Semantic Diversity-aware Prototype-based Learning for Unbiased Scene Graph Generation

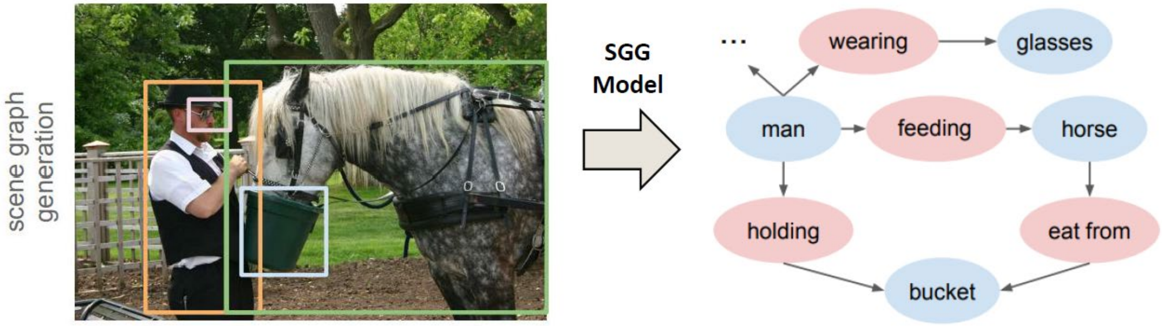
Jaehyeong Jeon, Kibum Kim, Kanghoon Yoon, Chanyoung Park

ECCV 2024

Presenter : Jaehyeong Jeon

MOTIVATION

Goal: Predict the relationships (predicates) between objects within an image and generate a structured graph



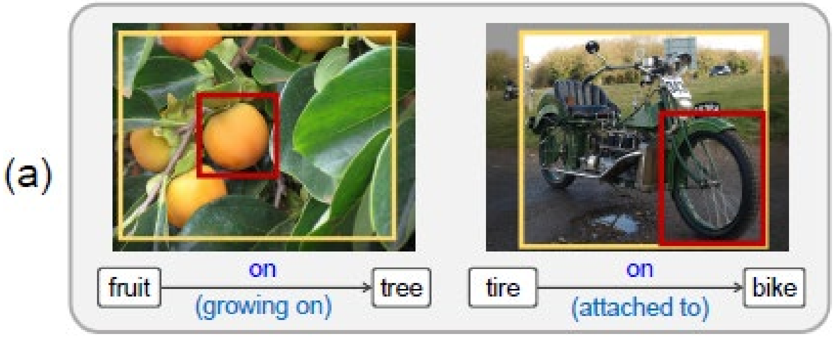
Characteristics of Datasets Used in Scene Graph Generation (SGG)

1. **Semantic Overlap in Predicates**: The predicates used to represent relationships often have semantic overlap.
    - on, sitting on, lying on, riding, above
    - attached to, mounted on, hanging from
2. **Single Predicate Annotation**: The relationship between object pairs is annotated with only a single predicate
    - <man – on – horse > / < man – riding – horse > / < man – sitting on – horse >
3. **Imbalance in Predicate Distribution**
    - Head class : on, has, near
    - Tail class : lying on, mounted on

MOTIVATION

Characteristics of Datasets

1. Semantic Overlap in Predicates
2. Single Predicate Annotation
3. Imbalance in Predicate Distribution



Example: <fruit, tree> and <tire, bike>

➢ Due to characteristics 2 and 3, relationships are usually labeled with the single head class predicate "on“

- *<fruit – on – tree>*
- *<tire – on – bike>*

➢ However, even though both relationships use the predicate “on”, they have different underlying meanings. Additionally, characteristic 1 suggests there are better predicates than "on" to represent these relationships.

- In *<fruit – on – tree>*, “on” is related to “growing on”
- In *<tire – on – bike>*, “on” is related to “attached to”

➢ But models trained heavily on the “on” fail to distinguish between *<fruit, tree>* and *<tire, bike>*, predicting both relationships as simply “on”.

MOTIVATION

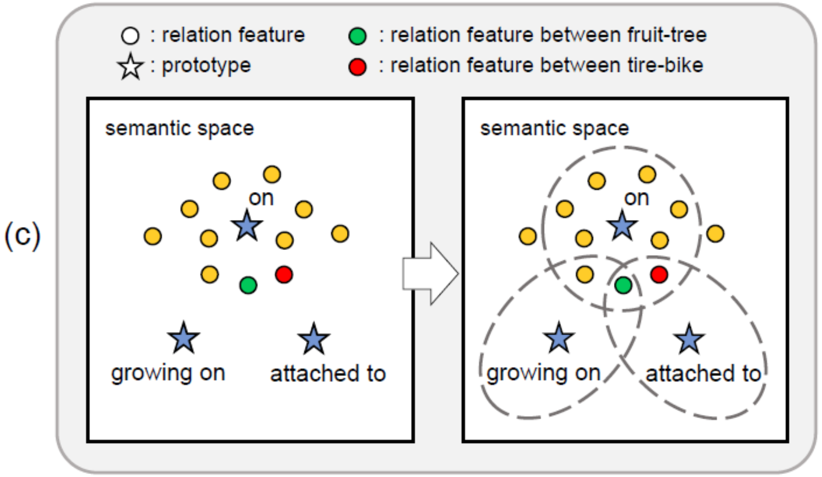
A single predicate can represent diverse semantics, which we define as **semantic diversity**.

Why Recognizing Semantic Diversity is Important in SGG?

If we can distinguish relationships that share the same predicate but have different semantics, we can:

- Represent *<fruit – on – tree>* as *<fruit – growing on – tree>*.
- Represent *<tire – on – bike>* as *<tire – attached to – bike>*.

This allows us to find more appropriate predicates for each relationship.



