



Research Article

Robust Visual Relationship Detection towards Sparse Images in Internet-of-Things

Yang He ¹, Guiduo Duan ^{1,2}, Guangchun Luo^{2,3} and Xin Liu³

¹School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China

²Trusted Cloud Computing and Big Data Key Laboratory of Sichuan Province, Chengdu 610000, China

³School of Information and Software Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China

Correspondence should be addressed to Guiduo Duan; duanguiduo@163.com

Received 21 April 2021; Accepted 5 July 2021; Published 20 July 2021

Academic Editor: Yan Huang

Copyright © 2021 Yang He et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Visual relationship can capture essential information for images, like the interactions between pairs of objects. Such relationships have become one prominent component of knowledge within sparse image data collected by multimedia sensing devices. Both the latent information and potential privacy can be included in the relationships. However, due to the high combinatorial complexity in modeling all potential relation triplets, previous studies on visual relationship detection have used the mixed visual and semantic features separately for each object, which is incapable for sparse data in IoT systems. Therefore, this paper proposes a new deep learning model for visual relationship detection, which is a novel attempt for cooperating computational intelligence (CI) methods with IoTs. The model imports the knowledge graph and adopts features for both entities and connections among them as extra information. It maps the visual features extracted from images into the knowledge-based embedding vector space, so as to benefit from information in the background knowledge domain and alleviate the impacts of data sparsity. This is the first time that visual features are projected and combined with prior knowledge for visual relationship detection. Moreover, the complexity of the network is reduced by avoiding the learning of redundant features from images. Finally, we show the superiority of our model by evaluating on two datasets.

1. Introduction

Visual relationship detection tries to simultaneously detect objects for an image and classify the predicate between each pair of these objects [1]. It has been considered as a bridge to semantically connect the low-level visual information [2–7] and high-level semantic information [8–11]. Generally, visual relationships indicate types of relations between objects in images and are usually represented by triplets (subject, predicate, and object), where the predicate can be a verb (person, ride, bicycle), spatial (cat, on, table), preposition (person, with, shirt), and comparative (elephant, taller, person) [1, 12]. The detection of these interactions can uncover diverse knowledge from images and significantly benefits the functionalities of IoT systems. Moreover, potential disclosure of sensitive information [13] can also be inferred with the autonomous relationship detection and provides guidelines for secure multimedia IoT data processing [14–16].

The early studies of visual relationship detection mainly rely on pure visual features capturing the complex visual variance of images [17, 18], suffering from the lack of diverse information for predicate classification. Considering the sparsity of IoT data, both the scale of the image dataset and the details within these images will be constrained. Sensing devices will be conservative on data publication [19, 20], especially when the image data contain abundant semantic information. Meanwhile, images maybe masked or obfuscated before publication due to privacy concerns [21]. Both constraints caused by sparsity of images have aggravated the difficulties for visual relationship detection, and purely visual-feature-based methods are not qualified.

Recently, additional sources of information, such as prior knowledge and semantic information, are incorporated into visual relationship detection [1, 22–24], as extra background information can be utilized to supply and refine the detection. Generally, two essential tasks are considered during

the incorporation of additional source of information: (1) How to apply the semantic associations among relationships [12, 25, 26] to refine the detection. For example, the relationship (person, ride, horse) is semantically similar to (person, ride, elephant) as the horse and elephant both belong to animals, even though horse and elephant are quite different in images. In this case, visual relationship detection models should be able to infer (person, ride, elephant) base on examples of (person, ride, horse). (2) How to alleviate the huge semantic space of possible relationships. Assume the category of objects to be N and the predicates to be K . Then, the number of possible relationships is $O(N^2K)$ as a relationship is composed of two objects [27]. Therefore, the size of semantic space in relationship detection increases by orders of magnitude, while many of relationships appear rarely in images. Visual relationship detection models should learn all relationship classes sufficiently.

Towards these tasks, extensive studies have been conducted. They mainly consider how to incorporate the additional source of information into the relationship detection. Initially, Lu et al. [1] introduced the additional language priors from semantic word embeddings to fine-tune the likelihood of a predicted relationship. Subsequently, Zhuang et al. [28] integrated the language representations of the subjects and objects as “context” to derive a better classification result for visual relationship detection. Then, Plummer et al. [8] use a large collection of linguistic and visual cues for the relationship detection in images, which contain attribute information and spatial relations between pairs of entities connected by verbs or prepositions. Furthermore, instead of using the pretrained and fixed language representations directly, Zhang et al. [29] tried to fine-tune the subject and object representations jointly and employ the interaction between visual branches to predict the relationship.

Although these methods achieve significant success, they still tend to focus on the word-level semantics [30] as the additional sources of information and lack in adopting the sophisticated knowledge and deep relations among objects. As for such kind of external knowledge, the knowledge graph is treated as a typical category of structural information providing abundant clues on relations between entities. It has been recently applied for many areas including computer vision and achieves dramatical improvements. Generally, a knowledge graph is a multirelational graph composed of entities (nodes) and relations (different types of edges). Each edge is a kind of relation in the form of triplets (head entity, relation, tail entity), indicating that two entities are connected by a specific relation. This type of additional information can provide more semantic association between objects and relations in an image and could be used for more rational reasoning to improve visual relationship detection. However, its application for visual relationship detection has not yet been properly considered, and neither of the above-mentioned tasks is investigated.

To take advantage of this type of information, this paper designs a deep neural network for visual relationship detection by considering the knowledge graph as an additional source of information. The input of the model includes the images and an external knowledge graph, and the outputs

are the relationships in images. The proposed model includes a visual module extracting the visual features of images, a knowledge module introducing the additional prior knowledge via the knowledge graph embedding [31], and a mapping module combining the visual features with prior knowledge. Finally, a new loss function based on the triplet loss [32] is designed in the mapping module to tune the projection of visual features into the knowledge space.

