One-Shot Learning for Long-Tail Visual Relation Detection

Weitao Wang,[1] Meng Wang,[1,2]* Sen Wang,[3] Guodong Long,[4] Lina Yao,[5] Guilin Qi,[1,2] Yang Chen[1,2]
[1]School of Computer Science and Engineering, Southeast University, Nanjing, China
[2]MOE Key Laboratory of Computer Network and Information Integration, China
[3]The University of Queensland, Brisbane, Australia
[4]The University of Technology Sydney, Sydney, Australia
[5]The University of New South Wales, Sydney, Australia
{wangweitao, meng.wang, gqi, chenyang.list}@seu.edu.cn,
sen.wang@uq.edu.au, guodong.long@uts.edu.au, lina.yao@unsw.edu.au

Abstract

The aim of visual relation detection is to provide a comprehensive understanding of an image by describing all the objects within the scene, and how they relate to each other, in <object-predicate-object> form; for example, <person-lean on-wall>. This ability is vital for image captioning, visual question answering, and many other applications. However, visual relationships have long-tailed distributions and, thus, the limited availability of training samples is hampering the practicability of conventional detection approaches. With this in mind, we designed a novel model for visual relation detection that works in one-shot settings. The embeddings of objects and predicates are extracted through a network that includes a feature-level attention mechanism. Attention alleviates some of the problems with feature sparsity, and the resulting representations capture more discriminative latent features. The core of our model is a dual graph neural network that passes and aggregates the context information of predicates and objects in an episodic training scheme to improve recognition of the one-shot predicates and then generate the triplets. To the best of our knowledge, we are the first to center on the viability of one-shot learning for visual relation detection. Extensive experiments on two newly-constructed datasets show that our model significantly improved the performance of two tasks PredCls and SGCls from 2.8% to 12.2% compared with state-of-the-art baselines.

Introduction

A common way to define the visual relationships in an image with a triplet, where two objects are connected by a predicate – for example, <person-adjacent to-bike> or <clock-attach to-building>. Identifying these relationships is useful for a wide range of image understanding tasks, such as captioning (Fang et al. 2015), retrieval (Johnson et al. 2015), reasoning (Shi, Zhang, and Li 2019; Wang et al. 2018), and visual question answering (Xiong, Merity, and Socher 2016). Conventional models for automatically detecting these relationships typically require a relatively large number of training instances to determine the predicates. However, as shown in Figure 1, visual relationships tend to have long-tailed distributions, which means that most predicates, such as "attach to", will have very few training samples, if any, compared to the most popular predicates, such as "on".

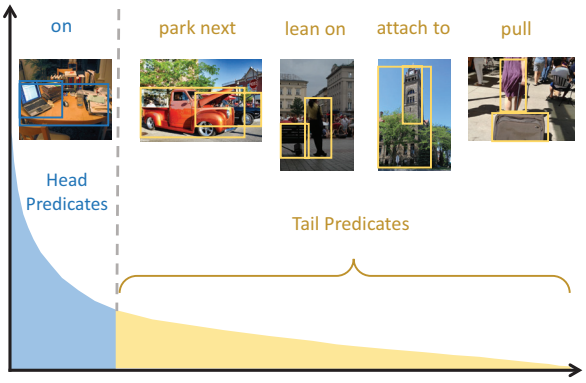

Figure 1: Frequencies of predicates in VRD dataset. There are a large portion of predicates that have few triplets.

For this reason, most current models (Lu et al. 2016; Zhang et al. 2017; 2019a; 2019b) only perform well with the most popular predicates. Dornadula et al. (Dornadula et al. 2019) recently introduced a new model that treats a predicate as a neural network transformation between two object representations. The proposed model can, rather effectively, detect visual relationships in few-shot learning scenarios. However, detecting these relationships with one-shot learning is yet to be mastered. Intuitively, the fewer the number of available training instances, the more a detection model is needed. Therefore, we have focused our research efforts on detecting visual relationships from one sample only, i.e., one-shot learning.

Our solution centers on a novel model for detecting the visual relationships in an image. The visual features of objects and predicates are extracted first, and embeddings are generated through the feature extraction network that includes a feature-level attention mechanism. Attention alleviates some of the problems with feature sparsity, and the resulting representations capture more discriminative latent features. A

*Corresponding Author.

dual graph neural network passes and aggregates the context information of predicates and objects in an episodic training scheme to improve recognition of the one-shot predicates and then generate the triplets.

The contributions of this research are summarized below.

