ViP-CNN: Visual Phrase Guided Convolutional Neural Network

Yikang Li[1], Wanli Ouyang[1,2], Xiaogang Wang[1], Xiao'ou 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

`ykli, wlouyang, xgwang@ee.cuhk.edu.hk  xtang@ie.cuhk.edu.hk`

Abstract

As the intermediate level task connecting image captioning and object detection, visual relationship detection started to catch researchers' attention because of its descriptive power and clear structure. It detects the objects and captures their pair-wise interactions with a subject-predicate-object triplet, e.g. $\langle person\text{-}ride\text{-}horse \rangle$. *In this paper, each visual relationship is considered as a phrase with three components. We formulate the visual relationship detection as three inter-connected recognition problems and propose a Visual Phrase guided Convolutional Neural Network (ViP-CNN) to address them simultaneously. In ViP-CNN, we present a Phrase-guided Message Passing Structure (PMPS) to establish the connection among relationship components and help the model consider the three problems jointly. Corresponding non-maximum suppression method and model training strategy are also proposed. Experimental results show that our ViP-CNN outperforms the state-of-art method both in speed and accuracy. We further pretrain ViP-CNN on our cleansed Visual Genome Relationship dataset, which is found to perform better than the pretraining on the ImageNet for this task.*

1. Introduction

Booted by the development of Deep Learning, letting the computer understand an image seems to be increasingly closer. With the research on object detection gradually becoming mature [37, 36, 32, 24, 22, 23], increasingly more researchers put their attention on higher-level understanding of the scene [21, 51, 2, 48, 49, 46, 9, 6, 7, 47]. As an intermediate level task connecting the image caption and object detection, visual relationship/phrase detection is gaining more attention in scene understanding [33, 41, 3].

Visual phrase detection is the task of localizing a $\langle subject\text{-}predicate\text{-}object \rangle$ phrase, where predicate describes the relationship between the subject and object. *Visual relationship* detection involves detecting and localizing pairs of interacting objects in an image and also classifying the predicate or interaction between them. For the example

Figure 1. Visual phrase (left) and visual relationship (right).

of $\langle person\text{-}hold\text{-}kite \rangle$ in Figure 1, visual relationship detection aims at locating the person, the kite and classifying the pair-wise relation 'hold', while visual phrase concentrates on describing the region as an integrated whole.

Compared with localized objects, visual relationships have more expressive power. It can help to build up the connection between objects within the image, which opens a new dimension for image understanding. In addition, unlike captions, the fixed structure of visual relationships makes it possible to explicitly design the architecture of the model for better use of the domain knowledge. Therefore, we can get a richer semantic understanding of the image through detecting visual relationships.

For visual relationship detection, there are two commonly used pipelines, the bottom-up and the top-down. The bottom-up design first detects objects and then recognizes the possible interactions among them, which is adopted by the state-of-art method [33]. The top-down design detects the $\langle subject\text{-}predicate\text{-}object \rangle$ phrase simultaneously by regarding the relationship as an integrated whole. In comparison, the sequential order of the bottom-up implementation cuts off the feature-level connection between the two steps, which is important for correctly recognizing the relationship. Thus, we follow the top-down design by viewing the visual relationship as a phrase and solve it as three closely-connected recognition problems. Because of the joint training for the subject, predicate and object, our model is capable of learning specific visual patterns for the interaction and taking the visual interdependency into consideration.

Another motivation in our model is that the predictions of different phrase components are dependent on each other

at the visual feature level. This corresponds to learning special features for the pair-wise relationships. For example, visual connection of the subject (person) appearing sitting on something and an object (sofa) with the appearance of human's legs on it help to enhance the evidence of the predicate "sit on". In return, the specific visual features for "sit on" also help to inference the subject (person) and object (sofa) as well. They collectively help to detect the relationship, $\langle person\text{-}sit\ on\text{-}sofa \rangle$. Therefore, how to use utilize such dependencies would be the core problem of visual relationship detections.

Based on the analysis above, we propose Phrase-guided Message Passing Strucvture (PMPS) to model the connection among relationship components. It can be implemented with convolutional layers or fully-connected layers. Deep networks typically pass information feedforward. In contrast, PMPS extracts useful information from other components in the phrase to refine the features before going to the next level. With the sense of other components introduced by PMPS, our model can predict the subject, object and predicate simultaneously as an integrated whole. Besides, unlike simply widening the network, PMPS makes good use of the domain knowledge of the problem and the message passing strategies are designed according to the specific structure of phrases. Such explicit design reduces the parameter number and makes the training easier.

Our main contributions of our work are summarized as three-fold:

First, we propose a phrase-guided visual relationship detection framework, which can detect the relationship in one step. Corresponding non-maximum suppression method and training strategy are also proposed to improve the speed and accuracy of our model.

Second, we propose Phrase-guided Message Passing Structure (PMPS) and corresponding training strategy to leverage the interdependency of three models for more accurage recognition. In PMPS, a new gather-broadcast message passing flow mechanism is proposed, which can be applied to various layers across deep models.

Third, we investigate two ways of utilizing Visual Genome Relationship dataset [27] to pretrain ViP-CNN, both of which improve the performance of our model when compared with the pretraining on ImageNet.

On the benchmark dataset [33], our approach is, respectively, 13.48% and 6.88% higher for visual phrase detection and visual relationship detection task compared with the state-of-art method.

2. Related Work

As the intermediate level task connecting image captioning and object detection, visual relationship detection is rooted in object detection, but shares many properties with image captioning. Also, our proposed model also involves the message passing structures. We review related works on these topics.