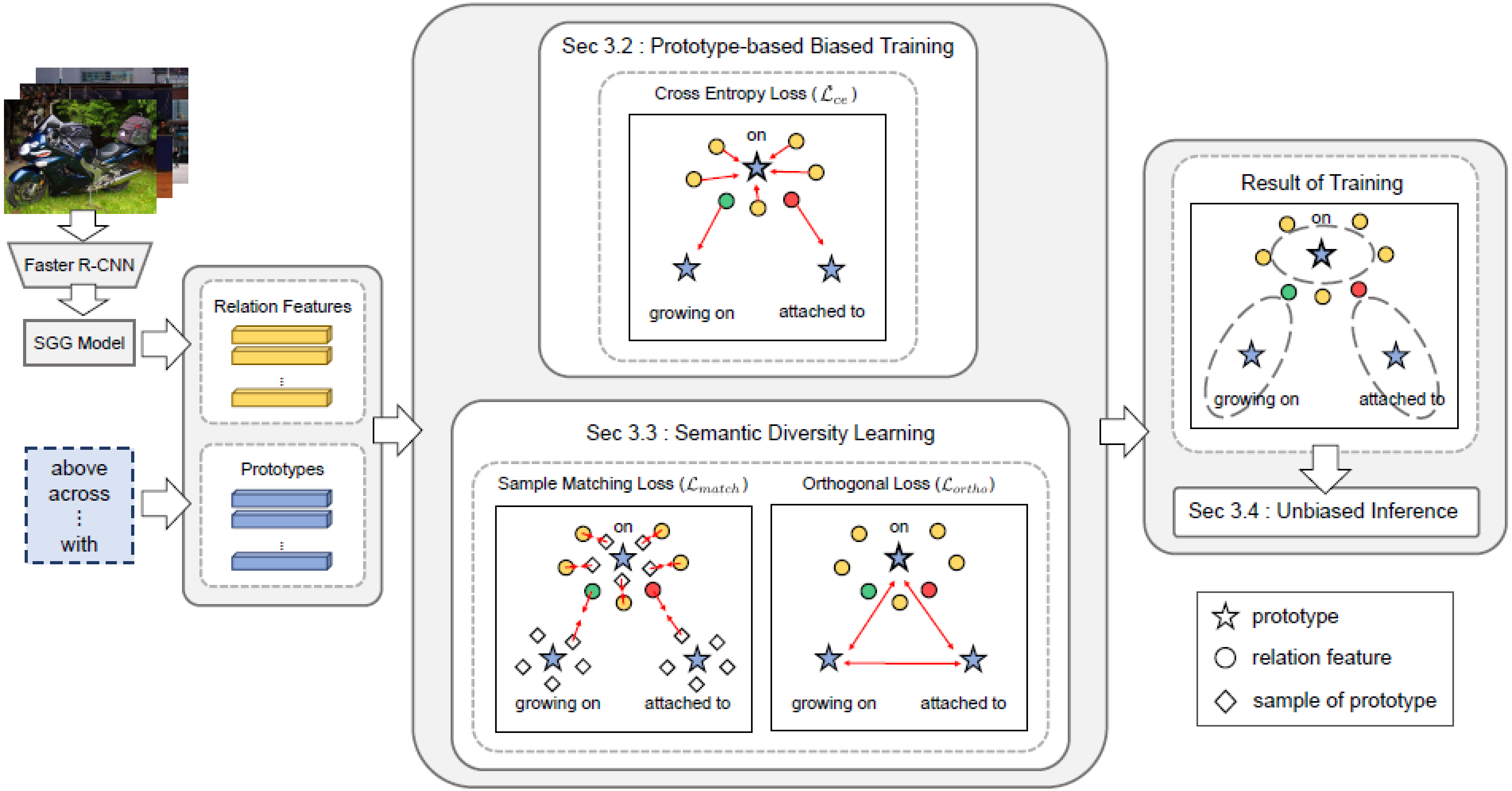
We propose a Semantic **D**iversity-aware **P**rototype-based **L**earning (**DPL**).

1. Create prototypes corresponding to each predicate.
2. Capture semantic diversity through probabilistic method and matching loss.
3. Perform unbiased predictions based on learned semantic diversity information.

METHOD - OVERALL FRAMEWORK

DPL is divided into three main parts.

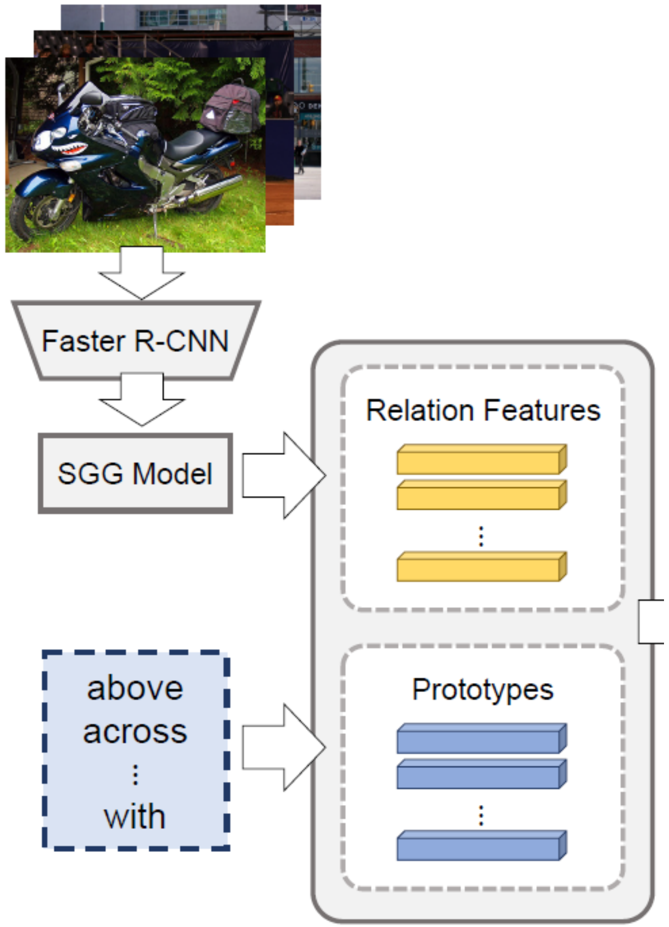
1. Protype-based Biased Training 2. Semantic Diversity Learning 3. Unbiased Inference



Faster R-CNN
SGG Model
above
across
⋮
with
Relation Features
Prototypes
Sec 3.2 : Prototype-based Biased Training
Cross Entropy Loss ($\mathcal{L}_{ce}$)
on
growing on
attached to
Sec 3.3 : Semantic Diversity Learning
Sample Matching Loss ($\mathcal{L}_{match}$)
Orthogonal Loss ($\mathcal{L}_{ortho}$)
on
growing on
attached to
on
growing on
attached to
Result of Training
on
growing on
attached to
Sec 3.4 : Unbiased Inference

☆ prototype
○ relation feature
◇ sample of prototype

METHOD - PROTOTYPE-BASED BIASED TRAINING



Proposal Generation

- Objects $O = \{o_i\}_{i=1}^{N_o}$ are detected using a pre-trained object detector.

