Deep Structured Learning for Visual Relationship Detection

Yaohui Zhu[1,2], **Shuqiang Jiang**[1,2]
[1]Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences (CAS),
Institute of Computing Technology, CAS, Beijing, 100190, China
[2]University of Chinese Academy of Sciences, Beijing, China
{yaohui.zhu}@vipl.ict.ac.cn,{sqjiang}@ict.ac.cn

Abstract

In the research area of computer vision and artificial intelligence, learning the relationships of objects is an important way to deeply understand images. Most of recent works detect visual relationship by learning objects and predicates respectively in feature level, but the dependencies between objects and predicates have not been fully considered. In this paper, we introduce deep structured learning for visual relationship detection. Specifically, we propose a deep structured model, which learns relationship by using feature-level prediction and label-level prediction to improve learning ability of only using feature-level predication. The feature-level prediction learns relationship by discriminative features, and the label-level prediction learns relationships by capturing dependencies between objects and predicates based on the learnt relationship of feature level. Additionally, we use structured SVM (SSVM) loss function as our optimization goal, and decompose this goal into the subject, predicate, and object optimizations which become more simple and more independent. Our experiments on the Visual Relationship Detection (VRD) dataset and the large-scale Visual Genome (VG) dataset validate the effectiveness of our method, which outperforms state-of-the-art methods.

Introduction

Although significant progress has been made on image recognition, including both global image classification (Szegedy et al. 2015; He et al. 2016; Szegedy et al. 2017) and local object detection (Ren et al. 2015; Dai et al. 2016), with the assistance of deep learning techniques (LeCun, Bengio, and Hinton 2015) and large scale training data (Deng et al. 2009; Xiao et al. 2010; Krishna et al. 2017), there still exists a huge gap in deep understanding of images. Recently, visual relationship detection has attracted more and more research attentions. Investigating on this problem can go one step further to understand images, and should be a potentially important topic in artificial intelligence. Furthermore, visual relationship, reflects relations between two objects such as spatial relations, action relations, is a description of the finer granularity of objects in the image. In addition, visual relationship is also useful for improving image retrieval (Lu et al. 2016), visual question answering (VQA) (Santoro et al. 2017), and object detection (Zhang et al. 2017).

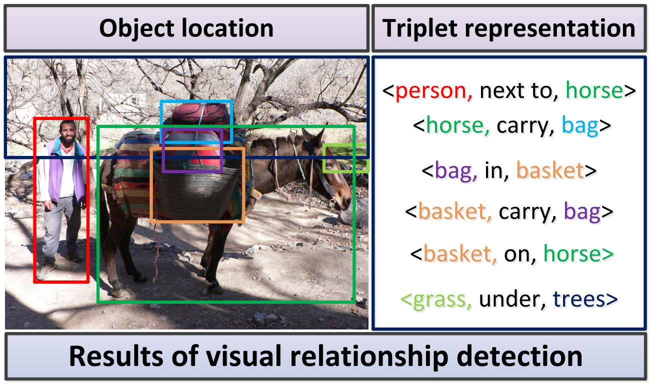

Figure 1: The results of visual relationship detection contain: 1) the location of objects and the concept of objects, 2) triplet representation of two objects.

Visual relationship detection is to detect all possible relations between objects. As shown in Figure 1, an object relationship in an image I can be represented by $R_{<sub,pre,obj>}$, where $sub = \{c_s, o_s\}, obj = \{c_o, o_o\}$, o_s, o_o are bounding boxes of sub, obj respectively, c_s, c_o belong to classification space of object, pre belongs to classification space of predicate. The goal of visual relationship detection is to extract a series of relationships $A = \{R_{<>}, R_{<>}, \cdots, R_{<>}\}$ from I.

The main challenges of visual relationship detection are: 1) The size of the classification space of possible relationships is huge. 2) The long-tail distribution of relationships lead to an extremely imbalanced dataset, where it is hard to collect enough training images for all the relationships, especially for infrequent relationships. For example, Visual Genome dataset (Krishna et al. 2017) contains over 75K distinct visual relationships, and the number of samples for each relationship ranges from just a handful to over 10K.

A natural approach to detect visual relationship is to treat it as a classification task. But this approach may work in a restricted context where the number of predicted space is moderate, such strategy would meet with a fundamental difficulty in a large number of imbalanced classes.

An alternative strategy is to learn the object and predicate respectively. This approach has been used in most of recent works (Lu et al. 2016; Liang, Lee, and Xing 2017; Dai, Zhang, and Lin 2017; Zhang et al. 2017) due to the following reasons: 1) The number of detectors declines drastically by using $N + K$ detectors (N is the number of object categories, and K is the number of predicate categories) to search the whole space of relationships. 2) Objects and predicates appear independently more frequently, which are more easy to learn compared with the infrequent relationships. However, this approach breaks up the structures of relationships between objects and predicates. For example, in the VRD training data (Lu et al. 2016), there is only 7,701 types of relationships with $100 \times 100 \times 70$ size space of relationships. To address this dilemma and obtain reasonable relationships, Lu *et al.* (Lu et al. 2016) leverage prior language knowledge, and Liang *et al.* (Liang, Lee, and Xing 2017) employ the graph of relationships.

In the strategy of learning the object and predicate respectively, the most difficult challenge is the low predicted accuracy of predicates. Although there appears some new approaches in recent works, such as using the union region of two objects (Lu et al. 2016; Zhu, Jiang, and Li 2017; Liang, Lee, and Xing 2017) and the positional information of objects (Zhang et al. 2017; Dai, Zhang, and Lin 2017; Zhu, Jiang, and Li 2017), these approaches exist a gap in precise representations of predicates. Firstly, the union region of two objects contains some noises, such as other objects. Secondly, the positional information of objects is exploited without embedding visual information that may lead to weak discrimination. A feasible approach to solve these problems is employing deep structured learning, which learns discriminative features, especially for predicates, and captures dependencies between objects and predicates directly.

