A Probabilistic Graphical Model Based on Neural-symbolic Reasoning for Visual Relationship Detection

A Probabilistic Graphical Model Based on Neural-symbolic Reasoning for Visual Relationship Detection

Dongran Yu[1, 2], Bo Yang[1, 3]‡, Qianhao Wei[1, 3] and Anchen Li[1, 3], Shirui Pan[4],
[1]Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, China
[2]College of Artificial Intelligence, Jilin university, China
[3]College of Computer Science and Technology, Jilin university, China
[4]Department of Data Science and AI, Faculty of IT, Monash University, Australia
yudran@foxmail.com, ybo@jlu.edu.cn, {weiqh20,liac20}@mails.jlu.edu.cn, shirui.pan@monash.edu

Abstract

*This paper aims to leverage symbolic knowledge to improve the performance and interpretability of the Visual Relationship Detection (VRD) models. Existing VRD methods based on deep learning suffer from the problems of poor performance on insufficient labeled examples and lack of interpretability. To overcome the aforementioned weaknesses, we integrate symbolic knowledge into deep learning models and propose a **bi**-level **p**robabilistic **g**raphical **r**easoning framework called* BPGR*. Specifically, in the high-level structure, we take the objects and relationships detected by the VRD model as hidden variables (reasoning results); In the low-level structure of* BPGR*, we use Markov Logic Networks (MLNs) to project First-Order Logic (FOL) as observed variables (symbolic knowledge) to correct error reasoning results. We adopt a variational EM algorithm for optimization. Experiments results show that our* BPGR *improves the performance of the VRD models. In particular,* BPGR *can also provide easy-to-understand insights for reasoning results to show interpretability.*

1. Introduction

The goal of Visual Relationship Detection(VRD) is to detect objects as well as their relationships with each other, representing as (subject, predicate, object) triplet. As shown Fig. 1 (a), a triplet is (person, hold, horse). As a foundation visual recognition task, VRD can benefit a wide range of high level image understanding tasks, such as scene graph generation [38, 43], image captioning [15, 17], and visual question answering [3, 4], etc. Consequently, VRD has emerged as an important research topic in the past few years. Most recent methods based on deep learning have been proposed, including LS-VRU [46] and GPS-Net [21], UVTransE [13], etc. However, these methods mainly rely on language prior (semantic information) of entities to help relationship detection, which suffers several limitations. First, these approaches need a lot of labeled examples to get decent performance, which goes against the characteristics of the dataset in Fig. 1 (b). Second, they are black boxes, lacking interpretability which is very important for many applications. An expected solution is neural-symbolic systems, which combines the excellent perceptual ability of neural networks and the cognitive ability of symbolic systems [41].

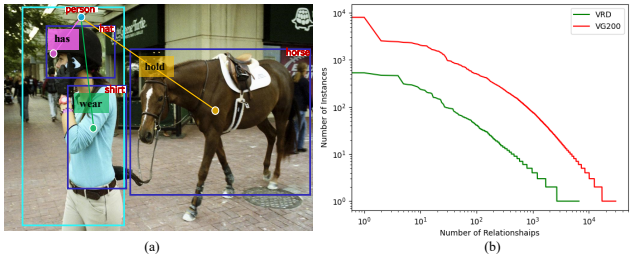

Figure 1. An example for visual relationship detection and statistic of datasets. (a) Different colored blocks with a line connecting subject and object mean different relationships. It is detected three triples (person, has, hat), (person, wear, shirt), and (person, hold, horse) in the image. (b)VRD and VG200 are long-tail distribution. The horizontal axis represents the number of relationships, and the vertical axis represents the number of instances of the relationship.

Recently some studies attempted to explore combining symbolic knowledge with the VRD models to enforce the performance of detection. LENSR [37] uses Conjunctive Normal Form (CNF) [8] or decision-Deterministic Decomposable Negation Normal Form (d-DNNF) [9] formulae to construct a graph for each propositional logic and adopts

‡Corresponding author.

graph neural network to encode them to models. DASL [33] encodes logic rules into the structure of the deep learning model for training. While these methods generally enrich the flexibility compared to the pure deep learning methods, they still have unnoticeable deficiencies. First, it is only capturing local information in LENSR, i.e., they construct an independent graph for each propositional logic and only encode interaction information between nodes in a propositional logic. Second, it is an implicit reasoning process in DASL. i.e., DASL encodes First-Order Logic as the neural network structure, and then the neural network will complete the next work.

To offset the above deficiencies, we adopt Markov Logic Networks (MLNs) [31] to represent First-Order Logic (FOL) and combine logic with the deep model in a probabilistic graphical model. MLN can build a global dependency graph for all FOLs and attain a joint probability distribution for all ground atoms. Furthermore, MLNs can be used as a general framework for joining logical AI and statistical AI, and can capture uncertainty. The probabilistic graphical model solves the model by way of probabilistic inference, reflecting an explicit reasoning process.

Therefore, we propose a bi-level probabilistic graphical reasoning framework (BPGR) to encode symbolic knowledge into the VRD model. BPGR includes two parts: the visual reasoning module and the symbolic reasoning module. The visual reasoning module extracts features of objects in images and reasons objects and relationships. The symbolic reasoning module uses symbolic knowledge to guide the reasoning of the visual reasoning module towards a good direction, which acts as an error correction. Specifically, the symbolic reasoning module is a double layer probabilistic graph and contains two types of nodes: one is the reasoning result of the VRD model (the visual reasoning module) in the high-level structure, and the other is ground atoms of logic rules in the low-level structure. When the probabilistic graphical model is constructed, the model can be trained efficiently end-to-end in the variational expectation–maximization (EM) framework. In particular, BPGR achieves superior performance on visual relationship detection dataset [23] and the scene graph dataset [38] and is also shown interpretable for reasoning results. An overall framework of our method is given in Fig. 2.

Our contributions can be summarized in threefold:

- We propose bi-level probabilistic graphical reasoning (BPGR) framework which is a novel VRD model based on neural-symbolic systems to improve the detection performance and provide interpretability of results. Our BPGR uses symbolic knowledge to guide the model towards improved performance and rectifies error reasoning results.