Object Class Prediction

- Obtain the object feature $x_i = f_{obj}(v_i, b_i, w_i)$ and predict the class.

Predicate Class Prediction

- Combine subject, object, and union features to get the relation feature, and then predict the class.

$$\hat{x}_i = f_{rel}(v_i, x_i, \hat{w}_i), \quad r_{i \to j} = f_p([\hat{x}_i, \hat{x}_j]) \odot u_{ij}, \quad p_{i \to j} = \phi_{rel}(r_{i \to j}),$$

Prototype & relation feature

- Learnable prototype $C = \{c_1, c_2, \ldots, c_{|\mathcal{P}|}\}$, where $c_i \in \mathbb{R}^d$
- Relation feature $z = \phi_{proj}(r)$, where $z \in \mathbb{R}^d$

METHOD - PROTOTYPE-BASED BIASED TRAINING

Prototype & relation feature

- Learnable prototype $C = \{c_1, c_2, \ldots, c_{|\mathcal{P}|}\}$, where $c_i \in \mathbb{R}^d$
- Relation feature $z = \phi_{proj}(r)$, where $z \in \mathbb{R}^d$

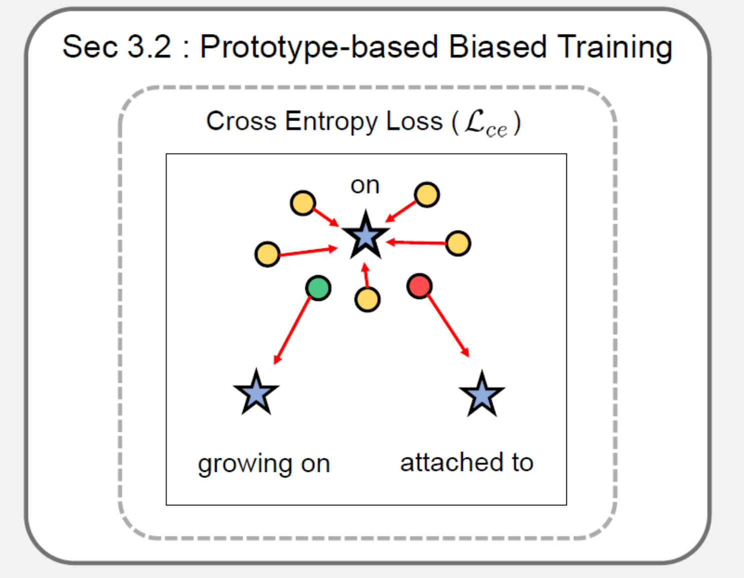
The probabilistic of relation feature z belonging to the i-th predicate class

$$p(i\text{-th class} \mid z) = \mathrm{Softmax}(-a \left\|z - c_i\right\|_2 + b),$$

Train with the cross-entropy loss

$$\mathcal{L}_{ce} = -\sum_{i=1}^{|\mathcal{P}|} y_i \log p(i\text{-th class} \mid z),$$

The training data is biased, leading to a concentration around the head class "on"


Sec 3.2 : Prototype-based Biased Training
Cross Entropy Loss ($\mathcal{L}_{ce}$)
on
growing on
attached to

METHOD - SEMANTIC DIVERSITY LEARNING

Recall that a single predicate may exhibit diverse semantics. Therefore, it is necessary to understand the range that each predicate can represent and identify which parts correspond to which semantics.

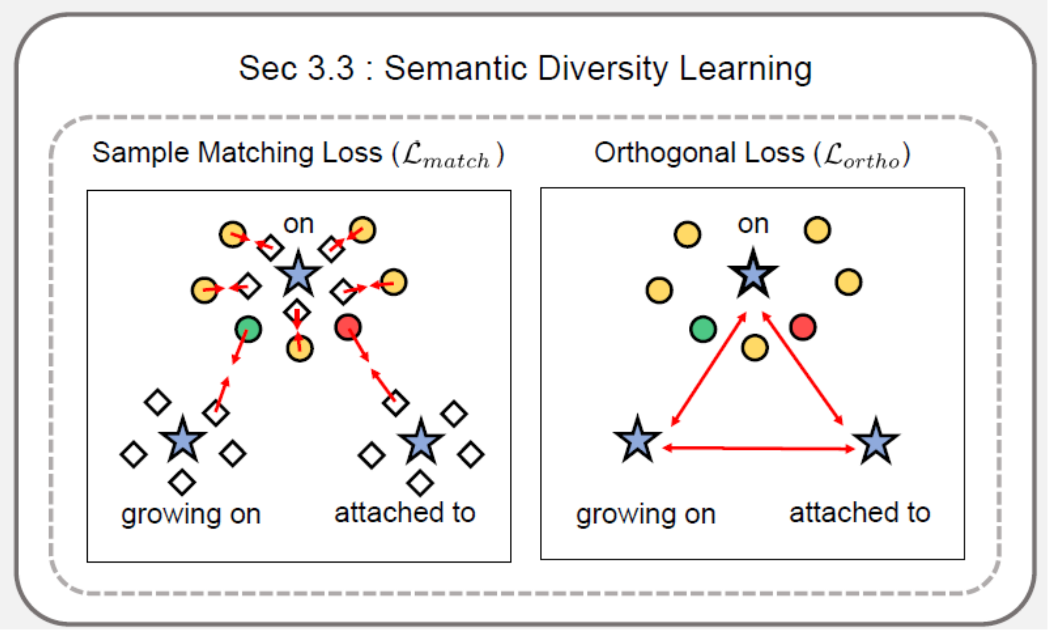
We generate samples from each prototype to capture this phenomenon

$$N \text{ samples } s_i = \{s_i^{(1)}, s_i^{(2)}, \ldots, s_i^{(N)}\} \qquad \mathrm{p}(s_i \mid c_i) \sim \mathcal{N}(\mu_i, \sigma_i^2), \quad \text{where } \mu_i = c_i, \, \sigma_i^2 = f_\sigma(c_i)$$

