MR-NET: Exploiting Mutual Relation for Visual Relationship Detection

Yi Bin,[1] Yang Yang,[1] Chaofan Tao,[1] Zi Huang,[2] Jingjing Li,[1] Heng Tao Shen[1]
[1]University of Electronic Science and Technology of China
[2]The University of Queensland

Abstract

Inferring the interactions between objects, *a.k.a visual relationship detection*, is a crucial point for vision understanding, which captures more definite concepts than object detection. Most previous work that treats the interaction between a pair of objects as a one way fail to exploit the mutual relation between objects, which is essential to modern visual application. In this work, we propose a mutual relation net, dubbed *MR-Net*, to explore the mutual relation between paired objects for visual relationship detection. Specifically, we construct a mutual relation space to model the mutual interaction of paired objects, and employ linear constraint to optimize the mutual interaction, which is called *mutual relation learning*. Our mutual relation learning does not introduce any parameters, and can adapt to improve the performance of other methods. In addition, we devise a semantic ranking loss to discriminatively penalize predicates with semantic similarity, which is ignored by traditional loss function (*e.g.*, cross entropy with softmax). Then, our MR-Net optimizes the mutual relation learning together with semantic ranking loss with a siamese network. The experimental results on two commonly used datasets (*VG and VRD*) demonstrate the superior performance of the proposed approach.

Introduction

Visual understanding is a challenging task, as it requires machines not only to recognize the concepts in the visual data (*e.g.*, image or video), but also to analyse the semantic meaning of the image/video. In the past decade, visual understanding, especially in the fields of recognition, classification and object detection (Ren et al. 2015; Li et al. 2018; Zhang et al. 2017b; 2018b), has achieved remarkable accomplishments thanks to the flourish of deep learning and CNN (Krizhevsky, Sutskever, and Hinton 2012; Zhu et al. 2018). Recently, it moves forward to further infer the interactions and relationships between concepts by going beyond object detection, which is termed as visual relationship detection (Lu et al. 2016; Zhang et al. 2017a; Zhu and Jiang 2018; Zhuang et al. 2017). Specifically, as shown in Figure 1, given an image, visual relationship detection aims to locate and recognize the visual entities, and then infer all the possible triplet phrases. Detecting such relationships enables the artificial intelligence to understand vision with more details and semantics, which is helpful to other visual understanding tasks, such as visual captioning (Bin et al. 2017; 2018; Gao et al. 2017), and visual question answering (Peng et al. 2018; Gao et al. 2018).

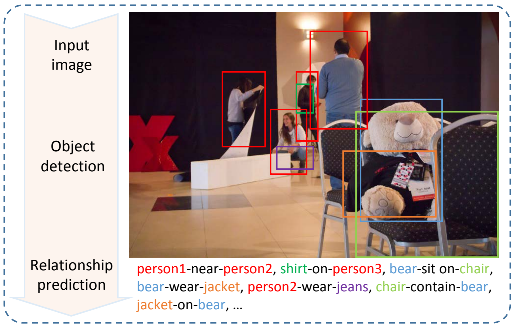
Input image
Object detection
Relationship prediction
person1-near-person2, shirt-on-person3, bear-sit on-chair,
bear-wear-jacket, person2-wear-jeans, chair-contain-bear,
jacket-on-bear, ...

Figure 1: A general process for visual relationship detection. Given an image, the model detects all the objects, and infers the relationships between object pairs.

Sadeghi and Farhadi (Sadeghi and Farhadi 2011) first define a triplet `subject-predicate-object` as a visual phrase, and train classifiers for every triplet phrase. Such pattern enables the model to predict `person-eating-bread` and `man-eating-bread` as two different phrases, though they share very similar entities and the same predicate. In other words, one may need $O(N^2K)$ classifiers for N unique object categories and K relationship categories. Moreover, because of the combination of infrequent subject, object and predicate, many triplet phrases cannot be well discriminated suffering from the long tail distribution. To address these problems, Lu et al. (Lu et al. 2016) propose to detect objects and predicates separately, and combine the results together to jointly learn the relationships. In this way, different triplet phrases share the same predicate, such as `person-eating-bread` and `man-eating-bread`, can be detected by the same classifier. Even more, `horse-eating-grass` with very different objects context can share the same classifier, and an unseen triplet (*e.g.*, `dog-ride-bike`) can be handled if the objects and predicate are in the training set. Therefore, $O(N^2K)$ classifiers reduce to $O(N+K)$ ones. Most of recent work (Lu et al. 2016; Zhang et al. 2018a;

Zhu and Jiang 2018; Han et al. 2018) choose to model visual relationship in the latter pattern.

To model the objects and predicates, existing work apply two paradigms: 1) two-stage learning (Lu et al. 2016; Zhuang et al. 2017), which first employs an object detector (*e.g.*, Faster RCNN (Ren et al. 2015)) to detect all the possible objects in an image, and then predicts the relationship between objects, a process known as *predicate prediction* at the second stage; 2) structural learning (Xu et al. 2017; Zhu and Jiang 2018) that takes an image as input, and predicts objects and fused relationships simultaneously. In two-stage learning, object detection is independent of the predicate prediction, and the word embeddings of object categories enable the relationship detection to bring in semantic information. Obviously, the drawback of two-stage learning is that bad object detection may mislead the predicate classifier. The latter paradigm, structural learning overcomes this issue by learning the object detector and predicate classifier simultaneously, which is capable of capturing the dependencies between objects and predicates (Zhu and Jiang 2018). However, limited to performance of current relationship detection (with about 0.2 of recall@50), the object detector may be harmed by the noise of predicates.