The proposed model uses the vector translation of the knowledge space for the first time, to capture the valuable structured information between objects and relations. By this mean, the structured semantic association in a knowledge graph can help improve the relationship detection. The proposed model also learns the objects and predicates and fuses them together to predict the relationship triplets [1]. This method can alleviate the impact of a huge semantic space of possible relationships, by reducing the space from $O(N^2K)$ to $O(N + K)$. Furthermore, the model achieves a reduced scale of parameters compared with state-of-the-art works [31], as it does not request the learning of visual features of predicates. The performance of the model is validated on two relation datasets: visual relationship detection (VRD) [1] with 5,000 images and 6,672 unique relations and visual genome (VG) [12] with 99,658 images and 19,237 unique relations. According to the comparison with several baselines, our model shows the superiority in visual relationship detection. In summary, the main contribution of this paper includes

- (1) We propose a novel framework for introducing the prior knowledge in visual relationship detection
- (2) Our model incorporates the priors in knowledge graph embedding for the first time to capture the valuable structured information between objects and relations
- (3) Our model reduces the parameters for extracting the visual features of predicates and designs a loss for combining the visual feature with the prior knowledge
- (4) Extensive evaluation shows that our model outperforms several strong baselines in visual relationship detection

This paper is organized as follows. The related works are introduced in Section 2. The proposed model is described in Section 3. The model is validated in VRD and VG datasets and compared with other methods in Section 4. The conclusion is described in Section 5.

2. Related Work

During the past years, there have been a number of studies in visual relationship detection. The earlier works regard visual relationships as an adminicle to improve the performance for other tasks, such as object detection [33, 34], image retrieval [12, 35, 36], and action recognition [37]. They focus on the specific types of relationships, such as spatial relationships

[2, 38], positional relationships [2, 35, 39], and actions (e.g., the interaction between objects) [40–42].

Lu et al. [1] first formalized the visual relationship as the (subject, predicate, object) triplet, defined the visual relationship detection task, and proposed a method by leveraging the language prior to model the more general correlation between objects. Afterwards, more studies on visual relationship detection have been developed, which can be divided into two categories: joint model and separate model.

For the joint model, it detects (subject, predicate, object) simultaneously by considering the relationship triplets as an integrated body [17, 22, 42–44], e.g., (person, ride, horse) and (person, ride, elephant) are of different classes. Vip-CNN [18] considers each visual relationship as a phrase with three components and formulates the visual relationship detection as three interconnected recognition problems. Plummer et al. [8] learned a Canonical Correlation Analysis (CCA) model on top of different combinations of the subject, object, and union regions and train a RankSVM to learn the visual relationship. However, it requires extremely large training data, because all possible combinations of predicates and entities (subject, object) are treated as independent classes. As a result, the general approaches usually pose the problem as a classification task in limited classes.

For a separate model, it first detects subjects and objects and then recognizes the possible interactions among them [1, 39, 45–47]. VtransE [48] uses the object detection output of a Faster-RCNN network and extracts features from every pair of objects to learn the visual translation embedding for relationship detection. Zhang et al. [29] embed the objects and relations of relationship triplets separately to the independent semantic spaces and then implicitly learn the connections between them via visual feature fusion and semantic meaning preservation in the embedding space.

The method proposed recently by Zhang et al. [29] is the most related one to ours. Compared with this work, instead of the word-level semantic embeddings, our work incorporates the knowledge graph and embeds it in a knowledge space as the additional sources of information. Due to the use of TransE [31] as the knowledge graph embedding, our work barely needs to model the large visual variance of relations in images.

Finally, our method adopts the additional semantic information to guide the visual recognition. This is consistent with the trend of using language information for visual recognition. For example, the language information has also been incorporated into visual question answering [49–52], few-shot learning [53–56], and image-sentence similarity task [57–60].

3. Method

3.1. Overview. The goal of the proposed model is to detect visual relationships from images which requires having discriminative power among a set of relationship categories. However, since object categories are often semantically associated, it is critical for a model to preserve semantic similarities, so as to benefit both frequent and rarely seen relationship categories.

The overview of the proposed model is shown in Figure 1. It consists of three modules, namely, visual module, knowledge module, and mapping module. The visual module detects a set of objects in images and extracts the visual features of the objects. The knowledge module consists of a knowledge graph, which is embedded in a low dimension vector space, so it can be used as the additional source of information. The mapping module considers the image and additional source of information comprehensively, which maps the visual features to the knowledge space for relationship detection. For any valid relationships, they are represented by the triplets (subject, predicate, object) in low dimension vectors \mathbf{s} , \mathbf{p} , and \mathbf{o} , respectively.

Note: in this paper, we use “relation” to refer to “predicate” in previous works and “relationship” to refer to the (subject, predicate, object) triplet. The detailed descriptions of notations can be found in Table 1.

3.2. Visual Module. The design of the visual module is based on the intuition that a relationship exists when its objects exist, but not vice versa. Therefore, to detect the visual relationships from images, the first step is to detect the objects and corresponding visual features in images.