- We present a joint framework for modeling symbolic knowledge and VRD models. Our framework can capture global symbolic knowledge in logic rules and maintain an explicit reasoning process than existing neural-symbolic methods because it applies Markov Logic Network (MLN) as knowledge representation and integrates by way of probabilistic inference.

- Experimental results show that BPGR performs better on two datasets of visual relationship detection, compared to state-of-the-art methods. We provide visualized results to show efficacy and interpretability.

2. Related work

Neural-symbolic systems. Recently, neural-symbolic reasoning has become a hot topic. It can combine the advantages of both neural network and symbol, not only reducing data requirements but also enabling explainable artificial intelligent, such as pLogicNet [30], ExpressGNN [48], DGP [14], CA-ZSL [24], VAI-SC [1]etc. These methods use logic rules or knowledge graphs to improve the ability of the knowledge graph reasoning or image classification or generate descriptions for video, which are quite different from the VRD task-focused in this paper.

Markov Logic Networks. Intelligent systems must be able to handle the complexity and uncertainty of the real world. MLN enables this by unifying FOL and probabilistic graphical models into a single representation. It has been widely studied due to the principle probabilistic models and effectiveness in a variety of reasoning tasks, including knowledge graph reasoning [30, 48], semantic parsing [29, 36], social networks analysis [47], etc. MLN can capture the complexity and uncertainty in relation data. However, inference and learning in MLN are computationally expensive due to the exponential cost of constructing the ground MLN and the NP-hard optimization problem. This hinders MLN to be applied to large-scale applications. Many works appear in the literature to improve original MLN in accuracy [25, 34], and efficiency [6, 16, 30, 35, 48]. For example, related works [30, 48] replace traditional inference algorithms with neural networks.

Visual relationship detection. Visual relationship detection involves detecting the objects that occur in an image as well as understanding the interactions between them. In other words, it requires recognizing relationships from the image. Most of these approaches can be divided into three broad categories. The first group of methods uses structured prediction techniques by message passing among the three triplet variables [7, 19, 38, 49]. These methods take into account triplet dependencies by message passing among object and predicate labels. The second group of methods applies rank-based loss functions to encourage similar relations to be close to each other in the learning feature

space [18, 46]. The third branch of approaches introduces extra information either in the form of word vector embeddings of the object labels or use knowledge from a large corpus or logic [2, 20, 33, 37, 42, 45].

Our method can be categorized into the third category, i.e., aiming to add extra information. In contrast to the above-mentioned methods, BPGR has a consistent probabilistic model built in the framework and can incorporate symbolic knowledge in logic rules. Further, BPGR provides interpretability while capturing rich external information.

3. Bi-level probabilistic graphical reasoning framework

In neural-symbolic systems, the final objective is to find a model F that can effectively map data I and symbolic knowledge R (logic rules) to ground truth Y. In this paper, the model is defined in Eq. (1).

$$\forall(I, Y)\ F(I, R) \rightarrow Y, \tag{1}$$

Based on the model's definition, our BPGR includes two main components: the visual reasoning module $P_{\theta_1}(y|I)$ and the symbolic reasoning module $P_{\theta_2,w}(y, R)$, where y is the preliminary reasoning result of the visual reasoning module, θ_1 and θ_2 are parameters, and w denotes the weights of logic rules. Fig. 2 shows BPGR's framework. The former aims to attain scores of objects and relationships from an image. The latter takes the result of the visual reasoning module as the high-level nodes and the logic rules as the low-level nodes of the probabilistic graphical model, respectively. During training, we minimize the loss of the visual reasoning module and maximize the joint probability distribution of the symbolic reasoning module by a variational EM algorithm. During testing, we feed an image to the visual reasoning module to infer results, then inferred results are spread from the high-level structure to the low-level structure to match logic rules as decision evidence. The following section describes the whole model in detail.

3.1. Visual reasoning module

In this section, based on LS-VRU [46], we develop our visual reasoning module (VRM). VRM's main idea is to minimize the distance between visual features and semantic features in objects and relations respectively. Specifically, an image is feed and outputs object score matrix $\mathbf{S_O} \in \mathbb{R}^{T\times O}$ and relation score matrix $\mathbf{S_R} \in \mathbb{R}^{M\times Re}$. Meanwhile, saving object's feature $\mathbf{M_O} \in \mathbb{R}^{T\times D}$ to as an input of the symbolic reasoning module. T represents the number of objects in an image. O means the number of the object category in the dataset. M is the number of object pairs. Re means the number of relation category in the dataset. D is the dimension of visual features. The above notations are used by the VRM part of Fig. 2. The visual reasoning module adopts word2vec [26] as a semantic feature in experiment.

3.2. Symbolic reasoning module

The symbolic reasoning module (SRM) is the key component that makes the model different from existing methods. VRM's reasoning results may be error, and we are inspired by methods in image de-noising [5] to design a probabilistic graphical model to combine VRM and SRM. Therefore, this probabilistic graphical model can correct error reasoning results and achieve end-to-end training.

Logic rules are a kind of commonsense knowledge that is easy to be understood by people. In this paper, we consider the FOL language that describes knowledge in the form of the logic rule, which has strong expression ability [10]. In Fig. 2, the SRM includes two types of nodes and cliques. Let y be the set of high-level nodes, and let A be the set of low-level nodes consisting of ground atoms in logic rules. Let $\{y_i, A_j\}$ be a clique expressing the correlation between levels. Let subset $A_r = \{A_1, \cdots, A_m\}$ be a clique consisting of the ground atoms in terms of a logic rule r, by assigning constants to its arguments. Let each node of the SRM represent a random variable, and the probabilistic graphical model represents a joint probability distribution over variables as a product of factors, the formula is as follows:

$$P(y, R) = \frac{1}{Z}\exp\{\sum_{y_i\in y, A_j\in A}\phi_b(y_i, A_j) + \sum_{r, A_r}\phi_l(A_r)\}, \tag{2}$$

where Z is a normalization constant known as the partition function, ϕ_b is the potential function between levels and implies a distribution that encourages the connected high-level nodes and low-level nodes to take the same values. ϕ_l is the potential function of low level.

