Hierarchical Graph Attention Network for Visual Relationship Detection

Li Mi, Zhenzhong Chen*
School of Remote Sensing and Information Engineering, Wuhan University, China
{milirs,zzchen}@whu.edu.cn

Abstract

Visual Relationship Detection (VRD) aims to describe the relationship between two objects by providing a structural triplet shown as <subject-predicate-object>. Existing graph-based methods mainly represent the relationships by an object-level graph, which ignores to model the triplet-level dependencies. In this work, a Hierarchical Graph Attention Network (HGAT) is proposed to capture the dependencies on both object-level and triplet-level. Object-level graph aims to capture the interactions between objects, while the triplet-level graph models the dependencies among relation triplets. In addition, prior knowledge and attention mechanism are introduced to fix the redundant or missing edges on graphs that are constructed according to spatial correlation. With these approaches, nodes are allowed to attend over their spatial and semantic neighborhoods' features based on the visual or semantic feature correlation. Experimental results on the well-known VG and VRD datasets demonstrate that our model significantly outperforms the state-of-the-art methods.

1. Introduction

Visual relationship detection serves as a middle-level task to bridge the gap between low-level image recognition task, such as object detection [24, 9], and high level image understanding tasks, such as image captioning[1], visual question answering[45, 15], visual reasoning [27] and scene graph generation [19, 43]. Based on single object detection, visual relationship detection aims to accurately localize a pair of objects and determine the predicate between them by providing several structural, comprehensive triplets, shown as <subject-predicate-object>.

Previous methods on VRD focus on modelling the relationship between a pair of objects independently, which ignore the global context information of an image scene. Recently, graph structures [2, 10] are introduced to capture the context information by an object-level graph where the nodes denote objects and edges represent predicates. However, long-dependencies on triplet-level are excluded. The long dependencies among triplets serve as important context information for VRD. For example, some triplets are more likely to co-occur with each other even if they do not contain the same objects: <person-ride-bike> is more likely to be associated with <car-on-street> than <elephant-on-grass>. Such dependencies among triplets cannot be modeled *explicitly* by object-level graph. To address this problem, a Hierarchical Graph Attention Network (HGAT) is proposed to model the dependencies on both object-level and triplet-level. As is shown in Figure 1, the task of predicting relation triplets is divided into two stages: object-level reasoning and triplet-level reasoning. The model joints the information to give a final prediction.

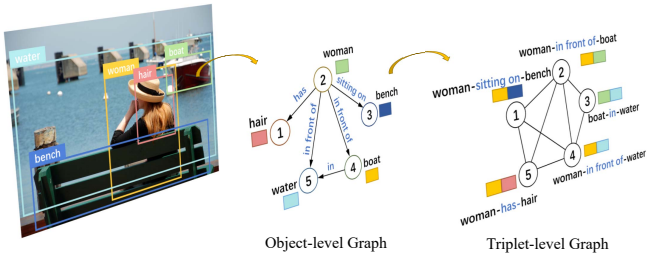


Figure 1. The illustration of the proposed Hierarchical Graph Attention Network (HGAT). The object-level graph captures the interactions among objects while the triplet-level graph models the interactions among relation triplets explicitly.

In addition, constructing the graph only based on spatial correlation brings some inappropriate edges, such as redundant edges or missing edges. For example, the two people in Figure 2 (a) are next to each other, resulting in a redundant edge (e.g. <person1-wearing-jacket2>) when establishing the graph. Another example in Figure 2 (b) shows a missing edge between pairwise objects (e.g. <boy-looking at-kite>). Because the distance between the two bounding boxes exceeds the threshold, the edge between these two bounding boxes will be considered non-existent. Furthermore, since the graph structure is fixed in the existing method, the errors in the original graph structure will be

*Corresponding author: Zhenzhong Chen.

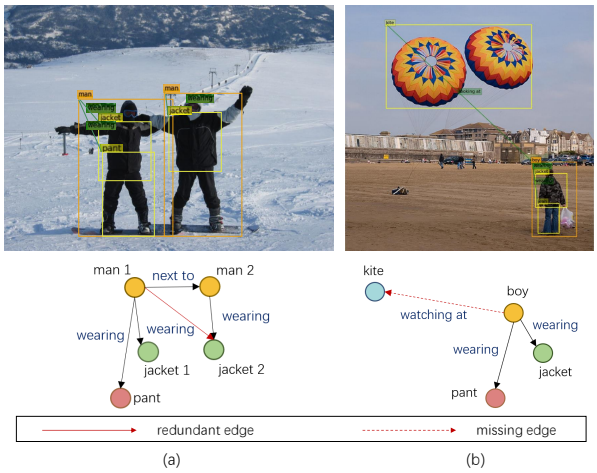


Figure 2. Problems of graph built on spatial correlation. (a) and (b) shows the redundant edge and the missing edge of object-level graph constructed based on spatial correlation, respectively.

accumulated during reasoning. To address these problems, prior knowledge and attention mechanism are introduced to the graph. Firstly, the graph is constructed based on the spatial correlation and semantic correlation which will connect some of the missing edges based on the prior knowledge. Then, with graph attention mechanism, the nodes are allowed to attend over their spatial and semantic neighbors' features by assigning learnable weights to different nodes based on the visual or semantic feature correlation. The detrimental effects of redundant edges can be alleviated by reducing the attention weights.

The main contribution of the paper can be summarized as:

- A Hierarchical Graph Attention Network (HGAT) is proposed to explore the relationship triplet on both object-level and triplet-level. By explicitly modeling the dependencies among triplets, more context information can be incorporated in the relationship reasoning.
- Prior knowledge and attention mechanism are introduced to the graph to alleviate the detrimental effects of inaccurate initial graph. With the attention mechanism, the nodes are allowed to attend over their spatial and semantic neighbors' feature by assigning learnable weights to these nodes based on the visual or semantic feature correlation.