In the visual module, the object detection is based on a Faster-RCNN [61] network with the VGG-16 [62] architecture, composed of a Region Proposal Network (RPN) and a classification layer. In the Faster-RCNN network, convolution does not change the size of the input image.

$$\text{output}_{\text{size}} = \frac{\text{input}_{\text{size}} - \text{kernel}_{\text{size}} + 2\text{pad}}{\text{stride}} + 1. \quad (1)$$

After that, the Feature Extraction Layer is proposed to extract \mathbf{x}_s and \mathbf{x}_o , when suppose $\mathbf{x}_s, \mathbf{x}_o \in \mathbb{R}^M$ are the M -dimensional visual features of the subject and object, respectively. The visual features \mathbf{x}_s and \mathbf{x}_o are obtained by concatenating the vector from the last convolution feature map in the Faster-RCNN network and the bounding box parameterization in [63].

3.3. Knowledge Module. A knowledge graph is represented by $G(V, E)$, while V is the set of nodes, which represents the entities (subjects, objects), and E is the set of edges, which represents the connections between entities. Hence, the relations between the subject and object can be represented by the connections between the entities in the knowledge graph, mainly describing real world entities and their interrelations organized in a graph. Compared with the word-level external information, this type of additional information can capture a more semantic association between objects and relations and be used for rational reasoning to improve the results of visual relationship detection.

The knowledge module introduces jointly a knowledge graph and projects it into an embedding space, to activate the rich prior knowledge in tuning the relationship detection. Translation embedding (TransE) [31] is a remarkable model that represents a valid relationship (subject, predicate, object) in the knowledge graph in low dimension

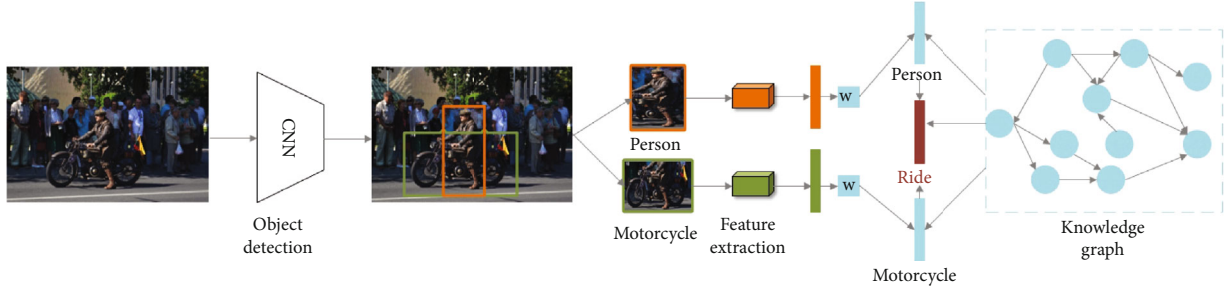


FIGURE 1: The overview of our visual relationship detection model. It consists of visual module, knowledge module, and mapping module. Visual module uses the CNN to detect a set of objects in images and extracts the visual features of the objects. Knowledge module consists of a knowledge graph, which is embedded in a low dimension vector space. Mapping module maps the visual features to the knowledge space for relationship detection.

vectors \mathbf{s} , \mathbf{p} , and \mathbf{o} , and $\mathbf{s}, \mathbf{p}, \mathbf{o} \in \mathbb{R}^r$, respectively. The relation is represented as a translation in the vector space:

$$\mathbf{y}_s + \mathbf{y}_p \approx \mathbf{y}_o, \quad (2)$$

when the relationship triplet holds, and $\mathbf{y}_s + \mathbf{y}_p \neq \mathbf{y}_o$ otherwise.

Since TransE offers a simple but effective method for representing the complex relationships in large knowledge graphs, it is adopted into the knowledge module for representing prior knowledge in the knowledge space. To learn such embeddings for the knowledge graph, we suppose a training set S of triplets (s, p, o) composed of two entities $s, o \in E$ (the set of entities) and a relation $p \in L$ (the set of relations). Since the relation is represented as a translation in the vector space, the energy of a triplet is defined by $d(\mathbf{y}_s + \mathbf{y}_p, \mathbf{y}_o)$, which regard the squared Euclidean distance as a dissimilarity function:

$$d(\mathbf{y}_s + \mathbf{y}_p, \mathbf{y}_o) = \|\mathbf{y}_s\|_2^2 + \|\mathbf{y}_p\|_2^2 + \|\mathbf{y}_o\|_2^2 - 2(\mathbf{y}_s^T \mathbf{y}_o + \mathbf{y}_p^T (\mathbf{y}_o - \mathbf{y}_s)). \quad (3)$$

To project the knowledge graph to knowledge space, we minimize a margin-based ranking criterion over the training set:

$$\mathcal{L} = \sum_{(s,p,o) \in S} \sum_{(s',p,o') \in S'} \left[\gamma + d(\mathbf{y}_s + \mathbf{y}_p, \mathbf{y}_o) - d(\mathbf{y}'_s + \mathbf{y}_p, \mathbf{y}'_o) \right]_+, \quad (4)$$

where $[x]_+$ denotes the positive part of x , $\gamma > 0$ is a margin hyperparameter, and

$$S'_{(s,p,o)} = \left\{ (s', p, o) \mid s' \in E \right\} \cup \left\{ (s, p, o') \mid o' \in E \right\}. \quad (5)$$

In the knowledge graph embedding, the loss function, constructed according to Equation (4), has lower values of the energy for training triplets than for wrong triplets, so the embeddings for the knowledge graph have the ability to distinguish wrong triplets. As for the wrong triplets, it is con-

structed according to Equation (5), which is composed of training triplets with either the subject or object replaced by a random entity (but not both at the same time).

