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.

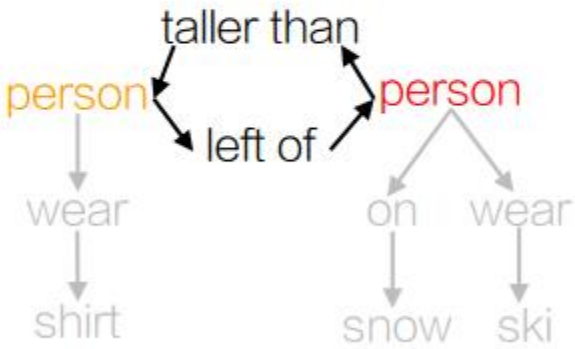
Task Definition

Task Definition

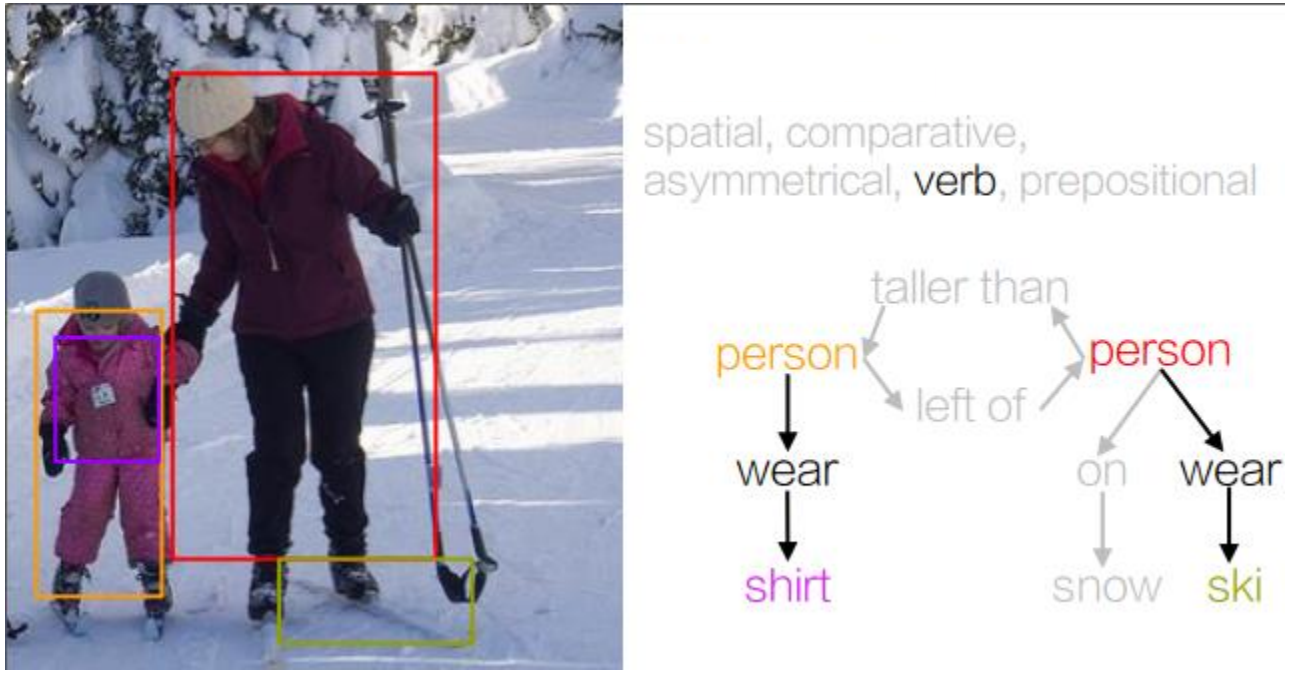
Task Definition

Task Definition

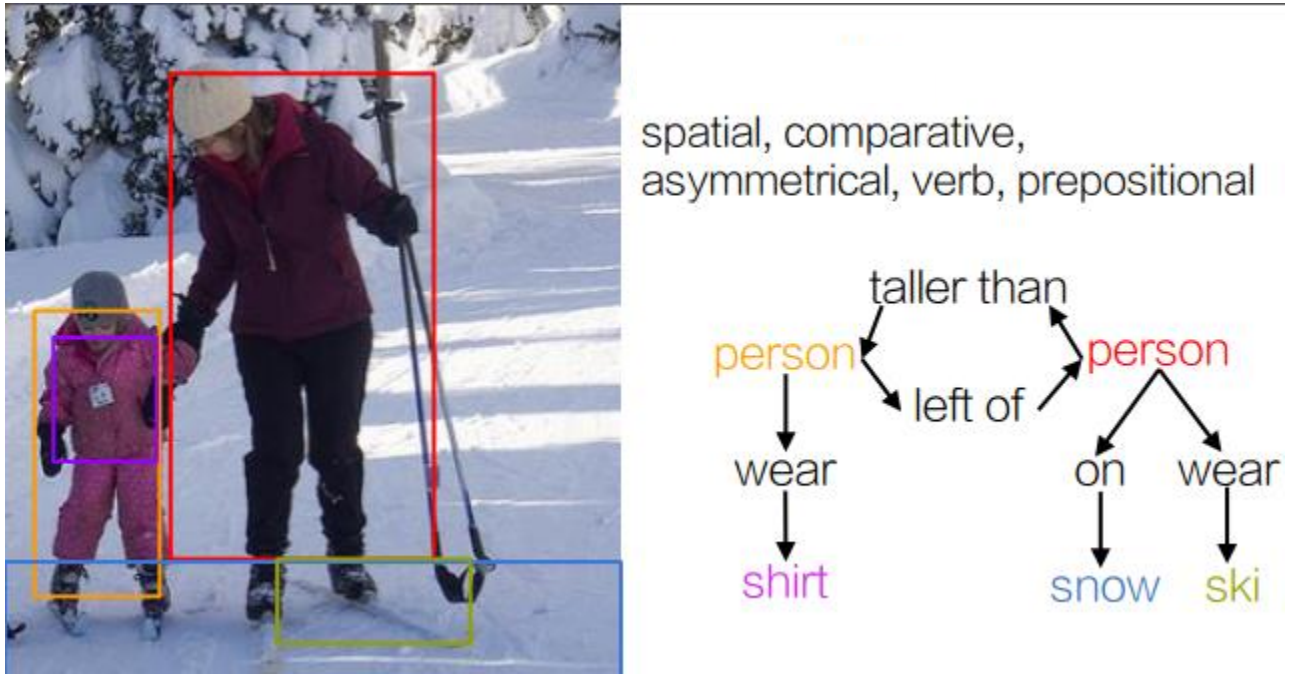
spatial, comparative,
asymmetrical, verb, prepositional