2. Related Work

2.1. Visual Relationships Detection

Visual relationship detection offers a comprehensive scene understanding of an image by providing several triplets of <subject-predicate-object>. Early work assigned a unique class to each relationship triplet [6, 26, 22], however, the search space is explosive. Assume that there are N object categories and K predicate categories. Then the search space of object detection is N and there will be N^2K relationship categories when representing relationship as <subject-predicate-object>. Previous work [16, 13, 44, 14, 42] tackled this problem by separating the prediction process or applying multiple features. Unlike directly taking the triplet <subject-predicate-object> as a whole learning task, the separate method predicts the objects and predicates separately. In that way, different relationships (e.g. <truck-on-street>, <car-on-street>) are merged into the same category if they share the same predicate, reducing the search space to $N + K$. The cost of separating prediction is that samples within the same predicate category are highly diverse. To better distinguish the predicates, researchers represented objects in visual, spatial and semantic cues which greatly improve the model performance [16, 13, 38]. In these methods, interactions between a pair of objects can be captured but global context information cannot be modeled explicitly. To tackle this problem, graph structures are utilized to explore the connections and constraints between objects [2, 15, 10].

2.2. Graph Structure in VRD

Graph Neural Networks (GNNs) were introduced in Gori *et al.* [8] and Scarselli *et al.* [28] as a generalization of recursive neural networks that can directly deal with a more general class of graphs [30]. Typical graph structures such as Graph Convolutional Network (GCN) [11] were used to learn representations for nodes. Nodes are able to attend over their semantic or spatial neighborhoods features in a pre-defined graph structure, which achieved significant success in various fields, such as link prediction [31], scene graph generation [34, 33] and human object interaction [21]. However, the graph convolution operation is restricted in the pre-defined graph structure [30, 18]. Velickovic *et. al* [30] proposed a Graph Attention Network (GAT) to specify arbitrary weights to the neighbors following self-attention strategy, which gets rid of the limitations of the fixed graph structure.

Graph structures have received an increasing amount of attention in VRD. Specifically, Cui *et al.* [2] proposed a context-dependent diffusion network to capture the interactions between different object instances through word semantic graph and visual scene graph. Yao *et al.* [36] explored the semantic and spatial relationship between objects by GCN for image captioning. Hu *et al.* [10] introduced a message-passing-style algorithm to propagate the contextual information. Object-level graph models the interactions among objects. However, the triplet-level dependencies are not fully exploited. In this work, a Hier-

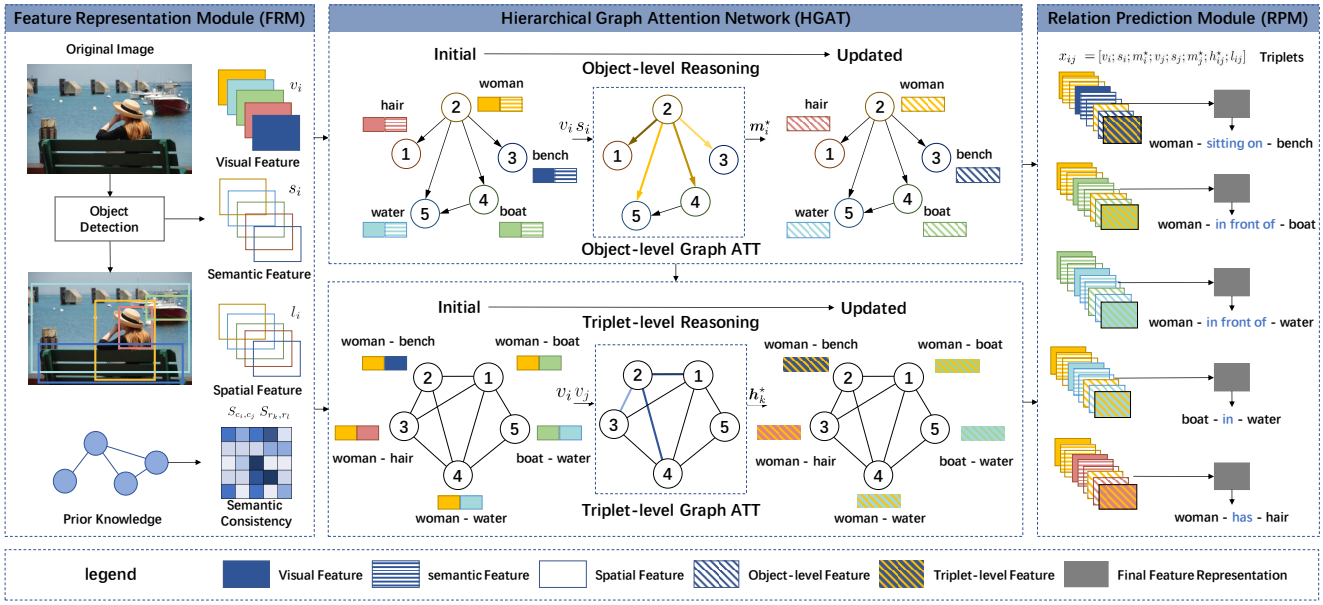


Figure 3. The framework of the Hierarchical Graph Attention Network (HGAT). The proposed method can be divided into three sub-modules: Feature Representation Module, Hierarchical Graph Attention Network and Predicate Prediction Module. In the feature representation module (Section 3.2), multi-cues are utilized to represent objects in an image. The proposed HGAT(Section 3.3) conducts object-level reasoning and triplet-level reasoning through the object-level graph and the triplet-level graph, respectively. The predicate prediction module (Section 3.4) in charge of predicting relationships.

archical Graph Attention Network (HGAT) is proposed to deeply exploit the dependencies on both object-level and triplet-level. With explicitly model the dependencies among triplets, more context information and global constraint can be incorporated in the relationship reasoning. In addition, graphs in previous work are constructed based on the spatial correlation of objects, which can be improved by taking semantic correlation into consideration.

