ViP-CNN: A Visual Phrase Reasoning Convolutional Neural Network for Visual Relationship Detection

Yikang LI  
ykli@ee.cuhk.edu.hk

Wanli Ouyang  
wlouyang@ee.cuhk.edu.hk

Xiaogang Wang  
xgwang@ee.cuhk.edu.hk

The Chinese University of Hong Kong  
Hong Kong, China

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 localizes the objects and captures their interactions with a subject-predicate-object triplet, e.g., \{person-ride-horse\}. In this paper, the visual relationship is considered as a phrase with three components. So we formulate the visual relationship detection as three inter-connected recognition problems and propose a Visual Phrase reasoning Convolutional Neural Network (ViP-CNN) to address them simultaneously. In ViP-CNN, we present a Visual Phrase Reasoning Structure (VPRS) to set up the connection among the 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 pre-train our model on our cleansed Visual Genome Relationship dataset, which is found to perform better than the pre-training 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 research on object detection gradually becoming mature [36, 35, 31, 23, 22], increasingly more researchers put their attention on higher-level understanding of the scene [21, 49, 2, 46, 45, 9, 6, 7]. As an intermediate level task connecting the image caption and object detection, visual relationship/phrase detection is gaining more attention in scene understanding [32, 40, 3].

Visual phrase detection is the task of simultaneously localizing pairs of interacting objects in an image and also classifying the predicate or interaction between them. For the example of \{person-hold-kite\} in Figure 1, visual relationship detection aims at locating the person, the kite and classifying the predicate ‘hold’, while visual phrase concentrates on detecting the region where a person holds a kite.

Compared with localized objects, visual relationships have more expressive power. It can help to build up the connection among the objects within the image. Unlike captions, the visual relationship has simple and fixed structure so that we can train different models for different parts. Therefore, we can get a richer semantic understanding of our visual world by detecting the visual relationships.

For visual relationship detection, there are two possible solutions, the bottom-up and the top-down. The bottom-up design follows the sequential order, first detecting the objects and then recognizing the possible interactions among them, which is adopted by the state-of-art method [32]. The top-down design detects the \{subject-predicate-object\} phrase simultaneously by regarding the relationship as a phrase. In comparison, the sequential order 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. As a result 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.

Under our proposed formulation, there are three kinds of
dependencies embedded in the phrase components, which are shown in Figure 2: 1) visual intra-dependency, the label depending on its visual appearance, which CNN is designed for; 2) language-level interdependency, which reflects the rationality of the combination of subject, object and predicate; 3) visual interdependency, the connection among the visual features of different phrase components. They help to solve the problem collectively.

Taking Figure 2 as an example, three CNN detectors help to generate large quantities of potential relationships. The inter-dependencies help to filter out the less possible ones. Language-level interdependency can easily rule out the weird combinations like ⟨shirt-hold-sofa⟩, because they strongly violate our common sense. However, there are still lots of the phrases that are reasonable at the language level but contradict to the visual appearance, e.g. ⟨person-beneath-sofa⟩ in Figure 2. The visual connection of the subject (person) and the object (sofa) help to enhance the evidence of the predicate “above” and suppress the predicate “beneath”. So the combination ⟨person-beneath-sofa⟩ can be ruled out by visual interdependency.

Based on the observation above, we propose the Visual Phrase Reasoning Structure (VPRS) to model the visual interdependency among relationship components. It can be implemented with convolutional layers or fully-connected layers. With VPRS, our model can predicate the subject, object and predicate simultaneously as a phase, for which we call our proposed model Visual Phrase Reasoning Convolutional Neural Network, denoted as VPR-CNN. It is faster and more accurate than previous methods.

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 weight sharing strategy are also proposed to improve the speed and accuracy of our model.

Second, we propose the visual phrase reasoning structure (VPRS) and corresponding training strategy to address the interdependency of three branches so the model can jointly consider the three problems.

Third, we investigate two ways of utilizing the dense-captioned Visual Genome dataset [26] to pretrain the model, both of which improve the performance of our model when compared with the pretraining on ImageNet.