In this paper, we focus on the exploration of mutual relation between paired objects. Specifically, mutual relationship of given objects can be either their parts, properties, or spatial locations (*e.g.*, `person-wear-shirt` *vs.* `shirt-on-person`, `person-has-arm` *vs.* `arm-belong to-person`, and `person-in front of-building` *vs.* `building-behind-person`). In (Zellers et al. 2018), Zellers et al. divided the relationships into four categories: geometric, possessive, semantic and misc. They also gave the distribution of each type of category, where there are more than 90% relationships that belong to geometric and possessive. It is reasonable to assume that such mutual relationships not only exist with explicit labels, but also implicitly occur in the visual feature. For example, one may say `man-hold-fork` rather than `fork-be holden-man`. But the fork is actually holden by the man in vision, it just miss an explicit relationship label to describe this scene. We also note that traditional loss functions for multi-classification (*e.g.*, cross entropy with softmax) fails to measure the semantic similarities among predicates. Therefore, we propose mutual relation net (MR-Net) to exploit the mutual interaction for visual relation detection. We construct a mutual relationship space, where paired relationship triplets are modelled to satisfy a linear constraint for exploiting mutual interaction. We call this process *mutual relation learning*. In addition, to help the classifier capture semantic similarity, we replace cross entropy with a semantic ranking loss, which penalizes similar predicates more mildly than others (*e.g.*, `near` receives less penalty than `in` under the ground-truth `man-next to-tree`). Finally, we model the mutual relation between objects by feeding the pair into a Siamese network (Chopra, Hadsell, and LeCun 2005), and optimize it with the proposed semantic ranking. More importantly, our mutual relation learning is parameter-free and can be inserted into all of other methods to boost the performance. To eliminate the influence of dependencies between objects and predicates, we employ the two-stage learning to verify our mutual relationship modelling.

In summary, the main contributions of our work are as follows: 1) to the best of our knowledge, our work is the first to explore mutual interactions between paired objects for visual relationship detection; 2) We propose to apply ranking loss with semantic embedding for predicate classification, which is optimized together with mutual relation learning in the siamese network; 3) To handle the noise introduced by the bounding boxes of detected objects, we randomly shift or scale a fraction of bounding boxes with a small range in training phase, which is also a way to augment samples for small dataset (*e.g., VRD*).

Related Work

Mutual Relation Exploration

Mutual relations widely exist in our world, *e.g.*, social relations (*wife and husband in a marriage*), spatial relations (*on and under*), and semantic relations (*hypernym and hyponym*). In the field of computer vision, most of existing work of mutual relation fall into multiple objects tracking. In (Duan et al. 2012), Duan et al. proposed to model the mutual relation between objects for multi-object tracking, which constructed a relational graph for mutually related objects, and group tracked via the mutual relation of objects. Helbing and Molnar (Helbing and Molnar 1995) proposed a social force model that consists of desired action, territorial effect, attractiveness to describe pedestrian dynamics in crowds, and proved that simulations of interaction of pedestrian in crowds are capable of describing pedestrian behavior very realistically. Zhou et al. (Zhou et al. 2018) employed asymmetric pairwise terms to model the inter-objects relations between object tracklets, which is able to constrain the displacement and control the directional influences between the pair of objects. Semantic relations are commonly exploited in word embedding (Mikolov et al. 2013; Pennington, Socher, and Manning 2014). Even more, in (Mikolov et al. 2013), the mutual relation between words can be represented by a linear vector arithmetic, *e.g.*, vec("Paris") - vec("France") + vec("Italy") $\approx$ vec("Rome"), and vice versa.

Visual Relationship Detection

Visual relationship detection is first introduced in (Sadeghi and Farhadi 2011), where a triplet of (subject, interaction, object) is defined as "visual phrase", and regarded as a single category for classification. While this way suffers from low scalability of model and lack of diversity of visual relationship tuple. To overcome these drawbacks, Lu et al. (Lu et al. 2016) redefined visual relationship as a combination of objects and predicates, which reduces the magnitude of classifier from $O(N^2K)$ to $O(N + K)$, and has capability to generate triplet phrase unseen. They integrated visual appearance and language priors, and optimized with a biconvex loss function. Subsequently, a large amount of work applied such two-stage learning to detect visual relationships in images. Dai et al. (Dai, Zhang, and Lin 2017) designed a

