MULTIMODAL LATENT FACTOR MODEL WITH LANGUAGE CONSTRAINT FOR PREDICATE DETECTION

Xuan Ma[12] *Bing-Kun Bao*[3] *Lingling Yao*[4] *Changsheng Xu*[12]

[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

ABSTRACT

Nowadays, visual relationship detection has shown an important utility in scene understanding. Predicate detection, which aims to detect the predicate between entities in an image, is an important part of visual relationship detection. In this paper, we propose Multimodal Latent Factor Model with Language Constraint (MMLFM-LC) for predicate detection with the novelty of integrating knowledge learned from multiple modalities, valid relationships and semantical similarities. Representations of visual and textual modalities are firstly input into the constructed model. Secondly, a bilinear structure is introduced to model the relationships using valid relationships, while a language constraint is also built utilizing semantical similarities. Lastly, visual and textual representations are fused in an embedded subspace for predicate detection. Experiments on both Visual Relationship and Visual Genome datasets show that our method outperforms other methods on predicate detection.

Index Terms— Predicate representation, Multimodal fusion, Valid relationships, Semantical similarities

1. INTRODUCTION

Visual relationship detection is a fundamental problem in computer vision and plays an important role in many visual tasks, such as action recognition [1] [2], visual phrase recognition [3] and visual question answering [4]. A visual relationship refers to a triplet of *(subject,predicate,object)*, where predicate describes the interaction between subject and object, such as "ride" in *(person,ride,bike)*. Relationship detection process can be decomposed into two parts: object detection and predicate detection [5]. Object detection aims to find a minimum bounding box for each entity and classify it. With the appearance of various deep neural networks, such as Faster-RCNN [6] and R-FCN [7], the precision of object detection has reached a relatively high level. The task of predicate detection is to represent and predict the predicates when bounding boxes and classes of related entities are given. Since predicates normally do not have stable visual appearance, the task of predicate detection remains challenging.

There are three aspects can be considered to improve the performance of predicate detection.

(1) Knowledge learned from multiple modalities. Recall that the given information includes the whole image and bounding boxes of entities in visual modality, as well as the classes of those entities in textual modality. And predicate also has terms of visual and textual modalities. Obviously, fully utilizing the multiple modalities of given information to seek a cross-modal representation of predicate is more discriminative than that on single modality.

(2) Knowledge learned from valid relationships. Inspired by one relevant study in natural language processing - Statistical Relational Learning (SRL) [8], whose basic assumption is that a valid predicate has higher probability of appearance in-between two given entities in real-world textual knowledge bases, we extract discriminative cross-modal representations of entities and predicate by assigning high probabilities to valid relationships and low probabilities to others. For example, the relationship (cow, ride, person) will be assigned low probability as rarely there is an image or text describing this relationship.

(3) Knowledge learned from semantical similarities. Besides valid relationships from real-world, semantical similarity is also beneficial for predicate detection. Intuitively, given a valid relationship, another relationship which is semantically similar to it is likely to be valid as well. For example, (person, ride, horse) and (person, ride, cow) are semantically similar because both horse and cow are animals. If we have the knowledge that (person, ride, horse) is a valid relationship, we might infer that (person, ride, cow) is also of high probability to be valid.

Some works have studied on above mentioned ideas but none is considering all these in a holistic way. [9] [10] on-

[1]This work is supported by the National Key Research & Development Plan of China (No. 2017YFB1002800), by the National Science Foundation of China under Grant 61872424, 61572503, 61720106006, 61432019, 61772287 and by NUPTSF (No. NY218001), also supported by the Key Research Program of Frontier Sciences, CAS, Grant NO. QYZDJ-SSW-JSC039, and the K.C.Wong Education Foundation.

ly use visual modality to represent relationship, which lose on more semantic information provided by textual modality. [11] [12] compute a conditional probability distribution of a predicate given a (subject, object) pair based on valid relationships, but neglect the correlations between relationships. [5] uses word vectors to measure semantical similarities between relationships, however, the interaction within single relationship is lost.

In this paper, we propose Multimodal Latent Factor Model with Language Constraint for predicate detection with novelty of integrating all three above mentioned aspects, knowledge learned from multiple modalities, valid relationships and semantical similarities. The framework is shown in Fig. 1. Firstly, visual and textual representations of each entity are extracted for cross-modal fusion into the model. Secondly, a bilinear structure is constructed to assign the probability to relationships using knowledge learned from valid relationships. Meanwhile, the semantical similarity between relationships is modeled as language constraint to enhance the probability assignment based on given valid ones. Lastly, we formulate an overall objective function to repeat the previous step iteratively to achieve a unified cross modal embedded subspace where the valid relationships and their similar ones are assigned with higher probabilities.

