JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Prototype-based Embedding Network for Scene Graph Generation

Chaofan Zheng*, **Xinyu Lyu***, Lianli Gao, Bo Dai, and Jingkuan Song

School of Computer Science and Engineering,
University of Electronic Science and Technology of China

THU-PM-208

Problem

Existing SGG methods fail to capture compact and distinctive relation representations.

- **Large intra-class variation:** arises from diverse appearance of entities and various subject-object combinations.
- **Severe inter-class similarity:** originates from similar-looking interactions shared among different relation categories.

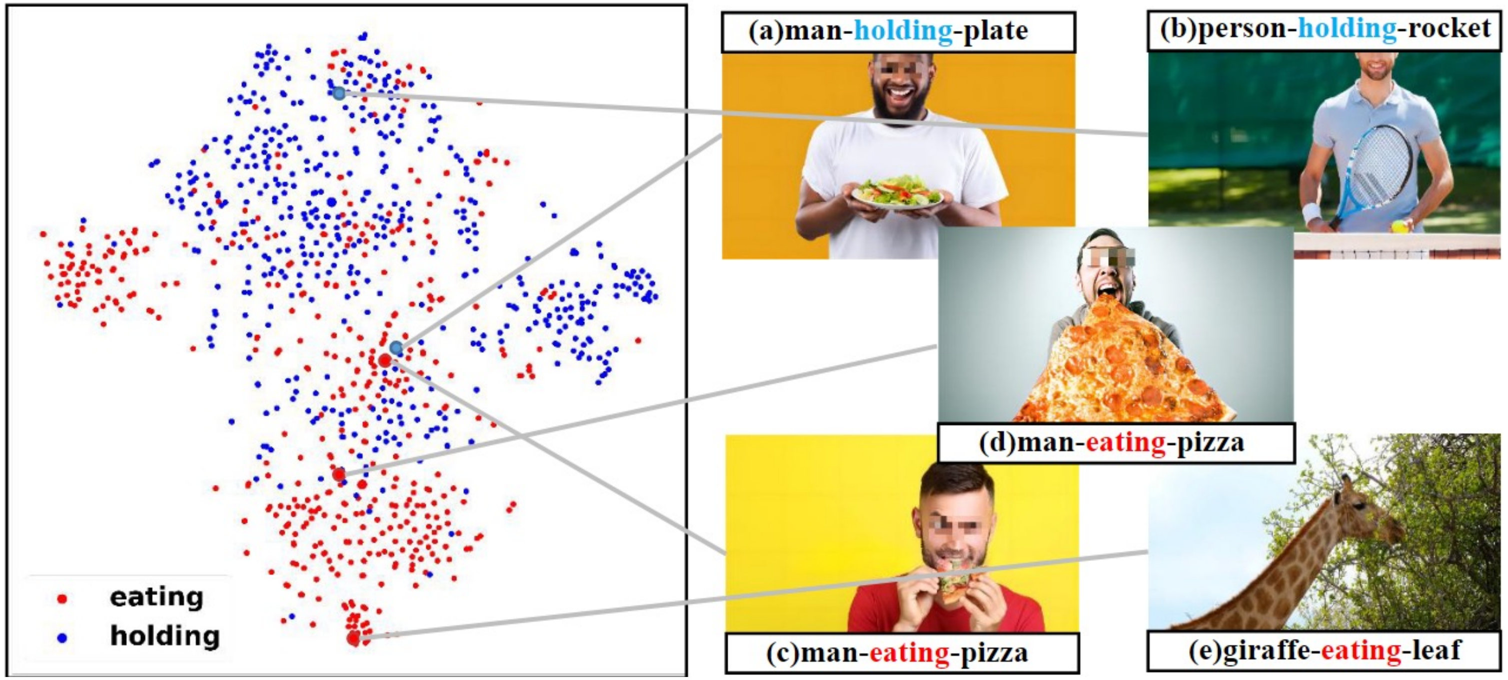

Fig. 1. The illustration of relation representations with large intra-class variation and severe inter-class similarity.

Motivation

Category-inherent Semantics is more reliable than visual appearance.

- **Intra-class variation:** Entities/predicates from each class share the same semantics, captured from class labels.
- **Inter-class similarity:** Class-inherent semantics is discriminative for visual-similar instances from different categories.

