Region-Object Relevance-Guided Visual Relationship Detection

Yusuke Goutsu
goutsu@nii.ac.jp

National Institute of Informatics
Tokyo, Japan

Abstract

In this paper, we address the problem of visual relationship detection, which requires joint image and language understanding to predict semantic connections between detected objects. Most of the previous works in this domain have applied visual classification methods based on extracted visual features to solve this problem. However, each type of relationship has various object combinations and each object pair has diverse interactions; thus, considering all possible relationships is difficult and expensive. In this paper, we mainly propose a region-object relevance-guided framework that identifies regions related to the labels of object pairs in images, thereby decreasing the number of possible relationships to be considered for efficiency. Furthermore, we construct a semantic space with an exploratory mechanism that weights informative regions as relevant. Our entire network is trained end-to-end using a multi-task loss function for estimating visual relationships. Finally, we evaluate our method on two public large-scale datasets. Our method achieves remarkably high performance levels on both datasets that are better than or comparable to those of state-of-the-art methods under conditions using only relationship annotations from the dataset. In addition, our proposed method facilitates prediction at speeds below 150 ms per image. Thus, our method is well-suited for real-time applications.

1 Introduction

In the history of computer vision, the capacity for feature learning and transfer learning of convolutional neural networks (CNNs) [18] has gradually improved image processing tasks. For example: *image classification* [14, 16, 34] labeled each image with a particular class, *object detection* [8, 23, 29, 30] labeled each region using a method that guides the search for object instances in an image, and *semantic or instance segmentation* [5, 15, 24, 26, 41] labeled each image pixel. As an important aspect of image processing, object detection has become a popular research field. This task involves not only classifying every object in an image, but also localizing each object by adding bounding boxes. This makes object detection significantly more difficult than image classification. With the development of high-speed GPU for parallel computing and crowd sourcing technology for collecting image annotations, fast and accurate object detection is becoming possible despite the large number of existing object categories.

More advanced except for semantic or instance segmentation is the task of *visual relationship detection (VRD)*. Visual relationships represent the visually observable interactions between the objects in images. Each relationship has three elements: *subject (sbj)*, *predicate (pred)*, and *object (obj)*. Thus, the relationship can be represented in a triplet form <*sbj-pred-obj*>, such as <*person-ride-horse*>. In addition, VRD involves classifying and localizing the objects of the <*sbj, obj*> pairs and the interactions between them as *preds*.

However, visual relationships are not a new concept. In previous research, Sadeghi *et al.* [31] proposed visual phrase detection (VPD) to recognize multiple objects that interact with each other as phrases. Visual phrases represent the semantic connections between interacting objects in bounding boxes, similar to the triplets described above. Each type of phrase is considered as a relationship category in the VPD framework. This is a natural approach for this task, but the increasing number of objects and predicates makes it difficult to treat VPD as a classification task. For example, consider a visual phrase dataset with $N = 100$ object categories and $K = 70$ predicates; there are $N^2K = 700k$ relationships in total. The task in [31] is conducted in a restricted context, where the number of possible relationships is moderate. However, the VRD frameworks can handle this problem by dividing the entire process into object classification and predicate classification, owing to the advent of object detection. Lu *et al.* [25] attempted to construct a linear model by optimizing the objective function that combines the visual appearance and language modules for relationship prediction. Furthermore, they introduced a visual relationship dataset (VR) to study the learning of a large number of visual relationships from given images using the linear model. The results in [25] suggest that predicates cannot be reliably predicted with a linear model using only visual modules ([25] reports Recall@100 of only 7.11% for their visual prediction). Although the visual features in [25] are extracted from the union of the *sbj* and the *obj* boxes, predicates are predicted without consideration of the relevance between object proposals and the labels of object pairs in images.

In this paper, we mainly propose a region-object relevance-guided framework that looks for regions related to the labels of object pairs in images. More specifically, we utilize the relevance between the visual features of object proposals and the language features of the labels of object pairs by constructing an inner product space. The relationships can be predicted by focusing on relevance regions, as unexpected relationships can largely be excluded. Figure 1 shows the illustration of predictions *preds* from relevance regions derived by weighting. High relevance regions (bright areas) are marked with bounding boxes and the corresponding *preds* are output from the weighted regions. In addition, the entire network for our method is trained in an end-to-end manner with a multi-task loss function for estimating visual relationships. This single network architecture reduces the complexity during training and testing while improving the overall relationship detection accuracy.

In summary, our main contributions can be summarized as follows: 1) We propose a region-object relevance guided framework; a simple but efficient end-to-end deep network using an inner product space with a weighting mechanism for relevance regions in images. To the best of our knowledge, such an architecture is the first of its kind for VRD; 2) Our method jointly optimizes object detection and relationship detection using a multi-task loss function; 3) When using only relationship annotations, our method outperforms most of existing methods for VRD in terms of detection accuracy and processing time.

