MULTIMODAL LATENT FACTOR MODEL WITH LANGUAGE CONSTRAINT FOR PREDICATE DETECTION

Xuan Ma[1,2], *Bing − Kun Bao*[3], *Lingling Yao*[4] *Changsheng Xu*[1,2]

[1] National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences

[2] University of Chinese Academy of Sciences

[3] College of Telecommunications & Information engineering, Nanjing University of Posts and Telecommunications

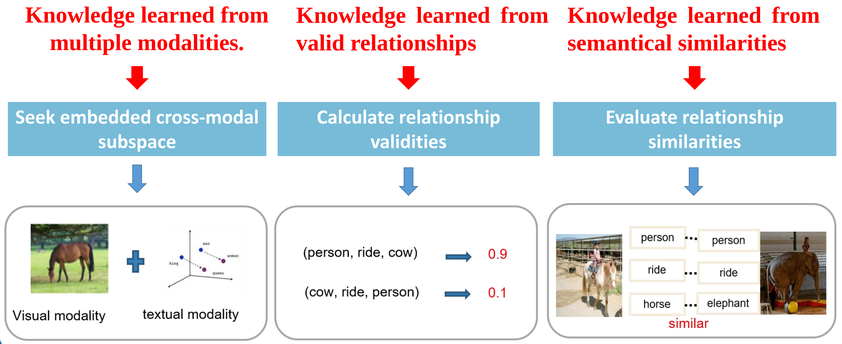
[4] Tencent, Shenzhen, China

Motivation

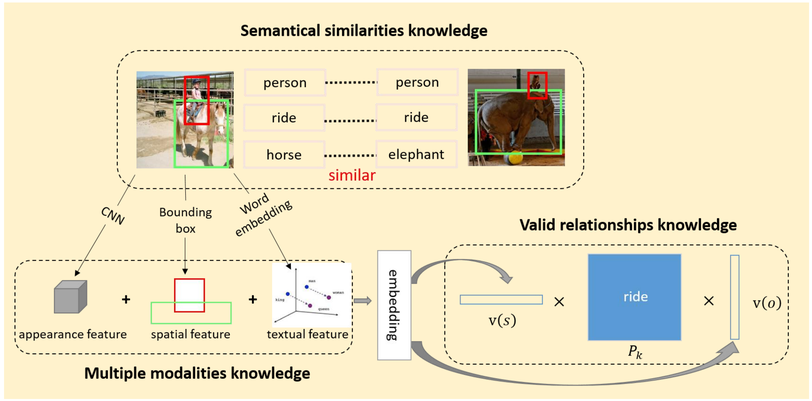
- In order to predict the predicate between entities, learning a proper representation for the predicate is of vital importance. However, predicate representation still has many challenges:
- ✓ Predicate can be represented in both visual and textual modalities, so a cross-modal representation is needed.
- ✓ Predicate is relevant to its subject and object, a predicate's representation should be considered combining its subject and object. That's to say:
  - valid/invalid triplets can be used to extract useful feature for predicates.
  - if subjects and objects are similar semantically, their predicates are similar probably.

Contributions

We propose Multimodal Latent Factor Model with Language Constraint (MMLFM-LC) for predicate detection with integrating three kinds of knowledge corresponding to three challenges:

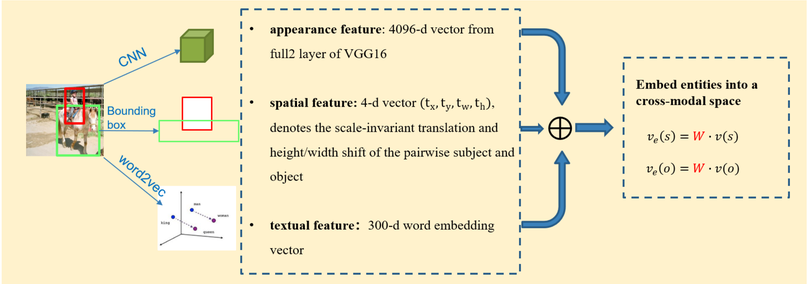


Framework



➢ Knowledge Learned from Multiple Modalities

We utilize visual and textual features to represent entities. Visual feature consists of appearance and spatial feature, while textual feature refers to word embedding vectors. Then these three kinds of features are concatenated and embedded into a cross-modal space.