In this paper, we introduce deep structured learning for visual relationship detection. In particular, we propose a deep structured model, which comprises feature-level relationship prediction and label-level relationship prediction. The feature-level prediction is to predict two objects and a predicate respectively by learning discriminative features, and the label-level prediction is to capture dependencies between objects and predicates based on the predicted relationship of feature level. The two predicted relationships determine the final relationships by weighted summing. Additionally, we use SSVM loss as our optimization goal, and decompose this goal into the subject, predicate, and object optimization goals. After decomposing, the optimizations become more simple and more independent.

To sum up, the main contributions of this work are as follows: 1) we propose a deep structured model for visual relationship detection, and this model captures dependencies between objects and predicates based on the predicted relationship at feature level. 2) SSVM loss function is exploited as our optimization goal, which is decomposed into the subject, predicate and object optimization goals for more simple and more independent optimizations.

Related work

Visual relationship detection. In earlier works, some base relationships are exploited to assist other computer vision tasks. For example, some spatial relations like “below”, “above” are exploited for object categorization (Galleguillos, Rabinovich, and Belongie 2008) and segmentation (Gould et al. 2008), and co-occurring relation is employed to assist scene classification (Izadinia, Sadeghi, and Farhadi 2014). Besides, some research tasks, such as human-object interactions (Gkioxari et al. 2017; Yao and Fei-Fei 2010), a few specific relationships (Sadeghi and Farhadi 2011), are also used by researchers.

Recently, the task of object relationship detection is proposed by Lu *et al.*(Lu et al. 2016), which combines appearance features and a language prior for relationship detection. To further improve the above work, Zhu *et al.* (Zhu, Jiang, and Li 2017) exploit spatial distributions of objects. Instead of calculating a language prior score of relationship, Liang *et al.* (Liang, Lee, and Xing 2017) employ a directed semantic action graph built on language priors, to detect visual relationships in the framework of deep enforcement learning. The above works are all using additional knowledge for relationship detection. Integrated into a single network that is learned in an end-to-end framework for relationship detection, Zhang *et al.* (Zhang et al. 2017) propose a VtransE model employing the key idea of transE, which learns embedding representations of triplets in natural language processing, and Dai *et al.* (Dai, Zhang, and Lin 2017) propose a deep relational networks. Compared with the V-transE, our method learns the structures of relationships in label level. Our method differs the deep relational networks (Dai, Zhang, and Lin 2017) in two aspects: 1) We learn discriminative feature of the predicate using the features of two objects. 2) The dependencies of objects and predicates are captured only by one layer.

Deep structured learning. There are two principal approaches to structured prediction: as a feed-forward function $y = f(x)$, and using an energy-based viewpoint $y = argmin_{y'} E(x, y')$ (LeCun et al. 2006). The feed-forward function models the structure only in feature levels, for example, using a fully convolution network models the structure of feature levels for image segmentation (Long, Shelhamer, and Darrell 2015). In contrast, the energy-based approach models the structure in both feature levels and label levels and can obtain more desirable results in most of cases, but this approach may involve non-trivial optimization, where the learning and prediction are more complex. Therefore, most of approaches explore an approximate learning procedure to solve this optimization (Chen et al. 2015; Belanger and McCallum 2016). A feasible approach is modeling the structure of feature levels and capturing dependencies of labels with a feed-forward function. For example, Liu *et al.* (Liu et al. 2015) propose a deep neural network, where the structure of feature levels is learnt by a deep convolution architecture, the dependency of labels is captured by additional two-layer convolution based on the prediction of feature levels, and the sum of two predictions decide the final prediction. Inspired by this work, we also capture the structure of relationship us-