2 Related Work

As the intermediate task connecting vision and language, VRD [25] is rooted in object detection, which has been investigated for several years. Convolutional neural networks [18]

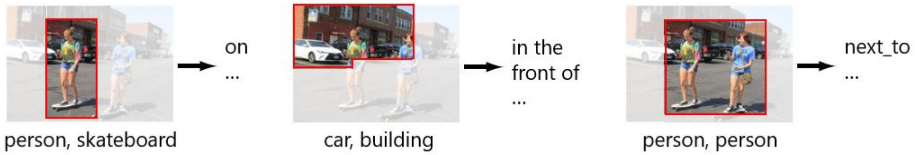

Figure 1: Illustration of predicting *preds* from relevance regions derived by weighting. High relevance regions (bright areas) are marked with bounding boxes and the corresponding *preds* are output from the weighted regions. The *preds* are arranged in descending order of likelihood.

were first introduced as the R-CNN [9] for object detection, which processes every object proposal without sharing computation, thereby resulting in time-consuming delays. Fast R-CNN [8] was proposed to share convolutional layers among object proposals and overcome this problem. A region of interest (RoI) pooling layer was then designed to shape convolutional features to the same size in [8]. Ren *et al*. proposed Faster R-CNN [30] by utilizing a region proposal network (RPN) to generate high-quality proposals and adopted [8] to perform RoI-wise classification and refinement. We use [30] as an object detection module due to its superior performance. Note that our method cannot be simply considered as appending a relationship prediction layer to Faster-RCNN. In fact, our method uses an inner product space with a weighting mechanism for relevance regions and jointly optimizes object detection and relationship detection using a multi-task loss function.

Modeling the relationships which represent interactions between multiple objects is not a new concept in the current literature. There have been a number of studies that explore the use of various relationships (e.g. co-occurrence statistics, spatial relationships, and action or human-object interactions). For example, [27] used object co-occurrence statistics derived from datasets to incorporate semantic context into object categorization tasks. [7, 11, 12] attempted to learn spatial relationships between objects to improve individual object categorization. [13, 32] also used spatial relationships in an operator's spoken instructions to control robots for object picking. Action is one of the most important relationships, considering that *sbj* is a human and *pred* is a verb. In recent years, action recognition in images [10] has been a popular research field, and various datasets [1, 2, 33, 37] have been constructed to focus on human-object interactions. However, most of these works were used for leveraging the relationships for other tasks, such as object segmentation [19], object detection and pose estimation [36], scene classification and object detection [3], action, pose, and object detection [6], and action retrieval [28]. Note that our work is essentially different from these works as we aim to provide a method dedicated to generic relationships, such as actions (e.g. “kick"), relative positions (e.g. “above"), functionals (e.g. “with"), and comparisons (e.g. “taller than").

Lu *et al*. [25] first formalized VRD as a task and provided a large-scale dataset that included generic relationships. Recent works in this field [20, 21, 22, 38, 40, 42] have followed this active research topic. Li *et al*. [20] proposed a CNN with a phrase-guided message passing structure (PMPS) to simultaneously predict relationship components within a single deep network. In contrast to this top-down design, we follow a bottom-up design, which detects objects and then predicts the possible relationships among them. Yu *et al*. [38] proposed an end-to-end deep neural network that absorbs internal (training annotations) and

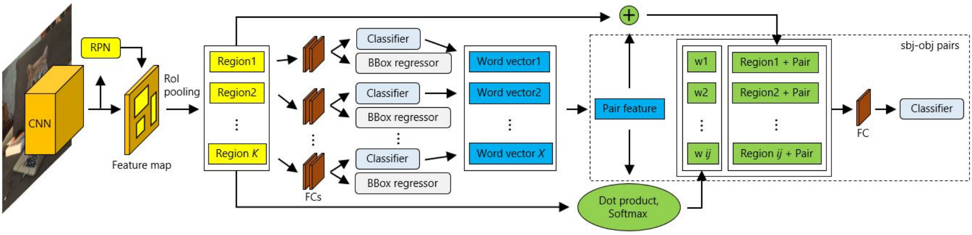

Figure 2: Overview of our system composed of the object detection module and visual relationship detection module. The former module output detected object labels and locations in an image using the classifier and the bounding box regressor. The object labels are converted into corresponding word-embedding vectors. The latter output predicted predicate labels from the classifier by constructing an inner product space between the image region vectors related to each object pair and the pair feature vector concatenating the paired word vectors and then using the weighted and summed vectors.

external (public text on the Internet, i.e. Wikipedia) linguistic knowledge using a teacher-student distillation framework to regularize the learning process. Finally, adding spatial features obtained from pairs of detected bounding boxes improved prediction performance. On the other hand, Zhang *et al.* [40] conducted research on weakly supervised learning and captured the spatial context of relationships by using a position-role-sensitive score map with pairwise RoI pooling. Compared with the above works, we aim to effectively and simply implement VRD using only relationship annotations from the dataset.