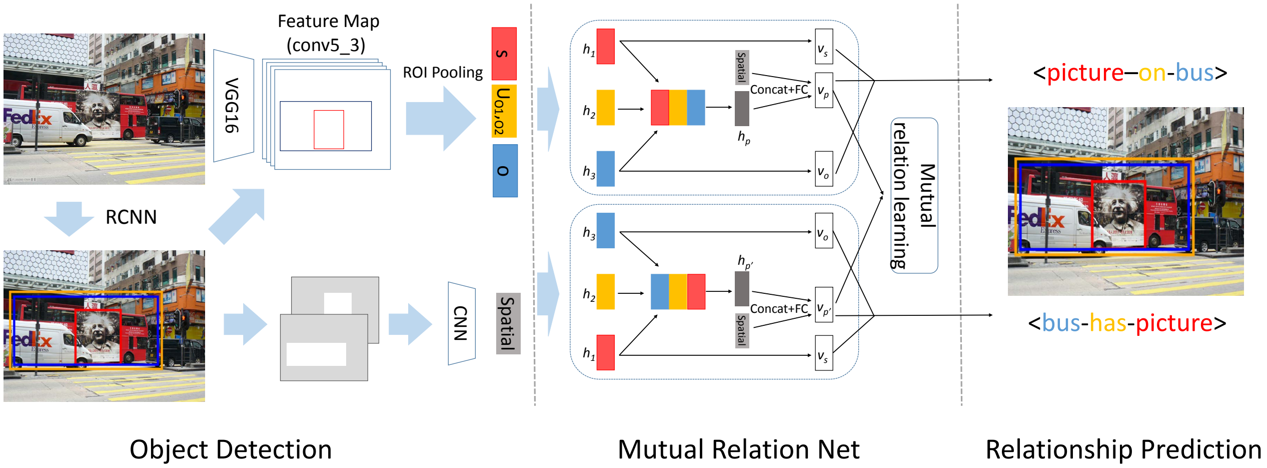

Figure 2: An overview for the flowchart of our mutual relation network. As the same of most two-stage paradigm, we first extract visual appearance representation and spatial configuration with VGG16, Faster-RCNN and custom CNN. Then the visual and spatial feature of object pair is injected to the mutual relation learning module, a siamese network, for mutual interaction exploration. We predict the relationships based on previous object detection and mutual relation learning in the final stage.

deep relational network to infer the predicate between object pairs by jointly exploiting their spatial information and statistical dependencies. They employed convolutional neural networks (CNNs) to process the binary masks for spatial configuration, and achieved the state-of-the-art. Zhang et al. (Zhang et al. 2017a) interpreted a relation triplet as a visual translation embedding manner (VTransE), and transferred knowledge between context and predicates. Liang et al. (Liang, Lee, and Xing 2017) implemented deep variation-structured reinforcement learning to predict relationships and attributes of objects together with global context cues. In (Zhuang et al. 2017), Zhuang et al. devised a framework to learn predicates classifier adapting to its context, which encourages the model to predict similar context with similar classifiers.

Another strategy to detect visual relationships is employing structural learning to predict a struct of objects and predicates simultaneously, and format to a triplet of relationship. Xu et al. (Xu et al. 2017) applied a dual graph to model the objects and predicates with node and edge, and employed Gated Recurrent Units (GRU) to iteratively update the graph along time, which passes messages between node GRU and edge GRU dually. Li et al. (Li et al. 2017) constructed a multi-level scene description network to generate phrase and captions, which consists of a module for dynamical graph construction, a feature refiner, and classifiers. Their framework constructed a dynamical graph to represent the scene and updated the graph via feature refining, then output the relation phrases and captions in the end. Zhu and Jiang (Zhu and Jiang 2018) proposed a deep structural learning framework to predict objects and relationships at feature and label level, then fused the predictions of both levels. They argued that their model enables the communication between objects and predicates at the label level, which helps the object detector and relationship classifier to be more accurate.

Method

From our statistical results, there are a considerable portion of images owning mutually paired relationship triplets in the visual relationship detection datasets. Naturally, when there is a relationship that can be expressed as subject-predicate-object, we observe that there is usually a mutual interaction between object and subject. Sometimes these interactions are explicit relationship pairs that are semantically relative, such as cloud-in-sky and sky-has-cloud. While some other mutual interactions are improper to be presented as an direct language expression. For instance, one may say man-hold-fork rather than fork-be holden-man, because the latter phrase is not a general expression in our grammar habits. Obviously, the mutual relations between objects in such phrases do exist as some kinds of latent representation, and we can infer any one from the mutually paired one (*e.g.*, from sky-has-cloud, one can easily predict the mutual one cloud-in-sky and vice versa). It means that visual relationships have great potential to be learned better with the knowledge from the mutual interactions between paired objects. Therefore, we construct a mutual relation space to present the visual relationship triplet, and learn to explore the mutual interaction between them. We call this process mutual relation learning.

For convenient description in following sections, we denote a subject-object pair as (s, o), and vice versa. The (s, o) and (o, s) are different instances for relationship detection. To explore the mutual interactions between paired objects, we propose to construct a siamese network (Chopra, Hadsell, and LeCun 2005) to model paired input instances and optimize the constraint of interaction in mutual relation space. Figure 2 visually summarizes the primary flowchart of our approach, which consists of *visual appearance representation*, *spatial module*, and *mutual relation learning*.