Task Definition



Task Definition



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
  - Subject BBox + Object BBox → Union BBox

- **Two boxes detection:**
  - Similar to 2, except treating Subject & Object individually

Visual Relationship Dataset(VGD)

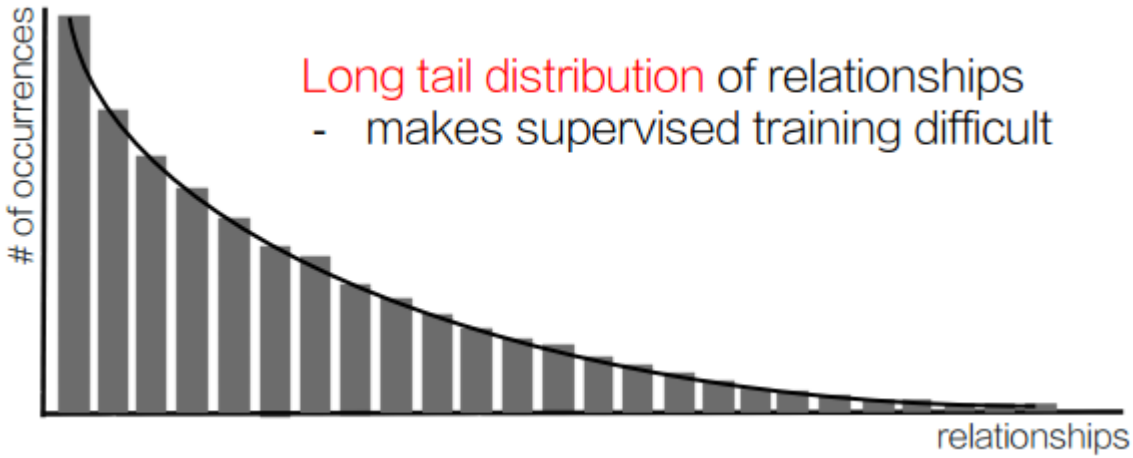
- Observation #1: Number of Relations

	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: Unbalanced 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



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

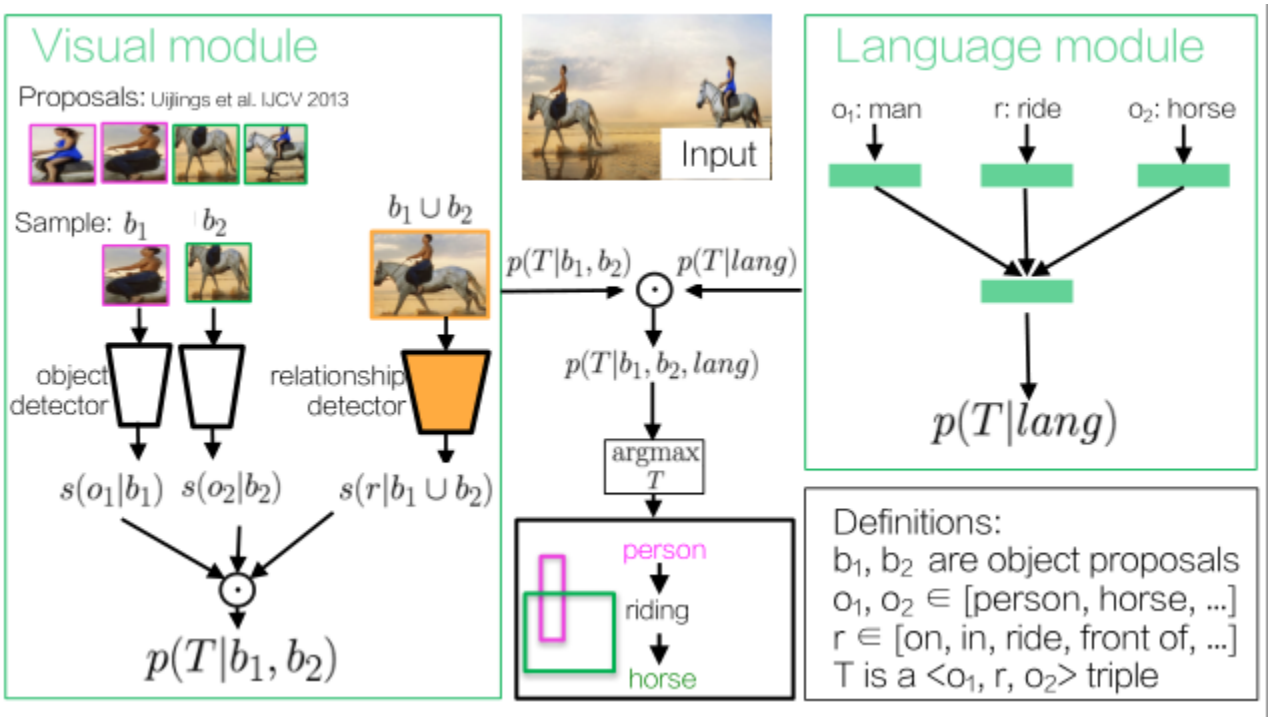
Visual Relationship Dataset(VGD)