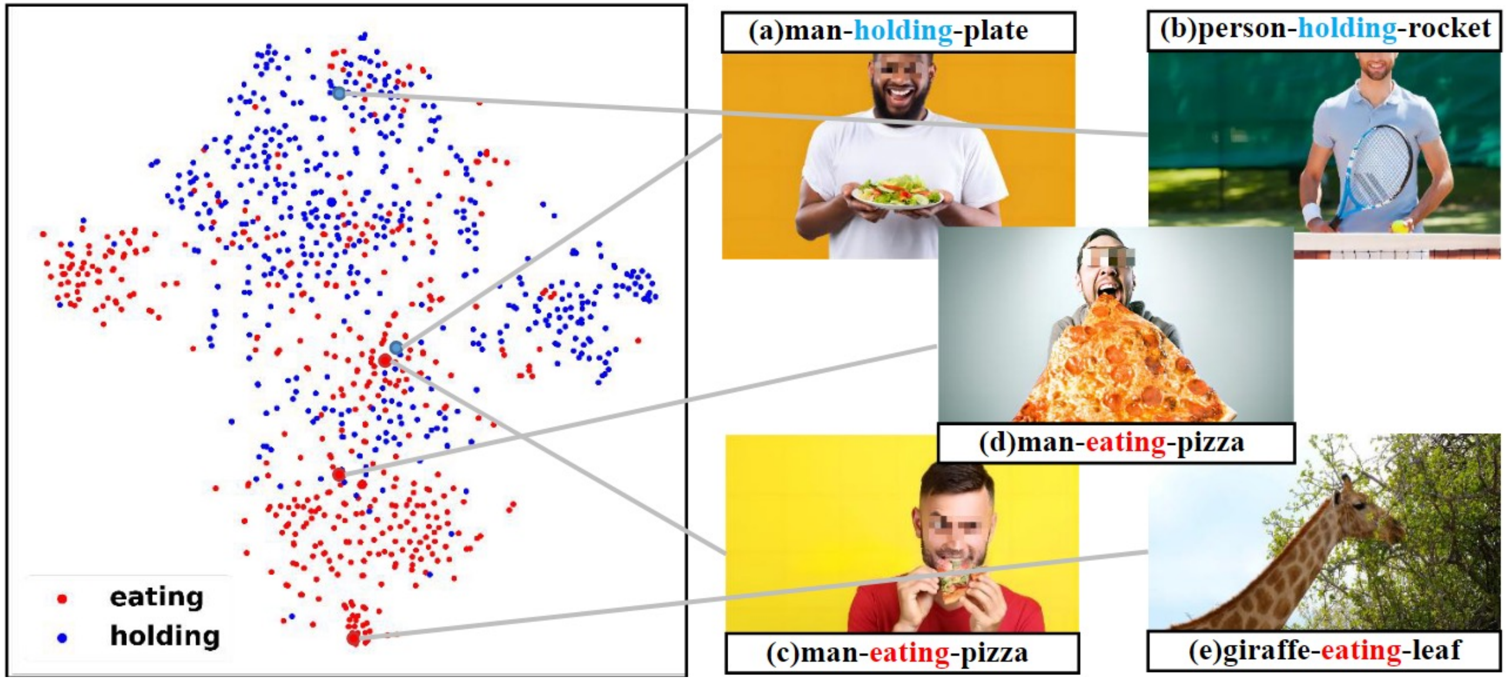

Fig. 1. The illustration of relation representations with large intra-class variation and severe inter-class similarity.

Method

Prototype-based Embedding Network (PE-Net):

- **Prototype-based Modeling**:

  Models entities/predicates with prototype-aligned representations in semantic space.

- **Prototype-guided Entity-Predicate Matching**:

  Match entity pairs to predicates in semantic embedding space for relation recognition.

- **Prototype-guided Learning**:

  Help PE-Net efficiently learn entity-predicate matching.

- **Prototype Regularization**:

  Relieve ambiguous entity-predicate matching caused by predicate's semantic overlap.

