ViP-CNN: Visual Phrase Guided Convolutional Neural Network

Yikang Li[1], Wanli Ouyang[2], Xiaogang Wang[1], Xiaoou Tang[3,1]

[1]The Chinese University of Hong Kong, Hong Kong SAR, China

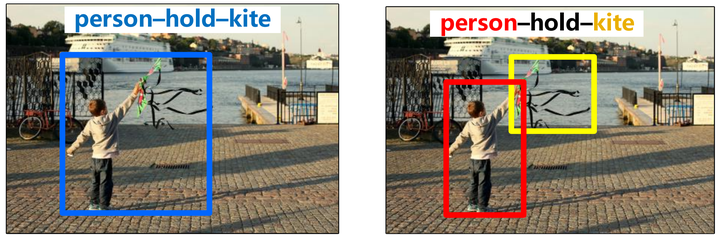
[2]University of Sydney, Australia [3]Shenzhen Institutes of Advanced Technology, China

香港中文大學 The Chinese University of Hong Kong

IEEE 2017 Conference on Computer Vision and Pattern Recognition

CVPR July 21-26 HONOLULU 2017

Motivation



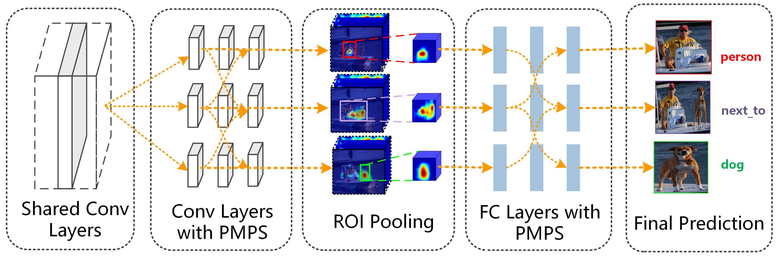
- Compared to visual relationships (right), visual phrases (left) considers the triplet as an integrated whole, which has less visual variances.
- The subject, predicate, and object are highly correlated at the feature level and label level.
  - Previous work only utilizes the language-level correlation which mainly captures the co-occurrence information.
  - Previous work divides the relationship detection process into two parts: first object detection and then predicate recognition. So predicate recognition cannot help the object classification.
- Previous work is done on a small dataset, including 4000 training images and 1000 testing ones.

Contribution

A Visual Phrase guided CNN (ViP-CNN) is proposed to detect visual relationships in an end-to-end manner.

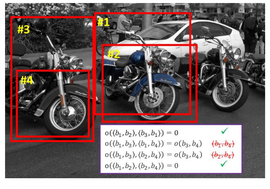
- A message passing structure is used to leverage the feature level interdependencies of subject, predicate and object for better recognition.
- A triplet NMS is proposed to reduce the redundant object pairs for efficiency.
- we investigate two ways of utilizing Visual Genome Relationship dataset to pretrain ViP-CNN, which both outperform ImageNet pretraining.

Framework



ViP-CNN first generates triplet proposals with RPN and then feeds them into corresponding models. A triplet NMS is proposed to reduce the number of triplets. ROI-pooling is used to generated the features of the fixed size. The three branches are connected at both conv and fc layers using Phrase-guided Message Passing Structure (PMPS). So the three models can make the decision with the sense of the each other. And the final recognition of subject, object and predicate is done simultaneously.

Triplet NMS



Triplet IOU $< t_1, t_2 >$:

$$o(b_{s,1}, b_{s,2}) \cdot o(b_{o,1}, b_{o,2})$$

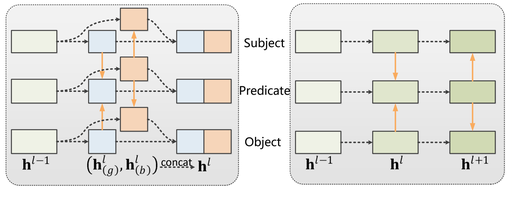
Objectiveness $< t_i >$:

$$s_i = s_{s,i} \cdot s_{o,i}$$

With well-defined IOU and objectiveness function, **Greedy NMS** can be used to reduce the redundant triplet proposals.

250 object proposals → grouping → 62,500 raw triplet proposals → triplet NMS → ~1600 after-NMS triplet proposals

Two kinds of PMPS



Phrase-guided Message Passing Structure (PMPS) is proposed to employ the complementary effects of simultaneously recognizing the objects and predicates.