On the benchmark dataset [32], our approach is, respectively, 13.48% and 6.88% higher for the visual phrase detection and the 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 [27] 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 [36]. YOLO [35] and SSD [31] shared more convolutional layers for region proposal and region classification and made detection even faster. Compared to object detection, visual relationship detection has more semantic and structural information.

Image Caption: Describing image with natural language have been explored for many years [4, 12, 19, 28, 30, 43]. Recently, using the visual features from CNN, Recurrent Neural Networks (RNNs) [44, 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 [45, 21, 5, 10, 11, 24]. 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...
3. ViP-CNN

3.1. Overview

An overview of our proposed model is shown in Figure 3. VGG-Net [41] is used as the basic building block for our ViP-CNN. Our model divides the whole procedure into two parts: triplet proposal and relationship detection. Inspired by Faster R-CNN [36], triplet proposal and phrase detection can share most of the convolutional layers to make the detection 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 to Conv4 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 as input, three convolutional layers are used for extracting the CNN features in this branch. Then features are used for proposing class-free regions of interest (ROIs) using the approach of RPN [36]. By grouping these ROIs, triplet proposal is obtained. A triplet, denoted by \((b_s, b_{p}, b_o)\), is made up of three ROIs, subject ROI \((b_s)\), object ROI \((b_{p})\) and predicate ROI \((b_o)\). The predicate ROI is the box that tightly covers both the subject and the object. The triplet non-maximum suppression (triplet NMS) is then applied to reduce the triplet number to about 1200. These triplets are then used for the relationship detection branch. For the example in Figure 3, 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 as input, subject, predicate, and object have their own branches because they have different visual appear-
also based on the overlap and the objectiveness score, which is introduced below.

Denote the ROI triplet by $t_i = \langle b_{s,i}, b_{p,i}, b_{o,i} \rangle$, where $b_{s,i}, b_{p,i}$ and $b_{o,i}$ are, respectively, the bounding boxes for subject, predicate and object. Denote $o(b_{s,i}, b_{s,j})$ as the area of intersection between $b_{s,i}$ and $b_{s,j}$ divided by the area of their union. Then, the triplet overlap $o(t_1, t_2)$ is the product of $o(b_{s,1}, b_{s,2})$ and $o(b_{o,1}, b_{o,2})$. 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, the usual 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. Visual Phrase Reasoning Structure

There are three branches corresponding to the relationship components, subject, predicate, and object, to learn specific visual features. To model the connection among the features, the visual phrase reasoning structure (VPRS) is proposed.

VPRS helps to transform the information from the source branch to the destination branch. Here we use the message passing from subject to the predicate as an example for illustration, which also applies to the others. Denote the feature at level $l$ of the subject, predicate, and object as $h_s^l$, $h_p^l$, and $h_o^l$ respectively. Without VPRS, $h_s^l$, $h_p^l$, and $h_o^l$ are obtained as follows:

$$
\begin{align*}
    h_s^l &= f \left( W_s^l \otimes h_{s}^{l-1} + b_s^l \right), \\
    h_p^l &= f \left( W_p^l \otimes h_{p}^{l-1} + b_p^l \right), \\
    h_o^l &= f \left( W_o^l \otimes h_{o}^{l-1} + b_o^l \right),
\end{align*}
$$

where $\otimes$ denotes the matrix-vector product for fully-connected (fc) layers and convolution for conv layers. $W_s^l$ and $b_s^l$ are parameters in the fc or conv layer. The predicate branch at layer $l$ only receives information from the predicate branch at layer $l-1$ in (1). When the message is passed from the subject to the predicate, the following formulation is used:

$$
    h_p^l = f \left( W_p^l \otimes h_{p}^{l-1} + W_{p\rightarrow s}^l \otimes h_{s}^l + b_p^l \right),
$$

where $W_{p\rightarrow s}^l$ denotes the parameters for passing message from the subject to the predicate.