Sample Matching Loss

To ensure that the samples fully cover the regions represented by each predicate, we introduce the following loss.

$$\mathcal{L}_{match} = \left(\max(0, \min_j \left\| z - s_k^{(j)} \right\|_2 - R) \right)^2,$$


Sec 3.3 : Semantic Diversity Learning
Sample Matching Loss ($\mathcal{L}_{match}$)
Orthogonal Loss ($\mathcal{L}_{ortho}$)
on
growing on
attached to
on
growing on
attached to

METHOD - SEMANTIC DIVERSITY LEARNING

Recall that a single predicate may exhibit diverse semantics. Therefore, it is necessary to understand the range that each predicate can represent and identify which parts correspond to which semantics.

We generate samples from each prototype to capture this phenomenon

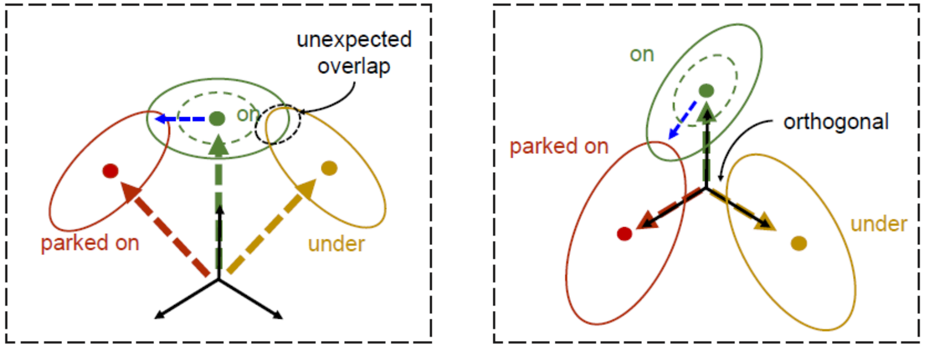
$$N \text{ samples } s_i = \{s_i^{(1)}, s_i^{(2)}, \ldots, s_i^{(N)}\} \qquad \mathrm{p}(s_i \mid c_i) \sim \mathcal{N}(\mu_i, \sigma_i^2), \quad \text{where } \mu_i = c_i, \, \sigma_i^2 = f_\sigma(c_i)$$

Orthogonal Loss

This is to prevent unexpected overlap during training caused by the symmetric nature of the Gaussian distribution

$$\mathcal{L}_{ortho} = \frac{1}{|\mathcal{P}|(|\mathcal{P}|-1)} \sum_{i \neq j} |c_i \cdot c_j^T|$$

Final loss $\mathcal{L} = \mathcal{L}_{ce} + \mathcal{L}_{ortho} + \alpha \mathcal{L}_{match},$



(a) Without $\mathcal{L}_{ortho}$ (b) With $\mathcal{L}_{ortho}$

METHOD - UNBIASED INFERENCE

Due to data imbalance, relation features still tend to cluster near the head class.

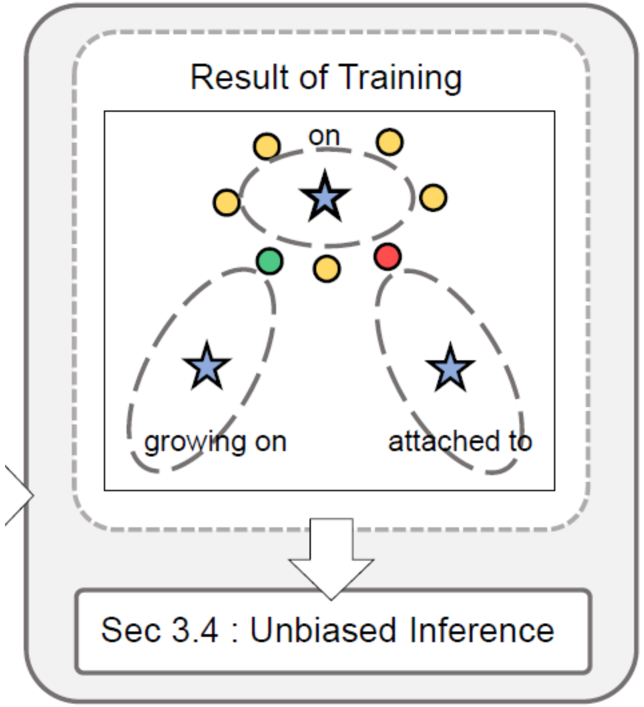
During inference, predictions are made based on normalized distance utilizing variance, which captures semantic diversity.

$$p(i\text{-th class} \mid z) = \text{Softmax}(-a' \underline{\left\|(z - c_j) \odot \sigma_j^{-1}\right\|_2} + b),$$

$$\Updownarrow$$

$$p(i\text{-th class} \mid z) = \text{Softmax}(-a \left\|z - c_i\right\|_2 + b),$$

