VISUAL RELATIONSHIP RECOGNITION VIA LANGUAGE AND POSITION GUIDED ATTENTION

Hao Zhou[1], *Chuanping Hu*[1,2], *Chongyang Zhang*[1,3]✉, *Shengyang Shen*[1]

[1]School of Electronic Information and Electrical Engineering,
Shanghai Jiao Tong University, Shanghai 200240, China
[2]Railway Police College, Zhengzhou 450053, China
[3]Shanghai Key Lab of Digital Media Processing and Transmission, Shanghai 200240, China
✉ Corresponding email: sunny_zhang@sjtu.edu.cn

ABSTRACT

Visual relationship recognition, as a challenging task used to distinguish the interactions between object pairs, has received much attention recently. Considering the fact that most visual relationships are semantic concepts defined by human beings, there are many human knowledge, or priors, hidden in them, which haven't been fully exploited by existing methods. In this work, we propose a novel visual relationship recognition model using language and position guided attention: language and position information are exploited and vectored firstly, and then both of them are used to guide the generation of attention maps. With the guided attention, the hidden human knowledge can be made better use to enhance the selection of spatial and channel features. Experiments on VRD [2] and VGR [1] show that, with language and position guided attention module, our proposed model achieves state-of-the-art performance.

Index Terms— Visual Relationship Recognition, Visual Attention, Deep Neutral Networks

1. INTRODUCTION

Visual relationship recognition is a key problem in image caption and image understanding. Given one image, traditional visual models only tell us the categories and positions of objects. For humans, we not only recognize objects but also catch the deep semantic information, especially the interaction of object pairs. Visual relationship recognition attempts to distinguish the different interactions of object pairs.

Generally, visual relationships can be expressed as triplets $\langle sub - pred - ob \rangle$ briefly, where sub, $pred$ and ob mean subject, predicate and object respectively. Based on the triplets expression, Lu et al.[2] evaluates visual relationship task in three conditions, including predicate detection, phrase detection and relationship detection. The most challenging and critical task in three conditions is the predictions of interactions of object pairs. Our work mainly focuses on distinguishing the interactions given the object pairs, which is most similar to the predicate detection task in [1].

Recently, the basic method [2] takes union regions of $subject$ and $object$ as inputs and adds language priors to preserve alignments with human perception. Visual features cannot distinguish complicate interactions of object pairs well. Inspired by many visual attention works [3, 4, 5, 6, 7, 8, 9], we believe visual attention module is beneficial to distinguish the variety of relationships within visual features. For human beings, we pay more attention to some specific areas in one image. With generated attention masks, visual attention module enhances the interest regions and suppresses the rest regions. However, most of these existing attention-based image recognition methods, use only visual features to learn the attention maps; the human knowledge, or priors, hidden in these human-defined relationship concepts, have not been fully exploited. Noteworthy, in some previous works, language [2, 10] and position [11, 10, 12] are also added into relation model. These previous models got combined features by concatenating visual features and other types of features directly, which means that the correlation and dependency between different modal features are ignored or exploited insufficiently.

In this work, rather than regarding position and language information as inputs equivalent to visual features, we exploit them generating attention masks as priors to guide the visual relationship recognition. Different from [6] generating attention masks only with visual features, language and position information are utilized as attention weights to guide visual attention generation, which called Language and Position Guided Attention module (LPGA). The intuitive motivation is that language and position information can provide specific clues to infer attention areas in relationships. For example, given an image, the position information, "$upper\ and\ lower$", guides the relation model to pay more attention to the object union region, and the language infor-

This work was partly funded by NSFC (No.61571297), the National Key Research and Development Program (2017YFB1002400), and STCSM (18DZ2270700, 18DZ1112300).

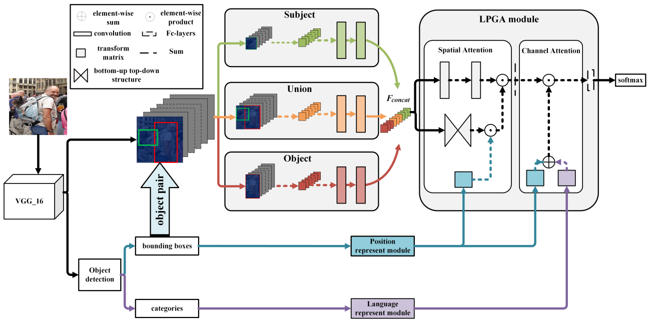

Fig. 1. Architecture of our proposed visual relation model.

mation, "$person-bike$", may enhance the region near person feet and hands. Further, we think different $predicates$ should have their own attention masks. For example, for "carry" and "ride", generated attention masks should be different even using the same language information (e.g., $person - bike$). Similarly, attention masks are different using similar position information. Thus, our LPGA module generates more accurate attention masks for each $predicate$ with language and position information.