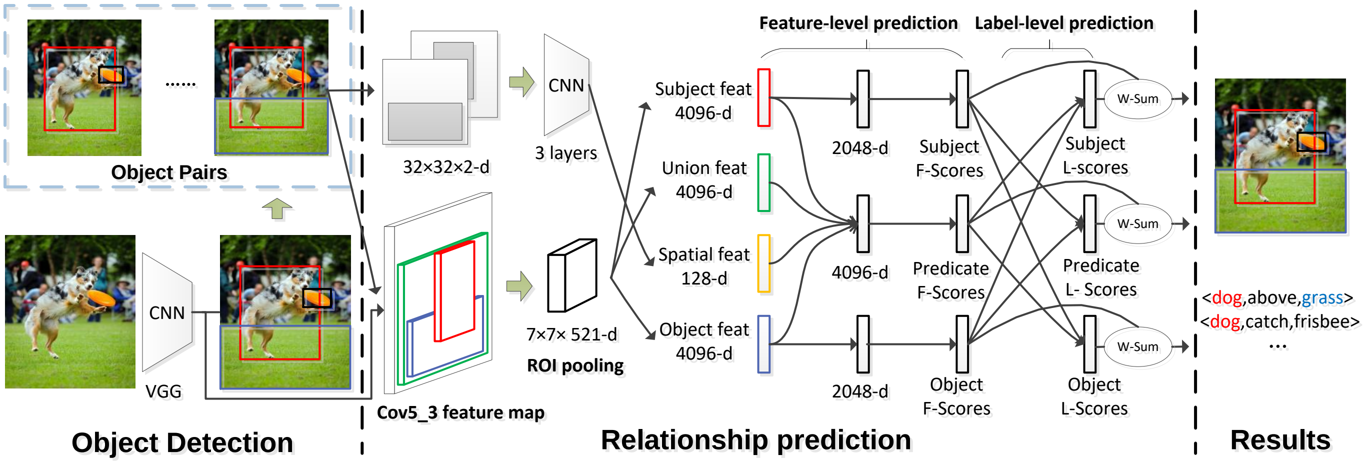

Figure 2: Network architecture of deep structured learning for visual relationship detection. An input image is first through the object detection module, which is a convolutional network that outputs a set of detected objects. Then, every pair of objects is fed into the relationship prediction module for relationship learning. In the relationship prediction module, the 4096-dimension feature of subject, object and union are obtained from the cov5_3 layer of the trained Faster R-CNN object detector, and the 128-dimension spatial feature is learnt by convolutional neural network (CNN) (Dai, Zhang, and Lin 2017), which is composed of three convolutional layers. These features are inputted into a multi-layer neural network to obtain feature-level prediction and label-level prediction. In the feature-level prediction, we share learnt weights for the prediction of subject and object.

ing additional layer based on the predictions of two objects and a predicate. In addition, deep structured learning is widely used for multi-label classification (Belanger and McCallum 2016), pose estimation (Song et al. 2017; Carreira et al. 2016), image segmentation (Liu et al. 2015; Long, Shelhamer, and Darrell 2015) and so on. However, deep structured learning has not been exploited for visual relationship detection.

Method

As shown in Figure 2, our proposed framework comprises object detection and relationship prediction. Object detection is used to locate regions of objects, and a series of object pairs are prepared for relationship prediction. In the relationship prediction, we firstly extract the visual features of two objects and their union, and employ a convolutional neural network (Dai, Zhang, and Lin 2017) to learn spatial features. The features are inputted into a neural network to learn feature-level scores of relationships, and simultaneously learn the label-level scores of relationships based on the predicted feature-level scores. The two predicted scores determine the final predicted relationships. The following paragraphs introduce the two stages.

Object detection. In this work, we use Faster R-CNN (Ren et al. 2015) to locate a set of objects, and generate a series of object pairs. Then each candidate object pairs with two bounding boxes and a cov5_3 layer feature map of the image are obtained for the relationship prediction, where the cov5_3 layer feature map can be reused to extract features of candidate objects.

Relationship prediction. The key components of relationship prediction module are feature-level prediction and label-level prediction. The feature-level predication is to obtain feature-level scores of relationships, and the label-level prediction is to calculate the label-level scores of relationships by capturing dependencies between objects and predicates. The final score of predicted relationship is the weighted sum of the two scores. We will discuss each component next, and finally give the details of the learning method.

Feature-level prediction

Given a pair of object bounding boxes $< o_s, o_o >$, the feature-level relationship is to predict two objects and a predicate respectively. The feature of objects and predicates for predicted relationship are obtained using a neural network, which is illustrated in Figure 2. The size of input feature of subject, object and the union is $7 \times 7 \times 512$ after ROI pooling, and each feature is inputted into a fully connected layer, where the weights come from fc6 layer of the trained Faster R-CNN model, with outputting a feature vector of 4096 dimensions, which is equal to the feature vector taken from the fc6 layer of the trained Faster R-CNN model. Our main goal is to reduce the size of learnt parameters to prevent overfitting. The two-channel input of binary value for learning the 128-D spatial feature with three convolutional layers, one for the subject and the other for the object, are down-sampled to the size 32×32. Finally, the feature vector of predicted relationship of objects and subjects is 2048 dimensions, the feature vector of predicted relationship of predicates is 4096 dimensions, and the corresponding predicted scores of the subject, the predicate, and the object in feature level are calculated as follows:

$$\begin{cases} S_{feat}(s_i) = w_i^T f_s + b_i \\ S_{feat}(p_k) = w_{p_k}^T f_p + b_{p_k} \\ S_{feat}(o_j) = w_j^T f_o + b_j \end{cases}$$

where f_s, f_p,f_o are feature vectors of subject, predicate, object respectively, $S_{feat}(s_i)$ is the feature-level score of i^{th} object category of subject, w_i, b_i are the learnt parameters to calculate $S_{feat}(s_i)$, $S_{feat}(p_k)$ is the feature-level score of k^{th} predicate category, w_{p_k}, b_{p_k} are the learnt parameters to calculate $S_{feat}(p_k)$, $S_{feat}(o_j)$ is the feature-level score of j^{th} object category of object, and w_j, b_j are the learnt parameters to calculate $S_{feat}(o_j)$.

Label-level prediction

In the structured learning, the common methods are learning a matrix to capture the interaction of pairwise labels, such as first-order Markov/Conditional random field, which may involve non-trivial optimization. Inspired by the work (Liu et al. 2015), we capture the dependencies of labels based on the feature-level prediction. The corresponding label-level scores are calculated as follows:

$$\begin{cases} S_{ld}(s_i) = w_{ps_i}^T S_{feat}(p) + w_{os_i}^T S_{feat}(o) \\ S_{ld}(p_k) = w_{sp_k}^T S_{feat}(s) + w_{op_k}^T S_{feat}(o) \\ S_{ld}(o_j) = w_{so_j}^T S_{feat}(s) + w_{po_j}^T S_{feat}(p) \end{cases}$$

where $S_{feat}(s)$, $S_{feat}(p)$, $S_{feat}(o)$ are three vectors, which are feature-level scores of the subject, the predicate and the object respectively, $S_{ld}(s_i)$ is the label-level score of i^{th} object category of subject, w_{ps_i},w_{os_i} are the learnt parameters to calculate $S_{ld}(s_i)$, $S_{ld}(p_k)$ is the label-level score of k^{th} predicate category, w_{sp_k}, w_{op_k} are the learnt parameters to calculate $S_{ld}(p_k)$, $S_{ld}(o_j)$ is the label-level score of j^{th} object category of object, and w_{so_j}, w_{po_j} are the learnt parameters to calculate $S_{ld}(o_j)$.