2.3. Prior Knowledge in VRD

Prior knowledge has been widely utilized as background information to assist the tasks in computer vision [5, 41, 45] and natural language processing [39, 4, 29, 32]. Rohrbach *et al.* [25] showed that external knowledge of attributes contributed to zero-shot learning by associating classes to attributes and recognizing instances of unseen classes. In VRD, Lu *et al.* [16] first leveraged language prior from semantic embeddings to finetune the likelihood of a predicted relationship. Yu *et al.* [40] proposed a teacher-student framework to incorporate predicate-object pair co-occurrences which are collected from both external and internal data. Plesse *et al.* [20] designed a framework to estimate the relevance if object pairs by incorporating prior knowledge. Unlike the previous methods that utilized prior knowledge to restrict the probability or adjust the prediction, prior knowledge contributes to the graph construction process and participates the relationship inference directly.

3. Hierarchical Graph Attention Network for VRD

3.1. Method Overview

3.1.1 Problem Formulation

For a given image I, visual relationship detection aims to provide several relation triplets shown as <subject-predicate-object>. Let O and P denote the object set and predicate set, respectively, then the relationship set can be defined as $R = \{r(s,p,o)|s,o \in O, p \in P\}$, where s, p and o are respectively the subject, predicate and object in a relationship triplet (s,p,o). The probabilistic model of visual relationship detection can be formulated as:

$$P(r) = P(p|s,o)P\left(s|b_s\right)P\left(o|b_o\right). \quad (1)$$

Here b_s and b_o are two individual bounding boxes for subject and object, which compose an object pair. $P\left(s|b_s\right)$ and $P\left(o|b_o\right)$ represent the subject confidence score and object confidence score with bounding boxes.

3.1.2 Framework

As is shown in Figure. 3, the proposed method can be divided into three sub-modules: Feature Representation Module (FRM), Hierarchical Graph Attention Network (HGAT) and Predicate Prediction Module (PPM). In the Fea-

ture Representation Module (Section 3.2), an object detector generates object proposals with bounding boxes and labels, then the visual, spatial and semantic cues of each object and the corresponding relative feature of pairwise objects are provided. Next, the proposed HGAT(Section 3.3) conducts object-level reasoning and triplet-level reasoning by a hierarchical graph structure. For each node, the graph attention mechanism assigns reasonable weights to its neighbors and obtains the final node representation. The predicate prediction module (Section 3.4) takes charge of predicting relationships based on the existing graph.

3.2. Feature Representation

The Feature Representation Module takes an image as input and outputs are bounding boxes with visual, spatial and semantic features.

3.2.1 Proposal Generation

Inspired by previous work in VRD [16, 13, 37], the Faster R-CNN [24] with VGG-16 backbone is utilized to locate and detect objects. Specifically, we first sample 300 proposal regions generated by the RPN with IoU>0.7. Then we perform the NMS with IoU>0.4 on the 300 proposals. The retained proposals with confidence score higher than 0.05 are kept as the detected objects in the image. After that, the locations and labels for all possible objects are collected. Note that, we choose Faster R-CNN with VGG-16 to compare our method with the previous methods fairly, however, the proposed method can be applied to any object detector such as Fast RCNN [7] and YOLO [23].

3.2.2 Feature Extraction

Single feature cannot represent the complex relationship between pairwise objects. Take the prediction of spatial interactions, such as 'near', 'under', and 'on', as an example. If we only use visual appearance to represent objects, the prediction will be challenging due to the lack of spatial information. In this paper, visual appearance, spatial feature and semantic embedding are considered in the feature extraction.

Visual Feature. Visual appearance plays an important role in distinguishing objects and understanding relations. For a relationship instance (s, p, o), b_s, b_{so} and b_o denote the bounding box of its corresponding subject, predicate and object. Note that b_{so} refers to the union of b_s and b_o with a small margin to capture the surrounding context. Following the previous work [16, 13], we adopt VGG-16 as a backbone and extract the RoI Pooling features of b_s, b_{so} and b_o from two fully connected layers. The visual feature can be denoted as v.

Spatial Feature. To complement the visual information, the spatial feature is regarded as an indispensable feature for visual relationship detection. To get the relative spatial feature of bounding boxes, we adopt the idea of box regression [10]. Assume $\Delta(b_i, b_j)$ denote the box delta that regresses the bounding box b_i to b_j. Then $\mathrm{dis}(b_i, b_j)$ and $\mathrm{iou}(b_i, b_j)$ denote the normalized distance and IoU between b_i and b_j. The union region of b_i and b_j is denoted as b_{ij}. The relative spatial location of the subject and object can be defined as:

$$l_{ij} = [\Delta(b_i, b_j); \Delta(b_i, b_{ij}); \Delta(b_j, b_{ij}); \mathrm{iou}(b_i, b_j); \mathrm{dis}(b_i, b_j)]. \quad (2)$$

Semantic Feature. Different relationships may exist between the same pair of objects (e.g. <person-near-car>, <person-drive-car>), meanwhile, the same predicate may be used to describe different types of object pairs (e.g. <person-ride-bike>, <person-ride-horse>). Language priors serve as a distinguishing feature to exploit the semantic context of an image. We adopt a semantic embedding layer to map the object category C into word embedding S. Then the embedding vectors of subject and object are jointed to learn the representation of object pair through a fully connected layer. Note that the parameters of object categories are initialized with the pre-trained word representations such as word2vec [17]. The semantic feature can be represented as s.