Object Detection: As the foundation of image understanding, object detection has been investigated for years. Convolutional Neural Networks [28] were first introduced by the R-CNN [14] for object detection. It processes the regions of interest independently, which is time-consuming. Then SPP-net [17], Fast R-CNN [13] were proposed to share convolutional layers among regions in classification. Ren, *et al.* proposed Faster R-CNN by utilizing CNN to do region proposal [37]. YOLO [36] and SSD [32] shared more convolutional layers for region proposal and region classification and made detection even faster. ViP-CNN is based on Faster R-CNN due to its superior performance. However, it is not trivial to arrange the subject, predicate and object into an end-to-end framework. We propose PMPS and corresponding training scheme that uses the entire phrase to guide learning. The entire design is based on careful analysis of the specific problem.

Image Caption: Describing image with natural language have been explored for many years [4, 12, 19, 29, 31, 44]. Recently, using the visual features from CNN, Recurrent Neural Networks (RNNs) [45, 18] have been adopted to generate captions because of its success on processing natural language. Combining the RNN and CNN becomes a standard pipeline on solving the Image Captioning problems [46, 21, 5, 10, 11, 25]. However, the pipeline does not fit for visual relationship detection due to the difference between the sentence for image captioning and phrase for visual relationship detection. Compared to sentences, the phrase has fixed structure of subject-predicate-object. In addition, most of the related works focus on the whole image or image region, while the visual relationship detection targets on the region and its subregions.

Visual Relationship Detection: Visual relationships are not a new concept. Sadeghi, *et al.* has proved the phrase, as a whole, can facilitate object recognition because of its special visual appearance [41]. Desai, *et al.* used the phrase that describes the interaction between a person and objects to facilitate actions, pose and object detection [8]. However, most of the existing works were done for leveraging the relationship for other tasks [16, 30, 40]. Rohrbach, *et al.* also investigated grounding the free-form phrases by reconstruction a given phrase using the attention mechanism [38]. Lu, *et al.* first formalized the visual relationship detection as a task and proposed the state-of-art method by leveraging the language prior to model the correlation between subject/object and predicate [41]. However, these methods do not consider the visual feature level connection among the subject, predicate and object, which is used in our approach to improve the detection accuracy.

Message Passing Structure: Deformable parts models are used for modeling the relationships among objects and their parts [15, 35, 8]. Conditional Random Field (CRF) is also proved a powerful tool for dependency learn-
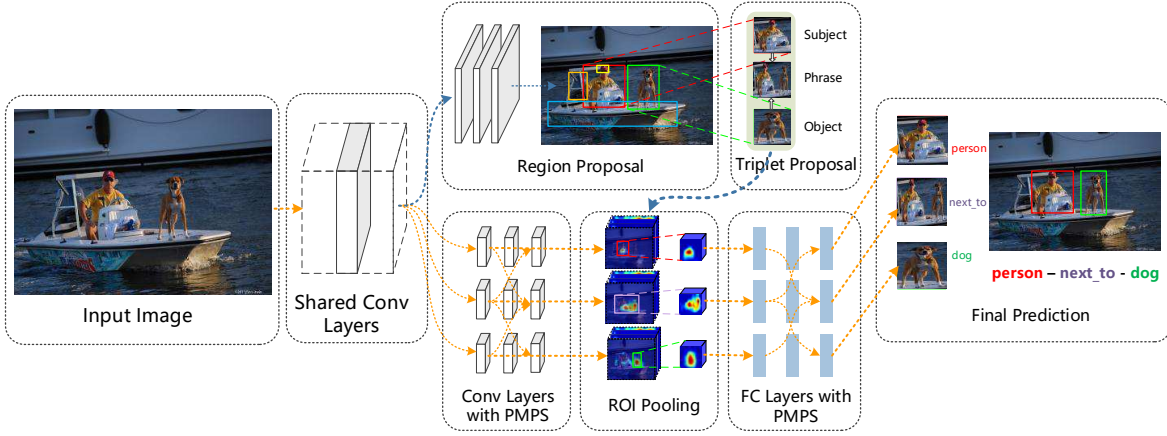
Figure 2. Overview of ViP-CNN. It generates triplet proposals with RPN and then feeds them into corresponding models. The three branches are interconnected at both conv layers and fc layers using our proposed Phrase-guided Message Passing Structure (PMPS). Results from three branches make the final prediction simultaneously.

ing [50, 26]. However, these works mainly focus on the correlations among predicted labels or object parts. Recurrent Neural Network (RNN) passes the information at feature level [45, 18] in sequential order. The fixed flow path and sharing of model parameters in RNN do not fit for the structural information in the phrase detection. Therefore, we propose a phrase-guided message passing structure with specific message passing flow to model the relationship among the visual features in the subject, predicate and object.

3. ViP-CNN

3.1. Overview

An overview of our proposed model is shown in Figure 2. VGG-Net [42] is used as the basic building block for our ViP-CNN. Our model divides the entire procedure into two parts: triplet proposal and phrase recognition. Inspired by Faster R-CNN [37], triplet proposal and phrase recognition can share most of the convolutional layers to make the inference much faster.

ViP-CNN takes an entire image as input, whose shorter side is scaled to 400 with aspect ratio kept. The image is fed into several convolutional (conv) and max-pooling layers to produce the feature map, which corresponds to Conv1_1 to Conv4_3 in the VGG-Net. Then the network is split into four branches. One for triplet proposal and three for phrase detection.