Training

The final score of relationship is $S(R_{<s_i,p_k,o_j>}| < o_s, o_o >) = S_{feat}(s_i) + S_{feat}(p_k) + S_{feat}(o_j) + S_{feat}(o_j) + \alpha S_{ld}(s_i) + \beta S_{ld}(p_k) + \gamma S_{ld}(o_j)$, where α, β and γ are hyperparameters. Our learning target is that given a pair of object bounding boxes $< o_s, o_o >$, the ground-truth relationship $R_{<>}$ has the highest score $S(R_{<>}| < o_s, o_o >)$. In this paper, we employ SSVM loss as our optimization objective due to SSVM has shown it's effectiveness to address the issue of involving complex outputs such as multiple dependent output variables and structured output spaces (Tsochantaridis et al. 2005). The SSVM minimizes:

$$\mathcal{L} = \sum_{<o_{is},o_{io}>,R_i} \max_R (\triangle(R_i,R) + S(R| < o_{is}, o_{io} >)) \\ - S(R_i| < o_{is}, o_{io} >) \quad (1)$$

where $\triangle(R_i,R)$ is an error function between a prediction R and the ground truth R_i. The $\triangle(R_i,R)$ can be decomposed into three terms: $\triangle(R_i,R) = \triangle(s_i,s) + \triangle(p_i,p) + \triangle(o_i,o)$, where $\triangle(s_i,s)$, $\triangle(p_i,p)$, $\triangle(o_i,o)$ are the error functions of subject, predicate, object respectively. Then Eq. 1 can be decomposed into the sum of the following three terms:

$$\begin{cases} \mathcal{L}_s = \sum_{s_i} \max_s (\triangle(s_i,s) + S_{feat}(s) + \alpha S_{ld}(s)) \\ \qquad - (S_{feat}(s_i) + \alpha S_{ld}(s_i)) \quad (2) \\ \mathcal{L}_p = \sum_{p_i} \max_p (\triangle(p_i,p) + S_{feat}(p) + \beta S_{ld}(p)) \\ \qquad - (S_{feat}(p_i) + \beta S_{ld}(p_i)) \quad (3) \\ \mathcal{L}_o = \sum_{o_i} \max_o (\triangle(o_i,o) + S_{feat}(o) + \gamma S_{ld}(o)) \\ \qquad - (S_{feat}(o_i) + \gamma S_{ld}(o_i)) \quad (4) \end{cases}$$

Where $\mathcal{L}_s$ is subject optimization, $\mathcal{L}_p$ is predicate optimization, and $\mathcal{L}_o$ is object optimization. For optimizing Eq. 1, the computational complexity of the maximum score of relationship is N^2K. But for the three terms, this computational complexity of the maximum score of relationship is down to $2N + K$. After decomposing, the optimizations become more simple and more independent.

Experiment

We will validate the effectiveness of the proposed deep structured model for visual relationship detection by answering the following two questions. Q1: Is the deep structured model effective for visual relationship detection? Q2: Is the learnt structure of relationship reasonable by this deep structured model?

Datasets and Metrics

Datasets. we evaluate our proposed method on Visual relationship detection (VRD) (Lu et al. 2016) and Visual Genome (VG) (Zhang et al. 2017) datasets. a) VRD: the dataset contains 5,000 images with 100 object categories and 70 predicate categories, and is annotated 37,993 relationships with 7,701 types. We follow the same train/test split as in (Lu et al. 2016), i.e., 4,000 training images containing 30,355 relationships with 6,672 types and 1,000 test images containing 7,638 relationships with 2,747 types, where 1,169 relationships with 1,029 types are only in the test data. b) VG: the dataset contains 99,652 images with 200 object categories and 100 predicates, resulting in 1,090,027 relationships with 19,561 types and 57 predicates per object category. We also follow the same train/test split as in (Zhang et al. 2017), 73,794 training images containing 803,276 relationships with 19,236 types, and 25,858 test images containing 286,751 relationships with 16,592 types.

Evalutaion. we evaluate our proposed method for the following tasks. **Phrase detection**: given an input image, output a phrase $< sub, pre, obj >$ and localize the entire phrase with one bounding box having intersection over union (IoU) > 0.5 with the ground-truth bounding box. **Relationship detection**: given an input image, output a relationship $< sub, pre, obj >$ and localize both the subject and the object with their bounding boxes having $IoU > 0.5$ with the ground-truth bounding boxes respectively. **Phrase prediction**: given an input image and two objects bounding boxes, output a phrase $< sub, pre, obj >$. Following (Lu et al. 2016), we use Recall@50 (R@50) and Recall@100 (R@100) as evaluation metrics for detection. $R@x$ computes the fraction of times of the correct relationship, which is predicted in the top x confident relationship predictions in

Table 1: Performances of phrase detection and relationship detection using various methods on VRD and VG datasets.