3.2.3 Prior Knowledge Distillation

To represent semantic consistency, an immediate approach is to utilize the cumulative number of co-occurrences for each pair of concepts from the prior knowledge data. Assume that there are N instances in the prior knowledge data in total. Let $n(c_i, c_j)$ denote the frequency of co-occurrences for concepts c_i and c_j, and $n(c_i)$ denote the frequency of c_i. Then, we define semantic consistency based on point-wise mutual information. Specifically, when c_i and c_j occur independently, or they co-occur less frequently than if they were to occur independently, the value would be zero; otherwise, the value is positive. Bounded by $\log N$ from the above, if the two concepts are more likely to occur together than appear independently, the value will get larger. This definition can be formulated as:

$$S_{c_i,c_j} = \max\left(\log \frac{n(c_i, c_j)\, N}{n(c_i) n(c_j)}, 0\right). \quad (3)$$

The semantic consistency among triplets denoted as S_{r_k,r_l} is calculated as the same way. The prior knowledge in our experiments is from relationship detection datasets.

3.3. Hierarchical Graph Attention Network

To model the dependencies on both object-level and triplet-level in an image, two types of the graph are consid-

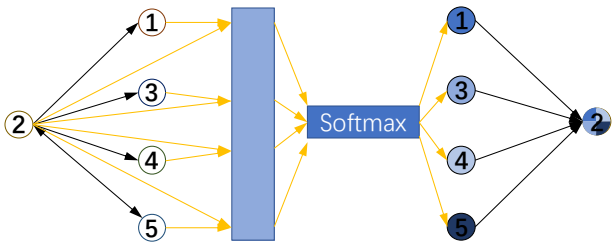


Figure 4. The structure of attention mechanism. With attention mechanism, nodes are allowed to attend over their spatial and semantic neighbors' feature by assigning learnable weights to these nodes based on the visual or semantic feature correlation.

ered. One is the object-level graph, which models the interactions among objects and conducts object-level reasoning. The other is triplet-level graph, which is constructed based on the interactions among triplets and conducts triplet-level reasoning. There are two types of attention according to these two graphs.

3.3.1 Object-level Reasoning

Object-level Graph Construction. Object-level graph is constructed to capture the interactions between pairwise objects. The object-level graph $\mathcal{G}_o = \{\mathcal{V}_o, \mathcal{E}_o\}$ contains a node set $\mathcal{V}_o$ and an edge set $\mathcal{E}_o$. Each node $n_i \in \mathcal{V}_o$ represents an object, which is composed of a bounding box b_i and a corresponding attribute embedding. Each edge $e^o_{ij} \in \mathcal{E}_o$ denotes the predicate between node n_i and n_j. The relationship triplet (n_i, e^o_{ij}, n_j) and (n_j, e^o_{ji}, n_i) represent two different instances, which are distinguished by a directed object-level graph.

Two factors are considered in establishing the graph: spatial correlation and semantic correlation. We use $\mathrm{dis}(b_i, b_j)$ and $\mathrm{iou}(b_i, b_j)$ to evaluate the spatial correlation of two object proposals. The spatial graph can be defined as:

$$e^{sp}_{ij} = \begin{cases} 1, & \mathrm{dis}\,(b_i, b_j) < t_1 \text{ or } \mathrm{iou}\,(b_i, b_j) > t_2 \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

where t_1 and t_2 are two thresholds which we set as 0.5 in our experiments. On the other hand, to evaluate the semantic correlation of pairwise objects, the semantic graph is established based on semantic consistency.

$$e^{se}_{ij} = \begin{cases} 1, & S_{c_i,c_j} > t_3 \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

where t_3 is set as 0 in our experiments. Finally, the object-level graph is constructed as:

$$e^o_{ij} = e^{sp}_{ij} \oplus e^{se}_{ij}, \tag{6}$$

where $\oplus$ denotes OR operation.

Object-level Attention. If we regard the joint feature of visual and semantic as the attribute of the node, the attribute vector $\boldsymbol{m_i}$ can be represented as $\boldsymbol{m_i} = \mathrm{concat}(v_i, s_i)$. Graph attention mechanism then can be formulated as:

$$\boldsymbol{m}^\star_i = \sigma\left(\sum_{j \in \mathcal{N}_i} \alpha_{ij} \cdot \left(\mathbf{W}^o_{dir(i,j)} \boldsymbol{m}_j + \boldsymbol{b}\right)\right), \tag{7}$$

where $\boldsymbol{m}^\star_i$ represents the generated hidden features. The definition of the attention coefficient α_{ij} is defined as:

$$\alpha_{ij} = \frac{\exp\left(\left(\mathbf{U}^o \boldsymbol{m}_i\right)^\top \cdot \mathbf{V}^o_{dir(i,j)} \boldsymbol{m}_j + \boldsymbol{c}\right)}{\sum_{j=1}^{K} \exp\left(\left(\mathbf{U}^o \boldsymbol{m}_i\right)^\top \cdot \mathbf{V}^o_{dir(i,j)} \boldsymbol{m}_j + \boldsymbol{c}\right)}, \tag{8}$$

where $\mathbf{U}^o, \mathbf{V}^o \in \mathbb{R}^{d_m \times (d_v + d_s)}$ are projection matrices and $\boldsymbol{b}$, $\boldsymbol{c}$ are bias terms. $dir(i,j)$ selects the transformation matrix based on the directionality of each edge.

3.3.2 Triplet-level Reasoning

Triplet-level Graph Construction. Some relationship is more likely to co-occur with each other. The other graph is constructed to capture such dependencies between relationship instances. Suppose there are node set $\mathcal{V}_t$ for possible relationship triplets and edge set $\mathcal{E}_t$ for the interactions between triplets. The triplet-level graph $\mathcal{G}_t = \{\mathcal{V}_t, \mathcal{E}_t\}$ can be defined as:

$$e^t_{kl} = \begin{cases} 1, & S_{r_k,r_l} > t_4 \\ 0, & \text{otherwise} \end{cases} \tag{9}$$

where t_4 is set to 0 in our experiments. Note that the triplet-level graph is an undirected graph.