In the high-level structure, nodes indicate reasoning results and there are no edges among nodes. To attain the clique $\{y_i, A_j\}$, we build connections between the high-level node and the low-level node according to their identifiers, which are defined in terms of both object region and predicate arity. Nodes with the same identifiers connect.

In the low-level structure, nodes are ground atoms A_j in FOL, and edges are constructed by MLN. MLN is an undirected graphical model, where nodes are generated by all ground atoms, and edge appears between two nodes if the two corresponding ground atoms cooccur in at least one ground FOL. Given the same MLN and different constant sets C, one can form different ground MLN. The scale of ground MLN is determined by the size of the constant set C. A ground MLN can be defined as a joint distribution as:

$$P(A) = \frac{1}{Z(w)}\exp\{\sum_{r\in R} w_r \sum_{A_r}\phi_l(A_r)\}, \tag{3}$$

where $Z(w)$ is the partition function summing overall ground atoms A. ϕ_l is a potential function in terms of the

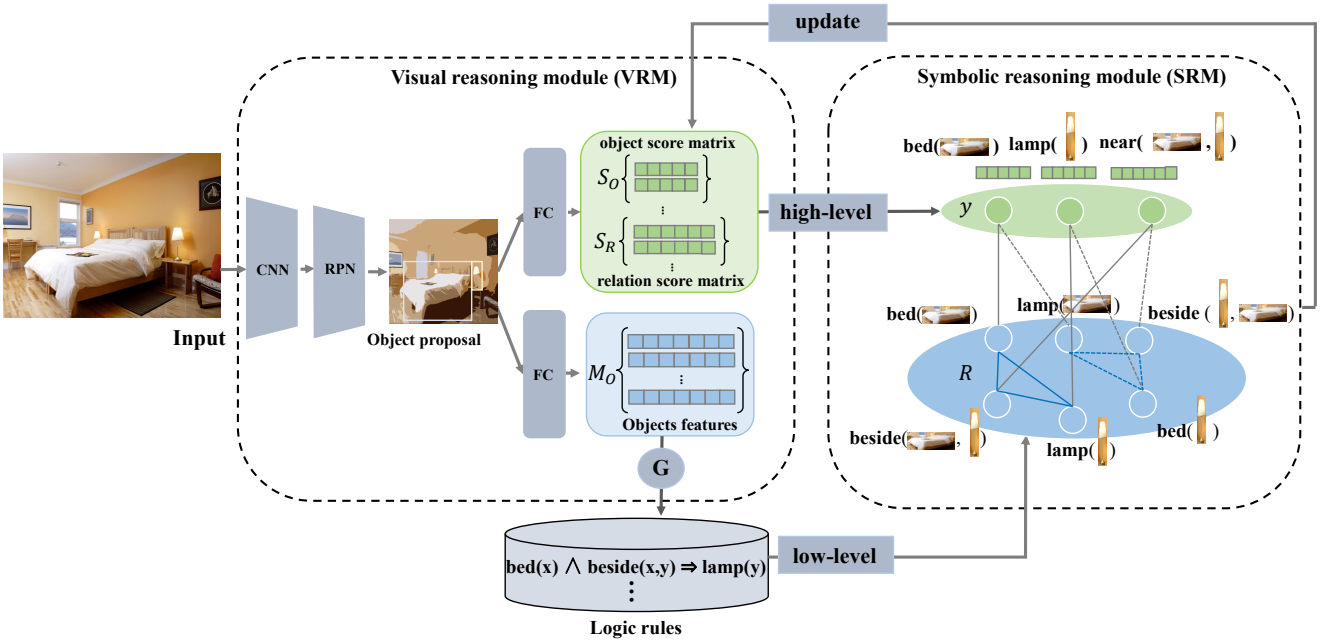

Figure 2. The proposed BPGR. To introduce logic rules and correct error reasoning results of the VRD model, we design a bi-level probabilical graphical reasoning framework, where the high-level structure is designed to take reasoning results of the visual reasoning module, while the low-level structure is the ground atom of logic rules to correct the error in the high-level structure, such as correcting "near" to "beside". The model is trained to output reasoning results of the visual reasoning module based on symbolic knowledge. Note, **G** means grounding operator. The solid line represents ground truth edges, the dotted line represents pseudo edges.

number of times the logic rule is true. w_r represents the weight of logic rule r. The greater the weight, the greater the confidence of the logic rule.

3.3. Training loss

Our model is trained end-to-end. The final training loss includes: VRM's loss $P_{\theta_1}(y|I)$ and SRM's loss $P_{\theta_2,w}(y,R)$, and cross-entropy of observed variables L_{cro}. Therefore, we train the model by the following objective:

$$\mathcal{L} = P_{\theta_1}(y|I) + P_{\theta_2,w}(y,R) - L_{cro}, \tag{4}$$

$P_{\theta_1}(y|I)$ includes three terms: triplet loss L^T, triplet softmax loss L^S and visual consistency loss L^C. Due to limited space, see [46] for reference.

The loss of SRM is Eq. (2), which can be rewritten in Eq. (5). The cross-entropy of the observed variable L_{cro} is given in Section 3.4.

$$P_{\theta_2,w}(y,R) = \frac{1}{Z(w)} \exp\{ \sum_{y_i \in y, A_j \in A} \phi_b(y_i, A_j) + \sum_{r \in R} w_r \sum_{A_r} \phi_l(A_r) \}, \tag{5}$$

3.4. Optimization

We need to maximize $\mathcal{L}$ to train the whole model. However, due to the requirement of computing the partition function $Z(w)$ of $P_{\theta_2,w}(y,R)$, it is intractable to directly optimize this objective function. Different from [22], we introduce the variational EM algorithm and optimize the variational evidence lower bound (ELBO):

$$L_{EBLO} = E_{Q_{\theta_2}}[\log P_w(y,R)] - E_{Q_{\theta_2}}[\log Q_{\theta_2}(y \mid R)], \tag{6}$$

where $Q_{\theta_2}(y|R)$ is the variational posterior distribution.