3.4. Mapping Module. To consider the image visual feature and extra knowledge feature comprehensively, the mapping module is adopted to learn the joint visual and knowledge embedding. In the mapping module, there is a projection matrix $\mathbf{W} \in \mathbb{R}^{r \times M}$ from the feature space to the knowledge embedding space:

$$\mathbf{y}'_s = \mathbf{W} \mathbf{x}_s, \quad (6)$$

$$\mathbf{y}'_o = \mathbf{W} \mathbf{x}_o, \quad (7)$$

$$\mathbf{y}'_o - \mathbf{y}'_s \approx \mathbf{y}_p, \quad (8)$$

where \mathbf{y}'_s and \mathbf{y}'_o are the vector representations after the projection of \mathbf{x}_s and \mathbf{x}_o . To guarantee that the corresponding entities are close to each other during the projection process, a modified triplet loss is employed, where the triplet loss [32] can encourage matched entities from the two modalities to be closer than the mismatched ones by a fixed margin. To this end, two sets of entity triplets for each positive visual-knowledge pair are denoted by $(\mathbf{y}'_E, \mathbf{y}_E)$:

$$\text{tri}_{\mathbf{y}'_E} = \left\{ \mathbf{y}'_E, \mathbf{y}_E, \mathbf{y}'_E^- \right\}, \quad (9)$$

$$\text{tri}_{\mathbf{y}_E} = \left\{ \mathbf{y}'_E, \mathbf{y}_E, \mathbf{y}_E^- \right\}, \quad (10)$$

where $s, o \in E$ and the set $\text{tri}'_{\mathbf{y}'_E}$ and $\text{tri}_{\mathbf{y}_E}$ correspond to triplets with negatives from the visual mapping and knowledge space, respectively. If the superscripts $s, o \in E$ are omitted for clarity, the triplet loss \mathcal{L}^{Tr} is the summation of two losses $\mathcal{L}'_{\mathbf{y}'_E}$ and $\mathcal{L}_{\mathbf{y}_E}$:

$$\mathcal{L}^{Tr} = \sum_{i=1}^N \max \left[0, \text{sim}(\mathbf{y}'_i, \mathbf{y}_i^-) - \text{sim}(\mathbf{y}'_i, \mathbf{y}_i) + m \right], \quad (11)$$

$$\mathcal{L}_y^{Tr} = \sum_{i=1}^N \max \left[0, \text{sim} \left(\mathbf{y}'_i, \mathbf{y}_i^- \right) - \text{sim} \left(\mathbf{y}'_i, \mathbf{y}_i \right) + m \right], \quad (12)$$

$$\mathcal{L}^{Tr} = \mathcal{L}_{y'}^{Tr} + \mathcal{L}_y^{Tr}, \quad (13)$$

where $\mathcal{L}_{y'}^{Tr}$ guarantees that entities in knowledge space can be close to the corresponding entities in the visual mapping space, \mathcal{L}_y^{Tr} guarantees that the entities in visual mapping space can be close to the corresponding entities in knowledge space, N is the number of entities, m is the margin between the distances of positive and negative pairs, and $\text{sim}()$ is a similarity function, which is the cosine similarity function:

$$\text{sim} \left(\mathbf{y}'_i, \mathbf{y}_i \right) = \frac{\mathbf{y}_i \cdot \mathbf{y}'_i}{\|\mathbf{y}_i\| \times \|\mathbf{y}'_i\|}. \quad (14)$$

4. Experiments

Datasets: the *visual relationship detection (VRD)* [1] dataset contains 5,000 images with 100 object categories and 70 relations. In total, VRD contains 37,993 relationship annotations with 6,672 unique relationships and 24.25 relationships per object category. We follow the same train/test split as in previous works [1] to get 4,000 training images and 1,000 test images. To demonstrate that the proposed method can work reasonably well on a dataset with small relationship space, experiments in terms of visual relationship detection task are performed in the VRD dataset.

The *visual genome (VG)* [12] dataset is the latest release version (VG v1.4) that contains 108,077 images with 21 relationships on average per image. Each relationship is of the form (subject, relation, object) with annotated subject and object bounding boxes. Since the VG dataset is annotated by crowd workers, the objects and relations are noisy. Therefore, we clean it by removing nonalphabet characters and stop words and use the autocorrect library to correct spelling. Finally, the data is split into 86,462 training images and 21,615 testing images. The statistics for datasets can be found in Table 2.

Knowledge graph: in order to take advantage of the effective background knowledge, the knowledge graph for visual relationship detection is constructed according to the processed image label information and the public knowledge graph, WordNet [64]. To build the accurate knowledge graph, the annotation noise in the dataset should be removed. Firstly, duplicate words are deleted, such as “apple apple” and “dog dog.” Secondly, phrases with the same meaning are merged, such as “surfboard” and “surf board.” Specifically, the one with more occurrences in the dataset is selected and replaces other phrases with identical meaning. Then, we build the knowledge graph by using the object-object relationship annotations in the dataset.

However, this kind of knowledge graph lacks some common sense information. For instance, it can be helpful to know that a horse is a kind of animal. But if images of horse labels miss the “animal” label, our constructed knowledge

TABLE 1: Notations used in this paper.

Notations	Descriptions
\mathbb{R}^M	m -dimensional Euclidean space
$x, \mathbf{x}, \mathbf{X}$	Scalar, vector, and matrix, respectively
$\mathbf{x}_s, \mathbf{x}_o$	Feature of subject and object in image, respectively
$\mathbf{y}_s, \mathbf{y}_o$	Knowledge embedding of subject and object, respectively
\mathbf{y}_p	Knowledge embedding of predicate
\mathbf{d}	Dissimilarity function
S	Set of relation triplets
E	Set of entities
R	Set of relations
sim	Similarity function

TABLE 2: Statistics for the datasets.

Datasets	Images	Object types	Predicate types	Relationship types
VRD [1]	5,000	100	70	6,672
VG [12]	108,077	200	100	19,237

TABLE 3: Results on the VRD dataset.