3 Visual Relationship Prediction

Figure 2 shows an overview of our system. We propose a probabilistic model to predict the *pred* jointly with the *sbj* and the *obj*. Let $\boldsymbol{s},\boldsymbol{o} = \{s_m, o_m; m = 1,\ldots,N\}$ and $\boldsymbol{p} = \{p_m; m = 1,\ldots,N\}$ denote a set of <*sbj, obj*> pairs and a set of *preds* representing their interactions. Object pairs and their interaction types are jointly predicted in triplet form, and we derive the optimal prediction by maximizing the following joint relationship probability.

$$< \boldsymbol{s}^*, \boldsymbol{p}^*, \boldsymbol{o}^* >= \arg\max_{\boldsymbol{s},\boldsymbol{p},\boldsymbol{o}} P(\boldsymbol{s},\boldsymbol{p},\boldsymbol{o}) \tag{1}$$

Next, $P(\boldsymbol{s},\boldsymbol{p},\boldsymbol{o})$ is divided into the probabilities estimated by the object detection module and the VRD module.

$$P(\boldsymbol{s},\boldsymbol{p},\boldsymbol{o}) = P(\boldsymbol{s},\boldsymbol{o})P(\boldsymbol{p}|\boldsymbol{s},\boldsymbol{o}) \tag{2}$$

where $P(\boldsymbol{s},\boldsymbol{o})$ is the joint probability of predicted *sbj* and *obj* obtained from the object detection module (i.e., Faster R-CNN), $P(\boldsymbol{p}|\boldsymbol{s},\boldsymbol{o})$ is the conditional probability of *preds*, given *sbjs* and *objs* obtained from the VRD module. More precise descriptions about each module are described below.

3.1 Object Detection Module

The left half of Figure 2 corresponds to the object detection module, which is utilized to localize objects (*sbj* and *obj*) and provide their classification labels in the images. Faster R-

CNN [30] is used for this task because of its high accuracy and efficiency. If the predictions of *sbj* and *obj* are independent of each other, the joint probability $P(\boldsymbol{s},\boldsymbol{o})$ can be rewritten as:

$$P(\boldsymbol{s},\boldsymbol{o}) = P(\boldsymbol{s})P(\boldsymbol{o}) \tag{3}$$

where $P(\boldsymbol{s})$, $P(\boldsymbol{o})$ are individual output probabilities of *sbj* and *obj* obtained from the object detection module.

3.2 Visual Relationship Detection Module

The right half of Figure 2 corresponds to the VRD module. In this process, the object pairs of <*sbj, obj*> are selected from among object labels obtained from the object detection module. Each selected label is converted into its word-embedding vector to represent the semantic meaning of each object. The word-embedding vectors of *sbj* and *obj* are concatenated to learn the joint representations of the object pairs (described as pair features in the figure). These operations can be written as:

$$\boldsymbol{l} = \mathrm{concat}(\mathrm{w2v}(s_m), \mathrm{w2v}(o_m)) \tag{4}$$

where concat$(\cdot, \cdot)$ indicates the concatenation of two target vectors, and w2v$(\cdot)$ indicates the conversion of the target label into the corresponding word-embedding vector. This semantic representation can be obtained by the learning model, which can predict surrounding words from the target within a context window.

Let $\boldsymbol{V} = \{\boldsymbol{v}_{ij} | \boldsymbol{v}_{ij} \in \mathbb{R}^M; i, j = 1, ..., K\}$ (i, j refers to the index of *sbj* and *obj*, K is the number of object proposals, and M is the dimension number of visual features) denote visual features reshaped into fixed size, the image regions related to object pairs are extracted by learning the semantic connection models between visual features corresponding to the integrated regions of the object pairs $\boldsymbol{v}_{ij}$ and the language features of their labels (i.e., the fixed-length word vector $\boldsymbol{l} \in \mathbb{R}^L$). The weight of the object proposals related to object pairs is expressed as:

$$\begin{aligned} g_{ij} &= (\boldsymbol{A}\boldsymbol{v}_{ij} + \boldsymbol{b}_A)^T(\boldsymbol{B}\boldsymbol{l} + \boldsymbol{b}_B) && (5) \\ w_{ij} &= \frac{\exp(g_{ij})}{\sum_m \sum_n \exp(g_{mn})}, \boldsymbol{W} = \{w_{ij} | i, j = 1, ..., K\} && (6) \end{aligned}$$