To summarize, the main contribution is that we propose a novel LPGA module, where language and position information are exploited to guide the generation of more efficient attention maps. With guided attention, hidden human knowledge can be made better use to enhance the selection of spatial and channel features. With the LPGA module, our model achieves the state-of-the-art performances on Visual Relationship Dataset [2], and keeps consistent performances on Visual Genome [1].

2. LANGUAGE AND POSITION GUIDED ATTENTION

2.1. Framework

In visual relationship recognition, object regions and the union region play different roles in relation models. Object regions catch more dedicated object features, and the union region mainly denotes the interactive features. In this work, we attempt to combine both the union region and object regions into rear networks.

The framework of our proposed model is shown in Fig.1. Firstly, given one image, we use a pre-trained object detection model to get the object regions, categories and their bounding boxes. In this paper, we only focus on distinguishing $predicates$, so object categories and their bounding boxes are regarded ground truth in the latter model. In the relation model, we use VGG-16 network [13], shared with object detection model, as a backbone to extract visual features, and remove the pooling layer after $Conv5\_3$ in the VGG-16 network. The visual features denoted F extracted from $Conv5\_3$ are fed into three branches. One branch applies max_pooling operation on F; two branches apply RoIAlign pooling [14] on F according to the bounding boxes of object pairs respectively. Then two convolutional layers are added following pooling layers to extract visual features. Finally, concatenation operation is applied to get F_{concat} which are fed into the latter LPGA module.

2.2. Vectorizating Representation of Language and Position

We represent the language information through concatenating the word2vec [15] embedding of object pairs, which are mapped into 300 dims each word. Then, L2-normalization is applied to each word2vec embedding. The language representations $R_l(sub, ob) \in \mathbb{R}^m$ are encoded as:

$$R_l(sub, ob) = \boldsymbol{w}_l[w2vec(l_{sub}), w2vec(l_{ob})] + \boldsymbol{b}_l, \qquad (1)$$

where sub (ob) means subject (object), l_* are the text words of objects, and $\boldsymbol{w}_l \in \mathbb{R}^{m\times 600}$, $\boldsymbol{b}_l \in \mathbb{R}^m$ are learnable weights.

The coordinates of two bounding boxes are represented as $[x_s, y_s, w_s, h_s]$ and $[x_o, y_o, w_o, h_o]$, where (x, y), (w, h) are the coordinates of the upper left corner, the width and height of the bounding box. Same with [16], position representations contain the respective position information and their mutual position information. Given a single bounding box, the respective position information is represented as $[\frac{x}{W_u}, \frac{y}{H_u}, \frac{x+w}{W_u}, \frac{y+h}{H_u}, \frac{S}{S_u}]$, where W_u, H_u and S_u are the width, height and area of the union region. The mutual position information is represented as $[\frac{x_s - x_o}{w_o}, \frac{y_s - y_o}{h_o}, \log\frac{w_s}{w_o}, \log\frac{h_s}{h_o}]$. Then, L2-normalization is applied on the position representation vector, which denotes as $P(p_{sub}, p_{ob}) \in \mathbb{R}^{14}$. The position representations $R_p(sub, ob) \in \mathbb{R}^n$ are encoded as:

$$R_p(sub, ob) = \boldsymbol{w}_p P(p_{sub}, p_{ob}) + \boldsymbol{b}_p, \qquad (2)$$

where p_* are the coordinates of bounding boxes, and $\boldsymbol{w}_p \in \mathbb{R}^{n\times 14}$, $\boldsymbol{b}_p \in \mathbb{R}^n$ are also learnable weights.

We observe that kinds of $predicates$ focus on different image regions. Thus, we set different classifiers C^i_{pred} for each $predicate$. In different classifiers, attention module guided by language and position representations is added following F_{concat}. In our paper, spatial and channel attention are applied separately to decrease model parameters. Spatial attention module generates spatial attention masks fusing visual features and position representations, and channel attention module generates masks fusing language and position representations. The reasons for different pieces of information fusion are mainly two: 1) it can decrease model parameters; 2) language information is not suitable for generating spatial attention masks comparing to visual features and position information.