Triplet proposal branch. Taking the output of the Conv4_3 as input, three convolutional layers are used for extracting CNN features. Then features are used for proposing class-free regions of interest (ROIs) using the approach of RPN [37]. By grouping these ROIs, triplet proposal is obtained. A triplet, denoted by $(\mathbf{b}_s, \mathbf{b}_p, \mathbf{b}_o)$, is made up of three ROIs, subject ROI ($\mathbf{b}_s$), object ROI ($\mathbf{b}_p$) and predi-

cate ROI ($\mathbf{b}_o$). The predicate ROI is the box that tightly covers both the subject and the object. Due to the sparsity of relationship annotations, triplet non-maximum suppression (triplet NMS) is proposed to reduce the redundancy. The remaining triplets are used for the phrase recognition branch. For the example in Figure 2, the triplet contains the bounding box for the subject (person), the object (dog), and their union.

Phrase detection branches. Taking the output of the Conv4_3 as input, subject, predicate, and object have their own models because they have different visual appearances and different statuses in the relationship. We use the subject branch for illustration, which applies for the other two. Three convolutional layers for the subject branch are used on the shared feature map. Then the subject proposal is fed into ROI pooling layer [13] to obtain corresponding features with fixed size. Features are processed by several subsequent fully-connected (fc) layers for category estimation and bounding box regression. The confidence of $\langle subject\text{-}predicate\text{-}object \rangle$ phrases is measured by the product of scores for subject, object and predicate. ViP-CNN is capable of detecting both visual phrases and visual relationships by adopting different bounding box regression values. Moreover, the phrase guided message passing structure is introduced to exchange the information among models, which helps to leverage the connection among the three components for better results.

3.2. Triplet Proposal with NMS

First, we use RPN [37] to generate object proposals based on the feature map. RPN is trained using both subject and object bounding boxes. Proposals are sorted in the order of objectiveness scores and top N proposals are selected. We group the proposals to construct N^2 $\langle subject \text{-} object \rangle$ pairs. The predicate ROI is the union
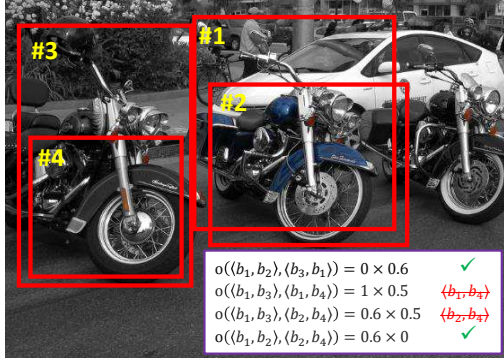
Figure 3. Illustration of our triplet NMS. Smaller index means higher objectiveness score. Some pairs are selected for illustration. $\mathbf{b}_i$ means the #i ROI, and $\langle \mathbf{b}_1, \mathbf{b}_2 \rangle$ represents the $\langle$subject, object$\rangle$ ROI pairs. Assume $o(\mathbf{b}_1, \mathbf{b}_2) = 0.6$, $o(\mathbf{b}_3, \mathbf{b}_4) = 0.5$ and the triplet NMS threshold is 0.25.

of the subject ROI and the object ROI. Thus we have N^2 $\langle subject\text{-}predicate\text{-}object \rangle$ triplet proposals.

Under this settings, 300 object proposals will generate 90,000 triplet proposals, which is a heavy load for the subsequent relationship detection step. It also deteriorates the final result measured by Recall because of the redundancy in the triplet proposals with high overlap. Therefore, we propose a triplet non-maximum suppression (NMS) method to reduce the redundant triplet proposals (figure 3). Triplet NMS is done previous to the detection procedure for the sake of inference speed.

Similar to the NMS in object proposal, triplet NMS is also based on the overlap and the objectiveness score, which is introduced below.

Denote the ROI triplet by $\mathbf{t}_i = \langle \mathbf{b}_{s,i}, \mathbf{b}_{p,i}, \mathbf{b}_{o,i} \rangle$, where $\mathbf{b}_{s,i}$, $\mathbf{b}_{p,i}$ and $\mathbf{b}_{o,i}$ are, respectively, the bounding boxes for subject, predicate and object. Denote $o(\mathbf{b}_{*,1}, \mathbf{b}_{*,2})$ as the area of intersection between $\mathbf{b}_{*,1}$ and $\mathbf{b}_{*,2}$ divided by the area of their union. Then, the triplet overlap $o(\mathbf{t}_1, \mathbf{t}_2)$ is the product of $o(\mathbf{b}_{s,1}, \mathbf{b}_{s,2})$ and $o(\mathbf{b}_{o,1}, \mathbf{b}_{o,2})$. The objectiveness score of the triplet is the product of subject objectiveness score and object objectiveness score.

With the triplet overlap and objectiveness score defined, greedy NMS [14] is done to remove redundant triplet proposals. In the experiment, we use 250 object ROIs to produce triplets and set 0.25 as the triplet NMS threshold. Under this setting, the number of triplets will be reduced from 62,500 to about 1,600, which increases the speed by more than 20 times.