Triplet-level Attention. Triplet graph is constructed to capture the interactions among triplets. The attribute feature of node k is the visual feature of triplet k denoted as $\boldsymbol{h}_k$. The generated hidden feature can be formulated as:

$$\boldsymbol{h}^\star_k = \sigma\left(\sum_{l \in \mathcal{N}_k} \alpha_{kl} \cdot \mathbf{W}^t \boldsymbol{h}_l\right), \tag{10}$$

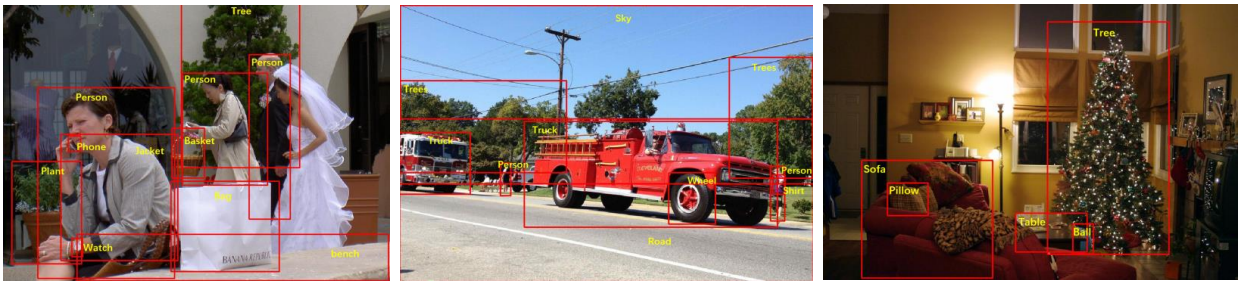
where $\boldsymbol{h}^\star_k$ denotes the hidden state representation of triplet k.

$$\alpha_{kl} = \frac{\exp\left(\left(\mathbf{U}^t \boldsymbol{h}_k\right)^\top \cdot \mathbf{V}^t \boldsymbol{h}_l\right)}{\sum_{l \in \mathcal{N}_k} \exp\left(\left(\mathbf{U}^t \boldsymbol{h}_k\right)^\top \cdot \mathbf{V}^t \boldsymbol{h}_l\right)}, \tag{11}$$

where $\mathbf{U}^t, \mathbf{V}^t \in \mathbb{R}^{d_h \times (d_v + d_v)}$ are projection matrices.

3.4. Predicate Prediction

The inputs of Predicate Prediction Module are features from both object-level reasoning and triplet-level reasoning, while the outputs are several relationship triplets shown



Baseline	person wear basket person wear phone person wear jacket person wear watch bag on bench	**Baseline**	wheel on road truck on road truck on truck truck on road person on truck	**Baseline**	tree on sofa sofa on tree table on tree ball on table table on sofa
HGAT	person wear jacket person has phone person near person plant behind person bag on bench	**HGAT**	truck on road truck on road truck has wheel trees behind truck wheel on truck	**HGAT**	sofa next to tree pillow on sofa table near tree table behind sofa ball on table

Figure 5. The comparison results of baseline model and HGAT. Green, yellow and red color denotes the correct triples, correct but unannotated triples and failed triples.

Feature	Predicate Det				Relationship Det			
	k=1		k=70		k=1		k=70	
	R@50	R@100	R@50	R@100	R@50	R@100	R@50	R@100
Baseline	48.86	48.86	86.21	94.32	17.90	21.54	19.27	25.15
Base+SC	50.21	50.21	87.55	95.43	19.15	22.82	21.41	26.10
Base+ATT	52.16	52.16	88.36	95.68	19.77	23.26	21.56	26.80
Base+TL	52.44	52.44	88.69	95.88	19.99	23.46	21.66	26.90
Base+SC+ATT	54.55	54.55	88.76	95.89	20.21	23.65	21.74	26.98
Base+SC+TL	54.89	54.89	89.04	96.21	20.56	23.45	22.12	27.00
Base+ATT+TL	58.42	58.42	89.44	96.37	20.92	23.92	22.23	27.16
HGAT (Base+SC+ATT+TL)	**59.54**	**59.54**	**90.91**	**97.02**	**22.52**	**24.63**	**22.90**	**27.23**

Table 1. Ablation study (%) on VRD dataset. The *Baseline* model is object-level reasoning without attention mechanism and semantic consistency. SC, ATT and TL represents with semantic consistency, with attention mechanism and with triplet-level reasoning, respectively.

as <subject-predicate-object>. The final representation of interactions between the i-th object and the j-th object is the concatenation of the aforementioned features: $x_{ij} = [v_i; s_i; m_i^\star; v_j; s_j; m_j^\star; h_{ij}^\star; l_{ij}]$. Then, the confidence of the predicate category between the i-th and the j-th objects is $y_{ij} = \text{softmax}\left(\mathbf{W}^f x_{ij}\right)$, where $\mathbf{W}^f$ is the embedding matrix that maps interaction embeddings to match the predicate categories. Multi-class cross entropy loss is used in our experiment.

4. Experiment

4.1. Experimental Details

We train and evaluate our models on the well-known Visual Relationship Detection (VRD) [16] and Visual Genome (VG) [12]. **VRD** dataset contains 5000 images with 100 object categories and 70 predicate categories. **VG** dataset provides human-annotated relationships for 100k images, which consists of over 1M instances of objects and 600k relations. In our experiments, we consider a simplified version named VG100K [44] which consists of 99658 images with 200 object categories and 100 predicate categories.