Visual Appearance Representation

Visual appearance directly presents the facets and relationships in an image, which is the key point to visual relationship detection. Given an image, human can easily recognize the entities and infer the interactions between them. Therefore, to obtain a good visual appearance representation is crucial to make machines have a holistic understanding of images like human, and then predict the relationship between recognized objects. Deep convolutional neural networks has been demonstrating excellent performance in visual representation (Ren et al. 2015; Simonyan and Zisserman 2014). In this work, we employ Faster R-CNN (Ren et al. 2015) to detect objects and infer the labels, and then extract $Conv5\_3$ feature map of VGG16 (Simonyan and Zisserman 2014) with ROI-pooling to obtain the visual appearance representation of subject and object according to the detected bounding boxes, *i.e.*, b_s and b_o respectively. These two branches further enter two fully-connected layers to be 4096-d vectors h_1 and h_3, which are then integrated with union feature h_2 of the union area of b_s and b_o as visual cues for relationship inferring. Following (Dai, Zhang, and Lin 2017), we represent the union feature by computing the union area with a small margin to capture surrounding information, extracting $fc7$ of VGG16 trained on ImageNet by resizing the shape to 224×224. To help learn the mutual relation, we make that the features of subject and object change symmetrically with the input order. As Figure 2 depicted, when (s, o) is fed into framework, we concatenate h_1, h_2 and h_3 and map it to a 4096-d feature vector h_p. When we input (o, s), representation h_1 and h_3 interchange positions, hence we obtain a different feature vector $h_{p'}$, which is concatenated with spatial configuration in the next step.

Spatial Module

Spatial configuration is supplementary to visual representation, which describes the position relationship between paired objects. In (Peyre et al. 2017), Peyre et al. divided the coordinate computation into k components, and combined these components by the Gaussian Mixture Model, which has verified that the spatial configuration is helpful to discriminate different types of relationships. There are two ways for spatial feature representation: mask convolutions and coordinate computation. The difference between two configurations is that coordinate computation presents representations for (s, o) and (o, s) independently, while the mask convolution exploits the locations of the paired objects simultaneously, which may help mine the interaction between objects. Therefore, in our work, we extract spatial representation with convolution of binary masks of paired objects. Following previous work (Dai, Zhang, and Lin 2017; Liang et al. 2018), we generate a spatial mask with a binary matrix with the shape of original image, and fill the area of object bounding box with 1 and set the other area as 0. The obtained masks of object pair are down-sampled to 32×32, and concatenated in channel to obtain the input with $2 \times 32 \times 32$. We then devise a 3-layers CNN to extract the compressed spatial configuration as the same with (Dai, Zhang, and Lin 2017).

Mutual Relation Learning

As aforementioned, relationships depend on the involved contexts, and very different contexts can share the same predicate, *e.g.*, `man-eating-bread` and `cow-eating-grass`. Therefore, the mutual interaction between paired objects cannot be well modelled in the discrete label space that ignores the contextual information of predicate. Moreover, one predicate may reasonably co-occur with other predicates, even with itself, in a relationship pair. Such as `near` can be paired with `near`, `next to`, or `on the left side` with different object contexts. To appropriately exploit the mutual relation between paired objects, we deem that the representation should contain all the facets (*objects and interaction*), and be regarded as a triplet phrase representation. We construct a mutual relation space that integrates visual appearance features and spatial configuration to model the latent interaction between object pair. In the mutual relation space, both explicitly paired triplets (`man-wear-shirt` and `shirt-on-man`) and implicitly paired triplets (`man-hold-fork` and `fork-be holden-man`) can be modelled to exploit the mutual relation, even `be holden` is not in the label set and will be predict as `background`. Motivated by the semantic relation transformation with linear vector calculation in (Mikolov et al. 2013), we propose that the latent representations in mutual relation space of paired triplets should satisfy $v_p \approx -v_{p'}$, and can be optimize with a surrogate form as:

$$\mathcal{M}(c_{so}, c_{os}) = \|v_p + v_{p'}\|, \tag{1}$$

where $\|\cdot\|$ is L2 norm of vector. c_{so} and c_{os} denote triplet relationship of (s, o) and (o, s) pairs. To avoid the latent representations of triplets shrinking to zero during the training phase, we add a constant margin to encourage them to keep away from 0 with:

$$\gamma = [\, m - \|v_p\| - \|v_{p'}\| \,]_+, \tag{2}$$

where $[\,\cdot\,]_+$ denotes $max(0, \cdot)$ and m is a constant margin.

We then predict the relationships based on the mutual representations, together with semantic presentation of objects. Specifically, given an image and an input pair (s, o), we denote the latent representations of subject, predicate and object as v_s, v_p and v_o. The compatibility representation of each instance $\mathcal{C}(p|s, o)$ is formulated as follow:

$$\mathcal{C}(p|s, o) = w_p^T [v_s, v_p, v_o], \tag{3}$$

where $[\cdot, \cdot]$ means vector concatenation, and w_p is the parameter to transform the latent representation to the compatible label space of p^{th} predicate, where the index of the highest score represents predicted category of predicate. Vector $[v_s,v_p,v_o]$ is the fused feature of a relationship instance.