2. RELATED WORK

This section introduces related work on textual relationship analysis, and visual relationship analysis.

Textual Relationship Analysis. Exploring the particular representation for each predicate is the main focus in textual relationship analysis. [13] proposes to map both entity and predicate into vectors to compute the"matching energy" of a relationship. [14] uses a single layer model(SLM) to evaluate a given triplet, where each relation is represented with two weight matrixes. To explain the role of predicate more intuitively, [15] proposes a translation model TransE, where the relation vector is regarded as a translation from head vector to tail vector. [16] proposes to use the predicate-based bilinear transformation to characterize the second-order correlation between entities and predicates. These methods provide a specific representation for predicates and can easily build the interaction between predicates and entities.

Visual Relationship Analysis. There are two kinds of models to analyze the visual relationship currently: joint model and separate model. Joint model [17] [18] trains a classifier for each relationship. However, the number of combinations is too large and the long-tailed distribution makes it hard to generalize. In contrast, separate model trains a classifier for each predicate without consideration of its subject and object [19] [5]. However, all these works do not learn a particular representation for the predicate. Zhang *et al.* [9] model the predicate as the translation vector between the subject and object only on visual modality by following TransE. Latter, Nian *et al.* [20] firstly study on cross-modal

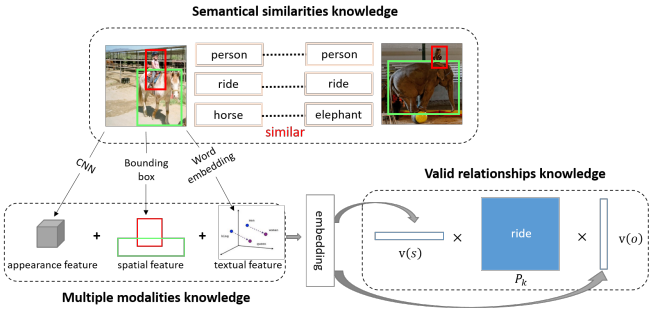

Fig. 1: Framework of our model. Multiple modalities mean the entities are represented in both visual and textual modalities. Valid relationships are modelled with a bilinear structure. Semantical similarities refer to the correlations between two relationships.

relationship representation, and get better performance than that on single visual modality. In order to make better use of semantic information of relationship, many recent works introduce language priors to guide relationship detection process. Yu *et al.* [11] compute a conditional probability distribution of a predicate given a (subject, object) pair, but neglect the semantic similarities between relationships. Lu *et al.* [5] calculate relationship similarities and project relationships into an embedding space where similar relationships are optimized to be close together. However, the similarities are measured only on the entirety of relationships. In this paper, we also take advantage of language information to guide the predicate detection process. By forcing similar relationships to have similar representations, we can infer the possible relationships based on the valid relationships.

3. MULTIMODAL LATENT FACTOR MODEL WITH LANGUAGE CONSTRAINT

This section details the proposed MMLFM-LC. We respectively introduce the method on how to learn knowledge from multiple modalities, valid relationships and semantical similarities. And at last, we formulate the overall objective function for predicate detection.

3.1. Knowledge Learned from Multiple Modalities

We utilize visual and textual features to represent entities, which not only describe the basic visual content, but also provide semantic correlations among different entities.

Visual Feature. Visual feature consists of appearance feature and spatial feature. We get appearance features in 4096-d from the full2 layer in VGG-16. As for spatial feature, we calculate a 4-d vector (t_x, t_y, t_w, t_z) to represent spatial information, with (x, y, w, h) and (x', y', w', h') being the coordinates of subject and object.

$$t_x = \frac{x - x'}{w'}, t_y = \frac{y - y'}{h'}, t_w = log\frac{w}{w'}, t_h = log\frac{h}{h'} \quad (1)$$

(t_x, t_y) denotes a scale-invariant translation and (t_w, t_y) specifies a height/width shift with respect to the pairwise subject and object.

Textual Feature. To obtain semantically similar expression, we use word embedding [21] to encode subject and object classes into N-d vectors in a semantic space, where similar words have a closer distance.

For a given image, we firstly get the entity segmentations according to the bounding boxes and resize them into 224×224. Then the entities are input into VGG-16 to get appearance feature. Spatial feature is calculated using the bounding box coordinates and textual feature is the word vector corresponding to the entity class. Finally, all these features are concatenated into a vector to represent entities.