We use two standard evaluation modes: (1) **Predicate Detection (Predicate Det)**: given a ground truth object location and categories, the network predicts relationships among objects. (2) **Relationship Detection (Relationship Det)**: the network predicts object location (bounding boxes), categories and relationships among objects at the same time. Following the standard evaluation in [16], R@n is used as the evaluation metric. R@n computes the fraction of true positive predicted relationships over the total annotated relationships among the top n confident pre-

k	Methods	Predicate Det		Relationship Det	
		R@50	R@100	R@50	R@100
k=1	VR-LP [16]	47.87	47.87	13.86	14.70
	VTransE [44]	44.76	44.76	14.07	15.20
	STA [35]	48.03	48.03	-	-
	CAI [46]	53.59	53.59	15.63	17.39
	VRL [14]	-	-	18.19	20.79
	Zoom-Net [38]	50.69	50.69	18.92	21.41
	NMP [10]	57.69	57.69	20.19	23.98
	HGAT (Ours)	**59.54**	**59.54**	**22.52**	**24.63**
k=70	DR-Net [3]	80.78	81.90	17.73	20.88
	Zoom-Net [38]	84.25	90.59	21.37	27.30
	VRD-DSR [13]	86.01	93.18	19.03	23.29
	CDDN [2]	87.57	93.76	21.46	26.14
	NMP [10]	90.61	96.61	21.50	27.50
	HGAT (Ours)	**90.91**	**97.02**	**22.90**	**27.73**

Table 2. Predicate and relationship detection results(%) on VRD Dataset. _ denotes the results are not reported in the original paper. k denotes the number of predicates associated with each object.

dictions. Let k be the number of predicates associated with each object. In our experiments, $n \in \{50, 100\}$ and $k \in \{1, 70, 100\}$.

In the experiments, we set the batch size as 32. The initial learning rate is set to 0.005 and the learning rate decay factor is 0.5. The dropout rate of the model is 0.5 and the hidden state dim is 512.

4.2. Ablation Study

We conduct ablation studies on VRD dataset to understand the importance of each component of our model. Specifically, the semantic consistency, attention mechanism and triplet-level reasoning are removed, respectively. The results are presented in Table 1. The *Baseline* model is object-level reasoning without attention mechanism and semantic consistency. SC, ATT and TL represents with semantic consistency, with attention mechanism and with triplet-level reasoning, respectively.

4.2.1 Semantic Consistency

Experiments show that semantic consistency promotes model performance by around 1.29% and 1.41% on the two tasks, respectively. Because the semantic consistency brings prior knowledge to fix the missing edges in the graph only based on spatial correlation.

4.2.2 Attention Mechanism

As is shown in the first row and the third row of Table 1, attention mechanism improves the performance of the model by around 2.53% and 1.77% on predicate detection and relationship detection, respectively, which indicates that the attention mechanism plays an important role in capturing the interactions among objects. It is worth noting that attention mechanism contributes more to the model performance than the semantic consistency, because the utilization of semantic consistency increases the redundant edges which can be adjusted by attention mechanism. Therefore, the combination of semantic consistency and attention mechanism improves the model performance by 3.88% and 2.18% on predicate detection and relationship detection, respectively.

k	Methods	Predicate Det	
		R@50	R@100
k=1	VTransE [44]	62.63	62.87
	STA [35]	62.71	62.94
	NMP [10]	67.03	67.29
	HGAT (Ours)	**68.11**	**68.32**
k=100	VRD-DSR [13]	69.06	74.37
	CDDN [2]	70.42	74.92
	DR-Net [3]	88.26	91.26
	NMP [10]	89.69	95.54
	HGAT (Ours)	**90.05**	**96.65**

Table 3. Predicate detection results (%) on VG dataset. The number of predicate categories is 100.

k	Methods	Predicate Det(ZS)	
		R@50	R@100
k=1	VR-LP [16]	8.45	8.45
	NMP [10]	27.50	27.50
	HGAT (Ours)	**29.12**	**29.12**
k=70	VRD-DSR [13]	60.90	79.81
	CDDN [2]	67.66	84.00
	NMP [10]	72.95	88.44
	HGAT (Ours)	**75.01**	**89.59**

Table 4. Zero-shot predicate detection results (%) in VRD dataset. The comparison only includes methods that reported the results on zero-shot setting.

4.2.3 Triplet-level Reasoning

Triplet-level reasoning significantly improves the model performance by providing triplet-level interactions. Furthermore, the combination of attention mechanism and triplet-level reasoning improves the recall stably by around 6.10% on predicate detection and around 2.59% on relationship detection, which indicates that incorporating attention mechanism to triplet-level reasoning reduces the influence of unrelated triplets.

4.3. Comparison with State-of-the-art Methods

To demonstrate the efficiency of the proposed model, we compare our methods to the state-of-the-art on VRD and