Most of existing work treat visual relationship detection as a task of multi-classification, and optimize the model with cross entropy loss with softmax. Zhang et al. (Zhang et al. 2018a) pointed out that visual relationships always associate with several similar classes of relationships. We hope similar classes to obtain the higher score than incorrect ones as well as ground-truth. Therefore, we design a semantic ranking loss to adaptively model the similarity between relationships as:

$$L_{so} = \sum_{p^- \in P^-, p^+ \in P^+} [\Delta + \mathcal{C}(p^-|s,o) - \mathcal{C}(p^+|s,o)]_+, \quad (4)$$

where L_{so} represents the ranking loss for input (s, o), and $\mathcal{C}(p|s,o)$ is the compatibility representation of prediction score for p^{th} class aforementioned. We use p^+ to denote ground-truth and P^+ to denote the set of all annotated relationships in a batch. p^- is an element of P^-, the set of negative relationship samples which do not appear in the annotations. $[\cdot]_+$ is $max(0, \cdot)$. As a result, the loss function does not stop training until the margin is larger than the margin threshold Δ. For most applications, the Δ is a hyper-parameter and hard to set empirically, because large threshold makes the loss function difficult to converge, while too small one may cause insufficient learning. Therefore, we aim to find an adaptive and reasonable margin that pushes the loss function to discriminate negative and positive instances according to the difference of semantic meaning. For example, the predicted instance `man-near-tree` is supposed to be penalized more mildly than `man-in-tree` under the ground-truth `man-next to-tree`. To achieve this purpose, we employ an adaptive threshold based on the semantic embedding of predicates with:

$$\Delta = 1.0 - \alpha(word(p^-), word(p^+)), \quad (5)$$

where $word(\cdot)$ is the word embedding of predicate and $\alpha(\cdot,\cdot)$ means cosine similarity. We use p^- and p^+ to denote the predicates of negative sample and ground-truth respectively. In our experiments, GloVe model (Pennington, Socher, and Manning 2014) is used as the off-the-shelf tool to embed words to vector space.

To exploit both mutual interaction and semantic relationship, we train our model with the combination of Eq. 1, Eq. 2, Eq. 4 as our final loss function as:

$$L = \frac{1}{2}(L_{so} + L_{os}) + \lambda \sum_{(so,os) \in \Gamma} \mathcal{M}(c_{so}, c_{os}) + \gamma, \quad (6)$$

where $\lambda = 0.005$ is a hyper-parameter and the mutual constraint $\mathcal{M}$ is accumulated over all paired objects Γ in a batch. By this model, knowledge from the mutual relations can be used to constrain the training process and possibly mine reasonable relationships which are not annotated.

For relationship triplet prediction, confidence scores obtained from object detector also provide discriminative information. Objects with low confidence usually fail to form relationships with other objects. Hence during the test phase, the final prediction score consists of confidence scores of object detection and predicate prediction as:

$$\begin{aligned} \mathcal{S}(r|s,o) &= \mathcal{S}(c_s, c_p, c_o|s,o) \\ &= \log[(\mathcal{S}(c_p|c_s, c_o)\mathcal{S}(c_s|s)\mathcal{S}(c_o|o))], \end{aligned} \quad (7)$$

where $\mathcal{S}(c_s|s)$ and $\mathcal{S}(c_o|o)$ represent the confidence of detected objects respectively, and are set as 1.0 for predicate detection task. $\mathcal{S}(r|s,o)$ is the score of a relationship instance r that is computed with softmax Φ as:

$$\mathcal{S}(c_p|c_s, c_o) = \Phi(\mathcal{C}(p|s,o)). \quad (8)$$

Notice that our proposed model possesses good transportability since the core of mutual relation learning is flexible. The performance of existing methods can be enhanced with our model if the latent representations of relationship triplets are extracted appropriately and relations between objects are learned in a mutual way. We will validate and analyze the results of proposed approach in subsequent section.

Experiment

Datasets and Metrics

We conduct experiments on two common datasets, the *Visual Relationship Detection* (VRD) and *Visual Genome* (VG), to evaluate our mutual relation network.

- VRD (Lu et al. 2016): VRD is the first benchmark for visual relationship detection, which contains $4,000$ and $1,000$ images for training and test, respectively. There are $7,701$ types of relationship tuple with 100 types of objects and 70 types of predicates in total. To sum up, VRD contains $37,993$ relationship instances, of which 24.3% $(9,224)$ instances involves in an explicit mutual relationship pair. For fair comparison, we follow the default data split (Lu et al. 2016) in all of our experiments.
- VG (Krishna et al. 2017): VG is a large scale knowledge dataset for VQA, region description, and relationship detection, *etc*. We use *VG v1.4*, the latest version of VG dataset. As relationships are annotated manually, some annotations may not conform to the fact. Hence extremely infrequent relationships are removed during preprocessing. Totally, we use $77,761$ images that contains 150 types of objects and 50 types of predicates. There are $774,167$ relationship instances with $196,238$ (25.3 percentage) instances that are with explicitly paired relationships. We randomly split the dataset with $62,253$ images for training and $15,508$ images for test.

In the experiments, we choose two kinds of tasks to evaluate the capability of our model: **Predicate Detection** and **Relationship Detection**. For predicate detection, we input an image and a set of object pairs with ground-truth labels and bounding-boxes, and then predict possible predicates between these known object pairs. For relationship detection, we firstly detect the locations and categories of objects with confidence in an image, then predict possible relationship triplets. This task can be regarded as a combination of object detection and predicate detection, the two-stage paradigm as aforementioned. Following previous work (Lu et al. 2016; Dai, Zhang, and Lin 2017; Liang et al. 2018), we use Recall@K as the performance metric (specifically, recall@50 and recall@100 are employed in our evaluation). Recall@K is the fraction of ground-truth relationships that are correctly recalled within the top K predictions. Metric *mean average precision* (mAP) is not employed because it pessimistically omit some positive predictions due to incompleteness of annotations.