where $\boldsymbol{A} \in \mathbb{R}^{S \times M}$, $\boldsymbol{B} \in \mathbb{R}^{S \times L}$ are the projection matrices embedding $\boldsymbol{v}_j$ and $\boldsymbol{l}$ into the inner product space, and $\boldsymbol{b}_A$, $\boldsymbol{b}_B \in \mathbb{R}^S$ (S is the dimension number of inner product space) are the bias terms of affine projection (Eq. (5)). Note that $\boldsymbol{A}$ and $\boldsymbol{B}$ control which visual features have high correlation with specific pair features. The weight w_{ij} is obtained by applying the softmax function to g_{ij} for normalization ($\sum_i \sum_j w_{ij} = 1$, $w_{ij} > 0$). The relevance weights of object proposals corresponding to object pairs are represented as $\boldsymbol{W}$ (Eq. (6)). This matrix contains the results of the inner products between each visual feature and each language feature; thus, each value in $\boldsymbol{W}$ measures the similarity between an integrated region and both labels of object pair. The visual and language features are then multiplied by the relevance weights w_{ij} and summed over all object pairs to generate a region-object relevance vector $\boldsymbol{S}_W$.

$$\boldsymbol{S}_W = \sum_{i=1}^{K} \sum_{j=1}^{K} w_{ij} \boldsymbol{v}_{ij}^l \tag{7}$$

where $\boldsymbol{v}^{l}_{ij}$ is the vector concatenating $\boldsymbol{v}_{ij}$ and $\boldsymbol{l}$. Finally, this region-object relevance vector is used to predict the *pred* for the given image regions and the object labels of *sbj* and *obj*. The final predictions that a set of p_m can be obtained from that of s_m and o_m are expressed as:

$$P(\boldsymbol{p}|\boldsymbol{s},\boldsymbol{o}) = \mathrm{softmax}(\boldsymbol{W_P} \cdot f(\boldsymbol{S_W}) + \boldsymbol{b_P}) \tag{8}$$

where f is the activation function called ReLU, and $\boldsymbol{b_P} \in \mathbb{R}^{C_p}$ is the bias term. (C_p is the number of possible relationships). $\boldsymbol{W_P} \in \mathbb{R}^{C_p \times M}$ converts so that the dimension of the relevance region vector is the same as C_p.

3.3 Multi-task Loss

We use a multi-task loss L to jointly train the object label, the bounding box position, and the relationship label. The prediction targets for each object proposal are a ground-truth object label u, a ground-truth bounding box offset $\boldsymbol{b}$, and a ground-truth relationship label v provided from the training dataset. More specifically, the loss function can be written as:

$$L(\boldsymbol{p}_o, u, \boldsymbol{b}, \boldsymbol{t}^u, \boldsymbol{p}_r, v) = L_{obj}(\boldsymbol{p}_o, u) + I[u \geq 1] L_{bbox}(\boldsymbol{b}, \boldsymbol{t}^u) + L_{rel}(\boldsymbol{p}_r, v) \tag{9}$$

where the first loss $L_{obj}(\boldsymbol{p}_o, u)$ is the multinomial cross entropy loss for the object classification, and the second loss $L_{bbox}(\boldsymbol{t}^u, \boldsymbol{b})$ is the smooth L_1 loss between the regressed box offset $\boldsymbol{t}^u$ (corresponding to the ground-truth object label u) and the ground-truth box offset $\boldsymbol{b}$. $I[u \geq 1]$ is an indicator function, which outputs 1 when $u \geq 1$ and 0 vice versa. Thus, we define the box location loss L_{bbox} on positive object proposals only; the object classification loss L_{obj} is defined on both positive and negative object proposals. The third loss $L_{rel}(\boldsymbol{p}_r, v)$ is the multinomial cross entropy loss for the relationship detection, and is defined as follows:

$$L_{rel}(\boldsymbol{p}_r, v) = -\log p^v_r \tag{10}$$

4 Experiment

4.1 Experimental Settings

We evaluated our approach on two large-scale public datasets, which contained the annotations of visual relationship in triplet form <*sbj-pred-obj*> and gave the ground-truth object labels with their bounding boxes. We used: (1) **VR** [25]: the Visual Relationship dataset, which contains 5,000 images with 100 objects and 70 predicates. In total, there are 37,993 visual relationship instances with 6,672 triplet types. We followed the same train/test split as in [25]; i.e., 4,000 images for training and 1,000 images for testing. This dataset includes a zero-shot testing set that contains relationships that never occur in the training set to evaluate generalization capabilities. (2) **VG** [17]: the Visual Genome Version 1.2 dataset contains 108K images and 998K relationship annotations that belong to 74,361 triplet types. This dataset is annotated by crowd workers; thus, the triplet labels are noisy (e.g., the spellings of nouns and verbs are inconsistent). Therefore, we followed the same filtering as in [21] to preprocess the dataset; specifically, we selected the top 150 frequent objects and top 50 predicates. After preprocessing, 95,952 images remained. We randomly divided these into two subsets: 5,000 images as the testing subset and the remaining images as the training subset.