- 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

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

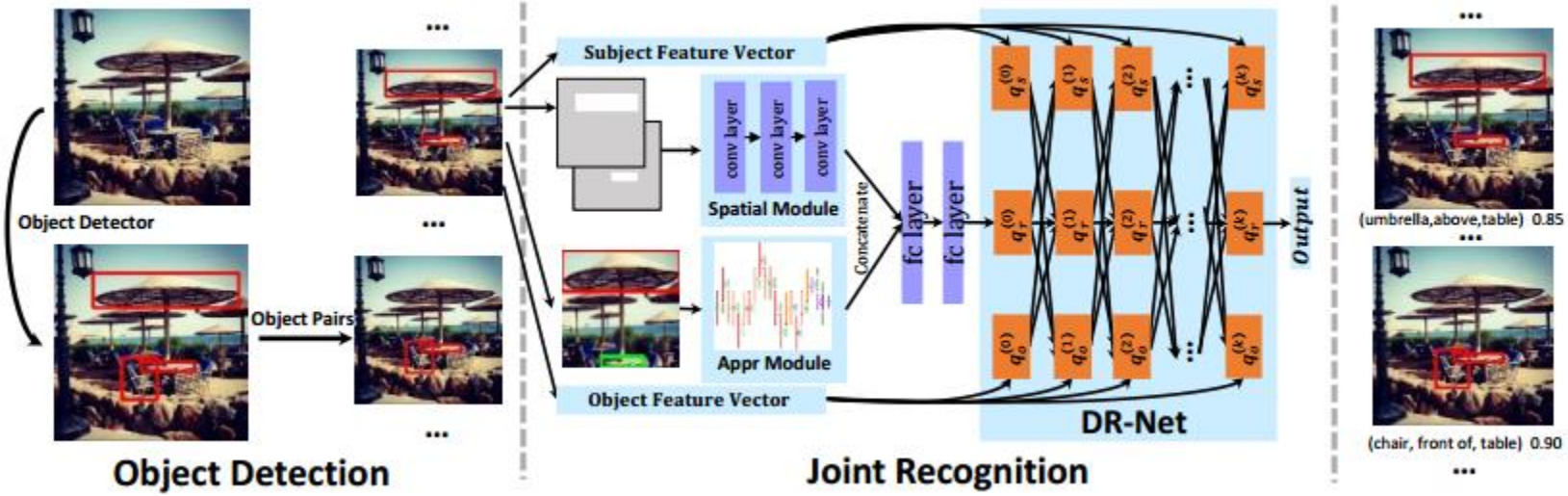
Pipeline:

- **Appearance Model**
  - Object detection: share feature to relation model
- **Relation Model**
  - Input: Three features (observed feature/raw prediction)
    - From Union Img; (Appearance)
    - From Spatial Mask: (Spatial Config)
    - From Faster-RCNN; (Appearance)
  - Output: three prediction vector(Class number fixed)

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

Pipeline

Pipeline:



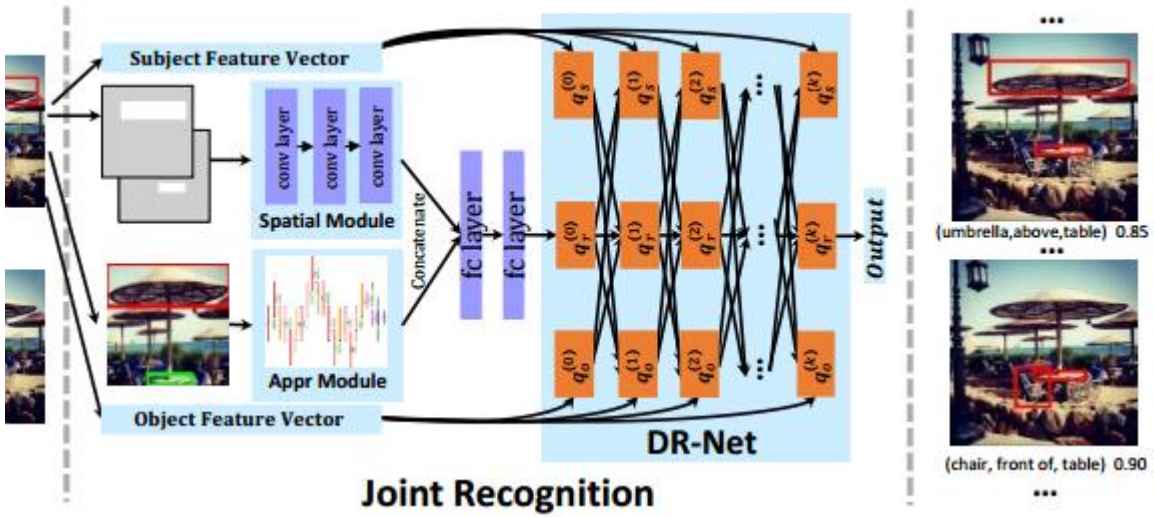
CVPR2017(Oral) Dai B, Zhang Y, Lin D. Detecting Visual Relationships with Deep Relational Networks[J]. arXiv preprint arXiv:1704.03114, 2017.