Implementation Details

We train 3 epochs for VG and 8 epochs for VRD respectively on a single GPU, GeForce GTX TITAN X. The constant

Methods	Predicate Det.		Relationship Det.	
	R@50	R@100	R@50	R@100
Language priors	48.87	48.87	13.86	14.70
VTransE	44.76	44.76	14.07	15.20
Context-Aware	53.59	53.59	15.63	17.39
VRDS	51.50	51.50	14.31	15.77
MR-NET (ours)	**61.19**	**61.19**	**16.71**	**17.58**

Table 1: Comparison with several state-of-the-art methods on VRD.

Methods	Predicate Det.		Relationship Det.	
	R@50	R@100	R@50	R@100
VTransE	62.63	62.87	5.52	6.04
VRDS	58.72	58.72	8.28	8.28
MR-NET (ours)	**65.27**	**66.45**	**12.64**	**14.32**

Table 2: Comparison with several state-of-the-art methods on VG.

margin term m in the mutual constraint is set as 0.5. We initialize the parameters of VGG16 with weights pre-trained on ImageNet (Deng et al. 2009). The weights of spatial model and subsequent fully-connected layers are initialized randomly. During visual relationship detection, only detected objects with the confidence greater than 0.5 are retained to be paired for further prediction. During training phase, all paired instances in an image are regarded as a mini-batch for one iteration. We choose Adam algorithm to optimize our model, and set the learning rate as 0.00001 initially that is decreased with the scale of 10 at the beginning of 3rd and 8th epoch respectively. In the last epoch, only mutual relationship instances are used to fine-tune the network.

Comparative Results

To verify the effectiveness of our model, we first compare our model with several state-of-the-arts. (1)**Language priors** (Lu et al. 2016) utilizes visual appearance feature to train models for objects and predicates separately, and then fine-tune the network with language priors. (2)**VTransE** (Zhang et al. 2017a) models relationship triplets as vector translation embedding in a low-dimensional space and trains the model end-to-end. (3)**Context-Aware** (Zhuang et al. 2017) integrates contextual information with predicate classifier, which makes the prediction more reasonable. (4)**VRDS** (Zhu, Jiang, and Li 2017) takes advantages of a variety of spatial distributions to infer visual relationships.

As shown on Table 1 and Table 2, we can observe that our proposed MR-NET outperforms other methods in both predicate detection and relationship detection tasks. We note that both Language Priors and VRDS utilize the language information, but in different ways, for relationship detection. The former fine-tunes pre-trained model with language priors, while the latter trains the entire framework that consists of visual, language and region models jointly and achieves remarkable improvements on both tasks. Context-aware models interactions adaptively with involved entities and outperforms other baselines, which means that explicitly integrating interaction and context makes predicate classifier more flexible and robust. Our MR-NET predicts relationship by considering the contextual information with implicit mutual relation learning, which exhibits superior performance compared to explicit context combination. We also note that though our approach employs much less modules (*i.e.*, less spatial configuration and no language module), which also surpasses VRDS with tremendous advance. It indicates that mutual relation between paired objects could provide abundant information for visual relationship detection.

Due to the long tail distribution of the dataset, many types of relationship have a few number of instances. Therefore, a well-designed model is supposed to learn the relationships efficiently with scarce annotations and even no annotations. In our model, knowledge from input pairs are utilized jointly for mutual relation learning. Besides, semantic similarities in the ranking loss help penalize different negative answers distinguishingly. As shown on the Figure 3, we visualize several samples on the VRD dataset. Our model can not only makes plenties of reasonable predictions (in the green boxes), but also mine some correct relationships (in the blue boxes) which do not appear in the ground-truth annotation set. Notice that the images that contain newly discovered relationships usually have mutual relationships among object pairs. It means mutual interaction between paired objects has been well explored for visual relationship detection.

Component Analyses

To investigate the effect of our semantic ranking loss and mutual constraint, we conduct predicate detection with our vanilla framework and its variants on VG.

- **Baseline** Our basic model with visual appearance and spatial configuration that removes the mutual constraint. Cross entropy with softmax is set as the loss function.
- **Sem-Rank** All the same as baseline, except that the loss function is replaced by our semantic ranking loss.
- **Mutual** Sem-rank with mutual relation learning (final).

From Table 3, we can see that Sem-Rank achieves better performance compared with Baseline. It demonstrates that semantic ranking is capable of capturing the similarity between predicates and optimizing with different penalties. In the third row, our semantic ranking jointly optimizes (s, o) and (o, s) pair in the siamese network with mutual relation learning, and improves the performance with a remarkable gap. This means that our mutual relation learning enables the model to infer the relationship learning the mutual interaction from each other as well as exploiting the predicate in single track.