Table 1: Component analysis of the proposed method on VR and VG using different factors. We used top 100 recall as the evaluation metric. RR, CF1, CF2, and SC denote whether to use region-object relevance, co-occurrence frequency of <*sbj-obj*> pair, co-occurrence frequency of <*sbj-pred*> and <*pred-obj*> pairs, and spatial configuration, respectively.

Dataset	Method				Predicate Det.	Phrase Det.	Relationship Det.
	RR	CF1	CF2	SC	R@100 (%)	R@100 (%)	R@100 (%)
VR	✓				71.85	16.38	11.73
	✓	✓			74.69	18.92	13.46
	✓	✓	✓		77.65	21.56	14.58
	✓	✓	✓	✓	**82.10**	**23.50**	**15.98**
VG	✓				67.52	8.39	6.40
	✓	✓			71.57	10.95	8.33
	✓	✓	✓		74.32	13.11	9.89
	✓	✓	✓	✓	**77.18**	**14.96**	**10.95**

The VRD involves detecting objects and predicting their relationships. We evaluated our approach using three conventional tasks [25]: (1) **Predicate detection**: The input was an image and a set of ground-truth boxes of *sbj* and *obj* with corresponding labels. The output was a set of *preds* representing the relationships between them. The purpose of this task was to predict *preds* without relying on object detection, where the labels and locations of the *sbj* and *obj* were given. (2) **Phrase detection**: The input was an image and a set of ground-truth boxes of *sbj* and *obj*. The output was a set of triplets and union bounding boxes, which covered the whole triplet. The purpose of this task was to predict <*sbj-pred-obj*> triplets and localize each whole triplet with a union bounding box. A prediction was considered as correct if *sbj*, *pred*, and *obj* were correctly classified and the IoU between the predicted union box and the ground-truth was greater than 0.5. (3) **Relationship detection**: Given an input image, a set of triplets and bounding boxes for *sbj* and *obj* was predicted. The purpose of this task was to predict <*sbj-pred-obj*> triplets and localize each *sbj* and *obj* with a bounding box. This is similar to the task above; however, the IoU between the predicted boxes and their ground-truth boxes were simultaneously greater than 0.5.

Following [25], we used Recall@100 (R@100) as the evaluation metric in our experiments. R@K computes the fraction of times a correct relationship was predicted in the top K confident relationship predictions for an image. As discussed in [25], we did not use the mean average precision (mAP) as it is not a proper metric and cannot exhaustively annotate all possible relationships. Thus, if some correct relationships are missing or incomplete, they will mistakenly penalize the detection as they will not be ground-truths.

We initialized our model on the ImageNet pretrained VGG-16 network provided by Caffe, and optimized the entire network parameters using the stochastic gradient descent (SGD) algorithm. We set the base learning rate at 0.001 and decreased it in stages. During training, the proposal having a minimum 0.5 IoU with ground-truth was regarded as the correct localization; the non-maximum suppression (NMS) threshold was set to 0.3. We used Gensim Word2Vec to convert object labels to word-embedding vectors. The word vector dimensions were set to 300. All experiments were run on a single GeForce GTX1070 GPU.

4.2 Experimental Results

We first tested various combinations of different factors of the proposed method to improve the performance. Table 1 summarizes the results. The RR factor is the baseline that utilizes

Table 2: Comparison of predicate, phrase, and relationship detections with various state-of-the-art approaches on VR (second row) and VG (third row). We used top 100 recall as the evaluation metric. “-" denotes that a result is not applicable.

Dataset	Method	Predicate Det.	Phrase Det.	Relationship Det.
		R@100 (%)	R@100 (%)	R@100 (%)
VR	LP [25]	47.87	17.03	14.70
	VtransE [39]	44.76	22.42	15.20
	PPR-FCN [40]	47.43	23.15	15.72
	VRL [22]	-	22.60	20.79
	DR-Net [4]	81.90	23.45	**20.88**
	Ours	**82.10**	**23.50**	15.98
VG	LP* [25]	33.32	12.64	0.14
	ISGG* [35]	62.74	20.23	9.91
	MSDN** [21]	66.41	**21.82**	10.51
	Ours	**77.18**	14.96	**10.95**

* These results of LP and ISGG were reported in [21].

** This result output from the method without using caption proposals and caption annotations.

Table 3: Comparison of the average processing time per image of the relationship prediction phase on VR (second row) and VG (third row). We excluded the proposal generation time cost by RPN.