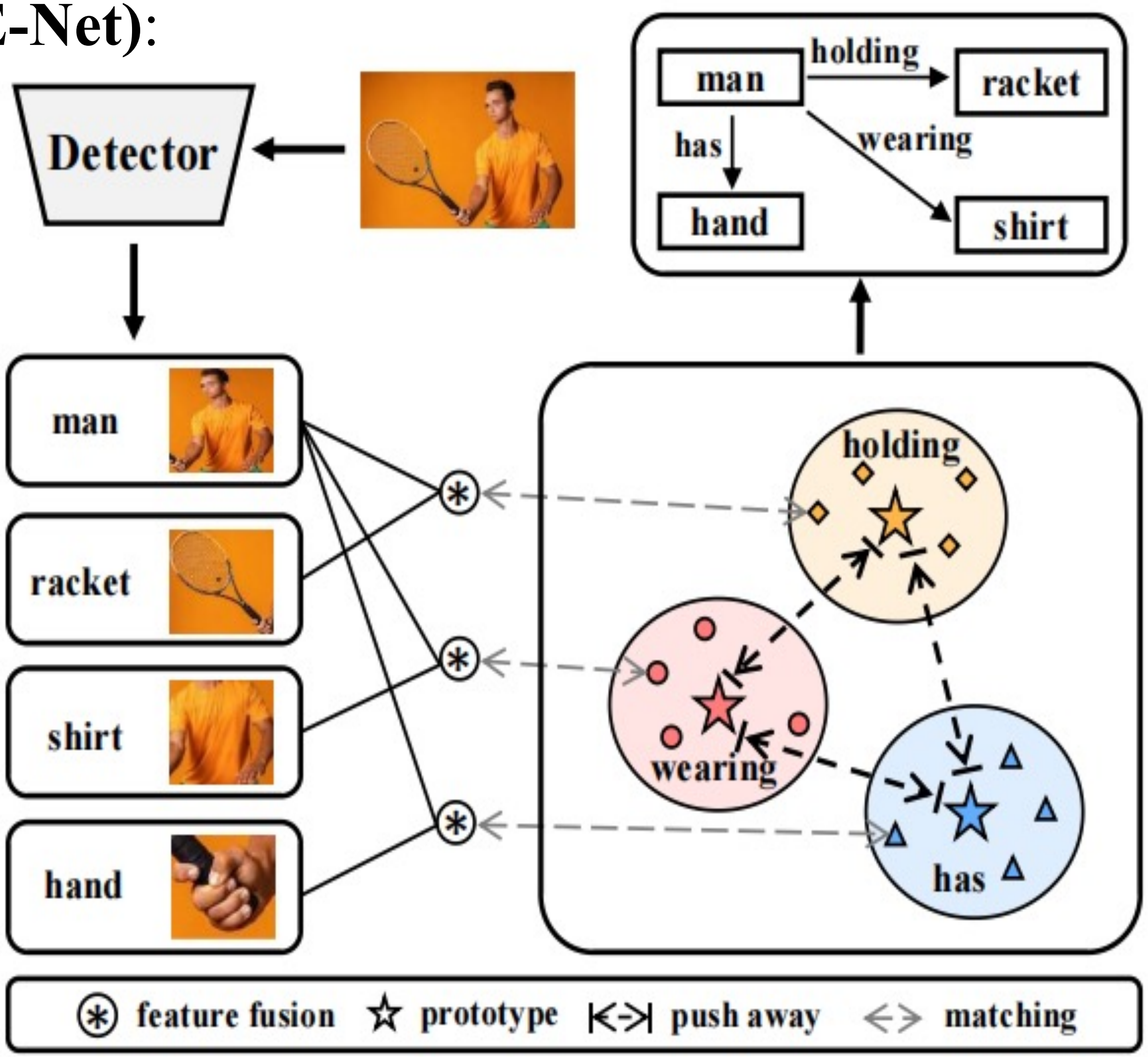

Fig. 2. The main process of our proposed PE-Net.

Method

Prototype-based Embedding Network (PE-Net):

- **Prototype-based Modeling**:

$$\boldsymbol{s} = \boldsymbol{W}_s \boldsymbol{t}_s + \boldsymbol{v}_s, \quad (1)$$

$$\boldsymbol{o} = \boldsymbol{W}_o \boldsymbol{t}_o + \boldsymbol{v}_o, \quad (2)$$

$$\boldsymbol{p} = \boldsymbol{W}_p \boldsymbol{t}_p + \boldsymbol{u}_p, \quad (3)$$

where t_s, t_o and t_p are class labels' word embedding, v_s, v_o, u_p are the instance-varied semantics contents, $W_s t_s$, $W_o t_o$, $W_p t_p$ are class-specific semantic prototypes.

$$\boldsymbol{g}_s = \boldsymbol{\sigma}(\boldsymbol{f}((\boldsymbol{W}_s \boldsymbol{t}_s) \oplus \boldsymbol{h}(\boldsymbol{x}_s))), \quad (4)$$

$$\boldsymbol{v}_s = \boldsymbol{g}_s \odot \boldsymbol{h}(\boldsymbol{x}_s), \quad (5)$$

where $h(\cdot)$ is visual-to-semantic function.

