Visual Relationship Recognition via Language and Position Guided Attention

Hao Zhou, Chuanping Hu, Chongyang Zhang, Shengyang Shen

School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, China

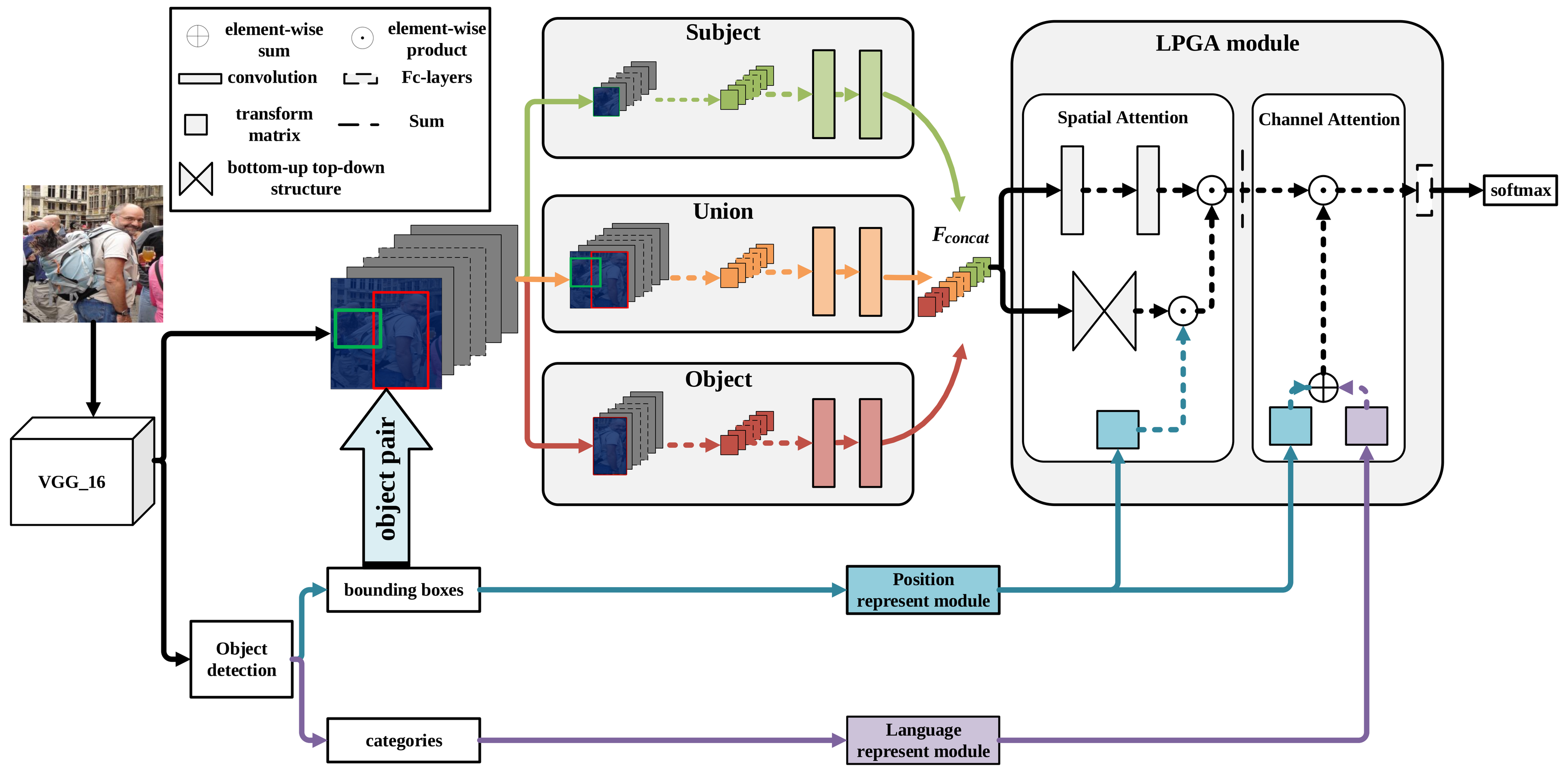
INTRODUCTION

- Visual relationship recognition focuses on distinguishing the interactions between object pairs.
- In this work, we propose a novel visual relationship recognition model using language and position guided attention(LPGA): 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.