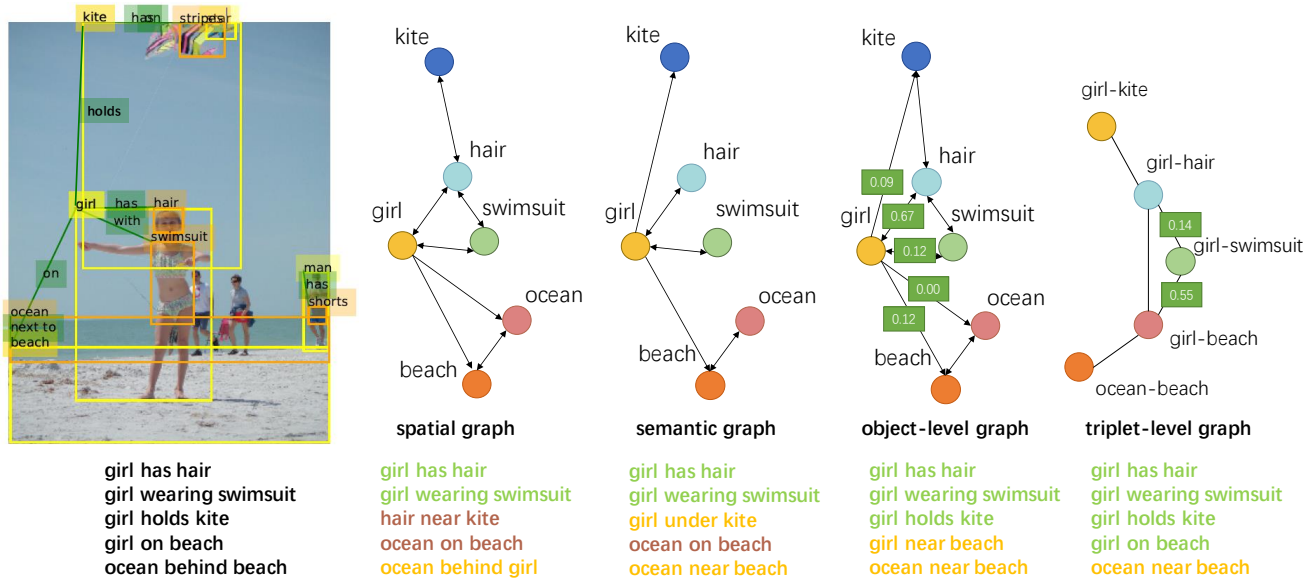


Figure 6. The visualization results of graphs. Spatial graph is constructed according to spatial correlation, semantic graph is based on semantic consistency. The attention weights in object-level graph are graph attention weights related to $girl$ while the attention weights in triplet-level graph are related to $girl - swimsuit$.

VG dataset. We list all the reported results and compare our method with two graph-based baselines: 1) CDDN [2]: CDDN designed a diffusion network to aggregate context information from both semantic graph and spatial graph. 2) NMP [10]: NMP modeled objects and interactions by an interaction graph and proposed a message-passing-style algorithm to propagate the contextual information.

4.3.1 VRD Dataset

As is shown in Table 2, the proposed method establishes a new state-of-the-art which is 97.02% on predicate detection and 27.73% on relationship detection, respectively. Different from NMP which models the interactions by message passing among edges and nodes, our method explicitly models the interactions on triplet-level. The experimental results demonstrate the effectiveness of triplet-level reasoning. Our model also outperforms CDDN which conducts a reasoning process based on the diffusion network. The improvements mainly come from the attention mechanism which takes the feature correlation into consideration.

4.3.2 VG Dataset

Similar to the previous work, we report the results on the predicate detection task for VG dataset in Table 3. We improve the state-of-the-art by 1.08% for R@50 and 1.03% R@100 compared to the previous best performance. The gain on VG dataset is smaller than VRD dataset because there are more annotation noises which weakened the effectiveness of attention mechanism and triplet-level reasoning.

4.3.3 Zero-shot Settings

The comparison results on zero-shot predicate detection are reported in Table 4. Those methods without reporting the results on zero-shot settings are excluded from the comparison. Our model achieves considerably superior performance compared with the previous work with the improvement of around 1.62% for k=1 and 1.61% for k=70.

5. Conclusion

In this paper, a novel framework named Hierarchical Graph Attention Network (HGAT) is proposed for visual relationship detection to exploit the dependencies on both object-level and triplet-level. With explicitly model the dependencies among triplets, more context information can be incorporated into the relationship reasoning process. In addition, prior knowledge and attention mechanism are introduced to alleviate the detrimental effects of the inappropriate edges of the graph. With the attention mechanism, the nodes are allowed to attend over their spatial and semantic neighbors' feature based on the visual or semantic feature correlation.

Acknowledgments

This work was supported by National Key R&D Program of China under contract No. 2017YFB1002202.

References

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, pages 6077–6086, 2018.

[2] Zhen Cui, Chunyan Xu, Wenming Zheng, and Jian Yang. Context-dependent diffusion network for visual relationship detection. In *ACM MM*, pages 1475–1482, 2018.

[3] Bo Dai, Yuqi Zhang, and Dahua Lin. Detecting visual relationships with deep relational networks. In *CVPR*, pages 3298–3308, 2017.

[4] Tim Dettmers, Pasquale Minervini, Pontus Stenetorp, and Sebastian Riedel. Convolutional 2d knowledge graph embeddings. In *AAAI*, pages 1811–1818, 2018.

[5] Yuan Fang, Kingsley Kuan, Jie Lin, Cheston Tan, and Vijay Chandrasekhar. Object detection meets knowledge graphs. In *IJCAI*, pages 1661–1667, 2017.

[6] Carolina Galleguillos, Andrew Rabinovich, and Serge Belongie. Object categorization using co-occurrence, location and appearance. In *CVPR*, pages 1–8, 2008.

[7] Ross Girshick. Fast r-cnn. In *ICCV*, pages 1440–1448, 2015.

[8] Marco Gori, Gabriele Monfardini, and Franco Scarselli. A new model for learning in graph domains. In *IJCNN*, pages 729–734, 2005.

[9] Han Hu, Jiayuan Gu, Zheng Zhang, Jifeng Dai, and Yichen Wei. Relation networks for object detection. In *CVPR*, pages 3588–3597, 2018.

[10] Yue Hu, Siheng Chen, Xu Chen, Ya Zhang, and Xiao Gu. Neural message passing for visual relationship detection. In *ICML Workshop on Learning and Reasoning with Graph-Structured Representations*, 2019.

[11] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *ICLR*, 2017.

[12] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li Jia Li, and David A. Shamma. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73, 2017.

[13] Kongming Liang, Yuhong Guo, Hong Chang, and Xilin Chen. Visual relationship detection with deep structural ranking. In *AAAI*, pages 7098–7105, 2018.

[14] Xiaodan Liang, Lisa Lee, and Eric P Xing. Deep variation-structured reinforcement learning for visual relationship and attribute detection. In *CVPR*, pages 4408–4417, 2017.

[15] Li Linjie, Gan Zhe, Cheng Yu, and Liu Jingjing. Relation-aware graph attention network for visual question answering. In *ICCV*, pages 10312–10321, 2019.

[16] Cewu Lu, Ranjay Krishna, Michael Bernstein, and Fei Fei Li. Visual relationship detection with language priors. In *ECCV*, pages 852–869, 2016.