Dataset	Method	Relationship Det.
		Time (ms)
VR	VtransE* [39]	270
	PPR-FCN* [40]	150
	DR-Net** [4]	113
	Ours	**108**
VG	MSDN** [21]	172
	Ours	**146**

* These results listed just for reference were reported in [40].

** These results of DR-Net and MSDN were output by ourselves on the same PC.

only the region-object relevance, which estimates the conditional probabilities defined by Eq. (8). This result indicates that the VRD task could not be effectively completed using visual appearances related to object pairs alone, as there are many object pairs that must be considered. It was necessary to select correct triplets among the object-pair candidates by excluding any unexpected triplets. The CF1 and CF2 factors excluded unexpected triplets from among the object-pair candidates by using co-occurrence frequencies for the <*sbj-obj*> pair and co-occurrence frequencies for the <*sbj-pred*> and <*pred-obj*> pairs. By adding these factors, we can see that statistical relationships among triplet components helped to improve recall by 2.85%∼5.80% in the VR dataset and 3.49%∼6.80% in the VG dataset. In addition, the SC factor leveraged the spatial configurations, i.e., the relative positions of bounding boxes between object pairs. This factor showed a 1.40%∼4.45% gain in the VR dataset and 1.06%∼2.86% gain in the VG dataset when compared to results where it was not added. This improvement indicates that visual appearance, statistical relationships, and spatial configurations are complementary to one another. Note that we used the best performance method in the following comparisons.

Next, we compared our method with previous works in terms of predicate, phrase, and relationship detections. Table 2 summarizes these comparisons. On the VR dataset, our method outperformed the state-of-the-art methods in both predicate and phrase detections,

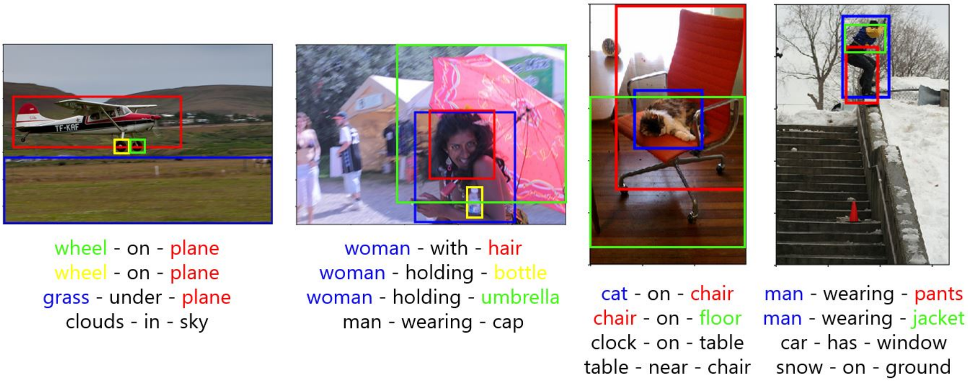

Figure 3: Qualitative examples of relationship detection on VG. The colored bounding boxes denote the detected objects in the image. The relationships under the images are colored when the relationship is correctly predicted in the top 5 most probable relationships to each object pair, while black indicates failed relationship detection owing to the failure of object detection. The colors of the object labels correspond to those of the bounding boxes in the image.

and achieved a comparable performance in relationship detection. However, the VRL and DR-Net methods performed slightly better than our method in relationship detection, which may owe to the rate of false positives and false negatives in object detection being much higher than those of the aforementioned methods. Furthermore, the heuristic exclusion method explained above could not effectively decrease the number of unexpected relationships. In addition, the ResNet-101 CNN network was used to improve performance in [4], while we used the standard VGG-16 network. Compared with the method in [22], this method uses attribute types describing color, shape, or poses of objects as well as relationship types from annotations of a large-scale dataset to build a semantic action graph. On the VG dataset, our method outperformed the state-of-the-art methods in both predicate and relationship detection when using only relationship annotations from the dataset. Note that we used the results of LP and ISGG reported in [21] and the results of MSDN without caption proposals and caption annotations as supplemental sources for fair comparisons. Figure 3 shows the qualitative examples of relationship detection on the VG dataset.

To ensure a multi-faceted analysis of our method, we also compared it with previous works in terms of the average processing time per image during the relationship prediction phase. Table 3 summarizes this comparison. Note that we ran DR-Net[1] and MSDN[2] under the same PC conditions for fair comparison, while we used the results of VtransE and PPR-FCN reported in [40]. In addition, we excluded the proposal generation time cost of RPN, etc., as the time of this phase (part of object detection phase) differs considerably depending on the applied method. From the results in Table 3, we can see that our method performed better than those from the described previous works on both VR and VG datasets. This achievement is most likely the result of the simple and efficient architecture of our method on a single network in an end-to-end manner.