- To the best of our knowledge, this is the first-ever one-shot learning approach to visual relationship detection.
- We introduce a novel model based on a dual-graph neural network that exploits intra-cluster similarities and inter-cluster dissimilarities between predicates and objects for one-shot visual relationship detection.
- We constructed two new datasets for evaluating one-shot visual relationship detection tasks.
- Extensive experiments show that our model significantly improves one-shot performance.

Related Work

This section discusses existing related research in the following aspects: visual relation detection, few-shot learning and graph neural network.

Visual Relation Detection. In recent years, there have been some works for visual relation detection (Lu et al. 2016; Zhang et al. 2017; 2019a; 2019b). Language prior is thought to provide useful information to detect visual relationships. In the earlier work, Lu *et al.* (Lu et al. 2016) fine-tuned the likelihood of the predicted relationship through the language. Inspired by TransE which has great success in representation learning on knowledge graph, Zhang *et al.* (Zhang et al. 2017) proposed the VTransE model which computes embedding of predicates by mapping the visual features into the predicate space. Some models also proposed to utilize the statistical information to improve detection performance. For instance, Dai *et al.* (Dai, Zhang, and Lin 2017) proposed a model to exploit the statistical dependencies among predicates, subjects, and objects. Context messages passing also plays a crucial role in recent researches, e.g., Xu *et al.* (Xu et al. 2017) predicted each visual relationship with joint inference by iterative message passing. Compared to previous models which focus on message passing in the same image, our work achieves messages passing from different images.

Few-shot Learning. Metric based approaches (Koch, Zemel, and Salakhutdinov 2015; Vinyals et al. 2016; Snell, Swersky, and Zemel 2017) and meta-learner based approaches (Finn, Abbeel, and Levine 2017; Ravi and Larochelle 2017; Sung et al. 2018) are the most important two ways to achieve few-shot learning. For metric-based approaches, the Siamese Network, which works in two shared weighted networks, is proposed by Gregory *et al.* (Koch, Zemel, and Salakhutdinov 2015) to compare two different images. In addition, Matching Networks (Vinyals et al. 2016) can also make predictions by comparing the input example with the small labeled support set and organize the train and test data in 'episodes'. Prototypical Network (Snell, Swersky, and Zemel 2017) usually recognizes the image by computing the prototypical of each class in metric space. Differently, meta-learner based approaches aim to learn the optimization of model parameters to achieve the few-shot learning. For example, MAML (Finn, Abbeel, and Levine 2017) aimed to meta-learn an initial condition in a model-agnostic way. Ravi *et al.* (Ravi and Larochelle 2017) used the LSTM-based meta-learner for replacing the stochastic gradient descent optimizer. Sung *et al.* (Sung et al. 2018) put forward a Relation Network that preforms few-shot recognition by learning to compare query images against few-shot labeled sample images. Most few-shot learning researches are on the task of image recognition. Unlike image recognition tasks, visual relationship detection is more complicated and cannot be recognized by only using visual features. Determining visual predicates usually needs both visual features and non-visual features such as context, geometrical layout, and semantics.

Graph Neural Network. Graph neural network is proposed to process graph structure data (Zhou et al. 2018). Recently, graph neural networks are applied to more scenarios such as social network mining (Hamilton, Ying, and Leskovec 2017), recommender systems (Xie et al. 2016), graph representation learning (Ying et al. 2018), as well as non-structural data scenarios like image classification (Kipf and Welling 2016), text classification (Peng et al. 2018), and program verification (Li et al. 2016). Besides, some approaches (Garcia and Bruna 2018; Liu et al. 2019) have also explored graph neural networks for few-shot learning. Garcia *et al.* (Garcia and Bruna 2018) cast few-shot learning as a supervised message passing task which is trained end-to-end by using graph neural networks. Liu *et al.* (Liu et al. 2019) proposed a transductive propagation network to propagate labels form labeled instances to unlabeled test instances. In this paper, our model will also use the dual graph structure to pass the message from different predicates or objects and improve the one-shot predicate recognition.

Method

Next, we will describe the one-shot setting and our visual relation detection model in detail. As illustrated in Figure 3, the proposed model includes the visual feature extraction module and the dual graph module.