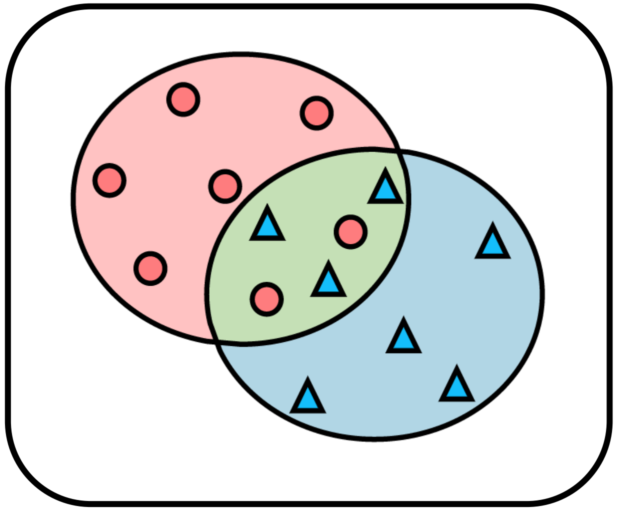

Fig. 3. Visual-based space Modeling.

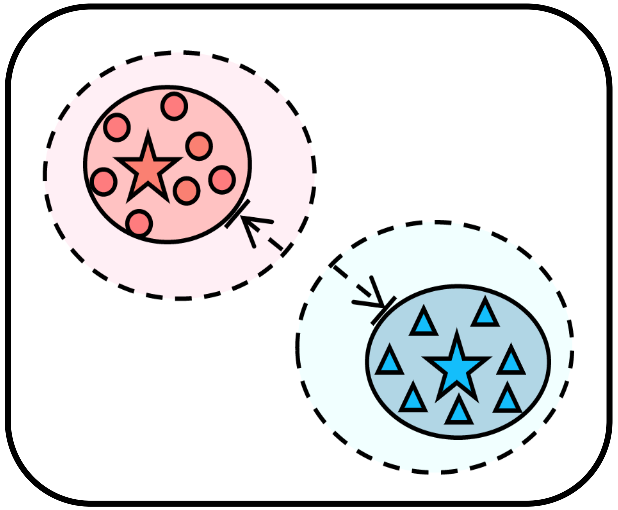

Fig. 4. Prototype-based space Modeling.

Method

Prototype-based Embedding Network (PE-Net):

- **Prototype-guided Entity-Predicate Matching**:

$$\mathcal{F}(\boldsymbol{s},\boldsymbol{o}) \rightarrow \boldsymbol{p} = \boldsymbol{W}_p \boldsymbol{t}_p + \boldsymbol{u}_p, \quad (6)$$

$$\mathcal{F}(\boldsymbol{s},\boldsymbol{o}) = \mathrm{ReLU}(\boldsymbol{s}+\boldsymbol{o}) - (\boldsymbol{s}-\boldsymbol{o})^2 \text{ [1]}, \quad (7)$$

where $\mathcal{F}(s,o)$ denotes the feature fusion function.

Equivalent transformation:

$$\mathcal{F}(\boldsymbol{s},\boldsymbol{o}) - \boldsymbol{u}_p \rightarrow \boldsymbol{W}_p \boldsymbol{t}_p, \quad (8)$$

where $\mathcal{F}(s,o) - u_p$ is defined as relation representation r, which should be matched to its corresponding predicate prototype $W_p t_p$. (represented as c in the following).