[1] https://github.com/doubledaibo/drnet_cvpr2017
[2] https://github.com/yikang-li/MSDN

5 Conclusion

This paper presents a region-object relevance-guided framework, which is a simple and efficient end-to-end deep network, using an inner product space with a weighting mechanism for relevance regions in images to implement VRD. Experimental results demonstrated that the VRD task could not be more effectively implemented using only relevance regions related to object pairs. However, visual appearance, statistical relationships, and spatial configurations are complementary to one another, which results in superior best performance. Finally, experiments on the VR and VG datasets showed improvements for VRD in terms of detection accuracy and processing time when compared with previous works using only relationship annotations from the dataset.

Acknowledgement

The author would like to thank Mr. Teppei Suzuki from the Keio University for much useful discussion and technical support. This work was conducted at the Computer Vision Research Group, National Institute of Advanced Industrial Science and Technology (AIST).

References

[1] Y. W. Chao, Z. Wang, Y. He, J. Wang, and J. Deng. HICO: A Benchmark for Recognizing Human-Object Interactions in Images. In *Proc. of IEEE ICCV*, pages 1017–1025, 2015.

[2] Y. W. Chao, Y. Liu, X. Liu, H. Zeng, and J. Deng. Learning to Detect Human-Object Interactions. In *Proc. of IEEE WACV*, pages 1–9, 2018.

[3] W. Choi, Y. W. Chao, C. Pantofaru, and S. Savarese. Understanding Indoor Scenes using 3D Geometric Phrases. In *Proc. of IEEE Conf. on CVPR*, pages 33–40, 2013.

[4] B. Dai, Y. Zhang, and D. Lin. Detecting Visual Relationships with Deep Relational Networks. In *Proc. of IEEE Conf. on CVPR*, pages 3298–3308, 2017.

[5] J. Dai, K. He, Y. Li, S. Ren, and J. Sun. Instance-Sensitive Fully Convolutional Networks. In *Proc. of ECCV*, pages 534–549, 2016.

[6] C. Desai and D. Ramanan. Detecting Actions, Poses, and Objects with Relational Phraselets. In *Proc. of ECCV*, pages 158–172, 2012.

[7] C. Galleguillos, A. Rabinovich, and S. Belongie. Object Categorization using Co-Occurrence, Location and Appearance. In *Proc. of IEEE Conf. on CVPR*, 2008.

[8] R. Girshick. Fast R-CNN. In *Proc. of IEEE ICCV*, pages 1440–1448, 2015.

[9] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In *Proc. of IEEE Conf. on CVPR*, pages 580–587, 2014.

[10] G. Gkioxari, R. Girshick, P. Dollár, and K. He. Detecting and Recognizing Human-Object Interactions. In *Proc. of IEEE Conf. on CVPR*, pages 8359–8367, 2018.

[11] S. Gould, J. Rodgers, D. Cohen, G. Elidan, and D. Koller. Multi-Class Segmentation with Relative Location Prior. *IJCV*, 80(3):300–316, 2008.

[12] A. Gupta and L. S. Davis. Beyond Nouns: Exploiting Prepositions and Comparative Adjectives for Learning Visual Classifiers. In *Proc. of ECCV*, pages 16–29, 2008.

[13] J. Hatori, Y. Kikuchi, S. Kobayashi, K. Takahashi, Y. Tsuboi, Y. Unno, W. Ko, and J. Tan. Interactively Picking Real-World Objects with Unconstrained Spoken Language Instructions. In *Proc. of IEEE ICRA*, pages 1–8, 2018.

[14] K. He, X. Zhang, S. Ren, and J. Sun. Deep Residual Learning for Image Recognition. In *Proc. of IEEE Conf. on CVPR*, pages 770–778, 2016.

[15] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask R-CNN. In *Proc. of IEEE ICCV*, pages 2980–2988, 2017.

[16] G. Huang, Z. Liu, K. Q. Weinberger, and L. van der Maaten. Densely Connected Convolutional Networks. In *Proc. of IEEE Conf. on CVPR*, pages 4700–4708, 2017.

[17] R. Krishna, Y. Zhu, O. Groth, et al. Visual Genome: Connecting Language and Vision using Crowdsourced Dense Image Annotations. *IJCV*, 123(1):32–73, 2017.

[18] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet Classification with Deep Convolutional Neural Networks. In *Proc. of Conf. on NIPS*, pages 1097–1105, 2012.

[19] L. Ladicky, C. Russell, P. Kohli, and P. H. S. Torr. Graph Cut Based Inference with Co-Occurrence Statistics. In *Proc. of ECCV*, pages 239–253, 2010.