[17] Tomas Mikolov, Kai Chen, Greg S Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *ICLR*, 2013.

[18] Federico Monti, Davide Boscaini, Jonathan Masci, Emanuele Rodola, Jan Svoboda, and Michael M Bronstein. Geometric deep learning on graphs and manifolds using mixture model cnns. In *CVPR*, pages 5425–5434, 2017.

[19] Alejandro Newell and Jia Deng. Pixels to graphs by associative embedding. In *NIPS*, pages 2171–2180, 2017.

[20] François Plesse, Alexandru Ginsca, Bertr Delezoide, and Françoise Prêteux. Visual relationship detection based on guided proposals and semantic knowledge distillation. In *ICME*, pages 1–6, 2018.

[21] Siyuan Qi, Wenguan Wang, Baoxiong Jia, Jianbing Shen, and Songchun Zhu. Learning human-object interactions by graph parsing neural networks. In *ECCV*, pages 407–423, 2018.

[22] Vignesh Ramanathan, Congcong Li, Jia Deng, Wei Han, Zhen Li, Kunlong Gu, Yang Song, Samy Bengio, Chuck Rossenberg, and Li Feifei. Learning semantic relationships for better action retrieval in images. In *CVPR*, pages 1100–1109, 2015.

[23] Joseph Redmon, Santosh Kumar Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *CVPR*, pages 779–788, 2016.

[24] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, pages 91–99, 2015.

[25] Marcus Rohrbach, Michael Stark, Gyorgy Szarvas, Iryna Gurevych, and Bernt Schiele. What helps where and why? semantic relatedness for knowledge transfer. In *CVPR*, pages 910–917, 2010.

[26] Mohammad Amin Sadeghi and Ali Farhadi. Recognition using visual phrases. In *CVPR*, pages 1745–1752, 2011.

[27] Adam Santoro, David Raposo, David G. T Barrett, Mateusz Malinowski, Razvan Pascanu, Peter Battaglia, and Timothy Lillicrap. A simple neural network module for relational reasoning. In *NIPS*, pages 4967–4976, 2017.

[28] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. *IEEE Transactions on Neural Networks*, 20(1):61–80, 2009.

[29] Cunchao Tu, Zhengyan Zhang, Zhiyuan Liu, and Maosong Sun. Transnet: Translation-based network representation learning for social relation extraction. In *IJCAI*, pages 2864–2870, 2017.

[30] Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. In *ICLR*, 2018.

[31] Hao Wang, Xingjian Shi, and Dityan Yeung. Relational deep learning: A deep latent variable model for link prediction. In *AAAI*, pages 2688–2694, 2017.

[32] Zhouxia Wang, Tianshui Chen, Jimmy Ren, Weihao Yu, Hui Cheng, and Liang Lin. Deep reasoning with knowledge graph for social relationship understanding. In *IJCAI*, pages 1021–1028, 2018.

[33] Sanghyun Woo, Dahun Kim, Donghyeon Cho, and In So Kweon. Linknet: Relational embedding for scene graph. In *NIPS*, pages 560–570, 2018.

[34] Jianwei Yang, Jiasen Lu, Stefan Lee, Dhruv Batra, and Devi Parikh. Graph r-cnn for scene graph generation. In *ECCV*, pages 690–706, 2018.

[35] Xu Yang, Hanwang Zhang, and Jianfei Cai. Shuffle-then-assemble: Learning object-agnostic visual relationship features. In *CVPR*, pages 36–52, 2018.

[36] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. Exploring visual relationship for image captioning. In *ECCV*, pages 684–699, 2018.

[37] Zhan Yibing, Yu Jun, Yu Ting, and Tao Dacheng. On exploring undetermined relationships for visual relationship detection. In *CVPR*, pages 5128–5137, 2019.

[38] Guojun Yin, Lu Sheng, Bin Liu, Nenghai Yu, Xiaogang Wang, Jing Shao, and Change Loy Chen. Zoom-net: Mining deep feature interactions for visual relationship recognition. In *ECCV*, pages 322–338, 2018.

[39] Mo Yu, Wenpeng Yin, Kazi Saidul Hasan, Cicero dos Santos, Bing Xiang, and Bowen Zhou. Improved neural relation detection for knowledge base question answering. In *ACL*, pages 571–581, 2017.

[40] Ruichi Yu, Ang Li, Vlad I Morariu, and Larry S Davis. Visual relationship detection with internal and external linguistic knowledge distillation. In *ICCV*, pages 1068–1076, 2017.

[41] Fang Yuan, Zhe Wang, Jie Lin, Luis Fernando D'Haro, Kim Jung Jae, Zeng Zeng, and Vijay Chandrasekhar. End-to-end video classification with knowledge graphs. *arXiv preprint arXiv: 1711.01714*, 2017.

[42] Vinicius Zambaldi, David Raposo, Adam Santoro, Victor Bapst, Yujia Li, Igor Babuschkin, Karl Tuyls, David Reichert, Timothy Lillicrap, and Edward Lockhart. Relational deep reinforcement learning. *arXiv preprint arXiv: 1806.01830*, 2018.

[43] Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. Neural motifs: Scene graph parsing with global context. In *CVPR*, pages 5831–5840, 2018.

[44] Hanwang Zhang, Zawlin Kyaw, Shih Fu Chang, and Tat Seng Chua. Visual translation embedding network for visual relation detection. In *CVPR*, pages 3107–3115, 2017.

[45] Yuyu Zhang, Hanjun Dai, Zornitsa Kozareva, Alexander J Smola, and Le Song. Variational reasoning for question answering with knowledge graph. In *AAAI*, pages 6069–6076, 2017.

[46] Bohan Zhuang, Lingqiao Liu, Chunhua Shen, and Ian Reid. Towards context-aware interaction recognition for visual relationship detection. In *ICCV*, pages 589–598, 2017.