➢ Knowledge Learned from Valid Relationships

We use a bilinear structure to model the complicated interactions among entities and predicates and calculate relationship validities in cross-modal space.

◆ **Predicate representation.**

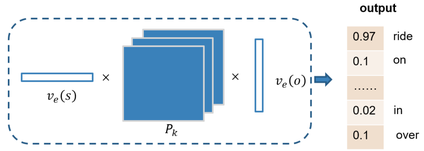
- decompose the predicate into a set of rank-one matrixes in order to reduce parameter number, also known as latent factors

$$\mathrm{P} = [\mathrm{P}_1, \cdots, \mathrm{P}_k, \cdots \mathrm{P}_N],\ \mathrm{P}_k \in \mathbb{R}^{K \times K}$$

$$P_k = \sum_{r=1}^{d} \alpha_r^k \Theta_r\ ,\ \alpha^k \epsilon \mathbb{R}^d,$$

◆ **Relationship validities.**

- use a bilinear structure to model the interactions among entities and predicates
- the score reflects the validity



◆ **Loss function**

- maximize possibilities of valid triplets and minimize possibilities of invalid triplets

$$C(\Theta, , \mathrm{A}) = -\sum_{(i,k,j)\epsilon\mathcal{P}} p(s_i, p_k, o_j) + \sum_{(i\prime,k\prime,j\prime)\epsilon\mathcal{N}} p(s_{i\prime}, p_{k\prime}, o_{j\prime})$$

◆Knowledge Learned from Semantical Similarities

- Compute the similarities between pairwise subjects, predicates, and objects respectively

$$sim_s = \cos\left(v_e(s_i), v_e(s_{i'})\right), sim_o = \cos\left(v_e(o_j), v_e(o_{j'})\right), sim_p = \cos(p_k, p_{k\prime})$$

- Evaluate similarities

$$f(sim) = \begin{cases} 1, sim \geq t \\ 0, sim < t \end{cases}$$

- Compute the semantic loss

$$L = f(sim_s)f(sim_p)(1 - f(sim_o) + f(sim_s)f(sim_o)(1 - f(sim_p) + f(sim_p)f(sim_o)(1 - f(sim_s)$$

➢ Objective function

$$\min_{\Theta,A,W} C + \lambda L$$

C is the loss according to the relationship validities, L is the loss related to the relationship similarities.

Experiment Results

Method	VR		VG	
	Recall@50	Recall@100	Recall@50	Recall@100
LP [5]	47.87	47.87	-	-
VTransE [9]	44.76	44.76	62.63	62.87
LK [11]	55.16	55.16	-	-
Zoom-Net [22]	50.69	50.69	67.25	77.51[1]
CAI+SCA-M [22]	55.98	55.98	-	-
DR-Net [23]	-	-	62.05	71.96
Vip-CNN [24]	-	-	63.44	74.15
Baseline: B+A	52.41	52.41	64.72	72.04
B+A+S	53.01	53.01	65.31	72.54
B+A+T	54.20	54.20	67.50	75.21
B+A+S+T	54.50	54.50	68.00	75.63
B+A+LC	52.98	52.98	66.74	74.35
B+A+S+LC	53.52	53.52	67.01	75.01
B+A+T+LC	56.30	56.30	69.89	77.90
B+A+S+T+LC	**56.65**	**56.65**	**70.30**	**78.25**

◆ Experiment setting
'B' is the bilinear structure,
'A' is appearance feature,
'S' is spatial feature,
'T' is textual feature,
'LC' is the language constraint.
'A'+ 'S' is equivalent to visual feature.

◆ Experiments on Visual Relationship and Visual Genome prove that our model gets the best performance.