3.3. Phrase-guided Message Passing Structure

For normal CNN, only visual intra-dependencies are considered, where features at level l of the subject, predicate and object as $\mathbf{h}_s^l$, $\mathbf{h}_p^l$, and $\mathbf{h}_o^l$ respectively, are obtained as below:

$$\mathbf{h}_*^l = f\left(\mathbf{W}_*^l \otimes \mathbf{h}_*^{l-1} + \mathbf{b}_*^l\right) \qquad (1)$$
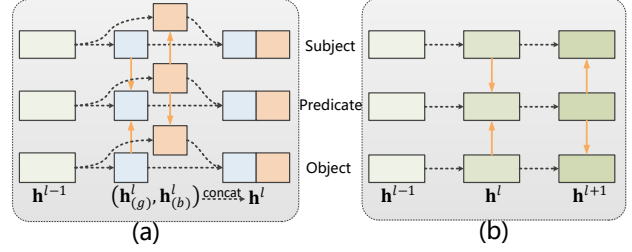

(a)   (b)

Figure 4. Two types of PMPS. (a) **Parallel**; (b) **Sequential**. Sequential implementation needs two layers while the parallel implementation only needs one. $\mathbf{h}_{(g)}^l$ denotes features for gather flow, and $\mathbf{h}_{(b)}^l$ for broadcast.

where $*$ can be $\{s, p, o\}$ and $\otimes$ denotes the matrix-vector product for fc layers and convolution for conv layers. $\mathbf{W}_*^l$ and $\mathbf{b}_*^l$ are parameters of fc or conv layers.

Under this settings, connections among subject, predicate and object are omitted. To employ the complementary information provided by other models, Phrase-guided Message Passing Structure (PMPS) is proposed PMPS extracts useful information from the source branch to refine features of the destination branch. Passing message horizontally across deep models helps them improve one another.

Subject-predicate-object triplet can be viewed as a simple graphical model. Predicate captures the general information about the phrase while subject and object focus on the details. To reflect the importance of the predicate within the triplet, we place the predicate at the dominant position and specifically design the gather-and-broadcast message passing flow. In the message passing flow, the predicate first gathers the messages from the subject and object as follows:

$$\mathbf{h}_p^l = f\left(\mathbf{W}_p^l \otimes \mathbf{h}_p^{l-1} + \mathbf{W}_{p\leftarrow s}^l \otimes \mathbf{h}_s^l + \mathbf{W}_{p\leftarrow o}^l \otimes \mathbf{h}_o^l + \mathbf{b}_p^l\right).$$
$$\text{— gather flow}$$
$$(2)$$

where $\mathbf{W}_{p\leftarrow s}^l$ and $\mathbf{W}_{p\leftarrow o}^l$ respectively denote the parameters for passing message from the subject and the object to the predicate. At the next layer, the predicate broadcasts message to the subject and the object as follows:

$$\mathbf{h}_s^{l+1} = f\left(\mathbf{W}_s^{l+1} \otimes \mathbf{h}_s^l + \mathbf{W}_{s\leftarrow p}^{l+1} \otimes \mathbf{h}_s^{l+1} + \mathbf{b}_p^l\right),$$
$$\mathbf{h}_o^{l+1} = f\left(\mathbf{W}_o^{l+1} \otimes \mathbf{h}_o^l + \mathbf{W}_{o\leftarrow p}^{l+1} \otimes \mathbf{h}_o^{l+1} + \mathbf{b}_p^l\right), \quad (3)$$
$$\text{— broadcast flow}$$

where $\mathbf{W}_{*\leftarrow p}^{l+1}$ denotes the parameter used for passing message from the predicate features.

In this PMPS, the gather flow collects the information from subject and object to refine the visual features of the predicate. Then, in the broadcast flow, the global visual information of interaction is broadcast back to the subject and object as context.
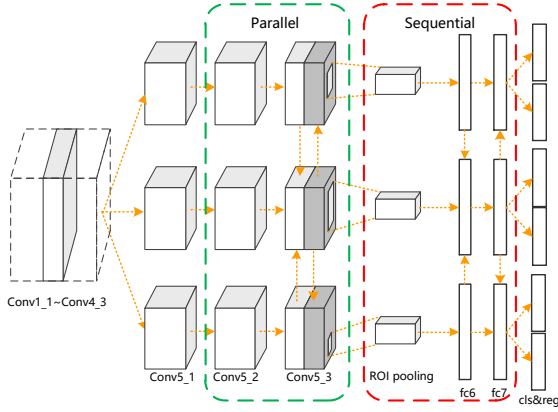
Figure 5. Our final proposed model with PMPS on VGG-Net. Parallel implementation is placed at conv5_3 layers while sequential one is placed at fc6 and fc7 layers.

The gather-and-broadcast flow has sequential and parallel implementations, as shown in Figure 4. The implementation in (2) and (3) is the sequential one, in which gathering is followed by broadcasting. In the parallel implementation, the original branch is divided into two parallel sub-branches, one for the gathering and the other for the broadcasting. The two sub-branches are concatenated after message passing. The sequential implementation is done on two adjacent layers while the parallel layer is done within the single layer.

Since different branches have different semantic meanings in the subject-predicate-object structure, their learned features are different. We can also apply message passing for the convolutional layers before ROI pooling for mutual improvement. We empirically find that the sequential implementation is better for the fc layers and the parallel implementation is better for convolutional layers.

4. Training the Model

The triplet proposal and visual relationship detection are accomplished in a single network. At the training stage, we train it in multiple steps.

4.1. Training procedure

For triplet proposal, we directly use the RPN proposed by He *et al.* in [37]. The model is initialized on ImageNet [39] pretrained VGG-16 model, and then trained using subject and object instances.

For detection, the training process has two stages.

At the first stage, we remove the message passing structure PMPS and treat three branches as three separate detectors. We initialize each branch with ImageNet pretrained VGG-16 model and train the three branches at the same time. During training, the ROI in the triplet is marked as foreground when its overlap with corresponding ground

truth is higher than a specific threshold (we use 0.5 for all three branches in the experiment), regardless of the other branches.