Dataset	VRD					
	Phrase det.		Relationship det.		Predicate det.	
Metric	R@50	R@100	R@50	R@100	R@50	R@100
Lu’s-V [1]	2.24	2.61	1.58	1.85	7.11	7.11
Lu’s-VLK [1]	16.17	17.03	13.86	14.70	47.87	47.87
VtransE [48]	19.42	22.42	14.07	15.20	46.99	46.99
Ours	23.67	25.01	16.56	18.13	48.64	48.64

TABLE 4: Results on the VG dataset.

Dataset	VG					
	Phrase det.		Relationship det.		Predicate det.	
Metric	R@50	R@100	R@50	R@100	R@50	R@100
Lu’s-V [1]	—	—	—	—	—	—
Lu’s-VLK [1]	—	—	—	—	—	—
VtransE [48]	9.46	10.45	5.52	6.04	61.45	61.70
Ours	9.59	10.52	5.63	6.16	62.52	62.73

graph will also lack in this common sense. Thus, it is necessary to combine our constructed knowledge graph with the semantic knowledge graph, WordNet. First, we collect the new nodes in WordNet which directly connect to the nodes in the constructed knowledge graph. Then, we add these new nodes to our knowledge graph. Finally, we take all of the WordNet edges between these nodes and add them to the knowledge graph.



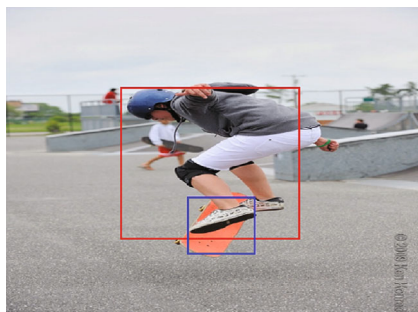
(a) Person, wear, skis



(b) Person, wear, skis



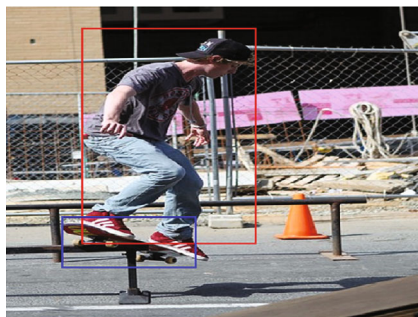
(c) Person, wear, skis



(d) Person, ride, skateboard



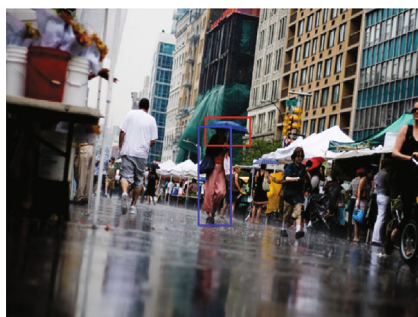
(e) Person, ride, skateboard



(f) Person, ride, skateboard

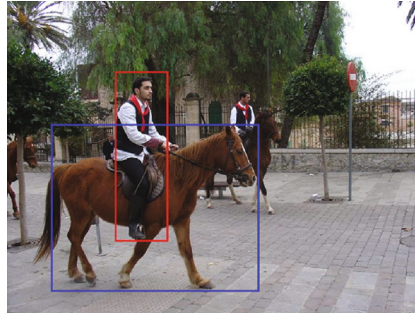


(g) Wheel, on, motorcycle



(h) Umbrella, cover, person

FIGURE 2: Continued.



(i) Person, ride, horse

FIGURE 2: Qualitative examples of relationship detection. The red rectangles are identified subjects, the blue rectangles are identified objects, and the captions below are identified visual relationships.

Detecting a visual relationship involves classifying both the objects, predicting the predicate, and localizing both the objects. To study the model’s performance for visual relationship detection, the visual relationship detection is measured in three tasks: (1) predicate detection: predict a set of possible predicates between pairs of objects, under the given ground truth object boxes and labels; (2) phrase detection: output a label (subject, predicate, object) and localize the entire relationship as one bounding box; and (3) relationship detection: output a set of (subject, predicate, object) and localize both subject and object in the image simultaneously.

Metrics: Recall@50 (R@50) and Recall@100 (R@100) are adopted as evaluation metrics for detection. R@K computes the fraction of times a correct relationship is predicted in the top K confident relationship predictions in an image. Note that precision and average precision (AP) are also widely used metrics, but they are not proper as visual relationships are labeled incompletely and they will penalize the detection if we do not have that particular ground truth.

Compared methods: we compare our model with three representative models. The three visual relationship detection models are as follows: (1) Lu’s-V (V-only in [1]): it is a two-stage separate model that first uses R-CNN [63] for object detection and then adopts a large-margin JointBox model for predicate classification; (2) Lu’s-VLK (V+L+K in [1]): a two-stage separate model that combines Lu’s-V and word2vec language priors [65]; (3) VtransE [48]: a fully convolutional visual architecture that draws upon the idea of knowledge embedding for predicate classification.

4.1. Comparison on VRD. The proposed model is first validated on the small VRD dataset with comparison to the similar methods using the metrics proposed above in Table 3. From the quantitative results, it can be found that the proposed model outperforms other methods in all tasks. Specifically, our proposed model improves performance by 4.25% in the phrase detection task and improves performance by 2.93% in the relationship detection task. These improvements validate the assumption that visual relationships might be helpful for object detection, which can be owed to the incorporation of the knowledge graph as an additional source of information. In addition, the improvement in predicate

detection shows that the incorporation of the knowledge graph can provide more meaningful information than the word-level text.