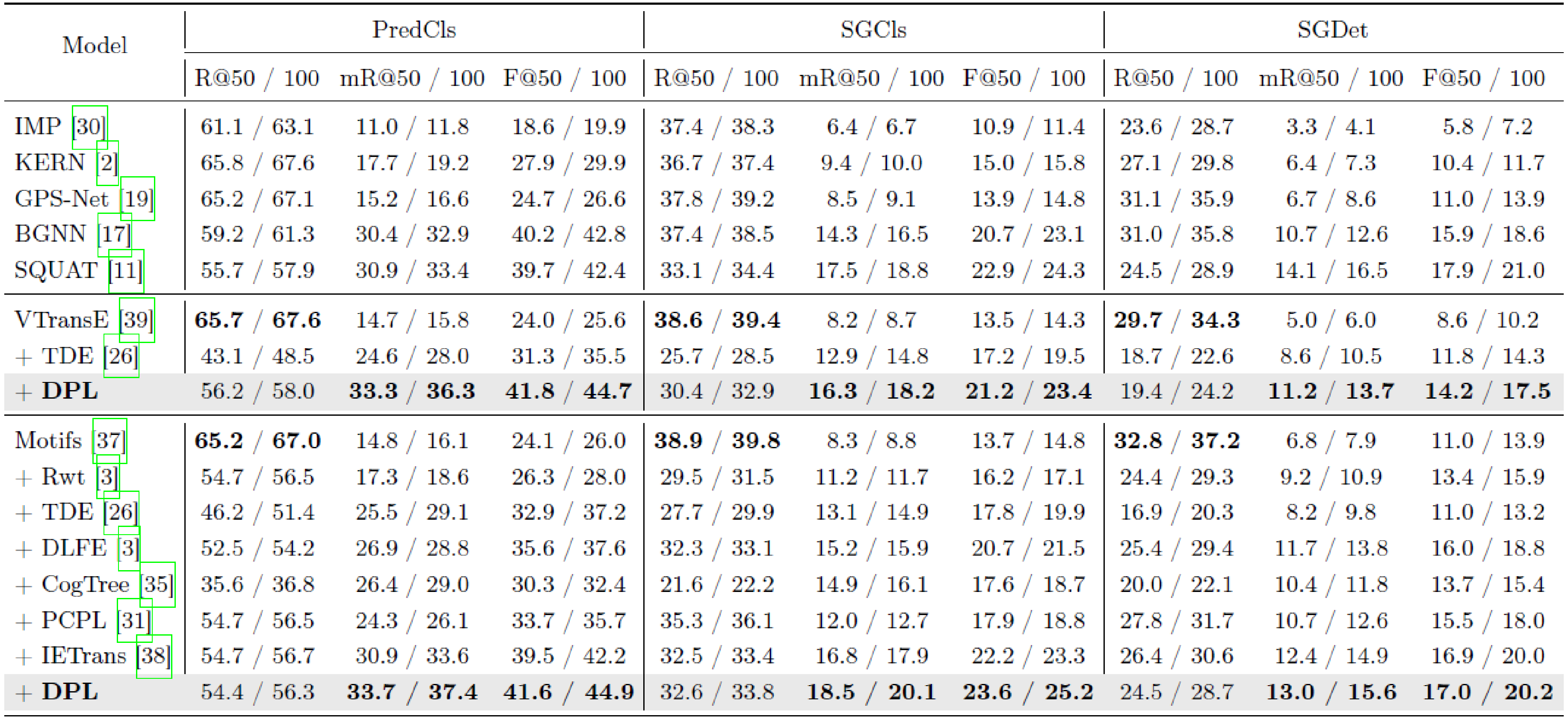
$$\text{where } \sigma_i^{-1} = \left[\frac{1}{\sigma_i^{(1)}}, \frac{1}{\sigma_i^{(2)}}, \ldots, \frac{1}{\sigma_i^{(d)}}\right] \text{ and } a' = a \cdot \frac{\max_j \|z - c_j\|_2}{\max_j \left\|(z - c_j) \odot \sigma_j^{-1}\right\|_2}$$



EXPERIMENT - MAIN EXPERIMENT

Results from experiments conducted on the Visual Genome dataset

Metric : R@K, mR@K, F@K



Model	PredCls			SGCls			SGDet		
	R@50 / 100	mR@50 / 100	F@50 / 100	R@50 / 100	mR@50 / 100	F@50 / 100	R@50 / 100	mR@50 / 100	F@50 / 100
IMP [30]	61.1 / 63.1	11.0 / 11.8	18.6 / 19.9	37.4 / 38.3	6.4 / 6.7	10.9 / 11.4	23.6 / 28.7	3.3 / 4.1	5.8 / 7.2
KERN [2]	65.8 / 67.6	17.7 / 19.2	27.9 / 29.9	36.7 / 37.4	9.4 / 10.0	15.0 / 15.8	27.1 / 29.8	6.4 / 7.3	10.4 / 11.7
GPS-Net [19]	65.2 / 67.1	15.2 / 16.6	24.7 / 26.6	37.8 / 39.2	8.5 / 9.1	13.9 / 14.8	31.1 / 35.9	6.7 / 8.6	11.0 / 13.9
BGNN [17]	59.2 / 61.3	30.4 / 32.9	40.2 / 42.8	37.4 / 38.5	14.3 / 16.5	20.7 / 23.1	31.0 / 35.8	10.7 / 12.6	15.9 / 18.6
SQUAT [11]	55.7 / 57.9	30.9 / 33.4	39.7 / 42.4	33.1 / 34.4	17.5 / 18.8	22.9 / 24.3	24.5 / 28.9	14.1 / 16.5	17.9 / 21.0
VTransE [39]	**65.7 / 67.6**	14.7 / 15.8	24.0 / 25.6	**38.6 / 39.4**	8.2 / 8.7	13.5 / 14.3	**29.7 / 34.3**	5.0 / 6.0	8.6 / 10.2
+ TDE [26]	43.1 / 48.5	24.6 / 28.0	31.3 / 35.5	25.7 / 28.5	12.9 / 14.8	17.2 / 19.5	18.7 / 22.6	8.6 / 10.5	11.8 / 14.3
+ **DPL**	56.2 / 58.0	**33.3 / 36.3**	**41.8 / 44.7**	30.4 / 32.9	**16.3 / 18.2**	**21.2 / 23.4**	19.4 / 24.2	**11.2 / 13.7**	**14.2 / 17.5**
Motifs [37]	**65.2 / 67.0**	14.8 / 16.1	24.1 / 26.0	**38.9 / 39.8**	8.3 / 8.8	13.7 / 14.8	**32.8 / 37.2**	6.8 / 7.9	11.0 / 13.9
+ Rwt [3]	54.7 / 56.5	17.3 / 18.6	26.3 / 28.0	29.5 / 31.5	11.2 / 11.7	16.2 / 17.1	24.4 / 29.3	9.2 / 10.9	13.4 / 15.9
+ TDE [26]	46.2 / 51.4	25.5 / 29.1	32.9 / 37.2	27.7 / 29.9	13.1 / 14.9	17.8 / 19.9	16.9 / 20.3	8.2 / 9.8	11.0 / 13.2
+ DLFE [3]	52.5 / 54.2	26.9 / 28.8	35.6 / 37.6	32.3 / 33.1	15.2 / 15.9	20.7 / 21.5	25.4 / 29.4	11.7 / 13.8	16.0 / 18.8
+ CogTree [35]	35.6 / 36.8	26.4 / 29.0	30.3 / 32.4	21.6 / 22.2	14.9 / 16.1	17.6 / 18.7	20.0 / 22.1	10.4 / 11.8	13.7 / 15.4
+ PCPL [31]	54.7 / 56.5	24.3 / 26.1	33.7 / 35.7	35.3 / 36.1	12.0 / 12.7	17.9 / 18.8	27.8 / 31.7	10.7 / 12.6	15.5 / 18.0
+ IETrans [38]	54.7 / 56.7	30.9 / 33.6	39.5 / 42.2	32.5 / 33.4	16.8 / 17.9	22.2 / 23.3	26.4 / 30.6	12.4 / 14.9	16.9 / 20.0
+ **DPL**	54.4 / 56.3	**33.7 / 37.4**	**41.6 / 44.9**	32.6 / 33.8	**18.5 / 20.1**	**23.6 / 25.2**	24.5 / 28.7	**13.0 / 15.6**	**17.0 / 20.2**