At the second stage, PMPS is enabled, and the model trained in stage one is used for initialization. In addition, to enforce the model considers the phrase as a whole, the subject/object/predicate ROI is regarded as foreground only if the overlaps of the subject, the object and the phrase proposals with their ground truth boxes are *all* higher than the threshold. That is to say, even if the triplet ROIs correctly localize the phrase and subject, wrong localization of the object will force the subject and the predicate to be the background. Under this training strategy, the model is forced to consider the interdependency of the entire phrase. Therefore, this training strategy provides a top-down guidance, where the top message from the entire phrase is used for guiding the learning of three branches.

To make the Conv1-Conv4 shared for triplet proposal and relationship detection, we take the four-step training like Faster R-CNN [37].

4.2. Training Loss

Each branch has two outputs. For the subject branch, the output is the probability over the $N + 1$ categories (N for targets and 1 for the background), $\mathbf{p}_s = (p_{s,0}, p_{s,1}, ..., p_{s,N})$, and the corresponding box regression offsets, $\mathbf{t}_s = (t_{s,1}, t_{s,2}, ..., t_{s,N})$. $t_{s,n} = \{t_{s,x}, t_{s,y}, t_{s,w}, t_{s,h}\}$, which denotes the scale-invariant translation and log-scale height/width shift on an given ROI [14]. Similarly, we denote the probabilities and offsets for the predicate and the object as $\mathbf{p}_p$, $\mathbf{p}_o$, $\mathbf{t}_p$, $\mathbf{t}_o$. We group the classification probabilities as $\mathbf{p} = \{\mathbf{p}_s, \mathbf{p}_p, \mathbf{p}_o\}$ and group the box regression offsets as $\mathbf{t} = \{\mathbf{t}_s, \mathbf{t}_p, \mathbf{t}_o\}$. The model employs multi-task loss for relationship detection[13]:

$$
\begin{aligned}
&L\left(\mathbf{p}, \mathbf{u}, \mathbf{t^u}, \mathbf{v}\right) \\
&= \sum_{\alpha \in \{s,p,o\}} L_{cls}(\mathbf{p}_\alpha, u_\alpha) + \lambda\left[u_\alpha \geq 1\right] L_{reg}\left(\mathbf{t}_\alpha^u, \mathbf{v}_\alpha\right),
\end{aligned}
$$
(4)

where L_{cls} is the loss for classification, implemented by the soft-max loss, and L_{reg} is the bounding box regression loss, implemented using Smooth L_1 loss [13]. $\mathbf{u} = \{u_s, u_p, u_o)$ denotes the targets for the subject, predicate and object. $\mathbf{v} = \{\mathbf{v}_s, \mathbf{v}_p, \mathbf{v}_o\}$ denotes the ground truth regression values for the three proposals. The indicator function $[u \geq 1]$ takes the value 1 for $u \geq 1$ and 0 otherwise, which means L_{reg} is ignored for the background. The hyper-parameter λ balances the weights of the two losses. $\lambda = 1$ in the experiment.

5. Experiment and Analysis

We use the pretrained VGG-16 [43] provided by Caffe [20] to initialize our model. The newly introduced

layers are randomly initialized. We set the base learning rate as 0.001 and fix the parameters from $Conv1\_1$ to $Conv2\_2$. When finetuning the model on Visual Genome pretrained model, more parameters are fixed, from $Conv1\_1$ to $Conv4\_1$. For all experiments, we select top-200 boxes from region proposal and group them into triplets. 0.25 is used as the triplet NMS threshold.

On single Titan X GPU, our model can finish the inference in about 0.6s per image. In comparison, the R-CNN-based state-of-art method [33] requires more than 10s per image.

Evaluation Criteria. We adopt the Recall [1] as the evaluation metric as in [33]. Top-N recall is denoted as Rec@N. For visual phrase detection, we aim to recognize ⟨*subject - predicate - object*⟩ and localize *one* bounding box for the entire phrase. For relationship detection, we should recognize ⟨*subject - predicate - object*⟩ and localize both subject and object. The proposal having at least 0.5 IOU with the ground-truth is regarded as correct localization.

5.1. Experiment on Visual Relationship dataset

Visual Relationship is proposed by Lu, *et al.* as a benchmark dataset for visual relationship detection. We will use the dataset to evaluate our proposed model and do the component analysis.

5.1.1 Comparison to existing approaches

We compare our model with the existing models [41, 33].

- **Visual Phrases.** 6,672 deformable parts models are trained for every relationship category in training set of Visual Relationship dataset.

- **Language Prior.** Lu, *et al.* first use R-CNN [14] to detect objects. A language model based on word vectors of the object categories and a visual model based on the CNN feature of the object pair are then trained to recognize the interactions.

- **Ours-ViP.** Full model of our proposed ViP-CNN as figure 5 with weight sharing.

Model	Phrase Det.		Relationship Det.	
	Rec@50	Rec@100	Rec@50	Rec@100
Visual Phrases [41]	0.04	0.07	-	-
Language Prior [33]	16.17	17.03	13.86	14.70
Ours-ViP	**22.78**	**27.91**	**17.32**	**20.01**

Table 1. Evaluation of different methods at Visual Relationship [33] on visual phase detection (Phrase Det.) and visual relationship detection (Relationship Det.) measured by Top-50 recall (Rec@50) and Top-100 recall (Rec@100).

Visual Phrases [41] performs poorly due to the lack of training instances. Compared with the state-of-art method

Model	Phrase Det.		Relationship Det.	
	Rec@50	Rec@100	Rec@50	Rec@100
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**

Table 2. Evaluation of different methods at Visual Relationship [33] on visual phase detection (Phrase Det.) and visual relationship detection (Relationship Det.) tasks measured by Top-50 recall (Rec@50) and Top-100 recall (Rec@100).

that uses Language Prior [33], our model increases Recall@100 by 9.04% and 4.38% on visual phrase detection and visual relationship detection task respectively. Since both models are based on VGG-Net [43], the gain is mainly from the better use of the interdependency within the phrase.