		Phrase Det (%)		Relationship Det (%)	
		R@100	R@50	R@100	R@50
VRD	F	18.44	16.98	13.90	12.96
	Fo	23.41	21.99	17.07	16.15
	F+L	23.12	21.62	17.37	16.28
	Fo+L	**23.92**	**22.61**	**18.26**	**17.27**
VG	F	9.07	7.46	4.73	3.90
	Fo	10.66	8.90	5.71	4.77
	F+L	**14.43**	11.77	7.33	5.96
	Fo+L	14.35	**12.07**	**7.50**	**6.37**

Table 2: Performances of phrase prediction using various methods on VRD and VG datasets. we use Precision (P) as the evaluation metric.

P(%)	F	Fo	F+L	Fo+L
VRD	26.43	35.94	34.32	**36.28**
VG	16.58	19.39	26.67	**26.86**

an image. In the task of phrase prediction, we use Precision (P) as the evaluation metric.

Experimental setup

Implementation details. The object detection architecture of Faster-RCNN is the VGG-16 network (Simonyan and Zisserman 2014). At the training time, we sample a mini-batch containing 256 region proposals generated by the RPN of Faster-RCNN, each of which is positive if it has an $IoU > 0.7$ with some ground-truth regions and it is negative if the $IoU < 0.3$. At the test time, we sample 300 region proposals generated by RPN with $IoU > 0.7$. After the classification layer, we perform non-maximum suppression (NMS) with $IoU > 0.5$ for every class on the 300 proposals. In the VRD dataset, the proposals with probability of a category > 0.5 are retained for NMS, resulting in 12.3 proposals per image on average. In the VG dataset, the proposals with the probability of a category> 0.3 are retained for NMS, resulting in 15.3 proposals per image on average. After object detection, each two objects is served for relationship prediction in the two cases: a) there is an intersection between the two bounding boxes of objects, b) the ratio of the sum area of two bounding boxes to the area of union bounding box is bigger than 0.4. In addition, due to the lack of training images for object detection in the VRD data, we select 32,715 images of the VG data to train Faster-RCNN model, and verify on the VRD training data obtaining 28.1% mAP. In the prediction of relationship, we empirically set $\triangle(s_i, s) = 3$, $\triangle(p_i, p) = 3$, $\triangle(o_i, o) = 3$, a momentum of 0.9, $\alpha = r = 0.005$, $\beta = 0.2$, a weight decay of 0.05 for the VRD dataset, and $\alpha = r = 0.1$, $\beta = 0.3$, a weight decay of 0.001 for the VG dataset.

Our Model. We perform ablation studies on our model and compare the results. Specifically, we consider the following variants of our model:

- **F.** Detecting visual relationship only uses feature-level prediction, but the prediction of predicate does not use the features of two objects.
- **Fo.** Detecting visual relationship only uses feature-level prediction, and the prediction of predicate uses the features of two objects.
- **F+L.** Detecting visual relationship uses feature-level prediction and label-level prediction, but the prediction of predicate does not use the features of two objects.
- **Fo+L.** Detecting visual relationship uses feature-level prediction and label-level prediction, and the prediction of predicate uses the features of two objects.

Experimental results

Table 1 compares the performance of our proposed method in both phrase detection and relationship detection, and the performance of our proposed method in phrase prediction is shown in Table 2. From the experimental results, we can obtain the following observations:

1). The method of 'Fo' have a significant advantage over the method of 'F' in the VRD dataset, but this advantage is weaker in the VG dataset. A possible reason is that there exists more noises in the VG dataset, which leads to learning a noisy feature of the predicate. On the whole, the experimental performance has been improved by using the features of two objects in feature-level relationship prediction. This illustrates that the method of using the features of two objects can learn a stronger discriminative feature of the predicate.

2). The method of 'F+L' thoroughly outperforms the method of 'F' with an improvement of more than 4% on the task of phrase detection and more than 3% on the task of relationship detection on two datasets. In addition, the method of 'Fo+L' have a significant improvement over the method of 'Fo' only on the VG dataset, but this improvement is smaller on the VRD dataset. The main reasons are: a). The method of 'F' gets more information of the structures of relationships from label-level relationship prediction, compared with the method of 'Fo'. b). The method of 'Fo+L' gets more information of the structures of relationships in the VG dataset from the label-level prediction, compared with this method on the VRD dataset. In a word, by adding the label-level prediction, we can obtain a performance improvement based on the same feature-level prediction. This validates that the structures of relationships have been captured in this label-level prediction and the deep structured model is effective for visual relationship detection (Q1). Furthermore, the average processing time per image of the relationship prediction phase is 0.08 seconds and 0.28 seconds for VRD dataset and VG dataset respectively.

To analyze what structures have been learnt in the deep structured model (Q2), we resort to the fully connected layer in label-level prediction. Figure 3 shows the weight information of fully connected layer of "subject - predicate" (w_{sp_*}) and "object - predicate" (w_{op_*}) in label-level prediction. We list 5 subjects and 5 objects, which are the top 5 of high frequency in the training data, and list another 5 subjects and 5 objects, which are the last 5 of low frequency in the training data. The corresponding predicates of top 3 and last 3