EXPERIMENT - ABLATION STUDY

Ablation studies on N and R in L_{match}

As N increases, the overall R@K increases, while mR@K decreases (Table 3)

- A large number of samples is required to capture the semantic diversity of the head classes,
- whereas a smaller number of samples is sufficient to capture the semantic diversity of the tail classes.

When R is set to a small or large value, a decrease in performance is observed. (Table 4)

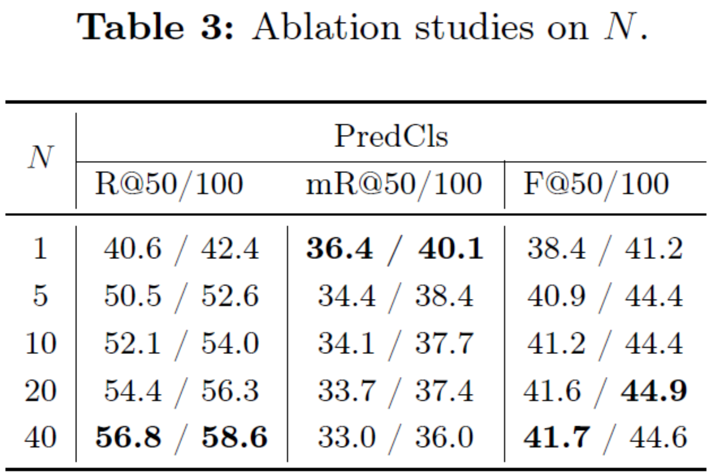


Table 3: Ablation studies on N.

N	PredCls		
	R@50/100	mR@50/100	F@50/100
1	40.6 / 42.4	**36.4** / **40.1**	38.4 / 41.2
5	50.5 / 52.6	34.4 / 38.4	40.9 / 44.4
10	52.1 / 54.0	34.1 / 37.7	41.2 / 44.4
20	54.4 / 56.3	33.7 / 37.4	41.6 / **44.9**
40	**56.8** / **58.6**	33.0 / 36.0	**41.7** / 44.6

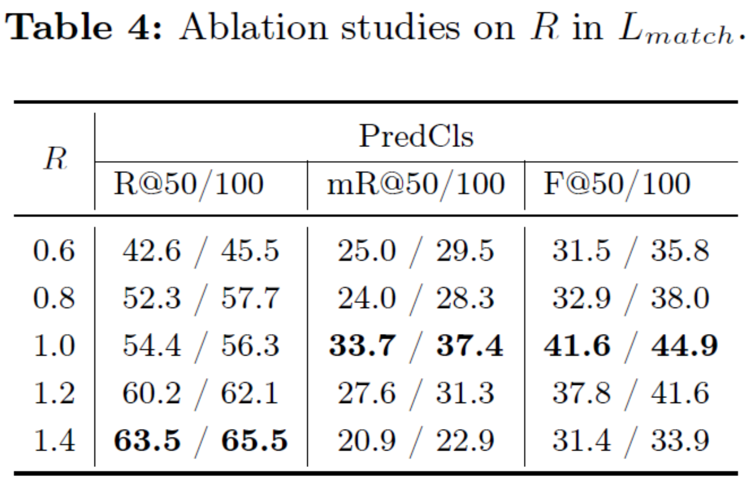


Table 4: Ablation studies on R in L_{match}.

R	PredCls		
	R@50/100	mR@50/100	F@50/100
0.6	42.6 / 45.5	25.0 / 29.5	31.5 / 35.8
0.8	52.3 / 57.7	24.0 / 28.3	32.9 / 38.0
1.0	54.4 / 56.3	**33.7** / **37.4**	**41.6** / **44.9**
1.2	60.2 / 62.1	27.6 / 31.3	37.8 / 41.6
1.4	**63.5** / **65.5**	20.9 / 22.9	31.4 / 33.9

EXPERIMENT - ABLATION STUDY

Ablation study on each component of DPL

1. Training the model with only cross-entropy leads to biased training.
   - Orthogonal loss and matching loss alone do not prevent biased training.

2. However, conducting unbiased inference based on the learned variance is crucial.
   - Reducing unexpected overlap through orthogonal loss also has a significant impact.