5.1.2 Component Analysis

There are many components that influence the performance of the proposed approach. Table 2 shows our investigation on the performance of different settings on the visual relationship dataset [33].

Phrase-guided Message Passing Structure. **Baseline** model disables PMPS and the three models make prediction separately. Compared to our full ViP-CNN (**ViP** in Figure 2), there is 10.66% and 6.33% Rec@100 drop on visual phrase detection and visual relationship detection respectively. **Baseline-Concat** directly concatenates the three branches to make them fully connected at feature level. Instead of manually designing the information flow, it allows the model to learn that by itself. However, it underperforms ViP by 7.17% and 4.71% Rec@100 on the two tasks, which shows that the use of domain knowledge can facilitate the model training. **RNN** model utilizes the feature-level interdependency with internal memory structure. But its weight sharing scheme and fixed flow path deteriorate the final results. In comparison, our ViP-CNN outperforms RNN by 8.06% and 5.83% Rec@100 on the two tasks respectively. Experimental results show that PMPS can help our proposed model better use the inter-connection among the phrase components for more accurate inference.

Triplet NMS. In our final model, triplet NMS is used before detection. We also investigate removing this NMS procedure and then randomly selecting 2000 triplets from 62,500 potential triplets to feed into detection network (**ViP-Rand. Select**). It can be seen that random sampling of triplets leads to 2.41% Rec@100 drop on visual phrase detection task when compared with our full ViP-CNN model

which uses the proposed triplet NMS. On the other hand, removing NMS from the pipeline performs worse than random selection, with 7.38% Rec@100 reduction for visual phrase detection (**ViP-No NMS**). Furthermore, when NMS is moved placed posterior to the detection, which is denoted by **ViP-Post NMS**, Rec@100 can increase by 1.14% and 0.74% on phrase detection and relationship detection. However, the execution time becomes more than 20 times as much as before. Thus, we adopt pre-detection NMS like Figure 2 to balance the performance and speed.

Weight Sharing in ViP-CNN. Object and subject branch share most of the properties. They detect the same set of objects, and they are of the similar status when we view the entire relationship as a graphical model. Therefore we can share the parameters for the subject and object branches except for their message passing parameters, denoted by **Ours-ViP+Param Sharing**. It can help the model to learn more general features and reduce parameter size. The experiment result shows that the parameter sharing scheme helps to improve Rec@100 of ViP-CNN by 1.84% and 0.93% for visual phrase detection and visual relationship respectively.

5.2. Experiments on Visual Genome

Newly-introduced dataset, Visual Genome [27], has several kinds of annotations, one of which is visual relationships. We denote the Visual Genome Relationship dataset as VGR. The dataset contains 108,077 images and 1,531,448 relationships.

However, we find that the annotations of VGR contain some misspellings and noisy characters (*e.g.* comma). And verbs and nouns are also in different forms. Therefore, by cleansing the Visual Genome Relationship dataset [27], we build up a new relationship dataset, denoted as VGR-Dense.

Due to the long-tail distribution of the categories in VGR-Dense, we further filter out the infrequent object and predicate categories by setting 200 as the frequency threshold for object categories and 400 for the predicate categories. Then we get another visual relationship dataset with more frequent categories, which is denoted as VGR-Frequent.

We will test our proposed model on these two datasets to further evaluate our proposed model. In addition, we will investigate different ways of pretraining methods on the two datasets for visual phrase/relationship detection task.

We randomly divide the dataset into three subsets, 70% as the training subset, 10% as the validation subset and 20% as the testing subset. Model training is done on the training subset. Additionally, the images overlapping with Visual Relationship [33] are removed from the training set.

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

Table 3. Statistics on the Visual Relationship [33] (VR), Visual Genome Relationship dataset [27] (VGR) and two cleansed dataset based on Visual Genome. The number of images (#Images), relationship triplets (#Rel), object categories (#Obj) and predicate categories (#Pred) are shown.

5.2.1 Model Evaluation on Visual Genome Relationship dataset

Since there are no existing works reported on Visual Genome dataset, we use our proposed model without PMPS as baseline model, denoted as **Baseline**, and compare our ViP-CNN with it to investigate the effect of our proposed PMPS and weight sharing strategy (**ViP** denotes our proposed ViP-CNN without parameter sharing in figure 5, **ViP-P.S.** denotes the ViP-CNN with parameter sharing).

Dataset	Model	Phrase Det.		Relationship Det.	
		Rec@50	Rec@100	Rec@50	Rec@100
VGR Dense	Basline	7.32	9.98	4.01	7.89
	ViP	13.71	16.75	8.12	11.81
	ViP-P.S.	**14.17**	**17.29**	**8.74**	**12.17**
VGR Frequent	Basline	8.17	13.89	5.71	8.94
	ViP	16.03	20.85	9.98	12.90
	ViP-P.S.	**16.58**	**21.54**	**10.67**	**13.81**

Table 4. Evaluation of different methods on Visual Genome [27] for visual phase detection (Phrase Det.) and visual relationship detection (Relationship Det.) tasks measured by Top-50 recall (Rec@50) and Top-100 recall (Rec@100).

The experiments on the two Visual-Genome-based relationship datasets both proves the effect of our proposed PMPS. For VGR-Dense, 6.77% and 3.92% Rec@100 increase on visual phrase detection and visual relationship detection come from the introduction of PMPS, while for VGR-Frequent, the increases are 6.96% and 3.96% correspondingly.