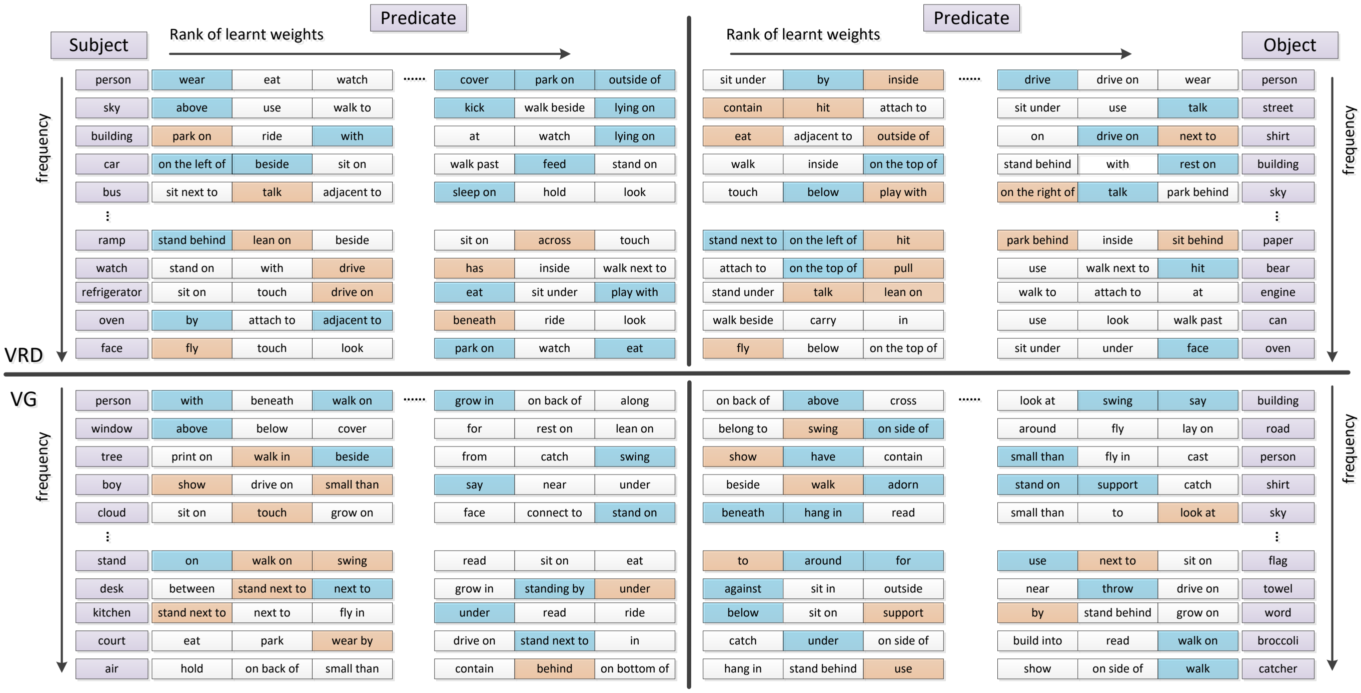
Subject
Rank of learnt weights
Predicate
Predicate
Rank of learnt weights
Object
frequency
frequency
VRD
person wear eat watch ...... cover park on outside of
sky above use walk to kick walk beside lying on
building park on ride with at watch lying on
car on the left of beside sit on walk past feed stand on
bus sit next to talk adjacent to sleep on hold look
ramp stand behind lean on beside sit on across touch
watch stand on with drive has inside walk next to
refrigerator sit on touch drive on eat sit under play with
oven by attach to adjacent to beneath ride look
face fly touch look park on watch eat
sit under by inside ...... drive drive on wear person
contain hit attach to sit under use talk street
eat adjacent to outside of on drive on next to shirt
walk inside on the top of stand behind with rest on building
touch below play with on the right of talk park behind sky
stand next to on the left of hit park behind inside sit behind paper
attach to on the top of pull use walk next to hit bear
stand under talk lean on walk to attach to at engine
walk beside carry in use look walk past can
fly below on the top of sit under under face oven
VG
person with beneath walk on ...... grow in on back of along
window above below cover for rest on lean on
tree print on walk in beside from catch swing
boy show drive on small than say near under
cloud sit on touch grow on face connect to stand on
stand on walk on swing read sit on eat
desk between stand next to next to grow in standing by under
kitchen stand next to next to fly in under read ride
court eat park wear by drive on stand next to in
air hold on back of small than contain behind on bottom of
on back of above cross ...... look at swing say building
belong to swing on side of around fly lay on road
show have contain small than fly in cast person
beside walk adorn stand on support catch shirt
beneath hang in read small than to look at sky
to around for use next to sit on flag
against sit in outside near throw drive on towel
below sit on support by stand behind grow on word
catch under on side of build into read walk on broccoli
hang in stand behind use show on side of walk catcher

Figure 3: The ranks of learnt weights of "subject - predicate" and "object - predicate" in VRD and VG dataset.

of subjects and objects, ranked by the weight values in descending order, are listed in Figure 3. We also compare this rank of learnt weights with the rank of frequency at the same "subject-predicate" or "object-predicate" in the training data. The blue predicates are the top-3 predicates appeared in the top 20 of frequency, or the last-3 predicates appeared in the last 20 of frequency. The brown predicates are the top-3 predicates appeared in the last 20 of frequency, or the last-3 predicates appeared in the top 20 of frequency. We can find that the blue predicates have a consistent rank of frequency, and the brown predicates have an adverse rank of frequency. From this figure, we have the following discussions:

1). The ranks of learnt weights of "subject - predicate" and "object - predicate" are basically satisfying. Most of ranks of the learnt weights are reasonable, especially for the objects of height frequency. For example, in the VRD data, when the subject is "person", the top 3 are "person - wear", "person - eat" and "person - watch" which are possible actions of the person, and the last 3 contains "person - cover" and "person - park" which are impossible actions of person. In the VG dataset, when the object is "building", the top 3 are "on back of - building" and "above - building" which are possible positional relations to the building, and the last 3 contains "swing - building" and "say - building" which are the impossible actions to the building.