Subject-predicate-object triplet can be viewed as a simple graphical model, so we propose the gather-broadcast message passing flow reflects the different statuses of subject/object and predicate. They are defined as below:

Gather Flow:

$$\boldsymbol{h}_p^l = f\left(\boldsymbol{W}_p^l \otimes \boldsymbol{h}_p^{l-1} + \boldsymbol{W}_{p\leftarrow s}^l \otimes \boldsymbol{h}_s^l + \boldsymbol{W}_{p\leftarrow o}^l \otimes \boldsymbol{h}_o^l + \boldsymbol{b}_p^l\right).$$

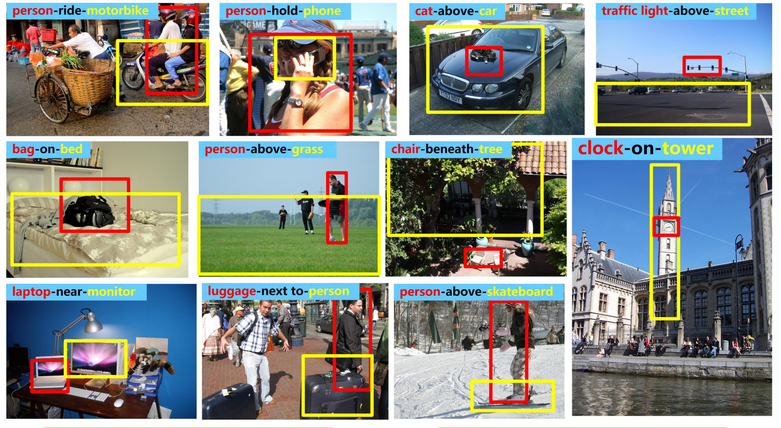
Broadcast Flow:

$$\boldsymbol{h}_s^l = f\left(\boldsymbol{W}_s^l \otimes \boldsymbol{h}_s^{l-1} + \boldsymbol{W}_{s\leftarrow p}^l \otimes \boldsymbol{h}_p^l + \boldsymbol{b}_s^l\right).$$

$$\boldsymbol{h}_o^l = f\left(\boldsymbol{W}_o^l \otimes \boldsymbol{h}_o^{l-1} + \boldsymbol{W}_{o\leftarrow p}^l \otimes \boldsymbol{h}_p^l + \boldsymbol{b}_o^l\right).$$

The gather-and-broadcast flow has parallel(left) and sequential(right) implementations.

Experiments



Quantitative Results

Model	Phrase Det. Rec@50	Phrase Det. Rec@100	Relationship Det. Rec@50	Relationship Det. Rec@100
Visual Phrases [41]	0.04	0.07	-	-
Language Prior [33]	16.17	17.03	13.86	14.70
Baseline	13.69	16.41	10.31	12.75
Baseline-Concat	15.71	18.90	11.98	14.33
RNN	14.08	18.01	11.08	13.25
ViP-No NMS	10.68	16.28	9.01	11.87
ViP-Rand. Select	17.71	23.66	13.96	17.04
ViP	21.24	26.07	16.57	19.08
ViP-Post NMS	22.31	27.24	16.95	19.81
ViP-Param Sharing	**22.78**	**27.91**	**17.32**	**20.01**

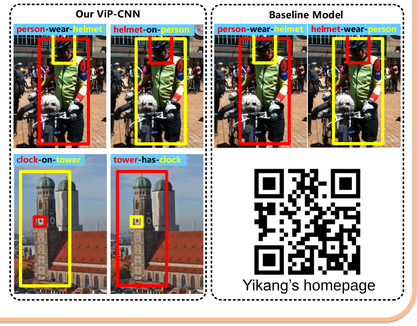
Dataset statistics

Dataset	#Images	#Rel	#Obj	#Pred
VR [33]	5,000	37,993	100	70
VGR [27]	108,077	1,531,448	-	-
ours-VGR-Dense	102,955	1,066,628	3407	142
ours-VGR-Frequent	101,649	900,739	507	92

Visual Genome pretraining

Pretrain Dataset	Target	Phrase Det. Rec@50	Phrase Det. Rec@100	Relationship Det. Rec@50	Relationship Det. Rec@100
baseline	-	22.78	27.91	17.32	20.01
Dense	vec	23.34	29.56	18.42	21.37
	class	23.98	30.01	19.01	21.96
Frequent	vec	23.29	29.61	18.39	21.32
	class	**24.21**	**30.51**	**19.44**	**22.28**

Swap experiments



Object detection

model	Faster R-CNN [9]	Ours-baseline	ViP-CNN
mean AP (%)	14.35	14.28	**20.56**

Yikang's homepage