One-shot Learning Settings

Following the standard one-shot learning settings (Vinyals et al. 2016; Ravi and Larochelle 2017), we split the visual relation dataset D into two sets T_{train} and T_{test} based on different predicates, i.e., the same predicate does not appear in the training set and the test set at the same time. T_{train} has a relatively large labeled dataset whereas T_{test} has only a few labeled examples. We employ the episodic learning paradigm for our one-shot classification task. During the episodic learning, a small subset of visual relations contains N predicates will be sampled from T_{train} to construct a *support set* S and a *query set* Q. The *support set* $S = \{(\mathrm{x}_1, r_1), (\mathrm{x}_2, r_2), \ldots, (\mathrm{x}_N, r_N)\}$ contains N predicates and each predicate has only one sample (i.e., *N*-way one-shot setting). For a sample $(\mathrm{x}_i, r_i) \in S$, $\mathrm{x}_i = \{x, x_h^{bbox}, x_t^{bbox}\}$ indicates the visual information about the

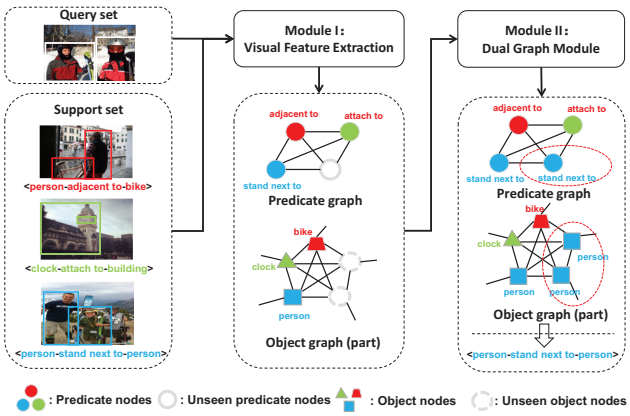

Figure 2: Illustration of 3-way 1-shot problem with one query example addressed by our model.

visual relation r_i, where x is the image where r_i comes from, x_h^{bbox} and x_t^{bbox} are bounding boxes of visual objects in x. Note that different relations may exist in the same image x due to the different of bounding boxes. The query set $Q = \{(x_1^{'}, r_1^{'}), (x_2^{'}, r_2^{'}), \ldots, (x_M^{'}, r_M^{'})\}$ contains M visual triples with the same predicates in S. The support set S is used to train the model to achieve minimize the loss of the predictions for the query set Q. The T_{test} uses the same approach to achieve visual relation detection, as shown in Figure 2. The idea of episodic training is making the training process imitate the testing phase. Because the distribution of training tasks is assumed to be similar to the test task, the performance of the test tasks can be improved by learning a model to work well on the training tasks.

Visual Feature Extraction Module

This module aims to extract visual features and obtain the embeddings of visual objects and predicates.

Inspired by VTransE (Zhang et al. 2017), we represent a visual relation $< h - p - t >$ as low-dimensional vectors $\mathbf{v}_h$, $\mathbf{v}_p$, and $\mathbf{v}_t$, where h and t indicate the head object and tail object, p is the predicate. Suppose $\mathbf{f}_h, \mathbf{f}_t \in \mathbb{R}^D$ are the D-dimensional features of h and t in the bounding box x_h^{bbox}, x_y^{bbox} respectively, we aim to learn projection matrices $\mathbf{W}_s$ and $\mathbf{W}_o$ from the feature space to the relation space, and obtain the $\mathbf{v}_h$ and $\mathbf{v}_t$, i.e., $\mathbf{W}_h\mathbf{f}_h$ and $\mathbf{W}_t\mathbf{f}_t$. Then, a visual relation can be represented as follows:

$$\mathbf{W}_h\mathbf{f}_h + \mathbf{v}_p \approx \mathbf{W}_t\mathbf{f}_t. \tag{1}$$

Labels of the head object, tail object, as well as spatial information are always useful to identify the predicate (Xu et al. 2017). For instance, some predicates can often be inferred by using spatial information, such as "on" and "under". Therefore, we utilize the information mixed with visual features as follows:

$$\mathbf{f}_h^{'} = f^{spa}(\mathbf{b}_h, \mathbf{b}_t) \oplus \mathbf{f}_h \oplus \mathbf{c}_h, \tag{2}$$

$$\mathbf{f}_t^{'} = f^{spa}(\mathbf{b}_h, \mathbf{b}_t) \oplus \mathbf{f}_t \oplus \mathbf{c}_t, \tag{3}$$

where f^{spa} is a 3-layers CNN to encode binary mask features, $\oplus$ denotes vectors connection. The $\mathbf{c}_h$ and $\mathbf{c}_t$ are the N-dimensional vectors of objects classification probabilities (i.e., N classes of objects) from the object detection network. We obtain the binary masks $\mathbf{b}_h$ and $\mathbf{b}_t$ by inputting bounding boxes of visual objects x_h^{bbox} and x_t^{bbox} .