Pipeline – Object Detection

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

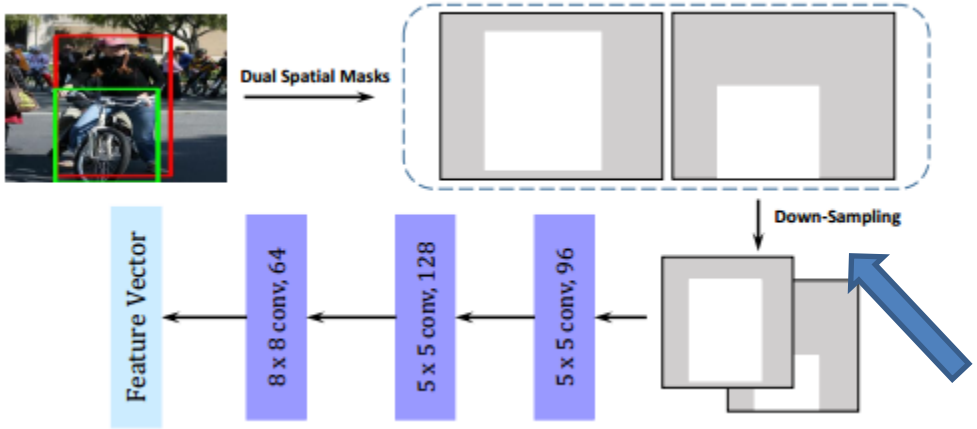
Pipeline – Object Detection

- **Relation Model**
  - Input: Three features (observed feature/raw prediction)
    - From Union Img; (Appearance)
    - From Spatial Mask: (Spatial Config)
    - From Faster-RCNN; (Appearance)
  - Output: three prediction vector(Class number fixed)



Spatial Mask

Hyper Param



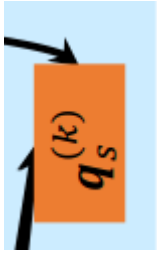
Mask Size	8	16	32	64	128
Recall	47.00%	48.00%	50.00%	51.00%	51.00%

- balance between fidelity and cost

Statistical Relation -- Implement

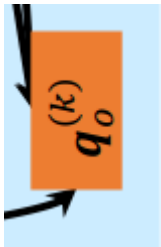
Inference Unit

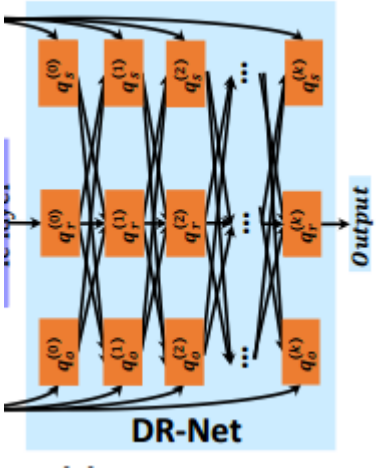
$$\mathbf{q}_s' = \sigma\left(\mathbf{W}_a\mathbf{x}_s + \mathbf{W}_{sr}\mathbf{q}_r + \mathbf{W}_{so}\mathbf{q}_o\right),$$



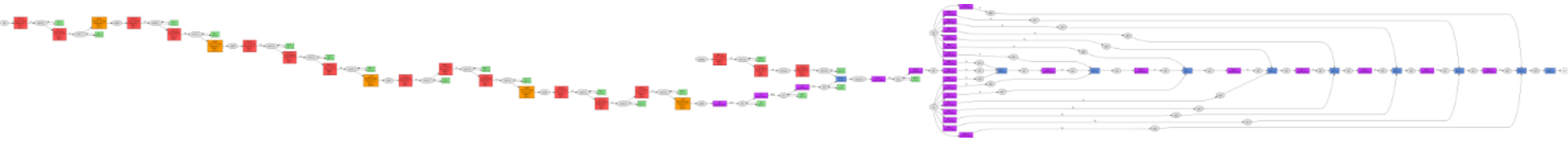
$$\mathbf{q}_r' = \sigma\left(\mathbf{W}_r\mathbf{x}_r + \mathbf{W}_{rs}\mathbf{q}_s + \mathbf{W}_{ro}\mathbf{q}_o\right),$$