In general, we can use the variational EM algorithm [11] to optimize the ELBO, that is to minimize KL divergence between the variational posterior distribution $Q_{\theta_2}(y|R)$ and the true posterior distribution $P_w(y|R)$ during the E-step. Due to the complicated graph structure among variables, the exact inference is computationally intractable. Therefore, we adopt a mean-field distribution to approximate the true posterior. In the mean-field variational distribution, between variables is independently inferred as follows:

$$Q_{\theta_2}(y|R) = \prod_{A_i \in A} Q_{\theta_2}(A_i), \tag{7}$$

Instead of traditional inference methods of $Q_{\theta_2}(A_i)$ with a MLP as inference network [48], we employ the logic tensor network (LTN) [32], which can learn representation of relation data. The inference process is illustrated in Fig. 3. In the E-step, our $L_{ELBO}(Q_{\theta_2}, P_w)$ in Eq. (6) can be

rewritten as follow:

$$L_{ELBO}(Q_{\theta_2}, P_w) = \sum_{r \in R} w_r \sum_{A_r} E_{Q_{\theta_2}}[\phi_l(A_r)] - \log Z(w) + \sum_{y_i \in y, A_j \in A} \phi_b(y_i, A_j) - E_{Q_{\theta_2}}[\sum_{A_i \in A} Q_{\theta_2}(A_i)], \tag{8}$$

$$L_{cro} = \sum_{A_i \in A} Q_{\theta_2}(A_i) \log Y, \tag{9}$$

Eq. (4) is rewritten:

$$\mathcal{L} = \alpha P_{\theta_1}(y|I) + \beta L_{ELBO}(Q_{\theta_2}, P_w) - \gamma L_{cro}, \tag{10}$$

where α, β and γ are trade-off factor whose domains are in the interval $[0, 1]$.

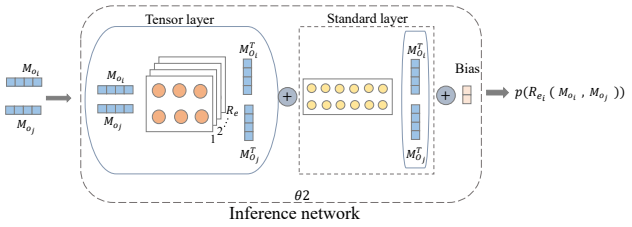

Figure 3. Inference network that inputs feature embeddings of objects pair and outputs the probability of affiliation relationships.

In the M-step, we are learning the weight of the FOL. As we need to optimize weights, the partition function $Z(w)$ in Eq. (5) is not a constant anymore. The partition function $Z(w)$ has an exponential number of terms, which makes it intractable to directly optimize ELBO. To solve the above problem, we use pseudo-log-likelihood [31], which is defined as:

$$P_w^*(y, R) := E_{Q_{\theta_2}}[\sum_{r, A_i \in A_r} \log P_w(A_i | MB_{A_i})], \tag{11}$$

where MB_{A_i} is Markov blanket of the ground atom A_i. For each rule r that connects A_i to its Markov blanket, we optimize the weights w_r by gradient descent, the derivative is following:

$$\bigtriangledown_{w_r} E_{Q_{\theta_2}}[\log P_w(A_i | MB_{A_i})] \simeq Y_{A_i} - P_w(A_i | MB_{A_i}), \tag{12}$$

where Y_{A_i} = 0 or 1 if A_i is an observed variable, and Y_{A_i} = $Q_{\theta_2}(A_i)$ otherwise.

3.5. Interpretability

Our proposed model can find corresponding logic rules for reasoning results to validate the results are credible. In the SRM, information is propagated from the low-level structure to the high-level structure during training, which uses the logic rule to guide VRD models learning. Inspired by production system, information is propagated from the high-level structure to the low-level structure, which can find corresponding logic rules for reasoning results. Specifically, according to reasoning results to match nodes of the low-level structure, if the match is successful, the logic rule containing nodes are triggered (that is, the clique consisting of the nodes is chosen). We will calculate the probability that the triggered rule is true according to t-norm fuzzy logic [28]. In other words, the whole process is the same as finding a path in the knowledge graph given head entity and tail entity. Implementing this process, we can attain evidence (logic rules) for reasoning results and choose the top pieces of evidence according to posterior $P(R \mid y)$ here. The equation as following:

$$P(R \mid y) = \prod_{r, A_i \in A_r} p(A_i \mid y). \tag{13}$$

where r is a triggered logic rule here. p is the probability of the ground atom is true, and the illustrative visual analysis is given in Section 4.6.

4. Experiments

In this section, we evaluate our model on two classical datasets: Visual Relationship Detection(VRD) [23] and Visual Genome with 200 categories (VG200) [38]. They are widely used in previous studies [21,46]. Next, we introduce two datasets in detail.

4.1. Datasets

The VRD [23] contains 5,000 images, with 4,000 as train sets and 1,000 as test sets. There are 100 object classes and 70 predicates (relations). The VRD includes 37,993 relation annotations with 6,672 unique relations and 24.25 relationships per object category. This dataset contains 1,877 relationships in the test set never occur in the training set, thus allowing us to evaluate the generalization of our model in zero-shot prediction.

The VG200 [38] contains 150 object categories and 50 predicates. Each image has a scene graph of around 11.5 objects and 6.2 relationships. 70% of the images is used for training and the remaining 30% is used for testing.