Transportability of Mutual Relation Learning

As aforementioned, the proposed mutual constraint is flexible, and can adapt to other visual relationships detection models without introducing any parameters. To verify the transportability of our mutual relation learning, we inject

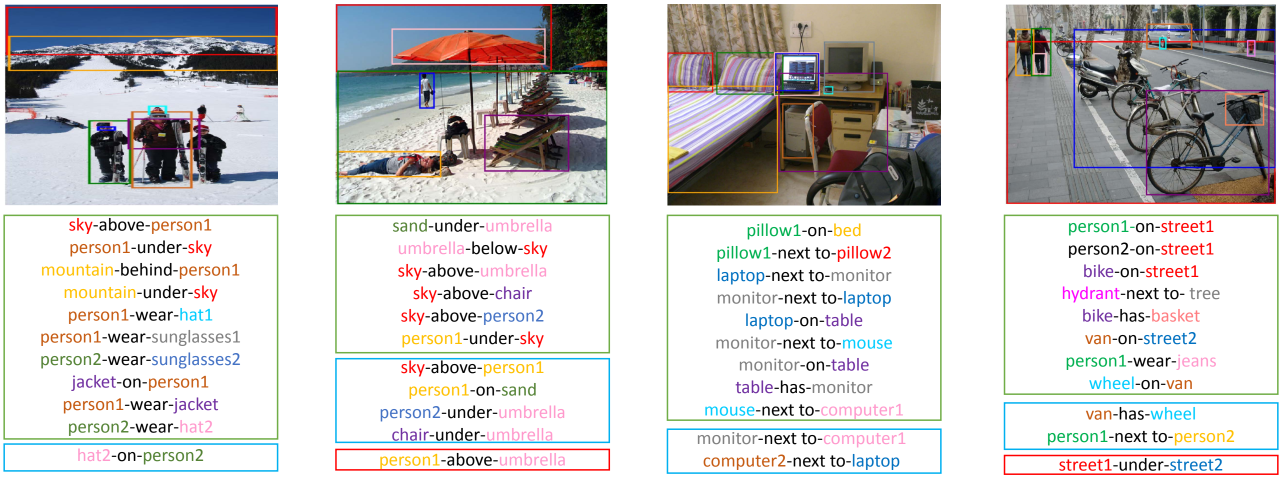

Figure 3: Exhibition of several examples of our MR-NET. Correct predictions and wrong predictions are marked in green and red boxes, respectively. Possibly correct predictions that miss in the ground-truth annotation set are marked in blue boxes.

	Predicate Det.	
	R@50	R@100
Baseline	61.78	62.53
Sem-Rank	62.59	63.74
Mutual	**65.27**	**66.45**

Table 3: Results for component analyses on VG

Models	Predicate Det.		Relationship Det.	
	R@50	R@100	R@50	R@100
DR1	**80.78**	81.90	**16.94**	20.20
DR2	77.73	79.84	16.05	19.82
DR2-A	78.21	79.76	16.06	20.12
Mutual	80.28	**81.98**	16.27	**20.52**

Table 4: Results for transportability analysis of mutual constraint on VRD. **DR1** is the results reported in (Dai, Zhang, and Lin 2017), which trained each module independently, and then fine-tuned all the modules jointly. **DR2** and **DR2-A** indicate DR-Nets that we trained all modules directly and augment the training set with random shift or scaling, respectively. **Mutual** means DR2-A with mutual constraint.

the mutual constraint to one of the state-of-the-arts, DR-Net (Dai, Zhang, and Lin 2017), and optimize together with the original softmax losses. We revise their code to a siamese network and optimize both relationships between a pair of objects simultaneously. Table 4 exhibits the experimental results on VRD. The authors mentioned that they trained each module in DR-Net individually and then fine-tuned jointly. To balance the performance and time consuming (about a quarter of theirs), we train all the modules of DR-Net simultaneously, and obtain comparable results (the second row). We then fine-tune siamese network for mutual relation constraint based on the weights of DR-Net that we trained. To retain more contextual information of relationships, we employ the first layer without activation in their relational network as the representation in mutual relation space. Our mutual relation networks achieve considerable improvements in both predicate and relationship detection tasks, even better than the pre-train and fine-tune training procedure in (Dai, Zhang, and Lin 2017). We also note that their training procedure performs well in $Recall@50$, which means that the pre-training may bring in subtle improvements for fine-tuning.

From our previous statistic, there are only about $10K$ relationship instances that are involved in mutual paired relations. To augment training samples, we randomly change the bounding boxes (including shifting and scaling) with 5 to 10 percent of the width or height, and no more than 20 pixels. We note this operation not only augments the training data, but also introduces reasonable noise, which makes the model more robust for relationship detection while does not weaken the predicate detection.

Conclusion

In this paper, we presented a novel approach for visual relationship detection, which employed a siamese network, and explored the mutual interaction between paired objects. Our model integrated visual and spatial information of objects to learn the mutual representations for mutual interaction exploration and predicates prediction. To discriminatively optimize the predictions by semantic information, we also applied word embeddings at the label level and ranked predicates with previous feature presentations. The experimental results compared with other methods on VRD and VG demonstrated the effectiveness and superiority of our approach. Moreover, the proposed mutual relation learning also exhibited flexible transportability to other framework.

Acknowledgments

This work was supported in part by the Natural Science Foundation of China (No. 61572108 and No. 61632007).

References

Bin, Y.; Yang, Y.; Zhou, J.; Huang, Z.; and Shen, H. T. 2017. Adaptively attending to visual attributes and linguistic knowledge for captioning. In *ACM Multimedia*, 1345–1353.