Because of the fixed structure, the subject-predicate-object triplet can be viewed as a simple graphical model. The predicate covers the whole region of subject and object and is the key that connects the subject and the object. 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:

\[
\mathbf{h}_p^l = f \left( \mathbf{W}_p^l \otimes \mathbf{h}_{s}^{l-1} + \mathbf{W}_{p-e-s}^l \otimes \mathbf{h}_{s}^l + \mathbf{W}_{p-e-o}^l \otimes \mathbf{h}_{o}^l + \mathbf{b}_p^l \right),
\]

— gather flow

(3)

where \(\mathbf{W}_p^l\) and \(\mathbf{W}_{p-e-o}^l\) 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:

\[
\begin{align*}
\mathbf{h}_{s}^{l+1} &= f \left( \mathbf{W}_{s}^{l+1} \otimes \mathbf{h}_{s}^{l} + \mathbf{W}_{s+p}^{l+1} \otimes \mathbf{h}_{s}^{l+1} + \mathbf{b}_s^{l+1} \right), \\
\mathbf{h}_{o}^{l+1} &= f \left( \mathbf{W}_{o}^{l+1} \otimes \mathbf{h}_{o}^{l} + \mathbf{W}_{o+p}^{l+1} \otimes \mathbf{h}_{o}^{l+1} + \mathbf{b}_o^{l+1} \right),
\end{align*}
\]

— broadcast flow

(4)

where \(\mathbf{W}_{s+p}^{l+1}\) denotes the parameter used for passing message from the predicate features.

In this VPRS, the gather flow collects the information from subject and object to the predicate, where the detailed visual features of the subject and the object are used for adjusting 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.

The gather-and-broadcast flow has sequential and parallel implementations, as shown in Figure 5. The implementation in (3) and (4) 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. We empirically find that the sequential implementation is better for the fc layers and the parallel implementation is better for convolutional layers. The comparison experiment results will be shown in supplementary materials.

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.

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 [36]. The model is initialized on ImageNet [38] pretrained VGG-16 model, and then trained using the subject and object instances.

For detection, the training process has two stages.

At the first stage, we remove the message passing structure VPRS and treat the three branches as three separate detectors. We initialize each branch with ImageNet pretrained VGG-16 model and then trained us-

At the second stage, the VPRS is included, 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 ROI correctly localizes 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, the top message from the entire phrase is used for
guiding the learning of three branches. This training strategy provides a top-down guidance.

To make the Conv1-Conv4 shared for triplet proposal and relationship detection, we take the four-step training like Faster R-CNN [36].

4.2. Training Loss

Each branch has two output. For the subject branch, the output is the probability over the \( N + 1 \) categories (\( N \) for targets and 1 for the background), \( \mathbf{p}_s = (p_{s,0}, p_{s,1}, \ldots, p_{s,N}) \), and the corresponding box regression offsets, \( \mathbf{t}_s = (t_{s,1}, t_{s,2}, \ldots, t_{s,N}) \). \( t_{s,u} = \{t_{s,x}, t_{s,y}, t_{s,w}, t_{s,h}\} \), which denotes the scale-invariant translation and log-scale height/width shift on an given ROI [14]. Similarly, we denote the probabilities and translation and log-scale height/width shift on an given object boxes. For relationship detection, we should recognize \((subject - object - predicate)\) and localize both subject and object. The proposal having at least 0.5 overlap 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 [40, 32].

- **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 6 with weight sharing.

| Model               | Phrase Det. | Relationship Det. |
|---------------------|-------------|-------------------|
|                     | Rec@50      | Rec@100           |
|                     | Rec@50      | Rec@100           |
| Visual Phrases [40] | 0.04        | -                 |
| Language Prior [32] | 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 [32] 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 [40] performs poorly due to the lack of training instances. Compared with the state-of-art method that uses Language Prior [32], 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 [42], 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 [32].
Visual Phrase Reasoning Structure. **Baseline** model adopts our proposed one-step pipeline for visual relationship detection without VPRS. So the three branches make prediction separately. Compared to our full ViP-CNN (ViP in Figure 2), there is 7.55% and 6.33% Rec@100 drop on visual phrase detection and visual relationship detection respectively. **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. Therefore, the experimental results show that VPRS can help our proposed model better use the inter-connection among the phrase components.