The logic rules. To generate logic rules, we use way of artificially constructed based on the training set. In this paper, logic rules encode relationship between a subject and multiple objects. They are constructed according to the label file in datasets, visual relationship together with its subject and object forms a logic rule. As shown in Fig. 4, if triplet including (person, wear, jacket) and (person, wear, skis), $\mathsf{person}(x) \wedge \mathsf{wear}(x, y) \Rightarrow \mathsf{jacket}(y) \vee \mathsf{skis}(y)$ is a logic rule and atom $\mathsf{person}(x)$ is true if x is a person in image. The numbers of logic rule is 1,642 and 3,435 on VRD and VG200 datasets respectively.

Table 1. Comparison with state-of-the-art on the VRD dataset. Table 1 comparative recall results for top 50/100 in "ReD" and "PhD" respectively on VRD dataset. The best result is highlighted in bold. The result of state-of-the-art methods is taken from the original papers. "—" denotes the corresponding result is not provided.

Methods	ReD		PhD		ReD				PhD			
	$free\ k$				$k=1$		$k=70$		$k=1$		$k=70$	
Recall@	50	100	50	100	50	100	50	100	50	100	50	100
Lk distilation [42]	22.7	31.9	26.5	29.8	19.2	21.3	22.7	31.9	23.1	24.0	26.3	29.4
Zoom-Net [40]	21.4	27.3	29.1	37.3	18.9	21.4	21.4	27.3	28.8	28.1	29.1	37.3
CAI+SCA-M [40]	22.3	28.5	29.6	38.4	19.5	22.4	22.3	28.5	25.2	28.9	29.6	38.4
MF-URLN [44]	23.9	26.8	31.5	36.1	23.9	26.8	—	—	23.9	26.8	—	—
LS-VRU [46]	27.0	32.6	32.9	39.6	23.7	26.7	27.0	32.6	28.9	32.9	32.9	39.6
GPS-Net [21]	27.8	31.7	33.8	39.2	—	—	27.8	31.7	—	—	33.8	39.2
UVTransE [13]	27.4	34.6	31.8	40.4	25.7	**29.7**	27.3	34.1	30.0	36.2	31.5	39.8
BPGR(E)	28.1	34.7	34.7	42.0	24.7	27.9	28.1	34.7	29.9	34.1	34.7	42.0
BPGR(E+M)	**29.4**	**35.3**	**36.2**	**43.0**	**26.2**	29.4	**29.4**	**35.3**	**32.3**	**36.4**	**36.2**	**43.0**

Table 2. Comparative results for top 50/100 in "SGCLS" and "PCLS" respectively on the VG200 dataset. The best result is highlighted in bold.

Metrics / Recall@	SGCLS			PCLS		
Methods	20	50	100	20	50	100
VRD [23]	—	11.8	14.1	—	27.9	35.0
Ass-Embedding [27]	18.2	21.8	22.6	47.9	54.1	55.4
Mess-Passing [38]	31.7	34.6	35.4	52.7	59.3	61.3
Graph-RCNN [39]	—	29.6	31.6	—	54.2	59.1
Per-Invariant [12]	—	36.5	38.8	—	65.1	66.9
Motifnet [43]	32.9	35.8	36.5	58.5	65.2	67.1
LS-VRU [46]	36.0	36.7	36.7	66.8	68.4	68.4
GPS-Net [21]	36.1	39.2	**40.1**	60.7	66.9	68.8
BPGR($k=1$)	**37.0**	**39.3**	39.3	**67.8**	**69.1**	**70.0**

4.2. Evaluation metrics

For VRD, we adopt evaluation metrics same as [46], which runs **Relationship detection (ReD)** and **Phrase detection (PhD)** and shows recall rates (Recall@) for the top 50 /100 results, with $k = 1, 70$ candidate relations per relationship proposal (or k relationship predictions for per object box pair) before taking the top 50/100 predictions. **ReD** is to input an image and output labels of triples and boxes of the objects. **PhD** is to input an image and output labels and boxes of triples.

For VG200, we use the same evaluation metrics used in [46], including 1) **Scene Graph Classification (SGCLS)**, which is to predict labels of the subject, object, and predicate given ground truth subject and object boxes; 2) **Predicate Classification(PCLS)**, where predict predicate labels are given ground truth subject and object boxes and labels. Recall@ under the top 20/50/100 predictions are reported.

For the logic rule, we compute the probability of a logic rule that is true to as evaluation of logic rules. Here, we adopt Łukaseiwicz of t-norm fuzzy logic [28].

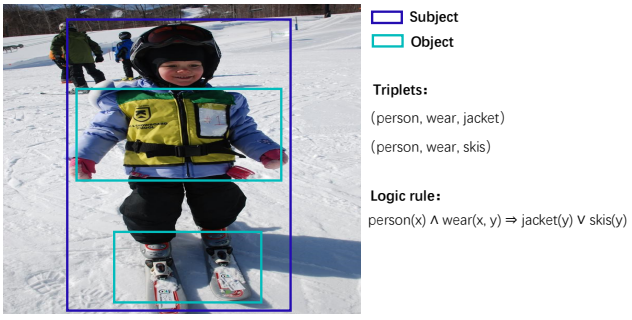

Figure 4. A generated logic rule. We obtain the subjects, the objects and relations directly from the annotations for the image. The body of the logic rule includes two atoms based on subject and relation. Two atoms are combined by "∧". The head of the logic rule consists of the object. All these atoms are combined by "∨".

4.3. Implementation details

In experiment, we adopt Faster-RCNN with the VGG16 backbone as an object detector and our model is trained for 8 epochs on a single NVIDIA TITAN RTX. The learning rate is 0.001 for the first 5 epochs and is 0.0001 for the rest 3 epochs. Dimension of the object feature is D = 512. The visual reasoning module is initialized with weights pretrained on the COCO dataset.

4.4. Results and analysis

We first show our experimental results and state-of-the-art methods in Table 1 for the VRD dataset. Note that variable k is the number of relation candidates when computing

Table 3. Ablation experiment of our model on the VRD dataset.