Bin, Y.; Yang, Y.; Shen, F.; Xie, N.; Shen, H. T.; and Li, X. 2018. Describing video with attention-based bidirectional lstm. *TCYB*.

Chopra, S.; Hadsell, R.; and LeCun, Y. 2005. Learning a similarity metric discriminatively, with application to face verification. In *CVPR*, volume 1, 539–546. IEEE.

Dai, B.; Zhang, Y.; and Lin, D. 2017. Detecting visual relationships with deep relational networks. In *CVPR*, 3298–3308. IEEE.

Deng, J.; Dong, W.; Socher, R.; Li, L. J.; Li, K.; and Li, F. F. 2009. Imagenet: A large-scale hierarchical image database. In *CVPR*, 248–255.

Duan, G.; Ai, H.; Cao, S.; and Lao, S. 2012. Group tracking: exploring mutual relations for multiple object tracking. In *ECCV*, 129–143. Springer.

Gao, L.; Guo, Z.; Zhang, H.; Xu, X.; and Shen, H. T. 2017. Video captioning with attention-based lstm and semantic consistency. *TMM* 19(9):2045–2055.

Gao, L.; Zeng, P.; Song, J.; Liu, X.; and Shen, H. T. 2018. Examine before you answer: Multi-task learning with adaptive-attentions for multiple-choice vqa. In *ACM Multimedia*, 1742–1750.

Han, C.; Shen, F.; Liu, L.; Yang, Y.; and Shen, H. T. 2018. Visual spatial attention network for relationship detection. In *ACM Multimedia*, 510–518.

Helbing, D., and Molnar, P. 1995. Social force model for pedestrian dynamics. *Physical review E* 51(5):4282–4286.

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. *IJCV* 123(1):32–73.

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

Li, Y.; Ouyang, W.; Zhou, B.; Wang, K.; and Wang, X. 2017. Scene graph generation from objects, phrases and region captions. In *ICCV*, 1261–1270.

Li, J.; Zhu, L.; Huang, Z.; Lu, K.; and Zhao, J. 2018. I read, i saw, i tell: Texts assisted fine-grained visual classification. In *ACM Multimedia*, 663–671.

Liang, K.; Guo, Y.; Chang, H.; and Chen, X. 2018. Visual relationship detection with deep structural ranking. In *AAAI*, 7098–7105.

Liang, X.; Lee, L.; and Xing, E. P. 2017. Deep variation-structured reinforcement learning for visual relationship and attribute detection. In *CVPR*, 4408–4417. IEEE.

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

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. In *NIPS*, 3111–3119.

Peng, L.; Yang, Y.; Bin, Y.; Xie, N.; Shen, F.; Ji, Y.; and Xu, X. 2018. Word-to-region attention network for visual question answering. *MTAP* 1–16.

Pennington, J.; Socher, R.; and Manning, C. 2014. Glove: Global vectors for word representation. In *EMNLP*, 1532–1543.

Peyre, J.; Laptev, I.; Schmid, C.; and Sivic, J. 2017. Weakly-supervised learning of visual relations. In *ICCV*, 5179–5188.

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

Sadeghi, M. A., and Farhadi, A. 2011. Recognition using visual phrases. In *CVPR*, 1745–1752. IEEE.

Simonyan, K., and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. In *ICLR*.

Xu, D.; Zhu, Y.; Choy, C. B.; and Fei-Fei, L. 2017. Scene graph generation by iterative message passing. In *CVPR*, 5410–5419.

Zellers, R.; Yatskar, M.; Thomson, S.; and Choi, Y. 2018. Neural motifs: Scene graph parsing with global context. In *CVPR*, 5831–5840.

Zhang, H.; Kyaw, Z.; Chang, S.-F.; and Chua, T.-S. 2017a. Visual translation embedding network for visual relation detection. In *CVPR*, 5532–5540.

Zhang, Z.; Shao, L.; Xu, Y.; Liu, L.; and Yang, J. 2017b. Marginal representation learning with graph structure self-adaptation. *TNNLS*.

Zhang, J.; Kalantidis, Y.; Rohrbach, M.; Paluri, M.; Elgammal, A.; and Elhoseiny, M. 2018a. Large-scale visual relationship understanding. *arXiv preprint arXiv:1804.10660*.

Zhang, Z.; Liu, L.; Shen, F.; Shen, H. T.; and Shao, L. 2018b. Binary multi-view clustering. *TPAMI*.

Zhou, H.; Ouyang, W.; Cheng, J.; Wang, X.; and Li, H. 2018. Deep continuous conditional random fields with asymmetric inter-object constraints for online multi-object tracking. *IEEE TCSVT*.

Zhu, Y., and Jiang, S. 2018. Deep structured learning for visual relationship detection. In *AAAI*, 7623–7630.

Zhu, F.; Liu, L.; Xie, J.; Shen, F.; Shao, L.; and Fang, Y. 2018. Learning to synthesize 3d indoor scenes from monocular images. In *ACM Multimedia*, 501–509.

Zhu, Y.; Jiang, S.; and Li, X. 2017. Visual relationship detection with object spatial distribution. In *ICME*, 379–384. IEEE.

Zhuang, B.; Liu, L.; Shen, C.; and Reid, I. 2017. Towards context-aware interaction recognition for visual relationship detection. In *ICCV*, 589–598. IEEE.