2). Despite there are some unreasonable ranks of weights of "subject - predicate" and "object - predicate" (For example, in the VRD data, when the subject is "building", the "building - park on" and "building - ride" are impossible actions to building, but appear in the top 3), the overall ranks are reasonable. To evaluate the rationality of the rank, we use the ratio of consistent ranks (blue) to inconsistent ranks (brown), and our hypothesis is that the rank of frequency is reasonable. The larger the ratio of the total of consistent ranks to the total of inconsistent ranks is, the more reasonable ranks have been learnt. In the VRD dataset, the ratio value is 33/24, and this ratio value is up to 21/10 in top-5 objects. In the VG dataset, the ratio value is 33/20, and this ratio value is up to 19/8 in top-5 objects. The ratio values of the two datasets in top-5 objects are larger than 2, which can be regarded that more than 2/3 of the ranks are reasonable. This validates that the learnt structure of relationship is reasonable by this deep structured model (Q2).

Our proposed method captures dependencies between objects and predicates, but has not learnt to evaluate the reasonability of detecting relationships. To evaluate this reasonability, we set a mask of relationships for the results of detection. A mask of relationships is a function $M(R_i)$, where $M(R_i) = \begin{cases} 0, & R_i \in B \\ -\infty, & R_i \notin B \end{cases}$, B is a set of relationships. After using the mask of relationships, the final score of relationship is $S(R_i) + M(R_i)$. In this work, the set B is all the types of relationships appeared in the training data. Table 3 and Table 4 are the results of the our proposed method by using the mask of relationships. In the two datasets, compared with the methods with label-level prediction, the methods of only using feature-level prediction have a relatively bigger improvement by using the mask of relationships, especially for the method of 'F' in the VRD dataset. A possible explanation is: the methods with label-level prediction learn more knowledge of structures of relationships, but achieve less knowledge of structures of relationships from the mask of

Table 3: Performances of phrase detection and relationship detection using various methods with the mask of relationships on VRD and VG datasets.

		Phrase Det (%)		Relationship Det (%)	
		R@100	R@50	R@100	R@50
VRD	F^m	21.01	19.62	15.72	14.85
	Fo^m	23.57	22.30	17.81	16.94
	F+L^m	23.41	22.06	17.43	16.51
	Fo+L^m	**23.95**	**22.67**	**18.33**	**17.40**
VG	F^m	10.46	8.54	5.39	4.42
	Fo^m	11.90	9.87	6.26	5.23
	F+L^m	15.59	13.05	7.98	6.58
	Fo+L^m	**15.61**	**13.08**	**8.00**	**6.82**

Table 4: Performances of phrase prediction using various methods on on VRD and VG datasets. we use Precision (P) as the evaluation metric.

P(%)	F^m	Fo^m	F+L^m	Fo+L^m
VRD	31.66	36.10	35.72	**37.13**
VG	17.77	20.74	28.01	**28.05**

relationships. In the VG dataset, the proposed four methods with the mask of relationships have a significant improvement. But in the VRD dataset only the method of ’F’ obtains a comparable improvement, especially for the task of phrase prediction. The main reason is that the mask of VG have a reasonable mask of relationships, which contains more reasonable relationships. For example, the mask of VRD only contains 6,672 types of relationships without 1029 types of relationships in the testing data, but the mask of VG contains 19,236 types of relationships without 325 types of relationships in the testing data.

We also compare our proposed method with the state-of-the-art models. Since the task of visual relationship detection is proposed, only VRD dataset is publicly released. All of proposed works conduct experiments in this dataset, and select data from the whole VG dataset (Krishna et al. 2017) by themselves. Recently, the work of VtransE (Zhang et al. 2017) has released their dataset. To evaluate our proposed method, we also conduct experiments, and implement some methods (Lu et al. 2016; Zhu, Jiang, and Li 2017) in the recently released dataset VG. The quantitative results in the two datasets are shown in Table 5. In the VRD dataset, our proposed method outperforms the state-of-the-art methods (Liang, Lee, and Xing 2017; Dai, Zhang, and Lin 2017) in phrase detection, and achieves a comparable performance in relationship detection. In the VG dataset, our proposed method outperforms the methods (Lu et al. 2016; Zhu, Jiang, and Li 2017; Zhang et al. 2017) significantly on both phrase detection and relationship detection.

To sum up, even without using the mask of relationships, our proposed method of ’Fo+L’ has outperformed existing state-of-the-art methods in some evaluation metrics on the two datasets. This validates the effectiveness of our proposed method (Q1), and also reflects that the dependencies between objects and predicates are useful to facilitate visual relationship detection. These dependencies are reasonable (Q2), which are validated by analyzing the learnt weight of “subject - predicate” and “predicate - object”.

Table 5: Comparison with state-of-the-art models. The second row and the third row are the performances on the VRD dataset and the VG dataset respectively. The following numbers of the first left column represent references. ([1]: (Lu et al. 2016), [2]: (Zhu, Jiang, and Li 2017), [3]: (Zhang et al. 2017), [4]: (Dai, Zhang, and Lin 2017), [5]: (Liang, Lee, and Xing 2017))

	Phrase Det (%)		Relationship Det (%)	
	R@100	R@50	R@100	R@50
VRD [1]	17.03	16.17	14.70	13.86
VRDS [2]	18.89	16.94	15.77	14.31
VtransE [3]	22.42	19.42	15.20	14.07
RDN [4]	23.45	19.93	**20.88**	17.73
DVSRL [5]	22.60	21.37	20.79	**18.19**
Fo+L our	23.92	22.61	18.26	17.27
Fo+L^m our	**23.95**	**22.67**	18.33	17.40
VRD [1]	11.85	9.85	5.62	4.77
VRDS [2]	12.19	10.54	5.75	5.08
VtransE [3]	10.45	9.46	6.04	5.52
Fo+L our	14.35	12.07	7.50	6.37
Fo+L^m our	**15.61**	**13.07**	**8.00**	**6.82**