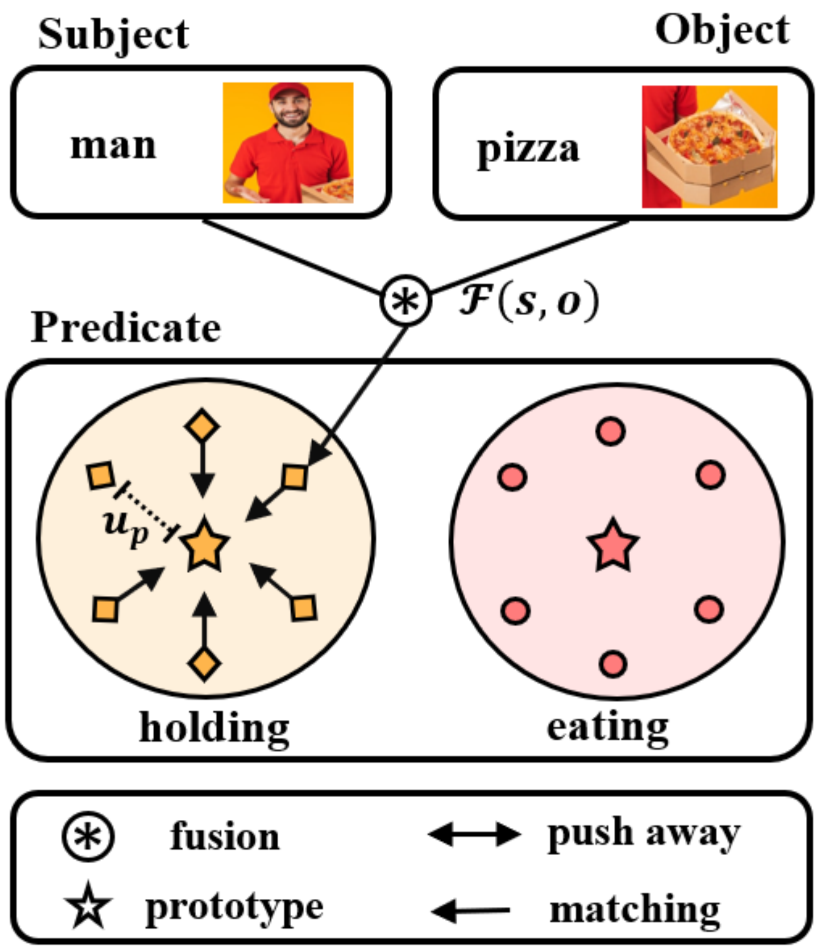

Fig. 5. Prototype-guided Entity-Predicate Matching.

[1] Zhang, Yan, et al. "Learning to count objects in natural images for visual question answering." arXiv preprint:1802.05766 (2018).

Method

Prototype-based Embedding Network (PE-Net):

- **Prototype-guided Learning**:

Cosine distance: Increasing the cosine similarity between the relation representation r, and its corresponding prototype c_t,

$$\mathcal{L}_{e\_sim} = -log \frac{exp(\langle \overline{r}, \overline{c_t} \rangle / \tau)}{\sum_{j=0}^{N} exp(\langle \overline{r}, \overline{c_j} \rangle / \tau)}. \quad (9)$$

Euclidean distance: Increasing the Euclidean distance between the relation representation r, and its corresponding prototype c_t,

$$g_j = \| r - c_j \|_2^2, \quad (10)$$

$$\mathcal{L}_{e\_euc} = max(0, g^+ - g^- + \gamma_1). \quad (11)$$

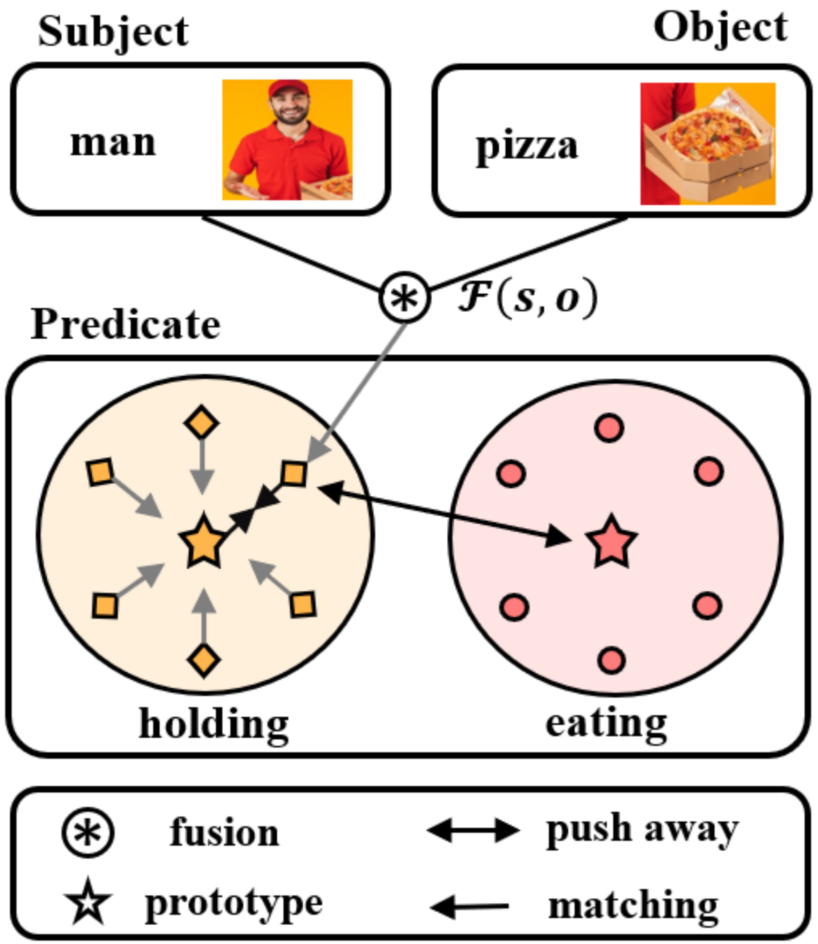

Fig. 6. Prototype-guided Learning.