Moreover, it has been verified that some dimensions of the learned latent representations are more discriminative for classifying in the low-dimensional space (Gao et al. 2019). We further propose an attention mechanism to enhance the representations of visual predicates. Then, the problem of feature sparsity can be alleviated by using the feature-level attention mechanism. Formally, Equation (1) becomes:

$$\mathbf{v}_p = f^{att}(\mathbf{f}_p) \cdot (\mathbf{W}_t\mathbf{f}_t^{'} - \mathbf{W}_h\mathbf{f}_h^{'}), \tag{4}$$

where f^{att} is the feature attention network, $\mathbf{f}_p$ is predicate visual features under the combination of the head object bounding box and the tail object bounding box.

Dual Graph Module

This module aims to achieve massages passing from predicates or objects in different images. As illustrated in Figure 3, the dual fully-connected graph is the core of our model. A predicate graph consists of nodes and edges. Each node is the embedding of the predicate. Each edge is a 2-dimensional vector which contains intra-cluster similarity and inter-cluster dissimilarity of predicates. The object graph is analogous to the predicate graph, except that nodes are replaced by object embeddings. Through this graph structure, intra-cluster similarity and inter-cluster dissimilarity can be simultaneously maximizing in node/edge updates. Particularly, a small neural network is used to aggregate and pass messages from object nodes to predicate nodes. The information transmission among different types of nodes can increase the amount of information on predicates and improve the recognition performance of predicates.

The predicate graph and object graph have different types of nodes but the same graph structure. Therefore, a graph can be denoted as $G = \{V, E\}$, where the V and E indicate the sets of nodes and edges of the graph. For each layer of dual graph network, the predicate graph G_p and object graph G_o update nodes and edges in the same way. We use $\mathbf{v}_i^l$ to represent the node in the predicate graph and object graph together in layer l as follows:

$$\mathbf{v}_i^l = \begin{cases} \mathbf{v}_p\,, & \text{in } G_p, \\ \mathbf{v}_h \text{ or } \mathbf{v}_t\,, & \text{in } G_o. \end{cases} \tag{5}$$

For edges E, each edge $\mathbf{e}_{ij} = \{e_{ijd}\}_{d=1}^2 \in [0,1]^2$ is used to connect node $\mathbf{v}_i$ and node $\mathbf{v}_j$. The more similar two nodes are, the greater the value of e_{ij1}. We initialize the edge $\mathbf{e}_{ij}$ by using the ground-truth node labels as:

$$\mathbf{e}_{ij} = \begin{cases} [1,0]\,, & \text{if } y_i = y_j \text{ and } i,j \le N, \\ [0,1]\,, & \text{if } y_i \ne y_j \text{ and } i,j \le N, \\ [0.5,0.5]\,, & \text{otherwise}, \end{cases} \tag{6}$$

where y_i is the label of node $\mathbf{v}_i^l$.

Our dual graph network contains L layers to process the graph. The nodes and edges are updated through each layer