Methods	ReD		PhD		ReD				PhD			
	$free\ k$				$k = 1$		$k = 70$		$k = 1$		$k = 70$	
Recall@	50	100	50	100	50	100	50	100	50	100	50	100
BPGR-**SRM**	27.0	32.6	32.9	39.6	23.7	26.7	27.0	32.6	29.0	32.9	32.9	39.6
BPGR-**VRM**	28.3	34.4	35.0	41.9	25.3	28.2	28.3	34.4	31.0	34.8	35.0	41.9
BPGR-**OI**	28.8	35.0	35.7	42.9	25.4	28.7	28.8	35.0	31.4	35.5	35.7	42.9
BPGR	**29.4**	**35.3**	**36.2**	**43.0**	**26.2**	**29.4**	**29.4**	**35.3**	**32.3**	**36.4**	**36.2**	**43.0**

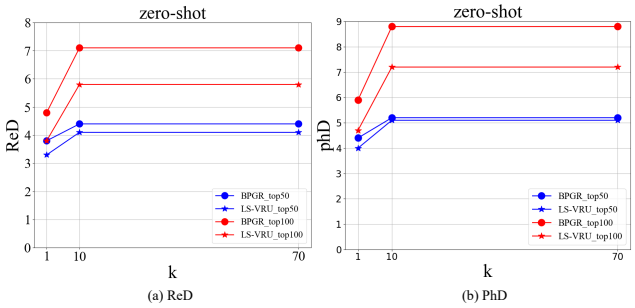

Figure 5. Zero-shot learning performance on VRD dataset. It shows results on relationship detection and phrase detection for zero-shot learning. As the hyperparameter k increases, the results show an upward trend.

Figure 6. Comparison of our BPGR with baseline LS-VRU on detection results. The first row is BPGR's detection results and the second row represents LS-VRU's detection results. "GT" is ground truth.

the top 50/100. Since not all state-of-the-art methods specified k in their experiment, we use the way presented in [46] to report results in the "$free\ k$" column when takes k as a hyper-parameter.

The state-of-the-art methods are based on language prior. Our BPGR(E+M) represents inference-and-learning and BPGR(E) is inference-only when the weight of all logic rules are fixed as 1. The result show that both BPGR(E) and BPGR(E+M) outperform state-of-the-art methods in most cases. With learning the weight of logic rules, BPGR(E+M) achieves the best performance. The reason is that BPGR can leverage the symbolic knowledge in logic rules to outperform those purely based on language prior. Compared to baseline LS-VRU, BPGR achieves much better performance and plays the function of correcting error overall.

Table 2 shows the result on VG200. It is not a clear value of k in state-of-the-art methods for VG200. Therefore, the result of our BPGR are reported for $k = 1$. We can see that our BPGR outperforms the state-of-art method in two metrics in three Recall@20/50/100. This clearly shows the benefit of leveraging symbolic knowledge in logic rules. We are noted that PCLS focus more on relationship recognition, our BPGR has a higher score on PCLS evaluation metric. This indicates that the logic rule are beneficial to relationship recognition in the model.

In practical scenarios, the relationship in the visual relationship detection task is long tail distribution. Therefore, it is significant to investigate the model generalization performance on relationships with insufficient training data. We have verified our BPGR and baseline LS-VRU in a zero-shot environment that the training and testing data are disjoint sets of relationships on the VRD dataset. Fig. 5 shows the result. As expected, the performance of BPGR outperforms LS-VRU on top 50/100. This shows the limitation of LS-VRU when coping with sparse relationships. In contrast, BPGR leverages both symbolic knowledge in logic rules and language prior for reasoning, which is much less affected by the sparse relationship.

4.5. Ablation experiments

To investigate how the model trade-off affects reasoning performance, we design three variants to verify the effect of individual components on BPGR. The three variants are as

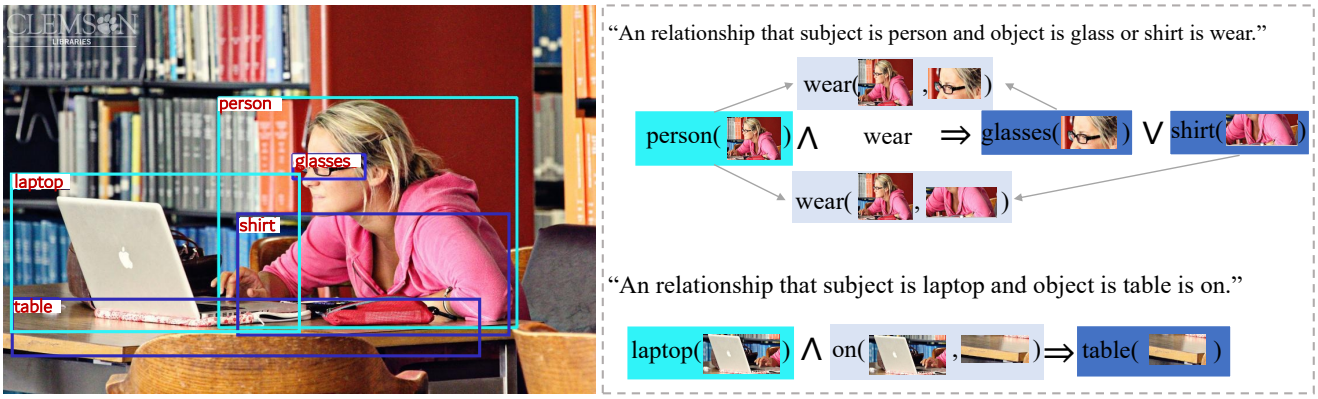

Figure 7. An example can describe the interpretability of reasoning results in an image. Our model can explain reasoning relationships between objects by finding the FOL that characterizes the commonsense knowledge. For example, why is a relation "wear" between "person" and "glasses" or "shirt"? According Eq. (13), the model can find the most confident logic rule is $\mathsf{people}(x) \wedge \mathsf{wear}(x, y) \Rightarrow \mathsf{glasses}(y) \vee \mathsf{shirt}(y)$. This indicates that the reasoning result of the model are consistent with commonsense.

follows: (1) BPGR-**SRM**($\alpha = 1$, $\beta = 0$, $\gamma = 0$): removing symbolic reasoning module. (2) BPGR-**VRM**($\alpha = 1/2$, $\beta = 1$, $\gamma = 1$): removing a half of visual reasoning module. (3) BPGR-**OI**($\alpha = 1$, $\beta = 1$, $\gamma = 0$): removing cross-entropy of observed variables.