3.2. Knowledge Learned from Valid Relationships

Inspired by Latent Factor Model [16], which is proposed to use a bilinear structure to model the complicated interactions among entities and predicates on textual modality, we further enhance the model by adding visual modality into it.

Entities s and o are firstly embedded into an cross-modal space using matrix $W \in \mathbb{R}^{K \times K}$:

$$v_e(s) = W \cdot v(s), v_e(o) = W \cdot v(o) \tag{2}$$

where $v(s)$ and $v(o)$ are entity vectors concatenated using both visual and texture features, $v_e(s)$ and $v_e(s)$ are embedded entity representations in K dimension. Let $\mathbf{P}_k \in \mathbb{R}^{K \times K}$ be the representation of the k-th predicate, which is unknown, K is the dimension of entity representation. Predicate set can be represented as $\mathbf{P} = [\mathbf{P}_1, \cdots, \mathbf{P}_N]$, where N is the number of predicates. For a relationship (s_i, p_k, o_j), the possibility of the subject s_i and object o_j being connected by predicate p_k is calculated as follows:

$$p(s_i, p_k, o_j) = v_e(s_i) \cdot \mathbf{P}_k \cdot v_e(o_j) \tag{3}$$

where $\mathbf{P}_k$ is the matrix corresponding to the k-th predicate.

However, leaning $\mathbf{P}$ directly from training data can easily cause overfitting due to the large amount of parameters. To avoid it, we build on the idea of LFM to reduce the parameter amount by decomposing the representation of each predicate into a set of rank-one matrixes $\mathbf{\Theta}$, also known as latent factors:

$$\mathbf{P}_k = \sum_{r=1}^{d} \alpha_r^k \mathbf{\Theta}_{\boldsymbol{r}}, \alpha^k \in \mathbb{R}^d \tag{4}$$

where d is the number of latent factors and α is a sparse vector to weight the contribution of each latent factor. With the assumption in SRL that valid predicate should have a high probability, we learn $\mathbf{\Theta}_{\boldsymbol{r}}$ and α by minimizing the following loss function:

$$C(\Theta, A) = -\sum_{(i,k,j)\in\mathcal{P}} p(s_i, p_k, o_j) + \sum_{(i',k',j')\in\mathcal{N}} p(s_{i'}, p_{k'}, o_{j'}) \tag{5}$$

where $\mathcal{P}$ is the set of valid relationships and $\mathcal{N}$ is the set of invalid relationships. The significance is that if the relationship is valid, its possibility should be as large as possible, otherwise it should be close to zero.

3.3. Knowledge Learned from Semantical Similarities

As discussed above, knowledge learned from valid relationships can well capture the interaction between entities and predicates. Different from Lu's work [5], which calculates similarities based on the representations of entirety of relationships, we mine the similarities with higher-level semantics. More specifically, we separate the relationships into a triplet *(s,p,o)*, and compute the similarities by integrating the similarities between subjects, predicates, and objects respectively. Similarities between entities and those between predicates are both measured by cosine distance:

$$\begin{aligned} sim_s =& cosine(v_e(s_i), v_e(s_{i'})), \quad sim_p = cosine(p_k, p_{k'}), \\ sim_o =& cosine(v_e(o_j), v_e(o_{j'})), \end{aligned}$$

where $cosine()$ calculates the cosine distance between the two items. We then set a threshold t to evaluate these similarities. If the similarity value is larger than t, we consider the two items to be similar.

$$f(sim) = \begin{cases} 1, & sim \geq t \\ 0, & sim < t. \end{cases} \tag{6}$$

For a pair of relationships, we calculate similarities between subjects, predicates, and objects respectively. Then the semantic loss of relationships can be written as:

$$\begin{aligned} L =& f(sim_s) f(sim_p)(1 - sim_o) + f(sim_p) f(sim_o) \\ & (1 - sim_s) + f(sim_o) f(sim_s)(1 - sim_p) \end{aligned} \tag{7}$$

The significance is that, if two pairs of items are similar, the two relationships are considered to be similar, and so are the remain items, which means the sim value of the remain items should be as close as possible to 1. By minimizing the above loss function, the semantic similarities of relationships can be captured thus benefits the entity and predicate representations.

3.4. Objective Function

Finally, we combine the above three kinds of knowledge to learn the cross-modal representations of entities and predicates with the following objective function:

$$\min_{\Theta, A, W} C + \lambda L \tag{8}$$

where C is the loss according to the valid relationships knowledge and L is the loss related to the semantical similarities knowledge. When testing, for a pair of entities, we calculate relationship probabilities for its combinations with every predicates. The predicted predicate is the one with highest probability.