Method

Prototype-based Embedding Network (PE-Net):

- **Prototype Regularization**:

  Cosine distance /Euclidean distance : Alleviates ambiguous matching caused by semantic overleap between predicates by enlarging distinction between predicate prototypes c_t.

$$S = \overline{C} \cdot \overline{C}^T = (s_{ij}), \quad (12)$$

$$\mathcal{L}_{r\_sim} = \|S\|_{2,1} = \sum_{i=0}^{N} \sqrt{\sum_{j=0}^{N} s_{ij}^2}, \quad (13)$$

$$d_{ij} = \| c_i - c_j \|_2^2, \quad (14)$$

$$\mathcal{L}_{r\_euc} = max(0, -d^- + \gamma_2). \quad (15)$$

- **Relation Inference:**

$$res_r = arg\,\max_i (q_i \mid q_i = \langle \overline{r}, \overline{c_i} \rangle / \tau). \quad (16)$$

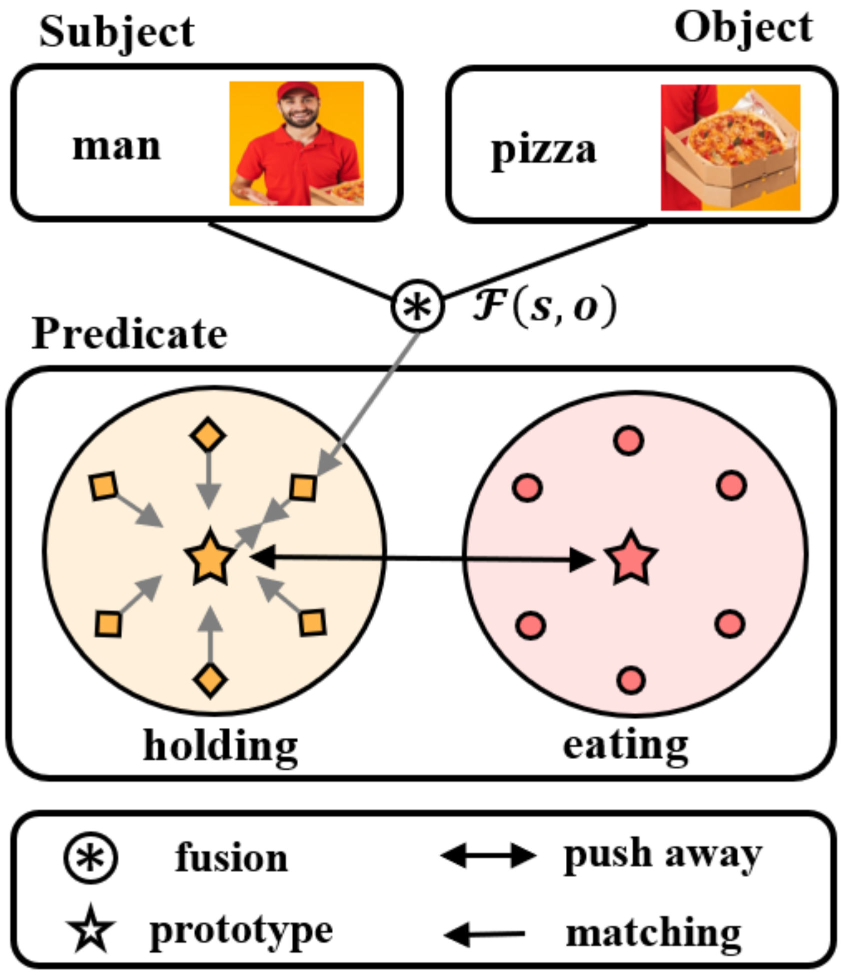

Fig. 7. Prototype Regularization.

Experiment

Compared with State of the Arts:

Model	PredCls			SGCls			SGDet		
	R@50/100	mR@50/100	M@50/100	R@50/100	mR@50/100	M@50/100	R@50/100	mR@50/100	M@50/100
Motifs$^\diamond$ [25,36]	65.3 / 67.2	14.9 / 16.3	40.1 / 41.8	38.9 / 39.8	8.3 / 8.8	23.6 / 24.3	32.1 / 36.8	6.6 / 7.9	19.4 / 22.4
VCTree$^\diamond$ [25,27]	65.5 / 67.4	16.7 / 17.9	41.1 / 42.7	40.3 / 41.6	7.9 / 8.3	24.1 / 25.0	31.9 / 36.0	6.4 / 7.3	19.2 / 21.7
G R-CNN* [11,33]	65.4 / 67.2	16.4 / 17.2	40.9 / 42.2	37.0 / 38.5	9.0 / 9.5	23.0 / 24.0	29.7 / 32.8	5.8 / 6.6	17.8 / 19.7
KERN* [1,11]	65.8 / 67.6	17.7 / 19.2	41.8 / 43.4	36.7 / 37.4	9.4 / 10.0	23.1 / 23.7	27.1 / 29.8	6.4 / 7.3	16.8 / 18.6
VTransE$^\diamond$ [25,40]	65.7 / 67.6	14.7 / 15.8	40.2 / 41.7	38.6 / 39.4	8.2 / 8.7	23.4 / 24.1	29.7 / 34.3	5.0 / 6.1	17.4 / 20.2
R-CAGCN [32]	66.6 / 68.3	18.3 / 19.9	42.5 / 44.1	38.3 / 39.0	10.2 / 11.1	24.3 / 25.1	28.1 / 31.3	7.9 / 8.8	18.0 / 20.1
GPS-Net* [11,15]	65.2 / 67.1	15.2 / 16.6	40.2 / 41.9	37.8 / 39.2	8.5 / 9.1	23.2 / 24.2	31.3 / 35.9	6.7 / 8.6	19.0 / 22.3
RU-Net [17]	67.7 / 69.6	- / 24.2	- / 46.9	**42.4 / 43.3**	- / 14.6	- / 29.0	**32.9 / 37.5**	- / 10.8	- / 24.2
PE-Net(P)	**68.2 / 70.1**	23.1 / 25.4	45.7 / 47.8	41.3 / 42.3	13.1 / 14.8	27.2 / 28.6	32.4 / 36.9	8.9 / 11.0	20.7 / 24.0
PE-Net	64.9 / 67.2	**31.5 / 33.8**	**48.2 / 50.5**	39.4 / 40.7	**17.8 / 18.9**	**28.6 / 29.8**	30.7 / 35.2	**12.4 / 14.5**	**21.6 / 24.9**
Motifs-TDE [26]	46.2 / 51.4	25.5 / 29.1	35.9 / 40.3	27.7 / 29.9	13.1 / 14.9	20.4 / 22.4	16.9 / 20.3	8.2 / 9.8	12.6 / 15.1
Motifs-CogTree [34]	35.6 / 36.8	26.4 / 29.0	31.0 / 32.9	21.6 / 22.2	14.9 / 16.1	18.3 / 19.2	20.0 / 22.1	10.4 / 11.8	15.2 / 17.0
Motifs-BPL-SA [5]	50.7 / 52.5	29.7 / 31.7	40.2 / 42.1	30.1 / 31.0	16.5 / 17.5	23.3 / 24.3	23.0 / 26.9	13.5 / 15.6	18.3 / 21.3
Motifs-NICE [10]	55.1 / 57.2	29.9 / 32.3	42.5 / 44.8	33.1 / 34.0	16.6 / 17.9	24.9 / 26.0	**27.8 / 31.8**	12.2 / 14.4	20.0 / 23.1
Motifs-PPDL [12]	47.2 / 47.6	32.2 / 33.3	39.7 / 40.5	28.4 / 29.3	17.5 / 18.2	23.0 / 23.8	21.2 / 23.9	11.4 / 13.5	16.3 / 18.7
Motifs-GCL [3]	42.7 / 44.4	36.1 / 38.2	39.4 / 41.3	26.1 / 27.1	20.8 / 21.8	23.5 / 24.5	18.4 / 22.0	**16.8 / 19.3**	17.6 / 20.7
Motifs-Reweight [2]	53.2 / 55.5	33.7 / 36.1	43.5 / 45.8	32.1 / 33.4	17.7 / 19.1	24.9 / 26.3	25.1 / 28.2	13.3 / 15.4	19.2 / 21.8
PE-Net-Reweight	**59.0 / 61.4**	**38.8 / 40.7**	**48.9 / 51.1**	**36.1 / 37.3**	**22.2 / 23.5**	**29.2 / 30.4**	26.5 / 30.9	16.7 / 18.8	**21.6 / 24.9**

Tab. 1. Performance comparison with the state-of-the-art SGG methods on VG dataset. PE-Net(P) refers to the PE-Net only trained with PL. PE-Net indicates PE-Net trained with both PL and PR.