[20] Y. Li, W. Ouyang, X. Wang, and X. Tang. ViP-CNN: Visual Phrase Guided Convolutional Neural Network. In *Proc. of IEEE Conf. on CVPR*, pages 7244–7253, 2017.

[21] Y. Li, W. Ouyang, B. Zhou, K. Wang, and X. Wang. Scene Graph Generation from Objects, Phrases and Region Captions. In *Proc. of IEEE Conf. on CVPR*, pages 1261–1270, 2017.

[22] X. Liang, L. Lee, and E. P. Xing. Deep Variation-Structured Reinforcement Learning for Visual Relationship and Attribute Detection. In *Proc. of IEEE Conf. on CVPR*, pages 4408–4417, 2017.

[23] W. Liu, D. Anguelov, D. Erhan, et al. SSD: Single Shot MultiBox Detector. In *Proc. of ECCV*, pages 21–37, 2016.

[24] J. Long, E. Shelhamer, and T. Darrell. Fully Convolutional Networks for Semantic Segmentation. In *Proc. of IEEE Conf. on CVPR*, pages 3431–3440, 2015.

[25] C. Lu, R. Krishna, M. Bernstein, and L. Fei-Fei. Visual Relationship Detection with Language Priors. In *Proc. of ECCV*, pages 852–869, 2016.

[26] P. O. Pinheiro, R. Collobert, and P. Dollár. Learning to Segment Object Candidates. In *Proc. of Conf. on NIPS*, pages 1990–1998, 2015.

[27] A. Rabinovich, A. Vedaldi, C. Galleguillos, E. Wiewiora, and S. Belongie. Objects in Context. In *Proc. of IEEE ICCV*, 2007.

[28] V. Ramanathan, C. Li, J. Deng, W. Han, Z. Li, K. Gu, Y. Song, S. Bengio, C. Rosenberg, and L. Fei-Fei. Learning Semantic Relationships for Better Action Retrieval in Images. In *Proc. of IEEE Conf. on CVPR*, pages 1100–1109, 2015.

[29] J. Redmon and A. Farhadi. YOLO9000: Better, Faster, Stronger. In *Proc. of IEEE Conf. on CVPR*, pages 7263–7271, 2017.

[30] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. on PAMI*, 39(6):1137–1149, 2017.

[31] M. A. Sadeghi and A. Farhadi. Recognition using Visual Phrases. In *Proc. of IEEE Conf. on CVPR*, pages 1745–1752, 2011.

[32] M. Shridhar and D. Hsu. Interactive Visual Grounding of Referring Expressions for Human-Robot Interaction. In *Proc. of RSS*, pages 1–9, 2018.

[33] K. Soomro, A. R. Zamir, and M. Shah. UCF101: A Dataset of 101 Human Actions Classes from Videos in The Wild. *arXiv preprint arXiv:1212.0402*, 2012.

[34] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, et al. Going Deeper with Convolutions. In *Proc. of IEEE Conf. on CVPR*, pages 1–9, 2015.

[35] D. Xu, Y. Zhu, C. B. Choy, and L. Fei-Fei. Scene Graph Generation by Iterative Message Passing. In *Proc. of IEEE Conf. on CVPR*, pages 5410–5419, 2017.

[36] B. Yao and L. Fei-Fei. Modeling Mutual Context of Object and Human Pose in Human-Object Interaction Activities. In *Proc. of IEEE Conf. on CVPR*, pages 17–24, 2010.

[37] B. Yao, X. Jiang, A. Khosla, A. L. Lin, L. Guibas, and L. Fei-Fei. Human Action Recognition by Learning Bases of Action Attributes and Parts. In *Proc. of IEEE ICCV*, pages 1331–1338, 2011.

[38] R. Yu, A. Li, V. I. Morariu, and L. S. Davis. Visual Relationship Detection with Internal and External Linguistic Knowledge Distillation. In *Proc. of IEEE ICCV*, pages 1974–1982, 2017.

[39] H. Zhang, Z. Kyaw, S. F. Chang, and T. S. Chua. Visual Translation Embedding Network for Visual Relation Detection. In *Proc. of IEEE Conf. on CVPR*, pages 5532–5540, 2017.

[40] H. Zhang, Z. Kyaw, J. Yu, and S. F. Chang. PPR-FCN: Weakly Supervised Visual Relation Detection via Parallel Pairwise R-FCN. In *Proc. of IEEE ICCV*, pages 4233–4241, 2017.

[41] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. H. S. Torr. Conditional Random Fields as Recurrent Neural Networks. In *Proc. of IEEE ICCV*, pages 1529–1537, 2015.

[42] Y. Zhu and S. Jiang. Deep Structured Learning for Visual Relationship Detection. In *Proc. of AAAI Conf. on Artificial Intelligence*, pages 7623–7630, 2018.