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





Statistical Relation -- Implement

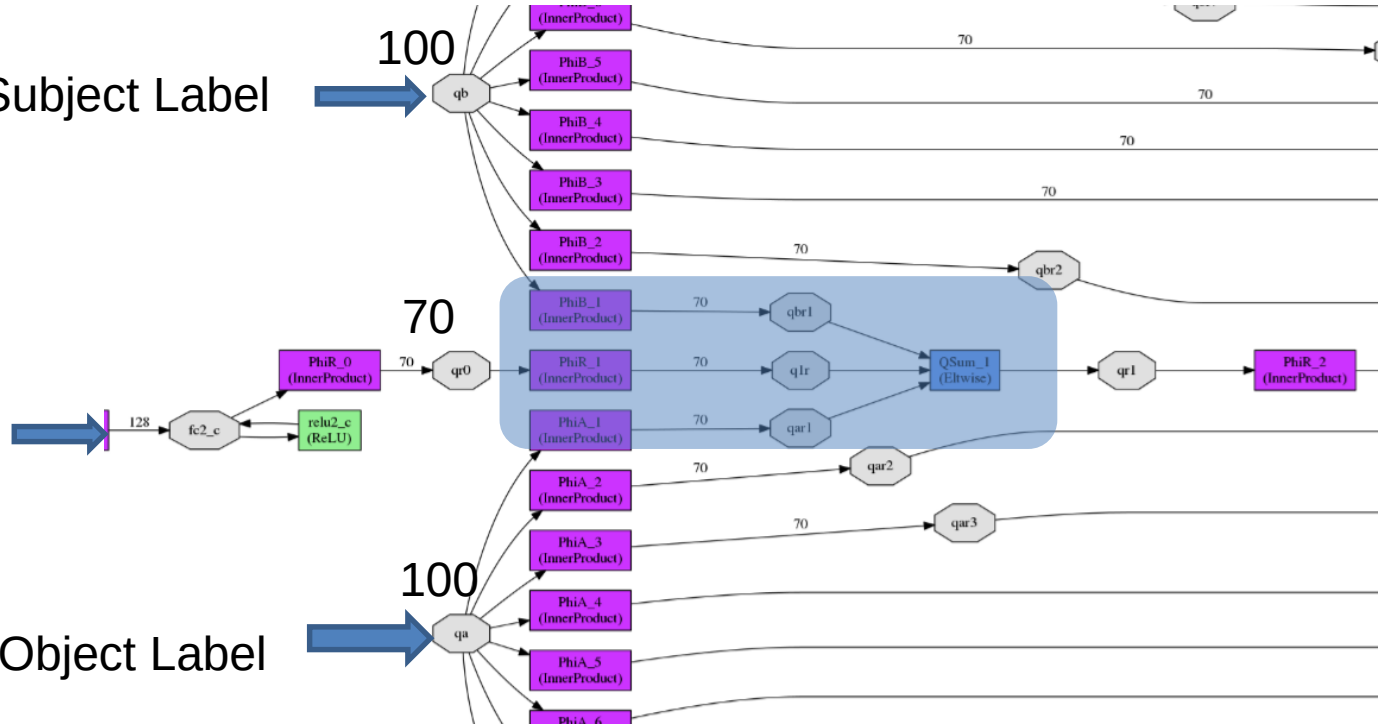


Prediction recognition Task

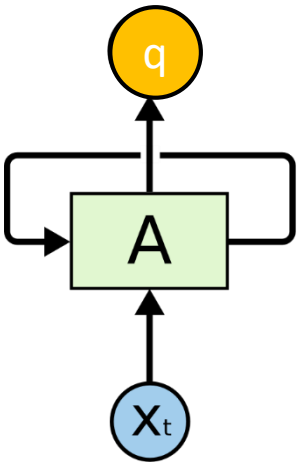
Subject Label

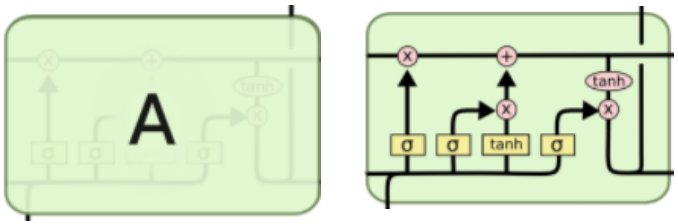
Union Features

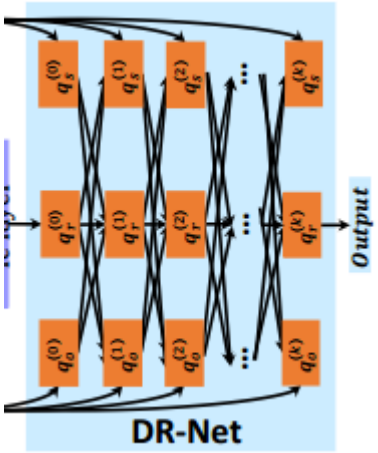
Object Label



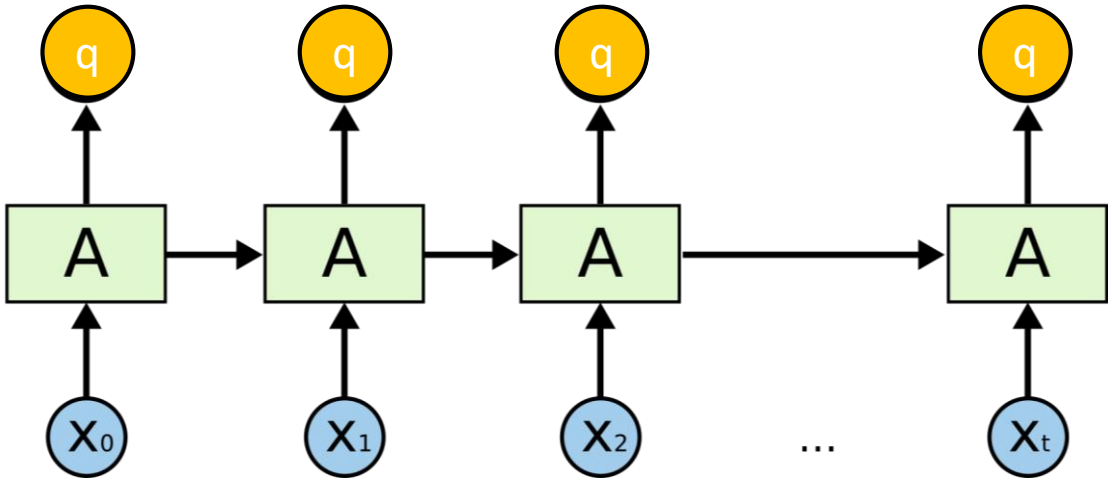
Statistical Relation -- RNN





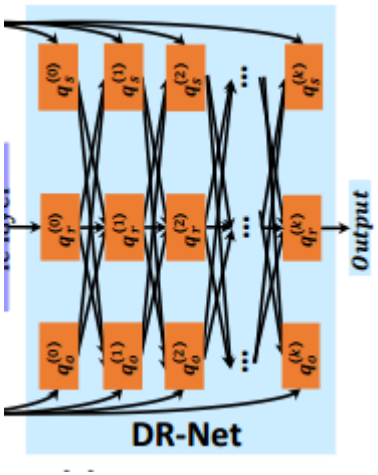


$$\mathbf{q}_s' = \boldsymbol{\sigma}\left(\mathbf{W}_a\mathbf{x}_s + \mathbf{W}_{sr}\mathbf{q}_r + \mathbf{W}_{so}\mathbf{q}_o\right),$$
$$\mathbf{q}_r' = \boldsymbol{\sigma}\left(\mathbf{W}_r\mathbf{x}_r + \mathbf{W}_{rs}\mathbf{q}_s + \mathbf{W}_{ro}\mathbf{q}_o\right),$$
$$\mathbf{q}_o' = \boldsymbol{\sigma}\left(\mathbf{W}_a\mathbf{x}_o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



Statistical Relation -- CRF

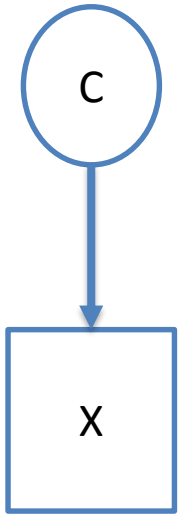
- Some Equivalent between Generative model and Discriminative model
- Inference can be unrolled into a forward Neural Network
- Inference Unit is a Computing layer