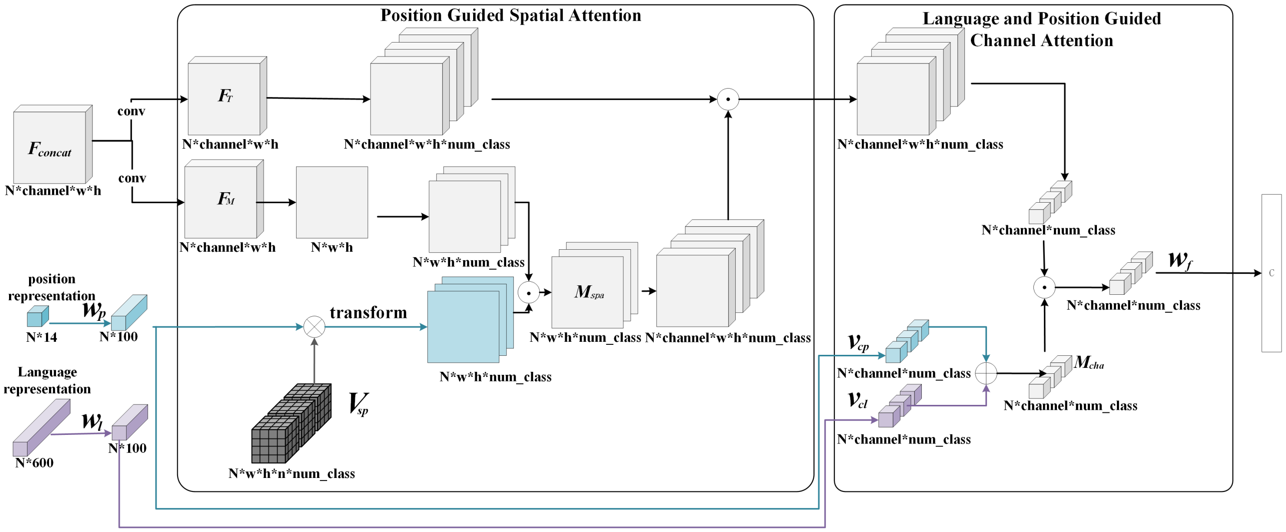

Fig. 2. The detail structure of LPGA module, left box is the spatial attention, right box is the channel attention. $\otimes$ denotes transform operation, $\odot$ and $\oplus$ denote the element-wise product and sum.

2.3. Position Guided Spatial Attention

Spatial attention module directs the attention in each pixel. Inspired by [6], our spatial attention module is split into two branches. Shown in the left box of Fig.2, one branch is a trunk of convolutional layers to extract visual features, and another one is a bottom-up top-down convolutional structure to generate spatial soft masks. Different from previous methods [6] only using visual features, the spatial mask for i^{th} $predicates$ classifier in our paper is the combination of visual features and position representations as below:

$$M^i_{spa}(sub, ob) = \left(\boldsymbol{V}^i_{sp} R_p\right) \odot \Sigma_{channel} F_M, \qquad (3)$$

where $\boldsymbol{V}^i_{sp} \in \mathbb{R}^{w \times h \times n}$ denotes the position projection matrix transforming the position representations vector to 2-D attention mask for the i^{th} $predicate$ classifier. $F_M \in \mathbb{R}^{channel \times w \times h}$ is the visual spatial mask generated by visual features. $\odot$ is the element-wise multiplication. Finally, normalization and duplication operations are applied to generate the final spatial mask $\widetilde{M^i_{spa}} \in \mathbb{R}^{channel \times w \times h}$.

2.4. Language and Position Guided Channel Attention

Channel attention module directs the attention in each channel. In our paper, the channel mask is generated using position and language representations. Similar with spatial attention module, channel attention module is trained with two matrices, $\boldsymbol{v}^i_{cl} \in \mathbb{R}^{channel \times m}$, $\boldsymbol{v}^i_{cp} \in \mathbb{R}^{channel \times n}$ denoting the language projection matrix and position projection matrix in channel attention for the i^{th} classifier. First, we combine language representations and position representations:

$$M^i_{cha}(sub, ob) = \boldsymbol{v}^i_{cl} R_l\left(sub, ob\right) + \boldsymbol{v}^i_{cp} R_p\left(sub, ob\right). \qquad (4)$$

Then normalization operation are applied on $M^i_{cha} \in \mathbb{R}^{channel}$ to generate the final channel mask $\widetilde{M^i_{cha}} \in \mathbb{R}^{channel}$ for the i^{th} $predicate$.

Finally, the output of the LPGA module for the i^{th} classifier is as below:

$$C^i_{pred} = \boldsymbol{w}^i_f \left[\widetilde{M^i_{cha}} \odot \Sigma_{w,h}\left(\widetilde{M^i_{spa}} \odot F_T\right)\right] + b^i_f, \qquad (5)$$

where $\boldsymbol{w}^i_f \in \mathbb{R}^{1 \times channel}$ and $b^i_f \in \mathbb{R}$ are the weights and bias term to produce final prediction C^i_{pred}, F_T are the feature maps generated by F_{concat} with two convolutional layers.