Conclusion and Future works

In this paper, we propose a deep structured model for visual relationship detection. Our proposed method not only predicts relationships on the feature level, but also captures dependencies between objects and predicates. Additionally, we use SSVM loss as our optimization goals, and decompose the optimization goal into multiple optimization goals. To evaluate our proposed method, we conduct experiment on VRD dataset and VG dataset and achieve state-of-the-art performance. With using the mask of relationships, the experimental performances are improved. This illustrates that evaluating reasonable relationships can facilitate visual relationship detection and the evaluation of reasonable relationship is to learn second-order relations of labels. Our proposed method learns the relations of labels only capturing first-order relations. In the future work, we will learn both first-order relations and second-order relations in label-level relationship prediction using a deep neural network.

Acknowledgement: This work was supported in part by the National Natural Science Foundation of China under Grant 61532018, in part by the Beijing Municipal Commission of Science and Technology under Grant D161100001816001, in part by the Lenovo Outstanding Young Scientists Program, in part by National Program for Special Support of Eminent Professionals and National Program for Support of Top-notch Young Professionals.

References

Belanger, D., and McCallum, A. 2016. Structured prediction energy networks. In *International Conference on Machine Learning*, 983–992.

Carreira, J.; Agrawal, P.; Fragkiadaki, K.; and Malik, J. 2016. Human pose estimation with iterative error feedback. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4733–4742.

Chen, L.-C.; Schwing, A.; Yuille, A.; and Urtasun, R. 2015. Learning deep structured models. In *Proceedings of the 32nd International Conference on Machine Learning*, 1785–1794.

Dai, J.; Li, Y.; He, K.; and Sun, J. 2016. R-fcn: Object detection via region-based fully convolutional networks. In *Advances in neural information processing systems*, 379–387.

Dai, B.; Zhang, Y.; and Lin, D. 2017. Detecting visual relationships with deep relational networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 248–255. IEEE.

Galleguillos, C.; Rabinovich, A.; and Belongie, S. 2008. Object categorization using co-occurrence, location and appearance. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1–8. IEEE.

Gkioxari, G.; Girshick, R.; Dollár, P.; and He, K. 2017. Detecting and recognizing human-object interactions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Gould, S.; Rodgers, J.; Cohen, D.; Elidan, G.; and Koller, D. 2008. Multi-class segmentation with relative location prior. *International Journal of Computer Vision* 80(3):300–316.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Izadinia, H.; Sadeghi, F.; and Farhadi, A. 2014. Incorporating scene context and object layout into appearance modeling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 232–239.

Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; Kravitz, J.; Chen, S.; Kalantidis, Y.; Li, L.-J.; Shamma, D. A.; et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision* 123(1):32–73.

LeCun, Y.; Bengio, Y.; and Hinton, G. 2015. Deep learning. *Nature* 521(7553):436–444.

LeCun, Y.; Chopra, S.; Hadsell, R.; Ranzato, M.; and Huang, F. 2006. A tutorial on energy-based learning. *Predicting structured data* 1:0.

Liang, X.; Lee, L.; and Xing, E. P. 2017. Deep variation-structured reinforcement learning for visual relationship and attribute detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Liu, Z.; Li, X.; Luo, P.; Loy, C.-C.; and Tang, X. 2015. Semantic image segmentation via deep parsing network. In *Proceedings of the IEEE International Conference on Computer Vision*, 1377–1385.

Long, J.; Shelhamer, E.; and Darrell, T. 2015. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3431–3440.

Lu, C.; Krishna, R.; Bernstein, M.; and Fei-Fei, L. 2016. Visual relationship detection with language priors. In *European Conference on Computer Vision*, 852–869. Springer.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, 91–99.

Sadeghi, M. A., and Farhadi, A. 2011. Recognition using visual phrases. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1745–1752. IEEE.

Santoro, A.; Raposo, D.; Barrett, D. G.; Malinowski, M.; Pascanu, R.; Battaglia, P.; and Lillicrap, T. 2017. A simple neural network module for relational reasoning. *arXiv preprint arXiv:1706.01427*.

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

Song, J.; Wang, L.; Van Gool, L.; and Hilliges, O. 2017. Thin-slicing network: A deep structured model for pose estimation in videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; and Rabinovich, A. 2015. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1–9.

Szegedy, C.; Ioffe, S.; Vanhoucke, V.; and Alemi, A. A. 2017. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Association for the Advancement of Artificial Intelligence*, 4278–4284.

Tsochantaridis, I.; Joachims, T.; Hofmann, T.; and Altun, Y. 2005. Large margin methods for structured and interdependent output variables. *Journal of machine learning research* 6(Sep):1453–1484.

Xiao, J.; Hays, J.; Ehinger, K. A.; Oliva, A.; and Torralba, A. 2010. Sun database: Large-scale scene recognition from abbey to zoo. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3485–3492. IEEE.

Yao, B., and Fei-Fei, L. 2010. Grouplet: A structured image representation for recognizing human and object interactions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 9–16. IEEE.

Zhang, H.; Kyaw, Z.; Chang, S.-F.; and Chua, T.-S. 2017. Visual translation embedding network for visual relation detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Zhu, Y.; Jiang, S.; and Li, X. 2017. Visual relationship detection with object spatital distribution. In *Proceedings of the IEEE International Conference on Multimedia and Expo*, 379–384. IEEE.