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

Toy Example

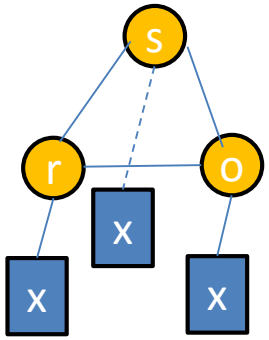
- Assume
  - Bayesian Net
  - C is label, x is raw feature/observation
  - C is binary, C1, C2
  - $x|C_1 \sim \mathcal{N}(\mu_1, \sigma) \quad x|C_2 \sim \mathcal{N}(\mu_2, \sigma)$
- Prove $p(C_1|x) = \sigma(w^T x + w_0)$



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

Statistical Relation -- CRF

Representation

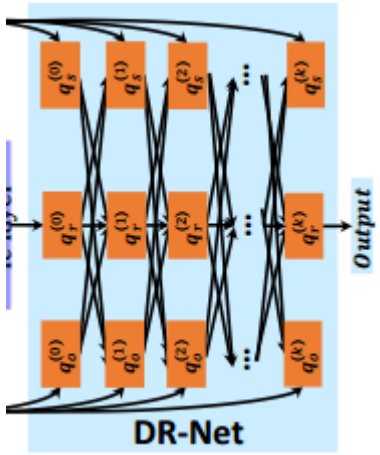


Inference

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

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

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



Unroll into a Network

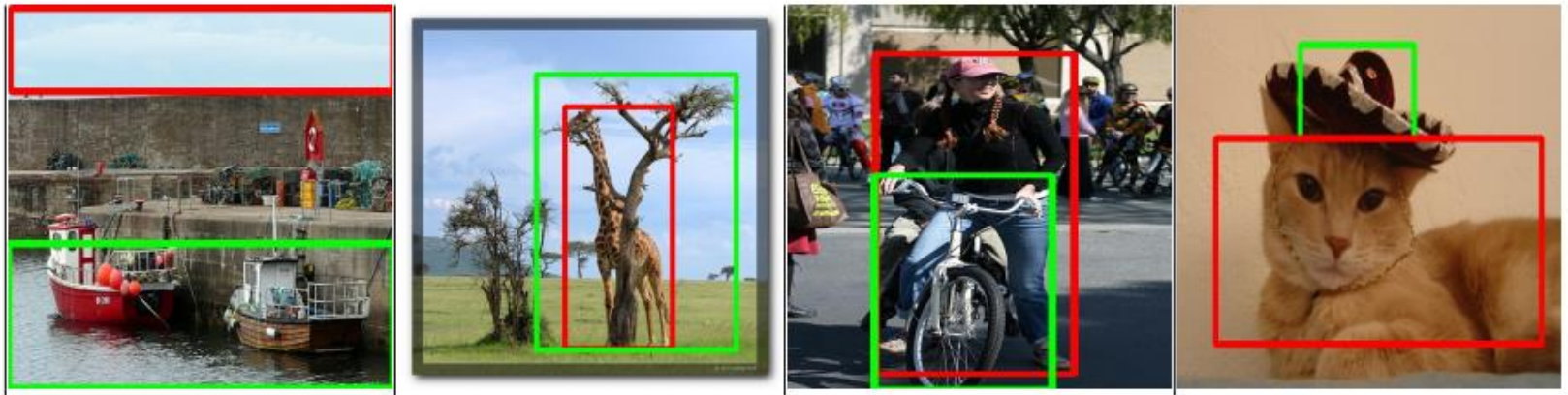
$$\mathbf{q}_r = \boldsymbol{\sigma}\left(\mathbf{W}_r\mathbf{x}_r + \mathbf{W}_{rs}\mathbf{q}_s + \mathbf{W}_{ro}\mathbf{q}_o\right).$$

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

Experiments

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)

Pair (F)ilter
(A)ppearance Module *A*1: based on VGG16; *A*2: based on ResNet101
(S)patial Module
(C)RF
(D)R-Net

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**

Experiments

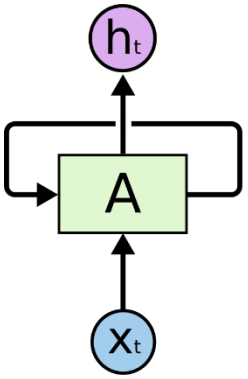
Hyper Param

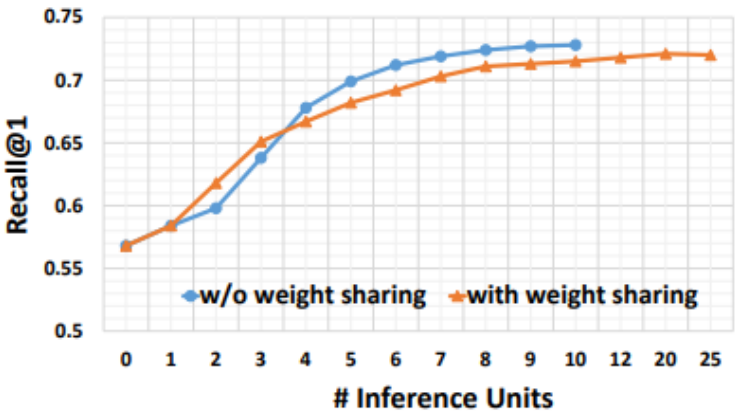
		A_1	A_2	S	A_1S	A_1SC	A_1SD	A_2SD	A_2SDF
VRD	Predicate Recognition	63.39	65.93	64.72	71.81	72.77	80.66	**80.78**	-
	Union Box Detection	12.01	12.56	13.76	16.04	16.37	18.15	**19.02**	**19.93**
	Two Boxes Detection	10.71	11.22	12.16	14.38	14.66	16.12	**16.94**	**17.73**
sVG	Predicate Recognition	72.13	72.54	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**