We are testing on the VRD dataset. The result are reported in Table 3. It is observed that correlations among the components of **SRM**, **VRM** and **OI** have a tremendously positive influence on visual relationship detection, and the **SRM** benefits the final results. This is consistent with our theoretical analysis result: symbolic knowledge in logic rules can rectify the result of VRM. Besides, BPGR-**VRM**'s performance drops when reducing the proportion of visual reasoning modules, which indicates that visual feature is an important factor for the model's performance.

4.6. Visual analysis

Detection result analysis. Fig. 6 shows the reasoning result of our BPGR and baseline LS-VRU respectively. Compared to LS-VRU, BPGR's results are better. We see that BPGR can play an important role in correcting error reasoning results. For example, LS-VRU shows some results "(motorcycle, behind, motorcycle)", "(bike, next to, person)" and "(sky, above, building)" are not matching ground truth. However, BPGR's results are matching ground truth. The above example is also shown that symbolic knowledge in logic rules can guide learning toward of the model to correct error results from reasoning performance.

Interpretability analysis. We show a visual image for interpretability in Fig. 7. For example, BPGR's reasoning result is (laptop, on, table) for an image. By Eq. (13), the model can provide some logic rules of the high score. we see that top 1 logic rule is $\mathsf{laptop}(x) \wedge \mathsf{on}(x, y) \Rightarrow \mathsf{table}(y)$. According to this logic rule, we know that when the subject is "laptop" and the object is "table", the relation is predicted as "on" in line with logic rule $\mathsf{laptop}(x) \wedge \mathsf{on}(x, y) \Rightarrow \mathsf{table}(y)$. Therefore, the logic rule can explain the reasoning results of the model to a certain extent.

5. Conclusion

To summarize, this paper contributes a novel framework for combining symbolic knowledge with the VRD model. Different from prior works, BPGR can utilize the probabilistic graphical model to encode logic rules into the VRD model to improve performance and provide interpretability. Further, to capture global information and uncertainty of symbolic knowledge in models, we model logic rules by MLN. Our empirical results show the effectiveness of the model over baselines. In the future, we will extend our idea of the neural-symbolic to other domains, such as recommended systems, etc. Further, we will design more general logic rules or introduce other symbolic knowledge, and design different combine ways.

Acknowledgements. This work was supported by the National Key R&D Program of China under Grant Nos. 2021ZD0112501 and 2021ZD0112502; the National Natural Science Foundation of China under Grant Nos. 62172185 and 61876069; Jilin Province Key Scientific and Technological Research and Development Project under Grant Nos. 20180201067GX and 20180201044GX; and Jilin Province Natural Science Foundation under Grant No. 20200201036JC.

References

[1] Sathyanarayanan Aakur, Fillipe DM de Souza, and Sudeep Sarkar. Going deeper with semantics: Video activity interpretation using semantic contextualization. In *WACV*, pages 190–199, 2019. 2

[2] Sherif Abdelkarim, Aniket Agarwal, Panos Achlioptas, Jun Chen, Jiaji Huang, Boyang Li, Kenneth Church, and Mohamed Elhoseiny. Exploring long tail visual relationship recognition with large vocabulary. In *ECCV*, pages 15921–15930, 2021. 3

[3] Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. Neural module networks. In *CVPR*, pages 39–48, 2016. 1

[4] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *ICCV*, pages 2425–2433, 2015. 1

[5] Yuichiro Anzai. *Pattern recognition and machine learning*. 2012. 3

[6] Stephen H Bach, Matthias Broecheler, Bert Huang, and Lise Getoor. Hinge-loss markov random fields and probabilistic soft logic. *arXiv preprint arXiv:1505.04406*, 2015. 2

[7] Bo Dai, Yuqi Zhang, and Dahua Lin. Detecting visual relationships with deep relational networks. In *CVPR*, pages 3076–3086, 2017. 2

[8] Adnan Darwiche. On the tractable counting of theory models and its application to truth maintenance and belief revision. *JANCL*, 11(1-2):11–34, 2001. 1

[9] Adnan Darwiche and Pierre Marquis. A knowledge compilation map. *JAIR*, 17:229–264, 2002. 1

[10] Herbert B Enderton. *A mathematical introduction to logic*. 2001. 3

[11] Zoubin Ghahramani, Matthew J Beal, et al. *Graphical models and variational methods*. 2000. 4

[12] Roei Herzig, Moshiko Raboh, Gal Chechik, Jonathan Berant, and Amir Globerson. Mapping images to scene graphs with permutation-invariant structured prediction. *NeurIPS*, 31:7211–7221, 2018. 6

[13] Zih-Siou Hung, Arun Mallya, and Svetlana Lazebnik. Contextual translation embedding for visual relationship detection and scene graph generation. *TPAMI*, 43(11):3820–3832, 2020. 1, 6

[14] Michael Kampffmeyer, Yinbo Chen, Xiaodan Liang, Hao Wang, Yujia Zhang, and Eric P Xing. Rethinking knowledge graph propagation for zero-shot learning. In *CVPR*, pages 11487–11496, 2019. 2

[15] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, pages 3128–3137, 2015. 1

[16] Tushar Khot, Sriraam Natarajan, Kristian Kersting, and Jude Shavlik. Learning markov logic networks via functional gradient boosting. In *ICDM*, pages 320–329, 2011. 2

[17] Girish Kulkarni, Visruth Premraj, Vicente Ordonez, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C Berg, and Tamara L Berg. Babytalk: Understanding and generating simple image descriptions. *TPAMI*, 35(12):2891–2903, 2013. 1

[18] Kongming Liang, Yuhong Guo, Hong Chang, and Xilin Chen. Visual relationship detection with deep structural ranking. In *AAAI*, 2018. 3