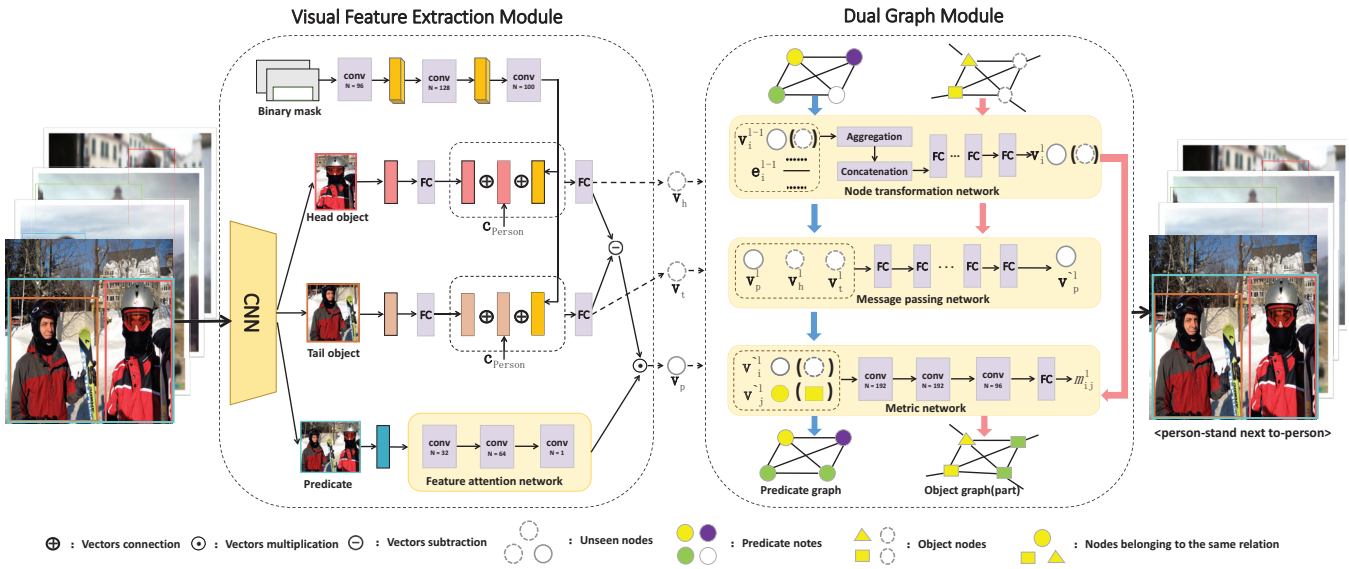

Figure 3: Our model architecture for a 3-way 1-shot problem with one query example. Different shapes represent different objects and different colors represent different predicates.

in the forward propagation. On each layer, nodes are updated by the neighborhood aggregation procedure both in predicate graph and object graph. In the aggregation procedure, the edges features are multiplied as coefficients in each dimension of the neighborhood nodes. The more similar neighborhood nodes are, the more information sends to the current node. For different label nodes, a higher dissimilarity coefficient can bring more dissimilarity information. The updating process of nodes features is defined as follows:

$$\mathbf{v}_i^l = f_v^l([\sum_j \widetilde{e}_{ij1}^{l-1}\mathbf{v}_j^{l-1} || \sum_j \widetilde{e}_{ij2}^{l-1}\mathbf{v}_j^{l-1}]), \tag{7}$$

where f_v^l is used to update node features, and $\widetilde{e}_{ijd} = \frac{e_{ijd}}{\sum_k e_{ikd}}$.

The labels of head object and tail object are always considered useful to identify the predicate in visual relation. However, it is impossible to always obtain a correct label through the object detection network. The object graph can improve the performance of object recognition by using explicit intra-cluster similarity and inter-cluster dissimilarity. Message passing network between different graphs can update the predicate representations. In this way, not only the information that may be wrong at first is compensated, but also more object information is integrated into the predicate representations. Only $\mathbf{v}_h^l$, $\mathbf{v}_t^l$ and $\mathbf{v}_p^l$ in the same visual relationship triplet can transmit information. The information transfer mechanism between different graphs is defined as follows:

$$\mathbf{v}^{'l} = \begin{cases} f_t^l(\mathbf{v}_p^l, \mathbf{v}_h^l, \mathbf{v}_t^l)\ , & \text{in } G_p, \\ \mathbf{v}_h^l \text{ or } \mathbf{v}_t^l, & \text{in } G_o, \end{cases} \tag{8}$$

where $\mathbf{v}^{'l}$ is a new node after the update in layer l. f_t^l is the message passing network between different graphs.

For the one-shot task, distance metric learning methods with Euclidean distance or cosine similarity are common ways to achieve it (Chen et al. 2019). However, they focus on learning shallow linear metrics for fixed feature representations which may be not discriminative. Inspired by (Sung et al. 2018) using Relation Network to learn a deep distance metric to compare a small number of images within episodes, we use a metric network as the distance metric. Metric networks f_m combines a 4-layer CNN to encode the difference between two nodes and a fully connected layer to out the similarity. The similarity is computed as follows:

$$m_{ij}^l = f_m^l(|\mathbf{v}_i^{'l} - \mathbf{v}_j^{'l}|), \tag{9}$$

where m_{ij}^l is the similarity between $\mathbf{v}_i^{'l}$ and $\mathbf{v}_j^{'l}$ in the layer l.