Contributions

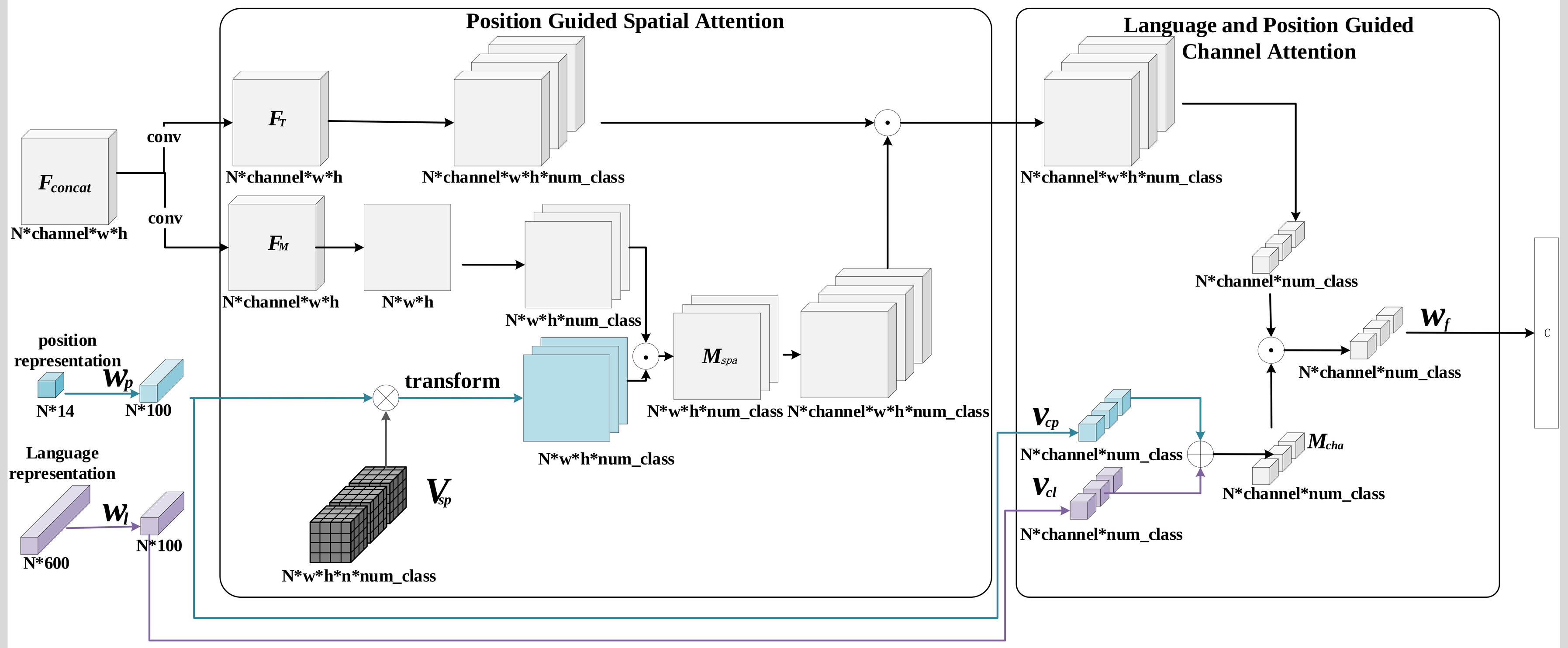
- We propose a novel LPGA module, where language and position information are exploited to guide the generation of more efficient attention maps.
- With guided attention, hidden human knowledge can be made better use to enhance the selection of spatial and channel features.
- With the LPGA module, our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

FRAMEWORK



- Firstly, given one image, we use a pre-trained object detection model to get the object regions, categories and their bounding boxes.
- Then, the visual features extracted from backbone are fed into three branches. One branch is fed into the union region of object-pairs; two branches are fed into two object areas respectively.
- Finally, concatenation operation is applied to get Fconcat which are fed into the latter LPGA module.

LPGA MODULE



- Language representations:

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

- Position representations:

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

$P(p_{sub}, p_{ob}) \in \mathbb{R}^{14}$ is position encoding vector

- Spatial attention:

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

- Channel attention:

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

- Outputs:

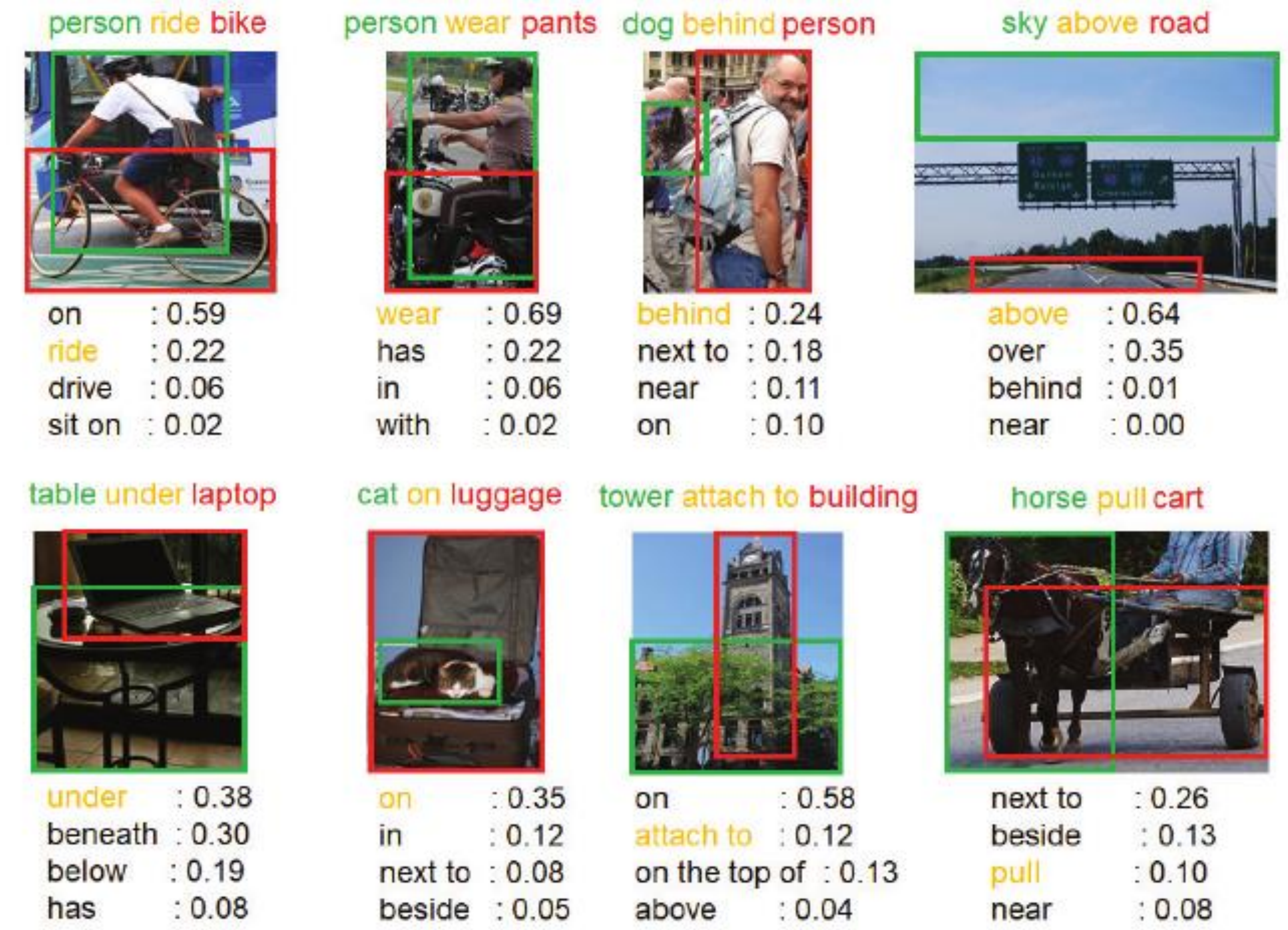
$$C^i_{pred} = \boldsymbol{w}^i_f \left[\widetilde{M}^i_{cha} \odot \Sigma_{w,h}\left(\widetilde{M}^i_{spa} \odot F_T\right)\right] + b^i_f$$

QUANTITATIVE RESULTS

Evaluation on VRD testing set.

Model	Entire set R@100/50 k=1	Entire set R@100 k=70	Entire set R@50 k=70	Zero-shot set R@100/50 k=1	Zero-shot set R@100 k=70	Zero-shot set 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**

QUALITATIVE RESULTS



DISCUSSION

- 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.
- Language and position information are better exploited as attention weights in our LPGA module.

REFERENCES

- 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,” IJCV, 2017.
- Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei, “Visual relationship detection with language priors,” in ECCV, 2016.
- Ruichi Yu, Ang Li, Vlad I Morariu, and Larry S Davis, “Visual relationship detection with internal and external linguistic knowledge distillation,” in ICCV, 2017.
- Bohan Zhuang, Lingqiao Liu, Chunhua Shen, and Ian Reid, “Towards context-aware interaction recognition for visual relationship detection,” in ICCV, 2017.

Corresponding Author: Chongyang Zhang(sunny_zhang@sjtu.edu.cn). http://ivlab.sjtu.edu.cn/