**Triplet NMS.** In our final model, triplet NMS is adopted 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, the NMS placed before detection can be removed and then placed after detection, which is denoted by ViP-Post NMS. Under this settings, 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 3 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. In addition, sharing parameters can help the model to learn more general features and reduce parameter size. Therefore we can share the parameters for the subject and object branches except for their message passing parameters, denoted by Ours-ViP+Param Sharing. 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 [26], 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 the verbs and nouns are also in different forms. Therefore, by cleansing the Visual Genome Relationship dataset [26], we build up a new relationship dataset, denoted as VGR-Dense. The detailed criteria will be shown in Supplementary Material.

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 [32] are removed from the training set.

#### 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 VPRS as baseline model, denoted as **Baseline**, and compare our ViP-CNN with it to investigate the effect of our proposed

| Model                  | Phrase Det. | Relationship Det. |
|------------------------|-------------|--------------------|
|                         | Rec@50     | Rec@100            | Rec@50 | Rec@100 |
| Baseline               | 13.69      | 16.41              | 10.31  | 12.75   |
| 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 Relation-ship detection without VPRS. So the three branches make prediction separately. Compared to our full ViP-CNN (ViP in Figure 2), there is 7.55% and 6.33% Rec@100 drop on visual phrase detection and visual relationship detection respectively. **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. Therefore, the experimental results show that VPRS can help our proposed model better use the inter-connection among the phrase components.

**Triplet NMS.** In our final model, triplet NMS is adopted 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, the NMS placed before detection can be removed and then placed after detection, which is denoted by ViP-Post NMS. Under this settings, 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 3 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. In addition, sharing parameters can help the model to learn more general features and reduce parameter size. Therefore we can share the parameters for the subject and object branches except for their message passing parameters, denoted by Ours-ViP+Param Sharing. 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 [26], 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 the verbs and nouns are also in different forms. Therefore, by cleansing the Visual Genome Relationship dataset [26], we build up a new relationship dataset, denoted as VGR-Dense. The detailed criteria will be shown in Supplementary Material.

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 [32] are removed from the training set.

| Model                  | Phrase Det. | Relationship Det. |
|------------------------|-------------|--------------------|
|                         | Rec@50     | Rec@100            | Rec@50 | Rec@100 |
| Baseline               | 13.69      | 16.41              | 10.31  | 12.75   |
| 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 3. Statistics on the Visual Relationship [32] (VR), Visual Genome Relationship dataset [26] (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.
VRPS and weight sharing strategy (ViP denotes our proposed ViP-CNN without parameter sharing in figure 6, ViP-P.S. denotes the ViP-CNN with parameter sharing).

| Dataset | Model       | Phrase Det. | Relationship Det. |
|---------|-------------|-------------|-------------------|
|         |             | Rec@50      | Rec@100           |
| VGR Dense | Basline     | 7.32        | 9.98              |
|         | ViP         | 13.71       | 16.75             |
|         | ViP-P.S.    | 14.17       | 17.29             |
| VGR Frequent | Basline     | 8.17        | 13.89             |
|         | ViP         | 16.03       | 20.85             |
|         | ViP-P.S.    | 16.58       | 21.54             |

Table 4. Evaluation of different methods on Visual Genome [26] 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 visual phrase reasoning structure. For VGR-Dense, 6.77% and 3.92% Rec@100 increase on visual phrase detection and visual relationship detection come from the introduction of VRPS, 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 [32].

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 [32]. During finetuning, the parameters from Conv1 to Conv4 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 [33], we can convert the label to their word vectors as the training targets to make full use of the rare categories for better features. 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           |
| baseline         | -      | 22.78       | 27.91             |
| Dense            | vec    | 23.34       | 29.56             |
| class            |        | 23.98       | 30.01             |
| Frequent         | vec    | 23.29       | 29.61             |
| class            |        | 24.21       | 30.51             |

Table 5. Comparison of different pretraining methods. The baseline is pretrained on ImageNet. All models are finetuned and tested on the dataset in [32]. 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 constructed datasets, VGR-Dense and VGR-Frequent, performs better than the baseline that pretrained 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. The 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 labels, 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.