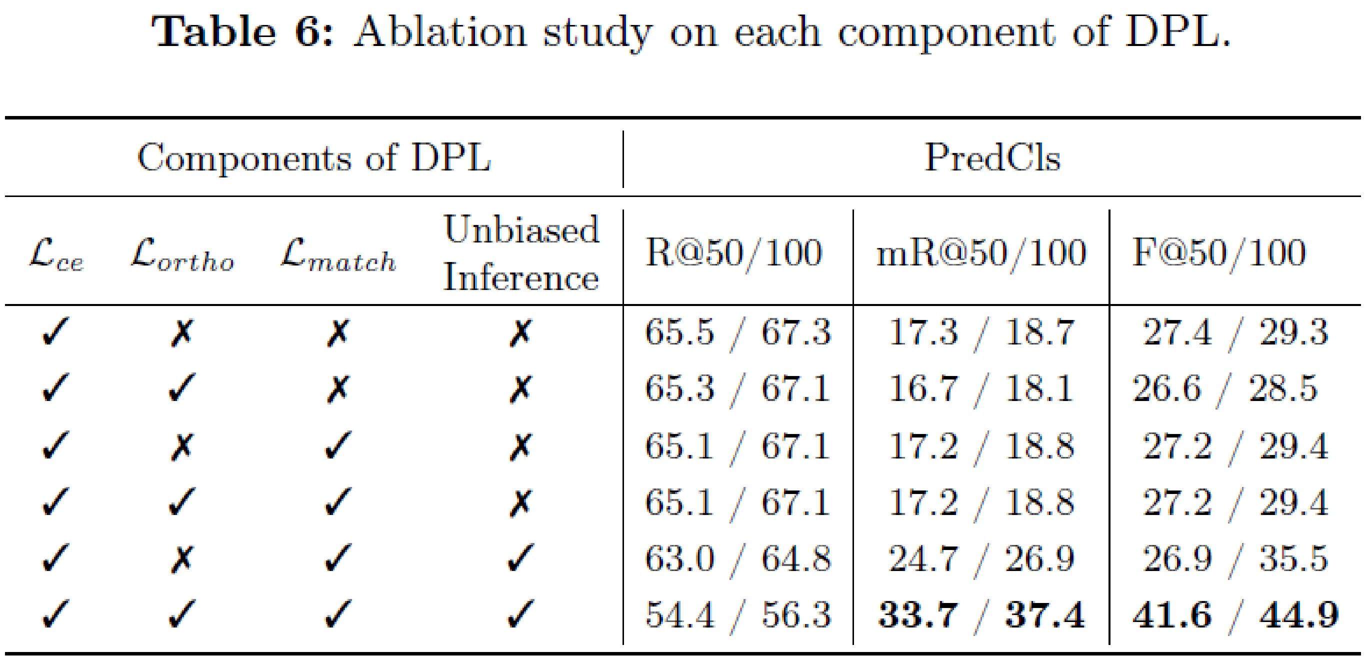
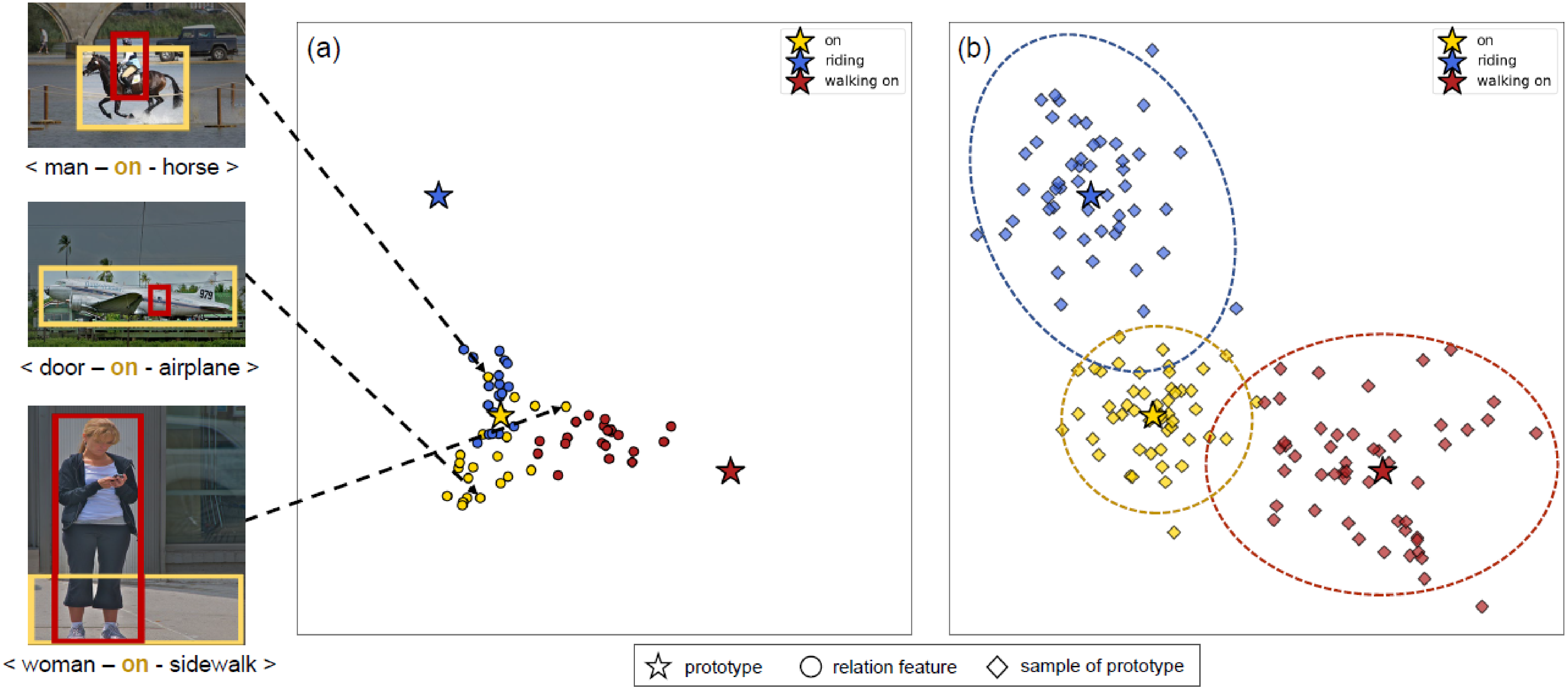


Table 6: Ablation study on each component of DPL.