The final loss function is shown below:

$$L = -\frac{1}{N}\sum_{k=1}^{N} \Delta\left(y^k_{pred}, C_{pred}(x_k)\right), \qquad (6)$$

where Δ is the softmax loss, y^k is the label of the k^{th} input, and N is the number of the mini-batch.

3. EXPERIMENTS

During the training stage, the learning rates are set to 0.01 in the rear relation layers, which are decayed by 0.3 every 15 epochs. To accelerate the convergence of training, Adam [17] is applied to optimize our relation model.

Because annotations of visual relationship are not exhaustive, mAP evaluation metrics will penalize positive predictions which are absent in ground truth. We follow [2] to use Recall@50 (R@50) and Recall@100 (R@100) as our evaluation metrics. R@n computes the Recall using the top n predictions in one image. Following [10], we also set a hyper-parameter k, which means to take the top k predictions into

Table 1. Evaluation on VRD testing set. ”Entire set” contains the whole testing set. ”Zero-shot set” only contains triplets which are not in the training set. ”spatial attention” / ”channel attention” only contains spatial / channel attention module. ”LR”/ ”PR” only uses language / position representations.

Model	Entire set R@100/50 k=1	 R@100 k=70	 R@50 k=70	Zero-shot set R@100/50 k=1	 R@100 k=70	 R@50 k=70
Visual Phr [18]	1.91	-	-	-	-	-
Joint CNN [13]	2.03	-	-	-	-	-
VTransE [11]	44.76	-	-	-	-	-
Language-Pri [2]	47.87	84.34	70.97	8.45	50.04	29.77
TCIR [16]	53.59	-	-	16.42	-	-
Weakly-sup [19]	52.6	-	-	23.6	-	-
DR-Net [12]	-	81.90	80.78	-	-	-
LKD [10]	55.16	94.65	85.64	16.98	74.65	54.20
Zoom-Net [20]	55.98	94.56	89.03	-	-	-
baseline	18.13	78.06	58.63	7.44	62.45	39.09
spatial attention	42.54	90.39	80.30	19.16	82.98	65.27
channel attention	55.70	96.41	**90.65**	22.33	86.57	71.26
LR	55.64	96.40	89.80	22.16	85.03	68.69
PR	45.26	92.34	82.95	23.61	83.75	69.12
Final Model	**56.60**	**96.66**	90.39	**26.52**	**86.66**	**72.63**

consideration per object pair. In visual relationship recognition, R@n,k=1 is equivalent to R@n in [2]. R@n,k=70 in VRD and R@n,k=130 in VGR are equivalent to take all *predicates* into consideration.

3.1. Experiments on Visual Relationship Dataset

In this section, we evaluate our model in Visual Relationship Dataset [2], which contains 70 predicates and 100 objects. We compare our model with some related methods [2, 18, 13, 11, 16, 10, 19, 12, 20]. Table 1 shows the results. To investigate different settings' influences on our proposed model, we list the performances of combinations of different components.

We explore the respective influence of language and position information in the LPGA module. ”Final Model” outperforms single ”LR” or ”PR”. It proves that language and position information are complementary in the LPGA module. ”PR” performs better in the zero-shot set than in the entire set, which indicates position information provides more powerful and explicit inferring in unseen relationships. We also add more ablation analysis to our attention module. With a single spatial or channel attention module, the model makes great gains compared to the baseline model. While channel attention performs relatively well, the final model still gets more gains combining spatial attention, especially in the zero-shot set. Further, while ”LKD” also combines language and position information with the external knowledge, our proposed model outperforms ”LKD” in all evaluation metrics, which proves language and position information are better exploited as attention weights in our attention module. Finally, our proposed model with the complete LPGA module (”Final Model”) reaches best results, especially in the zero-shot set. The visualization of some results on VRD can be seen in Fig.3.