4.2. Comparison on VG. The results of the proposed model on the VG dataset are presented in Table 4. Since VG is a relatively large and newer dataset, some representative models have not been validated on it. In addition, some methods have no public codes, and we can only mark the performance of these methods as a blank in Table 4. Even though the variety of possible relationships becomes more diverse, our proposed model still outperforms other methods in all tasks. Specifically, our proposed model improves performance by 1.07% in the predication detection task. Since the predicate detection isolates the factor of subject/object localization accuracy by using ground truth subject/object boxes and labels, it focuses more on the relationship recognition ability of a model. Therefore, the improvement of our model in this task shows that the incorporation of the knowledge graph is essentially effective for visual relationship detection. Besides, the performance of our proposed model has been improved to some extent, but it is not obvious in phrase detection task and relationship detection task. It may be due to the noise annotations in the large-scale VG dataset and the limited quality of the constructed knowledge graph.

4.3. Case Study. The VRD and VG datasets have densely annotated relationships for images with a wide range of types. From the qualitative results in Figure 2, it shows that our model can clearly detect a wide variety of visual relationship categories. Specifically, in Figures 2(a)–2(c) are the same interactive relationships (person, wear, skis). Figures 2(d)–2(f) are the same positional relationships (person, ride, skateboard). It shows that our model can detect different types of identical relationship, even though their visual representations are quite divergent. Moreover, there are more categories of relationships, such as Figure 2(g) (wheel, on, motorcycle), Figure 2(h) (umbrella, cover, person), and Figure 2(i) (person, ride, horse). It shows that the proposed model can be able to cover all kinds of relationships in (subject, predicate, object), where the predicate can be a verb, spatial, and preposition.

5. Conclusion

The visual relationship detection has been treated as a critical task in enhancing the functionalities of IoTs with CI tools. Considering the sparsity of multimedia IoT data, this work investigates the improvement of visual relationship detection with the knowledge graph as the additional structural semantic information. We proposed a new model for visual relationship detection incorporating the knowledge graph. In the proposed model, the Faster-RCNN and TransE models are used for feature learning from the image and knowledge graph, respectively. A third module is proposed to combine the two parts at the level of low dimensional vectors. Furthermore, a corresponding loss function is designed for the whole network. We validate the effectiveness of the proposed model on several datasets, both on the classification and detection task, and demonstrate the superiority of our approach over other similar methods. The proposed model can be applied for both the knowledge discovery and security analysis for sparse multimedia IoT data. Our future work includes the combination of other techniques like graph neural networks for visual relationship detection, as well as the privacy preservation towards these multimedia IoT data.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This research work was partly supported by the Sichuan Science and Technology Program (2019YFG0507 and 2020YFG0328) and the National Natural Science Foundation of China (NSFC) (U19A2059). The work was also supported in part by the Young Scientists Fund of the National Natural Science Foundation of China under Grant No. 61802050.

References

- [1] C. Lu, R. Krishna, M. Bernstein, and L. Fei-Fei, "Visual relationship detection with language priors," in *Computer Vision – ECCV 2016. ECCV 2016. Lecture Notes in Computer Science, vol 9905*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds., pp. 852–869, Springer, Cham, 2016.
- [2] C. Galleguillos, A. Rabinovich, and S. Belongie, "Object categorization using co-occurrence, location and appearance," in *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, Anchorage, AK, USA, June 2008.
- [3] W. Liu, D. Anguelov, D. Erhan et al., "Ssd: single shot multibox detector," in *Computer Vision – ECCV 2016. ECCV 2016. Lecture Notes in Computer Science, vol 9905*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds., pp. 21–37, Springer, Cham, 2016.
- [4] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: unified, real-time object detection," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 779–788, Las Vegas, NV, USA, June 2016.
- [5] Z. Xiong, W. Li, Q. Han, and Z. Cai, "Privacy-preserving auto-driving: a GAN-based approach to protect vehicular camera data," in *2019 IEEE International Conference on Data Mining (ICDM)*, pp. 668–677, Beijing, China, November 2019.
- [6] X. Fan, M. Dai, C. Liu et al., "Effect of image noise on the classification of skin lesions using deep convolutional neural networks," *Tsinghua Science and Technology*, vol. 25, no. 3, pp. 425–434, 2020.
- [7] G. Li, Y. Zhao, L. Zhang, X. Wang, Y. Zhang, and F. Guo, "Entropy-based global and local weight adaptive image segmentation models," *Tsinghua Science and Technology*, vol. 25, no. 1, pp. 149–160, 2020.
- [8] B. A. Plummer, A. Mallya, C. M. Cervantes, J. Hockenmaier, and S. Lazebnik, "Phrase localization and visual relationship detection with comprehensive image-language cues," in *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 1928–1937, Venice, Italy, October 2017.
- [9] H. Izadinia, F. Sadeghi, and A. Farhadi, "Incorporating scene context and object layout into appearance modeling," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 232–239, Columbus, OH, USA, June 2014.
- [10] M. Elhoseiny, A. Elgammal, and B. Saleh, "Write a classifier: predicting visual classifiers from unstructured text," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 12, pp. 2539–2553, 2017.
- [11] Y. Liang, Z. Cai, J. Yu, Q. Han, and Y. Li, "Deep learning based inference of private information using embedded sensors in smart devices," *IEEE Network*, vol. 32, no. 4, pp. 8–14, 2018.
- [12] R. Krishna, Y. Zhu, O. Groth et al., "Visual genome: connecting language and vision using crowdsourced dense image annotations," *International Journal of Computer Vision*, vol. 123, no. 1, pp. 32–73, 2017.
- [13] X. Zheng, Z. Cai, and Y. Li, "Data linkage in smart internet of things systems: a consideration from a privacy perspective," *IEEE Communications Magazine*, vol. 56, no. 9, pp. 55–61, 2018.
- [14] Z. Cai, X. Zheng, and J. Yu, "A differential-private framework for urban traffic flows estimation via taxi companies," *IEEE Transactions on Industrial Informatics*, vol. 15, no. 12, pp. 6492–6499, 2019.
- [15] J. Pang, Y. Huang, Z. Xie, J. Li, and Z. Cai, "Collaborative city digital twin for the covid-19 pandemic: a federated learning solution," *Tsinghua Science and Technology*, vol. 26, no. 5, pp. 759–771, 2021.
- [16] J. Pang, Y. Huang, Z. Xie, Q. Han, and Z. Cai, "Realizing the heterogeneity: a self-organized federated learning framework for IoT," *IEEE Internet of Things Journal*, vol. 8, no. 5, pp. 3088–3098, 2021.
- [17] M. A. Sadeghi and A. Farhadi, "Recognition using visual phrases," in *CVPR 2011*, pp. 1745–1752, Colorado Springs, CO, USA, June 2011.
- [18] Y. Li, W. Ouyang, X. Wang, and X. Tang, "ViP-CNN: visual phrase guided convolutional neural network," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1347–1356, Honolulu, HI, USA, July 2017.
- [19] Z. Cai and X. Zheng, "A private and efficient mechanism for data uploading in smart cyber-physical systems," *IEEE Transactions on Network Science and Engineering*, vol. 7, no. 2, pp. 766–775, 2018.
- [20] Z. Cai and Z. He, "Trading private range counting over big IoT data," in *2019 IEEE 39th International Conference on*