4. EXPERIMENTS

We conduct experiments on the Visual Relationship (VR) and Visual Genome (v-1.2) (VG) datasets. VR consists of 5000

images, including 100 entity classes and 70 predicates. We adapt the same split of training and test sets in [5], where 4000 images are used for training and 1000 images for test. VG contains 99658 images with 200 entity classes and 100 predicates. We randomly split it into 73801 for training and 25857 for test following [9]. Both datasets have the bounding boxes of all entities.

For baselines, we choose seven recently proposed methods: 1) LP. Visual Relationship Detection with Language Prior (LP) proposed in [5] considers knowledge from valid relationships only on visual modality. And the semantical similarities are measured on the entirety of relationships. 2) LK. Visual Relationship Detection with Internal and External Linguistic Knowledge Distillation (LK) proposed in [11] uses multiple modalities to represent entities and considers knowledge from valid relationships, but neglects the semantical similarities between relationships. 3) VTransE. VTransE [9] considers the predicate as a translation from subject to object, while it does not fully utilize the knowledge from valid relationships. 4) Zoom-Net. Zoom-Net [22] introduces an end-to-end visual relationship recognition model to mine feature-level interactions. However, the feature used is only from visual modality. 5) CAI+SCA-M. "CAI+SCA-M" [22] integrates both context-aware and spatiality-aware features to build interaction within a relationship, but does not consider semantic correlation between relationships. 6) DR-Net."DR-Net" [23] exploits both spatial configurations in visual modality and statistical dependencies among relationship predicates in textual modality. However, each modality is used singly without embedded into a unified space. 7) Vip-CNN. "Vip-CNN" [24] proposes a message passing structure to model the visual interdependency among relationship components, which does not consider semantic information from textual modality.

For better illustration, we also experiment on our method with different settings. The bilinear structure is to learn knowledge from valid relationships, which is referred as "B". To validate the effect of multiple modalities, we use different combinations of features in both with/without semantical similarities knowledge situation. "A" refers to appearance feature, "S" refers to spatial feature and "T" refers to textual feature. With appearance feature as the basis, "A", "A+S", "A+T", and "A+S+T" are four feature settings. Besides, we also explore the effect of language constraint('LC").

Following [5], we use recall @k as our evaluation metrics. Recall @k compute the fraction of times a true relation is predicted in the top K confident relation predictions in an image. Note that Recall@100 and Recall@50 are equivalent on Visual Relationship dataset because there are not enough objects in ground truth to produce over 50 pairs. The parameters are learned by 5-folds cross validation. The number of latent factors d is set to 200 and the dimension of entity vectors K is set to 600. Invalid relationships are generated by replacing the predicate in valid relationships with other invalid predicates.

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**

Table 1: Predicate detection result.'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.

Table 1 shows the compared results of our method with the baselines. Our model with multiple modalities and language constraint, that is "B+A+S+T+LC", outperforms on both VR and VG datasets. On VR, ours achieves 56.65 on recall@50. This is 0.67 higher than "CAI+SCA-M", which performs the best among the baselines. On VG, ours achieves 70.30 on recall@50 and 78.25 on recall@100.

Based on those results, we can make the following conclusions: (1) The usage of multiple modalities is more effective than that of single one. "B+A+S+T", with the combination of visual modality and textual modality, gets the result of 54.40 on VR while "B+A+S" that just uses visual modality only achieves 53.01. (2) The knowledge from valid relationships can improve the predicate detection. It can be demonstrated by the comparison of our baseline "B+A" and VransE, with 52.41 contrast to 47.87 on VR. (3) It is necessary to introduce semantical similarities on relationships to guide the predicate detection, and similarities measured on higher-level semantics are more efficient than that on low-level features. Shown in Table 4, "B+A+LC" is 2.02 and 2.31 higher than "B+A+S+T" on VG according to Recall@50 and Recall@100 respectively. Our "B+A+S+T+LC" with 56.65 is also better than LP with 47.87 on VR.

5. CONCLUSION

This paper proposes MMLFM-LC, which takes advantage of knowledge learned from multiple modalities, valid relationships and semantical similarities, to learn the cross-modality representation of entities and predicates. Experiments on Visual Relationship and Visual Genome prove that our model gets the best performance. Moving towards, we are going to tackle the zero-shot/one-shot relation learning problems, which are the most challenging tasks in predicate detection.

[1] [22] randomly split the VG dataset into training and test set with a ratio of 8 : 2, while our training/test sets are 73801 and 25857. So the training set size of [22] is bigger than ours.

