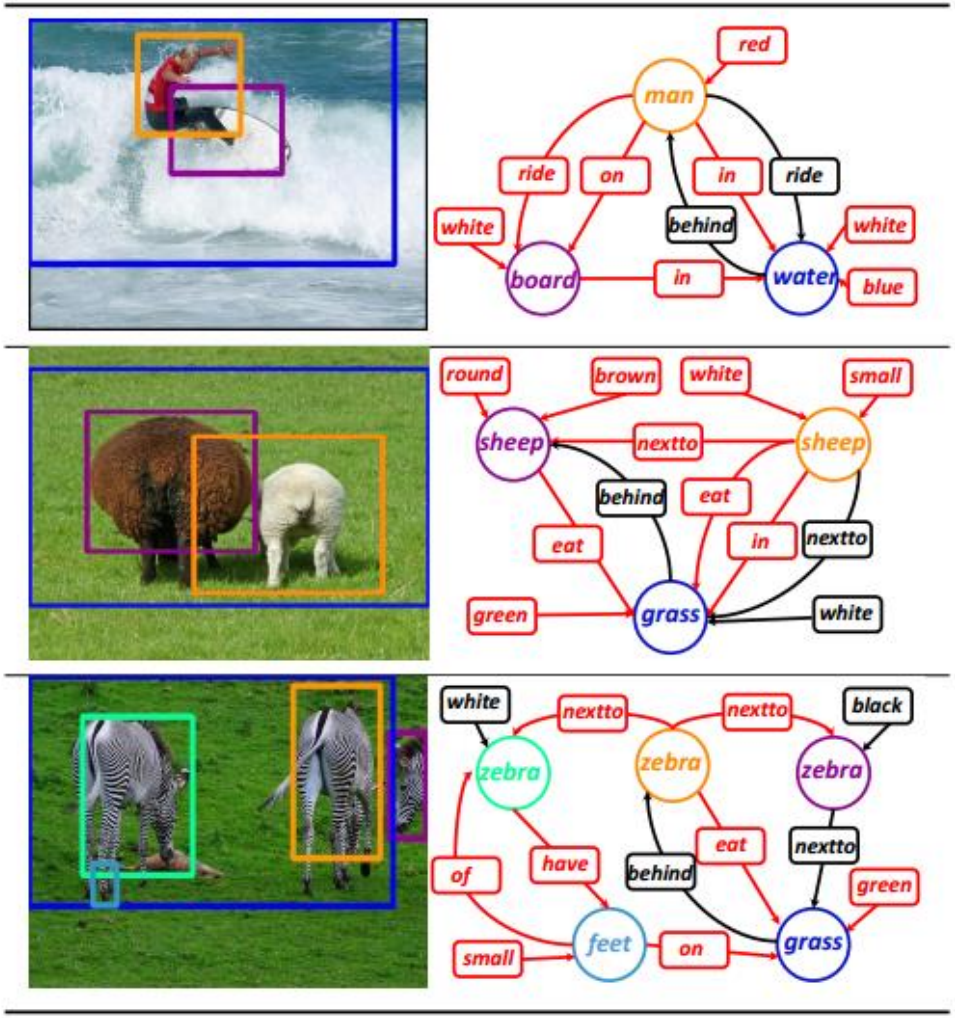
Experiments

Visualization



VR [1]	(sky, **in**, water)	(giraffe, **have**, tree)	(woman, ride, bicycle)	(cat, **have**, hat)
A_1	(sky, **on**, water)	(giraffe, **have**, tree)	(woman, **behind**, bicycle)	(cat, **on**, hat)
S	(sky, above, water)	(giraffe, **in**, tree)	(woman, **wear**, bicycle)	(cat, **have**, hat)
A_1S	(sky, above, water)	(giraffe, behind, tree)	(woman, **wear**, bicycle)	(cat, **have**, hat)
A_1SC	(sky, above, water)	(giraffe, behind, tree)	(woman, ride, bicycle)	(cat, **have**, hat)
A_1SD	(sky, above, water)	(giraffe, behind, tree)	(woman, ride, bicycle)	(cat, wear, hat)

Future Work


red
man
ride
on
in
ride
white
behind
white
board
in
water
blue
round
brown
white
small
sheep
nextto
sheep
behind
eat
eat
in
nextto
green
grass
white
white
nextto
nextto
black
zebra
zebra
zebra
have
eat
nextto
of
behind
green
small
feet
on
grass

Thank You!