- Distributed Computing Systems (ICDCS)*, pp. 144–153, Dallas, TX, USA, July 2019.
- [21] Z. Cai, Z. Xiong, H. Xu, P. Wang, W. Li, and Y. Pan, *Generative adversarial networks: a survey towards private and secure applications*, ACM Computing Surveys (CSUR), 2021.
- [22] Y. Atzmon, J. Berant, V. Kezami, A. Globerson, and G. Chechik, “Learning to generalize to new compositions in image understanding,” 2016, <http://arxiv.org/abs/1608.07639>.
- [23] A. Farhadi, M. Hejrati, M. A. Sadeghi et al., “Every picture tells a story: generating sentences from images,” in *Computer Vision – ECCV 2010. ECCV 2010. Lecture Notes in Computer Science, vol 6314*, K. Daniilidis, P. Maragos, and N. Paragios, Eds., pp. 15–29, Springer, Berlin, Heidelberg, 2010.
- [24] Y. Wu, X. Zhang, Y. Bian et al., “Second-order random walk-based proximity measures in graph analysis: formulations and algorithms,” *The VLDB Journal*, vol. 27, no. 1, pp. 127–152, 2018.
- [25] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “ImageNet: a large-scale hierarchical image database,” in *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255, Miami, FL, USA, June 2009.
- [26] J. Deng, N. Ding, Y. Jia et al., “Large-scale object classification using label relation graphs,” in *Computer Vision – ECCV 2014. ECCV 2014. Lecture Notes in Computer Science, vol 8689*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds., pp. 48–64, Springer, Cham, 2014.
- [27] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, “The PASCAL visual object classes (VOC) challenge,” *International Journal of Computer Vision*, vol. 88, no. 2, pp. 303–338, 2010.
- [28] B. Zhuang, L. Liu, C. Shen, and I. Reid, “Towards context-aware interaction recognition for visual relationship detection,” in *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 589–598, Venice, Italy, October 2017.
- [29] J. Zhang, Y. Kalantidis, M. Rohrbach, M. Paluri, A. Elgammal, and M. Elhoseiny, “Large-scale visual relationship understanding,” *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 9185–9194, 2019.
- [30] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” 2013, <http://arxiv.org/abs/1301.3781>.
- [31] A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, and O. Yakhnenko, “Translating embeddings for modeling multi-relational data,” in *Advances in neural information processing systems*, pp. 2787–2795, NIPS, 2013.
- [32] F. Schroff, D. Kalenichenko, and J. Philbin, “Facenet: a unified embedding for face recognition and clustering,” in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 815–823, Boston, MA, USA, June 2015.
- [33] M. J. Choi, J. J. Lim, A. Torralba, and A. S. Willsky, “Exploiting hierarchical context on a large database of object categories,” in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 129–136, San Francisco, CA, USA, June 2010.
- [34] M. P. Kumar and D. Koller, “Efficiently selecting regions for scene understanding,” in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 3217–3224, San Francisco, CA, USA, June 2010.
- [35] J. Johnson, R. Krishna, M. Stark et al., “Image retrieval using scene graphs,” in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3668–3678, Boston, MA, USA, June 2015.
- [36] S. Schuster, R. Krishna, A. Chang, L. Fei-Fei, and C. D. Manning, “Generating semantically precise scene graphs from textual descriptions for improved image retrieval,” in *Proceedings of the Fourth Workshop on Vision and Language*, pp. 70–80, Lisbon, Portugal, 2015.
- [37] A. Gupta, A. Kembhavi, and L. S. Davis, “Observing human-object interactions: using spatial and functional compatibility for recognition,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 10, pp. 1775–1789, 2009.
- [38] S. Gould, J. Rodgers, D. Cohen, G. Elidan, and D. Koller, “Multi-class segmentation with relative location prior,” *International Journal of Computer Vision*, vol. 80, no. 3, pp. 300–316, 2008.
- [39] A. Gupta and L. S. Davis, “Beyond nouns: exploiting prepositions and comparative adjectives for learning visual classifiers,” in *Computer Vision – ECCV 2008. ECCV 2008. Lecture Notes in Computer Science, vol 5302*, D. Forsyth, P. Torr, and A. Zisserman, Eds., pp. 16–29, Springer, Berlin, Heidelberg, 2008.
- [40] B. Yao and L. Fei-Fei, “Grouplet: a structured image representation for recognizing human and object interactions,” in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 9–16, San Francisco, CA, USA, June 2010.
- [41] G. Gkioxari, R. Girshick, and J. Malik, “Contextual action recognition with R* CNN,” in *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 1080–1088, Santiago, Chile, December 2015.
- [42] V. Ramanathan, C. Li, J. Deng et al., “Learning semantic relationships for better action retrieval in images,” in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1100–1109, Boston, MA, USA, June 2015.
- [43] X. Zheng and Z. Cai, “Privacy-preserved data sharing towards multiple parties in industrial IoTs,” *IEEE Journal on Selected Areas in Communications*, vol. 38, no. 5, pp. 968–979, 2020.
- [44] H. Zhao, X. Puig, B. Zhou, S. Fidler, and A. Torralba, “Open vocabulary scene parsing,” in *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 2002–2010, Venice, Italy, October 2017.
- [45] C. Desai, D. Ramanan, and C. C. Fowlkes, “Discriminative models for multi-class object layout,” *International Journal of Computer Vision*, vol. 95, no. 1, pp. 1–12, 2011.
- [46] F. Sadeghi, S. K. Kumar Divvala, and A. Farhadi, “Viske: visual knowledge extraction and question answering by visual verification of relation phrases,” in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1456–1464, Boston, MA, USA, June 2015.
- [47] B. Dai, Y. Zhang, and D. Lin, “Detecting visual relationships with deep relational networks,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3076–3086, Honolulu, HI, USA, July 2017.
- [48] H. Zhang, Z. Kyaw, S.-F. Chang, and T.-S. Chua, “Visual translation embedding network for visual relation detection,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5532–5540, Honolulu, HI, USA, July 2017.
- [49] J. Donahue, L. Anne Hendricks, S. Guadarrama et al., “Long-term recurrent convolutional networks for visual recognition and description,” in *2015 IEEE Conference on Computer Vision*