6. Conclusion

In this paper, our proposed Visual Phrase Reasoning Convolutional Neural Network is proved to be effective for the visual relationship detection. A triplet NMS procedure is proposed to remove redundant detection results for faster speed. In the ViP-CNN, a message passing structure called VRPS is proposed to model the visual interdependency in the visual phrase, which doubles the accuracy for visual relationship detection on the public available visual relationship dataset. 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 the current settings, the word vector training target is not comparable with the label target under the present settings.

References

[1] B. Alexe, T. Deselaers, and V. Ferrari. Measuring the objectness of image windows. TPAMI, 2012.
[2] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Lawrence Zitnick, and D. Parikh. Vqa: Visual question answering. In ICCV, 2015.
[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.
[4] K. Barnard, P. Duygulu, D. Forsyth, N. d. Freitas, D. M. Blei, and M. I. Jordan. Matching words and pictures. JMLR, 2003.
[5] X. Chen and C. L. Zitnick. Learning a recurrent visual representation for image caption generation. arXiv preprint arXiv:1411.5654, 2014.
[6] X. Chu, W. Ouyang, H. Li, and X. Wang. Structured feature learning for pose estimation. arXiv preprint arXiv:1603.09605, 2016.
[7] X. Chu, W. Ouyang, X. Wang, et al. Crf-cnn: Modeling structured information in human pose estimation. In NIPS, 2016.
[8] C. Desai and D. Ramanan. Detecting actions, poses, and objects with relational phraselets. In ECCV, 2012.
[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.
[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.
[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.
[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.
[13] R. Girshick. Fast r-cnn. In ICCV, 2015.
[14] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In CVPR, 2014.
[15] R. Girshick, F. Iandola, T. Darrell, and J. Malik. Deformable part models are convolutional neural networks. In CVPR, 2015.
[16] A. Gupta and L. S. Davis. Beyond nouns: Exploiting prepositions and comparative adjectives for learning visual classifiers. In ECCV, 2008.
[17] K. He, X. Zhang, S. Ren, and J. Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. In CVPR, 2014.
[18] S. Hochreiter and J. Schmidhuber. Long short-term memory. Neural computation, 1997.
[19] Y. Jia, M. Salzmann, and T. Darrell. Learning cross-modality similarity for multimodal data. In ICCV, 2011.
[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.
[21] J. Johnson, A. Karpathy, and L. Fei-Fei. Densecaps: Fully convolutional networks for dense captioning. arXiv preprint arXiv:1511.07571, 2015.
[22] 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.
[23] K. Kang, W. Ouyang, H. Li, and X. Wang. Object detection from video tubelets with convolutional neural networks. In CVPR, 2016.
[24] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In CVPR, 2015.
[25] V. Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. NIPS, 2011.
[26] 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.
[27] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In NIPS, pages 1097–1105, 2012.
[28] 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.
[29] M. P. Kumar and D. Koller. Efficiently selecting regions for scene understanding. In CVPR, 2010.
[30] P. Kuznetsova, V. Ordonez, A. C. Berg, T. L. Berg, and Y. Choi. Generalizing image captions for image-text parallel corpus. In ACL, 2013.
[31] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, and S. Reed. Ssd: Single shot multibox detector. arXiv preprint arXiv:1512.02325, 2015.
[32] C. Lu, R. Krishna, M. Bernstein, and L. Fei-Fei. Visual relationship detection with language priors. In ECCV, 2016.
[33] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781, 2013.
[34] S. Yang, Z. Wang, C.-C. Loy, et al. Deepid-net: Deformable deep convolutional neural networks for object detection. In CVPR, 2015.
[35] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. arXiv preprint arXiv:1506.02640, 2015.
[36] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In NIPS, 2015.
[37] 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. 3

[38] 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

[39] 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. 3

[40] M. A. Sadeghi and A. Farhadi. Recognition using visual phrases. In CVPR, 2011. 1, 3, 6

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

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

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

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

[45] 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

[46] 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

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

[48] 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

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