6. REFERENCES

[1] Hakan Bilen, Basura Fernando, Efstratios Gavves, Andrea Vedaldi, and Stephen Gould, "Dynamic image networks for action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3034–3042.

[2] Arun Mallya and Svetlana Lazebnik, "Learning models for actions and person-object interactions with transfer to question answering," in *European Conference on Computer Vision*. Springer, 2016, pp. 414–428.

[3] Ronghang Hu, Marcus Rohrbach, Jacob Andreas, Trevor Darrell, and Kate Saenko, "Modeling relationships in referential expressions with compositional modular networks," in *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*. IEEE, 2017, pp. 4418–4427.

[4] Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher, "Knowing when to look: Adaptive attention via a visual sentinel for image captioning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, vol. 6, p. 2.

[5] Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei, "Visual relationship detection with language priors," in *European Conference on Computer Vision*. Springer, 2016, pp. 852–869.

[6] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99.

[7] Jifeng Dai, Yi Li, Kaiming He, and Jian Sun, "R-fcn: Object detection via region-based fully convolutional networks," in *Advances in neural information processing systems*, 2016, pp. 379–387.

[8] Alexandrin Popescul and Lyle H Ungar, "Statistical relational learning for link prediction," in *IJCAI workshop on learning statistical models from relational data*. Citeseer, 2003, vol. 2003.

[9] Hanwang Zhang, Zawlin Kyaw, Shih Fu Chang, and Tat Seng Chua, "Visual translation embedding network for visual relation detection," .

[10] Bohan Zhuang, Lingqiao Liu, Chunhua Shen, and Ian Reid, "Towards context-aware interaction recognition for visual relationship detection," in *Computer Vision (ICCV), 2017 IEEE International Conference on*. IEEE, 2017, pp. 589–598.

[11] Ruichi Yu, Ang Li, Vlad I Morariu, and Larry S Davis, "Visual relationship detection with internal and external linguistic knowledge distillation," in *IEEE International Conference on Computer Vision (ICCV)*, 2017.

[12] Hanwang Zhang, Zawlin Kyaw, Jinyang Yu, and Shih-Fu Chang, "Ppr-fcn: Weakly supervised visual relation detection via parallel pairwise r-fcn," *arXiv preprint arXiv:1708.01956*, 2017.

[13] Maximilian Nickel, Volker Tresp, and Hans-Peter Kriegel, "A three-way model for collective learning on multi-relational data.," in *ICML*, 2011, vol. 11, pp. 809–816.

[14] Richard Socher, Danqi Chen, Christopher D Manning, and Andrew Ng, "Reasoning with neural tensor networks for knowledge base completion," in *Advances in neural information processing systems*, 2013, pp. 926–934.

[15] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko, "Translating embeddings for modeling multi-relational data," in *Advances in neural information processing systems*, 2013, pp. 2787–2795.

[16] Rodolphe Jenatton, Nicolas L Roux, Antoine Bordes, and Guillaume R Obozinski, "A latent factor model for highly multi-relational data," in *Advances in Neural Information Processing Systems*, 2012, pp. 3167–3175.

[17] Yuval Atzmon, Jonathan Berant, Vahid Kezami, Amir Globerson, and Gal Chechik, "Learning to generalize to new compositions in image understanding," *arXiv preprint arXiv:1608.07639*, 2016.

[18] Mohammad Amin Sadeghi and Ali Farhadi, "Recognition using visual phrases," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 1745–1752.

[19] Fereshteh Sadeghi, Santosh K. Divvala, and Ali Farhadi, "Viske: Visual knowledge extraction and question answering by visual verification of relation phrases," in *Computer Vision & Pattern Recognition*, 2015.

[20] Fudong Nian, Bing-Kun Bao, Teng Li, and Changsheng Xu, "Multi-modal knowledge representation learning via webly-supervised relationships mining," in *Proceedings of the 2017 ACM on Multimedia Conference*. ACM, 2017, pp. 411–419.

[21] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[22] Guojun Yin, Lu Sheng, Bin Liu, Nenghai Yu, Xiaogang Wang, Jing Shao, and Chen Change Loy, "Zoom-net: Mining deep feature interactions for visual relationship recognition," *arXiv preprint arXiv:1807.04979*, vol. 2, 2018.

[23] Dai Bo, Yuqi Zhang, and Dahua Lin, "Detecting visual relationships with deep relational networks," 2017.

[24] Yikang Li, Wanli Ouyang, Xiaogang Wang, and Xiao'ou Tang, "Vip-cnn: Visual phrase guided convolutional neural network," in *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*. IEEE, 2017, pp. 7244–7253.