Experiment

Measuring Representation Modeling of PE-Net:

- **Calculation of IV and IIVR**:

Intra-class Variance (IV): measure the intra-class compactness of entity's or predicate's representations,

$$\sigma_{within}^2 = \frac{1}{Mn}\sum_{i=0}^{M}\sum_{j=1}^{n}|\phi_{i,j} - \mu_i|_2^2, \qquad \textbf{(17)}$$

Intra-class to Inter-class Variance (IIV): measure the inter-class distinctiveness of the representations.

$$\frac{\sigma_{within}^2}{\sigma_{between}^2} = \frac{1}{n}\frac{\sum_{i=0}^{M}\sum_{j=1}^{n}|\phi_{i,j} - \mu_i|^2}{\sum_{i=0}^{M}|\mu_i - \mu|_2^2}. \qquad \textbf{(18)}$$

Models	IV-O ↓	IIVR-O ↓	IV-R ↓	IIVR-R ↓
Motifs [26, 38]	9.73	1.93	1.41	2.72
VCTree [26, 28]	8.31	2.11	1.50	2.78
Transformer [26, 30]	9.08	2.05	1.44	2.76
G-RCNN [12, 35]	8.76	1.99	1.46	2.81
GPS-Net [12, 16]	9.36	2.07	1.53	2.69
PE-Net	**0.74**	**0.24**	**1.06**	**1.67**

Tab. 2. Quantitative results on representation quality.

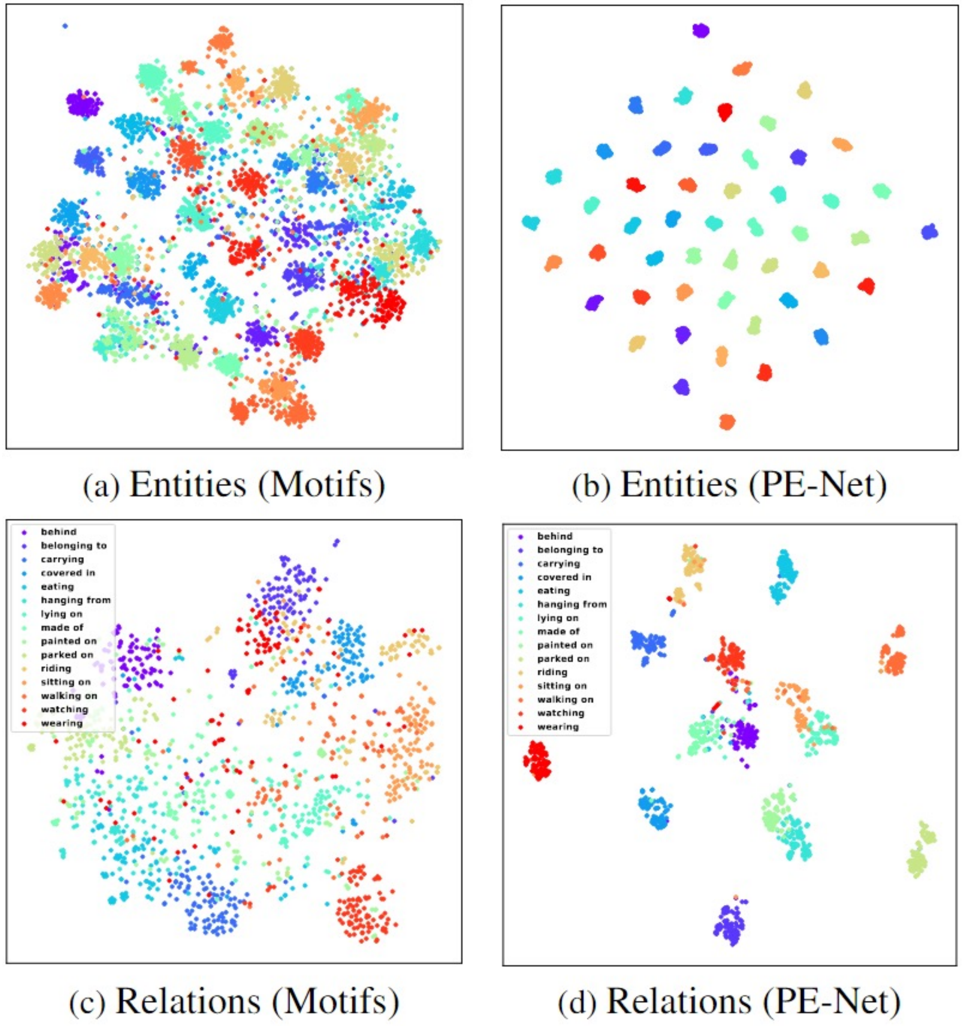

Fig. 8. The comparison of t-SNE visualization results on entity and predicate feature distributions.

Thanks

If you have any questions, please contact me at :

xinyulyu68@gmail.com

Codes: https://github.com/VL-Group/PENET