Furthermore, we evaluate our proposed parameter sharing strategy. On VGR-Dense, it brings 0.54% and 0.36% Rec@100 gains on the two tasks respectively. On VGR-Frequent, the gains are 0.68% and 0.91%. The results show that parameter sharing still works on the large-scale relationship dataset, although the gain is not as much as on the small set like Visual Relationship [33].

5.2.2 Investigation on pretraining settings

We further investigate how the training data and training targets influence the pretraining on our two datasets. The results tested on the visual relationship dataset are shown in Table 5.

For pretraining dataset, the **baseline** is pretrained on the ImageNet Classification data, the others are pretrained on our VGR-dense and VGR-frequent. The pretrained models are used for initialization and then finetuned on the Visual Relationship dataset [33]. During finetuning, the parameters from $Conv1_1$ to $Conv4_1$ are all fixed, because it is found not influencing results in [13].

The categories in the two datasets follow long-tail distribution. Especially for the VGR-Dense dataset, most of the classes have few instances for training. Therefore, we can use some structured targets to make full use of the instances of rare categories for pretraining. Because of the amazing properties of word vector [34], we can convert the label to their word vectors as the training targets to make full use of the rare categories. For word vector, we employ the Smooth L_1 loss function [13], which is denoted by *vec* in Table 5. In addition, we also use the original multi-class targets as comparison, which is denoted by *class*.

Pretrain Dataset	Target	Phrase Det.		Relationship Det.	
		Rec@50	Rec@100	Rec@50	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**

Table 5. Comparison of different pretraining methods. *Baseline* is pretrained on ImageNet. All models are finetuned and tested on the dataset in [33]. Target denotes the training targets for classification, class label (*class*) or word vector (*vec*).

From the experimental result, we can see that pretraining using our cleansed datasets, VGR-Dense and VGR-Frequent, performs better than the pretraining on ImageNet. When the class label is used as the target, pretraining using the VGR-Frequent dataset has 22.28% Rec@100 on visual relationship detection, is the best choice, with 2.27% gain when compared with the baseline. Pretraining using the VGR-frequent dataset slightly outperforms pretraining using the VGR-Dense dataset, with 0.32% Rec@100 improvement. Category labels perform better than word vector labels for pretraining, with 0.96% Rec@100 improvement on visual relationship detection when using the VGR-frequent dataset. Besides, with word vector supervision, the two datasets have similar performance.

Based on the results, too many low-frequency categories will deteriorate the pretraining gain when utilizing class label as training targets. With implicitly embedded structure, word vector is expected to be a solution to the problem. However, the result reveals that under current experiment settings, it still cannot surpass the widely used multi-class label targets. But the word vector target is still a possible way to utilize the large quantities instances of rare classes, which is worthwhile for other potential applications.

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

Table 6. **Object detection** result on the dataset in [33]. Ours-baseline has removed message passing.

5.3. Further investigation on Object Detection

To separately investigate the improvement on recognizing objects, we further compare ViP-CNN with Faster R-CNN [37] and the baseline model without PMPS on the intersection of the subject and object sets. ViP-CNN only uses the *subject* detection results without any specific tuning. Our proposed model outperforms Faster R-CNN and our baseline model by 6.21% and 6.28% respectively measured by mAP (Table 6). By comparing the three results, we can see that the gain mainly comes from the extra context information introduced by PMPS. Therefore, how to leverage the context information with explicit semantic meaning based on the visual relationship dataset could be a possible direction in the future work.

6. Conclusion

In this paper, our proposed Visual Phrase Guided Convolutional Neural Network is proved to be effective for visual relationship detection. A triplet NMS procedure is proposed to remove redundant detection results for more efficient inference. In ViP-CNN, a message passing structure called PMPS is proposed to model the visual interdependency among relationship components, which setting up a horizontal information flow among different deep models at the same layer. PMPS helps deep models for s/o/p improve one another through mining their connections. Evaluated on the Visual Relationship dataset, our model outperforms the state-of-art model in both speed and accuracy. Experimental results of the pretrained model on Visual Genome Relationship dataset are also presented. It performs better than the ImageNet pretrained model on the visual phrase/relationship detection task. In comparison, the word vector training target is not comparable with the label target under the present settings. Future work can be done in this direction for mining the structural information among the labels.

Acknowledgement

References

[1] B. Alexe, T. Deselaers, and V. Ferrari. Measuring the objectness of image windows. *TPAMI*, 2012. 6

[2] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Lawrence Zitnick, and D. Parikh. Vqa: Visual question answering. In *ICCV*, 2015. 1

[3] Y. Atzmon, J. Berant, V. Kezami, A. Globerson, and G. Chechik. Learning to generalize to new compositions in image understanding. *arXiv preprint arXiv:1608.07639*, 2016. 1

[4] K. Barnard, P. Duygulu, D. Forsyth, N. d. Freitas, D. M. Blei, and M. I. Jordan. Matching words and pictures. *JMLR*, 2003. 2

[5] X. Chen and C. L. Zitnick. Learning a recurrent visual representation for image caption generation. *arXiv preprint arXiv:1411.5654*, 2014. 2

[6] X. Chu, W. Ouyang, H. Li, and X. Wang. Structured feature learning for pose estimation. *arXiv preprint arXiv:1603.09065*, 2016. 1

[7] X. Chu, W. Ouyang, X. Wang, et al. Crf-cnn: Modeling structured information in human pose estimation. In *NIPS*, 2016. 1