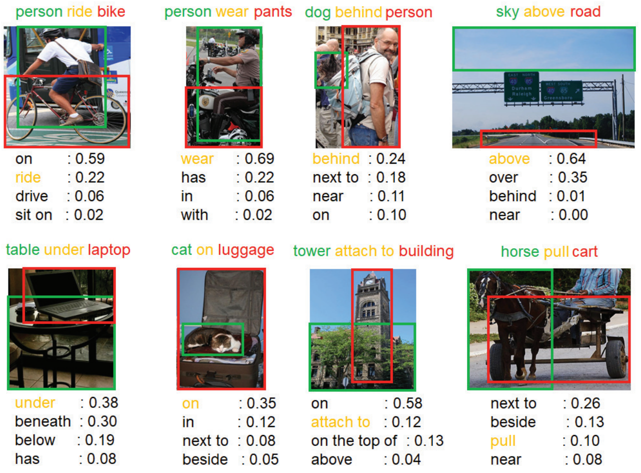

Fig. 3. Visualization of some results on VRD. Images are union regions of object pairs. The top rows above images are ground truths. Green boxes correspond to *subjects*, and red boxes correspond to *objects*. We list the top four predictions per object pair, where yellow indicates the positive prediction.

3.2. Experiments on Visual Genome Relationship Dataset

We also evaluate our model in Visual Genome-based Relationship (VGR). Visual Genome is a large-scale dataset and is annotated with much noise. We construct a clean subset VGR, which contains 130 predicates and 200 objects. There are 75697 images in the training set, and 32552 images in the testing set. The results are shown in Table 2.

Table 2. Evaluation on Visual Genome Relationship Dataset.

Model	Entire set R@100 k=1	 R@50 k=1	 R@100 k=130	 R@50 k=130	Zero-shot set R@100 k=1	 R@50 k=1	 R@100 k=130	 R@50 k=130
baseline	38.24	38.10	86.02	74.26	13.08	13.07	50.10	35.55
LR	74.58	74.34	**96.21**	92.27	15.76	15.74	69.38	51.89
PR	59.45	59.24	91.50	84.97	17.18	17.18	61.02	46.97
Final Model	**75.00**	**74.76**	96.13	**92.34**	**18.52**	**18.51**	**70.69**	**55.48**

From Table 2, we can see the final model achieves better results comparing to ”LR” and ”PR”. It proves that the fusion of language and position information as attention weights in LPGA module can boost the performances of relation model, which is consistent with the results in VRD.

4. CONCLUSIONS

In this work, we propose a novel visual relationship recognition model using language and position guided attention: language and position information are exploited to guide the generation of more accurate attention maps, and thus the selection efficiency of spatial and channel features can be increased. Experiments on VRD and VGR show that, with language and position guided attention module, our proposed model achieves state-of-the-art performance.

5. REFERENCES

[1] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al., "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *International Journal of Computer Vision*, vol. 123, no. 1, pp. 32–73, 2017.

[2] 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.

[3] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *International conference on machine learning*, 2015, pp. 2048–2057.

[4] Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei, "Visual7w: Grounded question answering in images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4995–5004.

[5] Huijuan Xu and Kate Saenko, "Ask, attend and answer: Exploring question-guided spatial attention for visual question answering," in *European Conference on Computer Vision*. Springer, 2016, pp. 451–466.

[6] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang, "Residual attention network for image classification," *arXiv preprint arXiv:1704.06904*, 2017.

[7] Long Chen, Hanwang Zhang, Jun Xiao, Liqiang Nie, Jian Shao, Wei Liu, and Tat-Seng Chua, "Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning," *arXiv preprint arXiv:1611.05594*, 2016.

[8] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al., "Spatial transformer networks," in *Advances in neural information processing systems*, 2015, pp. 2017–2025.

[9] Liang-Chieh Chen, Yi Yang, Jiang Wang, Wei Xu, and Alan L Yuille, "Attention to scale: Scale-aware semantic image segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3640–3649.

[10] 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.

[11] Hanwang Zhang, Zawlin Kyaw, Shih-Fu Chang, and Tat-Seng Chua, "Visual translation embedding network for visual relation detection," in *CVPR*, 2017, vol. 1, p. 5.

[12] Bo Dai, Yuqi Zhang, and Dahua Lin, "Detecting visual relationships with deep relational networks," in *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*. IEEE, 2017, pp. 3298–3308.

[13] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[14] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick, "Mask r-cnn," in *Computer Vision (ICCV), 2017 IEEE International Conference on*. IEEE, 2017, pp. 2980–2988.

[15] R Yin, Q Wang, R Li, P Li, and B Wang, "Distributed representations of words and phrases and their compositionality," EMNLP, 2016.

[16] 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.

[17] DP Kingma and Ba J Adam, "A method for stochastic optimization. cornell university library," *arXiv preprint arXiv:1412.6980*, 2017.

[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] Julia Peyre, Ivan Laptev, Cordelia Schmid, and Josef Sivic, "Weakly-supervised learning of visual relations," in *ICCV 2017-International Conference on Computer Vision 2017*, 2017.

[20] 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*, 2018.