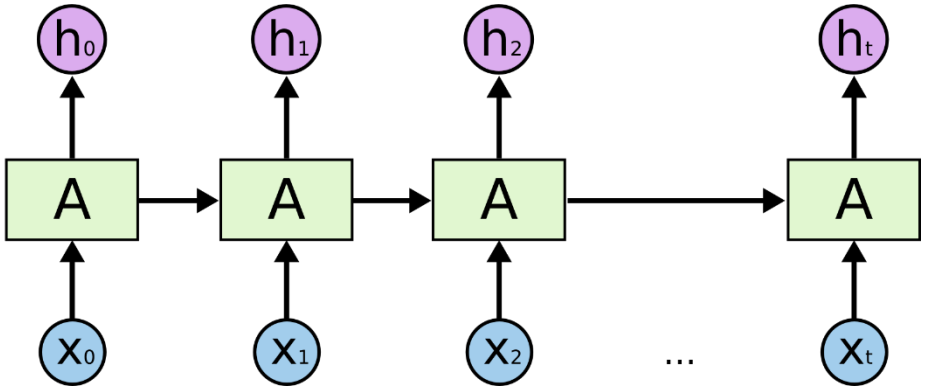
Pair (F)ilter
(A)ppearance Module *A*1: based on VGG16; *A*2: based on ResNet101
(S)patial Module
(C)RF
(D)R-Net