[8] C. Desai and D. Ramanan. Detecting actions, poses, and objects with relational phraselets. In *ECCV*, 2012. 2

[9] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *CVPR*, 2015. 1

[10] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *CVPR*, 2015. 2

[11] H. Fang, S. Gupta, F. Iandola, R. K. Srivastava, L. Deng, P. Dollár, J. Gao, X. He, M. Mitchell, J. C. Platt, et al. From captions to visual concepts and back. In *CVPR*, 2015. 2

[12] A. Farhadi, M. Hejrati, M. A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth. Every picture tells a story: Generating sentences from images. In *ECCV*, 2010. 2

[13] R. Girshick. Fast r-cnn. In *ICCV*, 2015. 2, 3, 5, 8

[14] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014. 2, 4, 5, 6

[15] R. Girshick, F. Iandola, T. Darrell, and J. Malik. Deformable part models are convolutional neural networks. In *CVPR*, 2015. 2

[16] A. Gupta and L. S. Davis. Beyond nouns: Exploiting prepositions and comparative adjectives for learning visual classifiers. In *ECCV*, 2008. 2

[17] K. He, X. Zhang, S. Ren, and J. Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. In *CVPR*, 2014. 2

[18] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 1997. 2, 3

[19] Y. Jia, M. Salzmann, and T. Darrell. Learning cross-modality similarity for multinomial data. In *ICCV*, 2011. 2

[20] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *ACM MM*, 2014. 5

[21] J. Johnson, A. Karpathy, and L. Fei-Fei. Densecap: Fully convolutional localization networks for dense captioning. *arXiv preprint arXiv:1511.07571*, 2015. 1, 2

[22] K. Kang, H. Li, T. Xiao, W. Ouyang, J. Yan, X. Liu, and X. Wang. Object detection in videos with tubelet proposal networks. In *CVPR*, 2017. 1

[23] K. Kang, H. Li, J. Yan, X. Zeng, B. Yang, T. Xiao, C. Zhang, Z. Wang, R. Wang, X. Wang, et al. T-cnn: Tubelets with convolutional neural networks for object detection from videos. *arXiv preprint arXiv:1604.02532*, 2016. 1

[24] K. Kang, W. Ouyang, H. Li, and X. Wang. Object detection from video tubelets with convolutional neural networks. In *CVPR*, 2016. 1

[25] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, 2015. 2

[26] V. Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. *NIPS*, 2011. 3

[27] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *arXiv preprint arXiv:1602.07332*, 2016. 2, 7

[28] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, pages 1097–1105, 2012. 2

[29] G. Kulkarni, V. Premraj, V. Ordonez, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L. Berg. Babytalk: Understanding and generating simple image descriptions. *TPAMI*, 2013. 2

[30] M. P. Kumar and D. Koller. Efficiently selecting regions for scene understanding. In *CVPR*, 2010. 2

[31] P. Kuznetsova, V. Ordonez, A. C. Berg, T. L. Berg, and Y. Choi. Generalizing image captions for image-text parallel corpus. In *ACL*, 2013. 2

[32] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, and S. Reed. Ssd: Single shot multibox detector. *arXiv preprint arXiv:1512.02325*, 2015. 1, 2

[33] C. Lu, R. Krishna, M. Bernstein, and L. Fei-Fei. Visual relationship detection with language priors. In *ECCV*, 2016. 1, 2, 6, 7, 8

[34] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013. 8

[35] W. Ouyang, X. Wang, X. Zeng, S. Qiu, P. Luo, Y. Tian, H. Li, S. Yang, Z. Wang, C.-C. Loy, et al. Deepid-net: Deformable deep convolutional neural networks for object detection. In *CVPR*, 2015. 2

[36] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. *arXiv preprint arXiv:1506.02640*, 2015. 1, 2

[37] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, 2015. 1, 2, 3, 5, 8

[38] A. Rohrbach, M. Rohrbach, R. Hu, T. Darrell, and B. Schiele. Grounding of textual phrases in images by reconstruction. *arXiv preprint arXiv:1511.03745*, 2015. 2

[39] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 2015. 5

[40] B. C. Russell, W. T. Freeman, A. A. Efros, J. Sivic, and A. Zisserman. Using multiple segmentations to discover objects and their extent in image collections. In *CVPR*, 2006. 2

[41] M. A. Sadeghi and A. Farhadi. Recognition using visual phrases. In *CVPR*, 2011. 1, 2, 6

[42] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 3

[43] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 5, 6

[44] R. Socher and L. Fei-Fei. Connecting modalities: Semi-supervised segmentation and annotation of images using unaligned text corpora. In *CVPR*, 2010. 2

[45] P. J. Werbos. Generalization of backpropagation with application to a recurrent gas market model. *Neural Networks*, 1988. 2, 3

[46] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. S. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. *arXiv preprint arXiv:1502.03044*, 2015. 1, 2

[47] L. Yang, K. Tang, J. Yang, and L.-J. Li. Dense captioning with joint inference and visual context. *arXiv preprint arXiv:1611.06949*, 2016. 1

[48] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola. Stacked attention networks for image question answering. *arXiv preprint arXiv:1511.02274*, 2015. 1

[49] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo. Image captioning with semantic attention. *arXiv preprint arXiv:1603.03925*, 2016. 1

[50] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. H. Torr. Conditional random fields as recurrent neural networks. In *ICCV*, 2015. 3

[51] Y. Zhu, O. Groth, M. Bernstein, and L. Fei-Fei. Visual7w: Grounded question answering in images. *arXiv preprint arXiv:1511.03416*, 2015. 1