Components of DPL				PredCls		
$\mathcal{L}_{ce}$	$\mathcal{L}_{ortho}$	$\mathcal{L}_{match}$	Unbiased Inference	R@50/100	mR@50/100	F@50/100
✓	✗	✗	✗	65.5 / 67.3	17.3 / 18.7	27.4 / 29.3
✓	✓	✗	✗	65.3 / 67.1	16.7 / 18.1	26.6 / 28.5
✓	✗	✓	✗	65.1 / 67.1	17.2 / 18.8	27.2 / 29.4
✓	✓	✓	✗	65.1 / 67.1	17.2 / 18.8	27.2 / 29.4
✓	✗	✓	✓	63.0 / 64.8	24.7 / 26.9	26.9 / 35.5
✓	✓	✓	✓	54.4 / 56.3	**33.7** / **37.4**	**41.6** / **44.9**

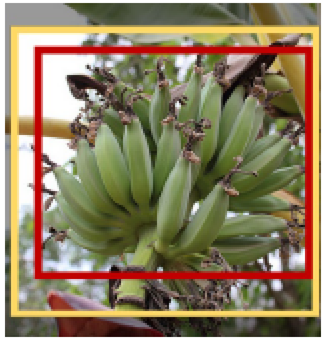
EXPERIMENT - QUALITATIVE ANALYSIS

Visualization of prototypes, sample and relation features.

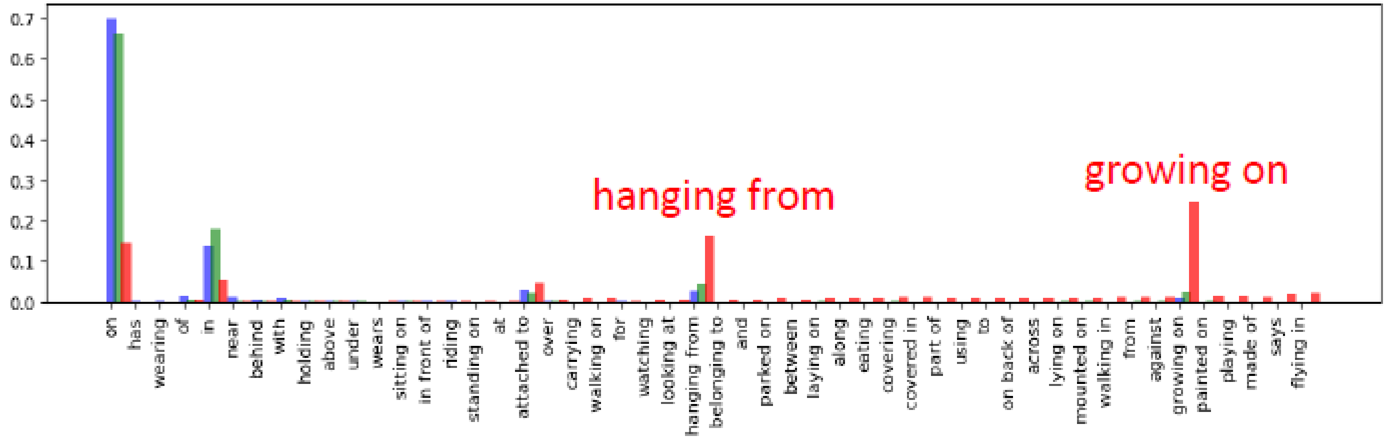


EXPERIMENT - PREDICTED DISTRIBUTION COMPARISION

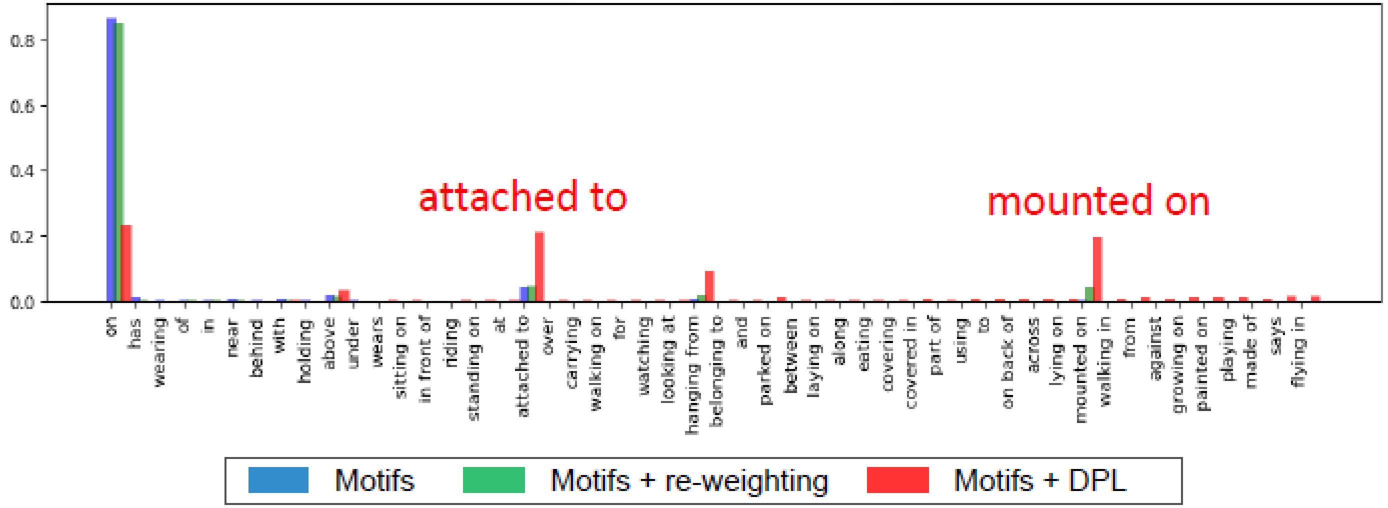
Comparisons with baselines (Motifs, Motif + re-weighting, Motif + DPL)



< banana, hanging from, tree >



< sign, attached to, pole >



Thank You!