- and Pattern Recognition (CVPR)*, pp. 2625–2634, Boston, MA, USA, June 2015.
- [50] A. Karpathy and L. Fei-Fei, “Deep visual-semantic alignments for generating image descriptions,” in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3128–3137, Boston, MA, USA, June 2015.
 - [51] S. Antol, A. Agrawal, J. Lu et al., “Vqa: visual question answering,” in *2020 IEEE International Conference on Image Processing (ICIP)*, pp. 2425–2433, Abu Dhabi, UAE, October 2015.
 - [52] P. Wang, Q. Wu, C. Shen, A. Dick, and A. van den Hengel, “FVQA: fact-based visual question answering,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 10, pp. 2413–2427, 2018.
 - [53] A. Frome, G. S. Corrado, J. Shlens et al., “Devise: a deep visual-semantic embedding model,” in *Advances in neural information processing systems*, pp. 2121–2129, NIPS, 2013.
 - [54] O. Vinyals, C. Blundell, T. Lillicrap, D. Wierstra, and K. Kavukcuoglu, “Matching networks for one shot learning,” in *30th Conference on Neural Information Processing Systems (NIPS 2016)*, pp. 3630–3638, Barcelona, Spain, 2016.
 - [55] M. Norouzi, T. Mikolov, S. Bengio et al., “Zero-shot learning by convex combination of semantic embeddings,” 2013, <http://arxiv.org/abs/1312.5650>.
 - [56] M. Elhoseiny, S. Cohen, W. Chang, B. Price, and A. Elgammal, “Sherlock: scalable fact learning in images,” *Thirty-First AAAI Conference on Artificial Intelligence*, vol. 31, no. 1, 2017.
 - [57] R. Kiros, R. Salakhutdinov, and R. S. Zemel, “Unifying visual-semantic embeddings with multimodal neural language models,” 2014, <http://arxiv.org/abs/1411.2539>.
 - [58] I. Vendrov, R. Kiros, S. Fidler, and R. Urtasun, “Order-embeddings of images and language,” 2015, <http://arxiv.org/abs/1511.06361>.
 - [59] Y. Gong, Q. Ke, M. Isard, and S. Lazebnik, “A multi-view embedding space for modeling internet images, tags, and their semantics,” *International Journal of Computer Vision*, vol. 106, no. 2, pp. 210–233, 2014.
 - [60] F. Faghri, D. J. Fleet, J. R. Kiros, and S. Fidler, “VSE++: improving visual-semantic embeddings with hard negatives,” 2017, <http://arxiv.org/abs/1707.05612>.
 - [61] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: towards real-time object detection with region proposal networks,” *Advances in Neural Information Processing Systems*, vol. 28, pp. 91–99, 2015.
 - [62] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” 2014, <http://arxiv.org/abs/1409.1556>.
 - [63] R. Girshick, J. Donahue, T. Darrell, and J. Malik, “Rich feature hierarchies for accurate object detection and semantic segmentation,” in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 580–587, Columbus, OH, USA, June 2014.
 - [64] A. Bordes, X. Glorot, J. Weston, and Y. Bengio, “A semantic matching energy function for learning with multi-relational data,” *Machine Learning*, vol. 94, no. 2, pp. 233–259, 2014.
 - [65] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, “Distributed representations of words and phrases and their compositionality,” in *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2, NIPS’13, Curran Associates Inc.*, pp. 3111–3119, Red Hook, NY, USA, 2013.