[19] Xiaodan Liang, Lisa Lee, and Eric P Xing. Deep variation-structured reinforcement learning for visual relationship and attribute detection. In *CVPR*, pages 848–857, 2017. 2

[20] Bingqian Lin, Yi Zhu, and Xiaodan Liang. Atom correlation based graph propagation for scene graph generation. *PR*, 122:108300, 2022. 3

[21] Xin Lin, Changxing Ding, Jinquan Zeng, and Dacheng Tao. Gps-net: Graph property sensing network for scene graph generation. In *CVPR*, pages 3746–3753, 2020. 1, 5, 6

[22] Xueyan Liu, Bo Yang, Hechang Chen, Katarzyna Musial, Hongxu Chen, Yang Li, and Wanli Zuo. A scalable redefined stochastic blockmodel. *TKDD*, 15(3):1–28, 2021. 4

[23] Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. Visual relationship detection with language priors. In *ECCV*, pages 852–869, 2016. 2, 5, 6

[24] Ruotian Luo, Ning Zhang, Bohyung Han, and Linjie Yang. Context-aware zero-shot recognition. In *AAAI*, volume 34, pages 11709–11716, 2020. 2

[25] Lilyana Mihalkova and Raymond J Mooney. Bottom-up learning of markov logic network structure. In *ML*, pages 625–632, 2007. 2

[26] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *NeurIPS*, pages 3111–3119, 2013. 3

[27] Alejandro Newell and Jia Deng. Pixels to graphs by associative embedding. *arXiv preprint arXiv:1706.07365*, 2017. 6

[28] Vilém Novák, Irina Perfilieva, and Jiri Mockor. *Mathematical principles of fuzzy logic*, volume 517. 2012. 5, 6

[29] Hoifung Poon and Pedro Domingos. Unsupervised semantic parsing. In *EMNLP*, pages 1–10, 2009. 2

[30] Meng Qu and Jian Tang. Probabilistic logic neural networks for reasoning. *arXiv preprint arXiv:1906.08495*, 2019. 2

[31] Matthew Richardson and Pedro Domingos. Markov logic networks. *ML*, 62(1-2):107–136, 2006. 2, 5

[32] Luciano Serafini and Artur d'Avila Garcez. Logic tensor networks: Deep learning and logical reasoning from data and knowledge. *arXiv preprint arXiv:1606.04422*, 2016. 4

[33] Karan Sikka, Andrew Silberfarb, John Byrnes, Indranil Sur, Ed Chow, Ajay Divakaran, and Richard Rohwer. Deep adaptive semantic logic (dasl): Compiling declarative knowledge into deep neural networks. *arXiv preprint arXiv:2003.07344*, 2020. 2, 3

[34] Parag Singla and Pedro Domingos. Discriminative training of markov logic networks. In *AAAI*, volume 5, pages 868–873, 2005. 2

[35] Parag Singla and Pedro Domingos. Memory-efficient inference in relational domains. In *AAAI*, volume 6, pages 488–493, 2006. 2

[36] Son D Tran and Larry S Davis. Event modeling and recognition using markov logic networks. In *ECCV*, pages 610–623, 2008. 2

[37] Yaqi Xie, Ziwei Xu, Mohan S Kankanhalli, Kuldeep S Meel, and Harold Soh. Embedding symbolic knowledge into deep networks. *arXiv preprint arXiv:1909.01161*, 2019. 1, 3

[38] Danfei Xu, Yuke Zhu, Christopher B Choy, and Li Fei-Fei. Scene graph generation by iterative message passing. In *CVPR*, pages 5410–5419, 2017. 1, 2, 5, 6

[39] Jianwei Yang, Jiasen Lu, Stefan Lee, Dhruv Batra, and Devi Parikh. Graph r-cnn for scene graph generation. In *ECCV*, pages 670–685, 2018. 6

[40] Guojun Yin, Lu Sheng, Bin Liu, Nenghai Yu, Xiaogang Wang, Jing Shao, and Chen Change Loy. Zoom-net: Mining deep feature interactions for visual relationship recognition. In *ECCV*, pages 322–338, 2018. 6

[41] Dongran Yu, Bo Yang, Dayou Liu, and Hui Wang. A survey on neural-symbolic systems. *arXiv preprint arXiv:2111.08164*, 2021. 1

[42] Ruichi Yu, Ang Li, Vlad I Morariu, and Larry S Davis. Visual relationship detection with internal and external linguistic knowledge distillation. In *ICCV*, pages 1974–1982, 2017. 3, 6

[43] Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. Neural motifs: Scene graph parsing with global context. In *CVPR*, pages 5831–5840, 2018. 1, 6

[44] Yibing Zhan, Jun Yu, Ting Yu, and Dacheng Tao. On exploring undetermined relationships for visual relationship detection. In *CVPR*, pages 5128–5137, 2019. 6

[45] Hanwang Zhang, Zawlin Kyaw, Shih-Fu Chang, and Tat-Seng Chua. Visual translation embedding network for visual relation detection. In *CVPR*, pages 5532–5540, 2017. 3

[46] Ji Zhang, Yannis Kalantidis, Marcus Rohrbach, Manohar Paluri, Ahmed Elgammal, and Mohamed Elhoseiny. Large-scale visual relationship understanding. In *AAAI*, volume 33, pages 9185–9194, 2019. 1, 3, 4, 5, 6, 7

[47] Weizhe Zhang, Xiaoqiang Li, Hui He, and Xing Wang. Identifying network public opinion leaders based on markov logic networks. *Sci. World J.*, 2014, 2014. 2

[48] Yuyu Zhang, Xinshi Chen, Yuan Yang, Arun Ramamurthy, Bo Li, Yuan Qi, and Le Song. Efficient probabilistic logic reasoning with graph neural networks. *arXiv preprint arXiv:2001.11850*, 2020. 2, 4

[49] Yaohui Zhu and Shuqiang Jiang. Deep structured learning for visual relationship detection. In *AAAI*, volume 32, 2018. 2