Experiments

Hyper Param

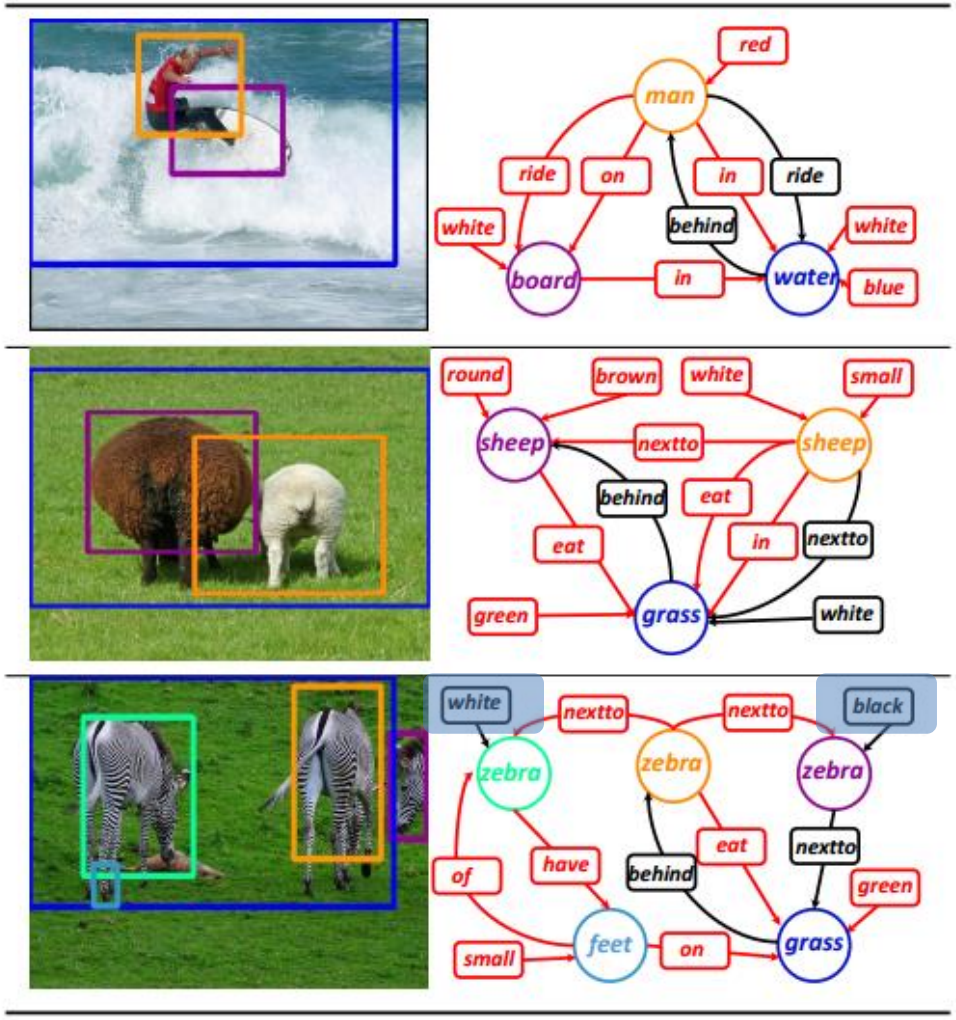






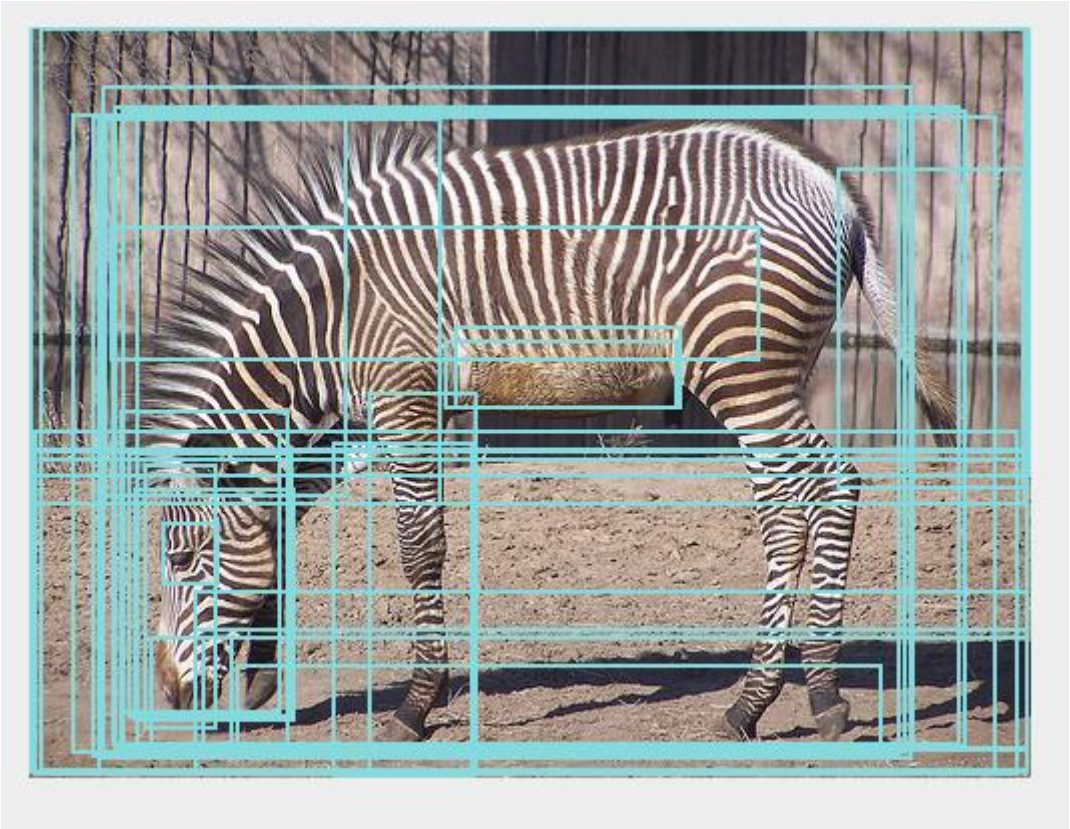
Future Work

On sVG Dataset



Future Work

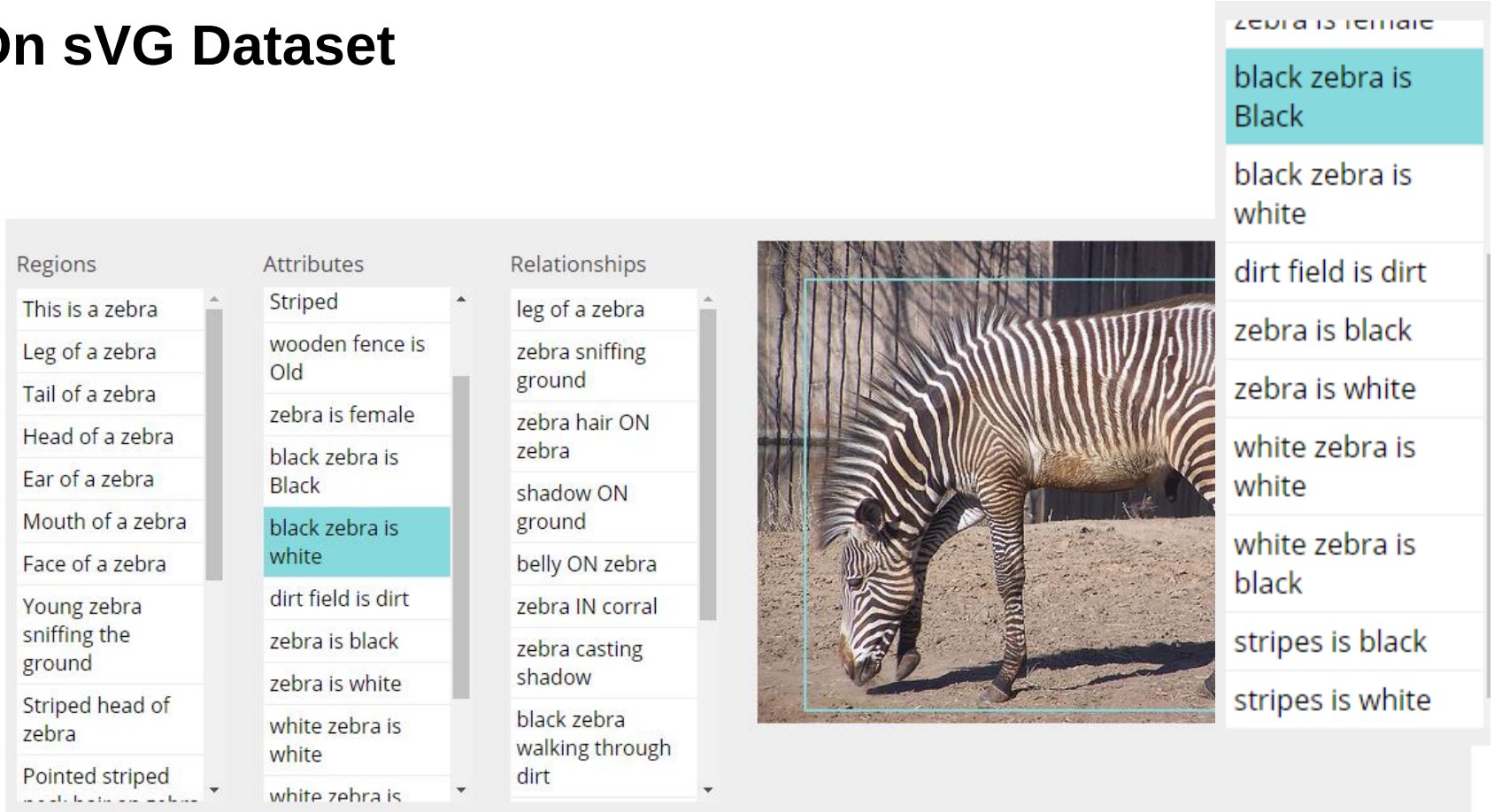
On sVG Dataset



https://visualgenome.org/VGViz/explore?query=zebra

Future Work

On sVG Dataset



Sumup

- Multi Feature
  - Spatial configurations
  - Statistical dependencies
- New Fuse Method
  - Relational modeling
  - End-to-end
- Multi Task

Thank You!