Edge features containing intra-cluster similarity and inter-cluster dissimilarity is updated by re-obtained nodes. New edge features can be computed by using the current layer edge information and the node information connected by this edge. The updating process of new edge features is:

$$\overline{e}_{ij1}^l = \frac{m_{ij}^l e_{ij1}^{l-1}}{\sum_k m_{ik}^l e_{ik1}^{l-1} / \sum_k e_{ik1}^{l-1}}, \tag{10}$$

$$\overline{e}_{ij2}^l = \frac{(1-m_{ij}^l) e_{ij2}^{l-1}}{\sum_k (1-m_{ik}^l) e_{ik2}^{l-1} / \sum_k e_{ik2}^{l-1}}, \tag{11}$$

$$\overline{\mathbf{e}}_{ij}^l = \mathbf{e}_{ij}^l / \|\overline{\mathbf{e}}_{ik}^l\|_1, \tag{12}$$

where $\|\overline{\mathbf{e}}_{ik}^l\|_1$ is the normalization of $\overline{\mathbf{e}}_{ik}^l$.

The prediction of predicate nodes can be obtained by edge features. For each edge, e_{ij1} is denoted the probability that two nodes belong to the same class. We predict query set nodes by nodes belong to support sets. For the confidence of

the predicate node belonging to each class can be computed as follows:

$$c_i^{(n)} = \text{softmax}(\sum_{\{j:j\neq i \bigwedge(\mathrm{x}_j,r_j)\in S\}} (\hat{p}_{ij}\delta(p_j = \mathcal{C}_n))), \quad (13)$$

where $\hat{p}_{ij} = e_{ij1}$ is the edge feature in the last layer. $c_i^{(n)}$ is the probability that $\mathbf{v}_i$ belong to the class $\mathcal{C}_n$. $\delta(p_j = \mathcal{C}_n)$ is equal to one when $p_j = \mathcal{C}_n$ and zero otherwise. p_j is the ground-true of the predicate in visual relation triplet r_j in the support set.

For the prediction of object nodes, we use a fully connected layer to compute the confidence of each class. Because the support set and the query set constructed by different predicates, it can not guarantee that objects in the query set also appear in the support set. The probability of the object node belongs to each class can be computed as follows:

$$\hat{\mathbf{o}} = fc(\mathbf{v}^{'l}), \quad \text{in } G_o. \quad (14)$$

Loss

The set of ground-truth edge labels Y depends on the labels of the nodes connected by the edge. Each ground-truth edge label is defined by the ground-truth node label:

$$y_{ij} = \begin{cases} 1, \text{ if } y_i = y_j, \\ 0, \text{ otherwise.} \end{cases} \quad (15)$$

For the predicate graph, $\hat{Y}_p^l$ is the set of similarity in the first dimension of edge features in layer l. Similarly, $\hat{Y}_o^l$ is the set of similarity for the object graph. The loss functions of the visual relation prediction are defined as follows:

$$\mathcal{L}_o = -\lambda_o \sum_{i=1}^{2\times N+2\times M} \mathbf{o}_i \log(\hat{\mathbf{o}}_i) + \sum_{l=1}^{L} \mathcal{L}_e(Y_o, \hat{Y}_o^l), \quad (16)$$

$$\mathcal{L}_p = \sum_{l=1}^{L} \lambda_p \mathcal{L}_e(Y_p, \hat{Y}_p^l), \quad (17)$$

where $\mathcal{L}_e$ is the binary cross-entropy loss. λ_p and λ_o are the hyperparameters to control the loss. $\mathbf{o}_i$ is the ground true label of the object node.

The total loss function is defined as follows:

$$\mathcal{L}_{total} = \mathcal{L}_p + \mathcal{L}_o, \quad (18)$$

in which the total loss combines all losses that are computed in all layers to improve the gradient flow in the lower layers.

Experiments

We constructed two new datasets for the task of one-shot visual relationship detection and compared our model with state-of-the-art approaches. Our codes and datasets are available at https://github.com/Witt-Wang/oneshot.

Experiments Settings

Datasets. Compared with previous works, we focus on the recognition of tail data in the long-tail distribution datasets. Existing datasets consider the same set of visual predicates during training and testing and often include sufficient training visual relations for every predicate. To construct datasets for one-shot learning, we go back to the original datasets and select those visual predicates that do not have too many visual relations as one-shot task predicates.

Table 1: Statistics of VRD-One and VG-One.

Dataset	Objects	Predicates	Visual Relations
VRD-One	99	29	2554
VG-One	127	164	27170

The detailed statistics of two new datasets are summarized in Table 1. In our datasets, each sample contains a visual relation label, ground truth bounding boxes of objects and the image. For dataset VRD-One, we selected visual relations to construct our datasets. The visual relations formed by the visual predicates are less than 300 but more than 20 in the original dataset VRD. We follow a similar process to build another larger dataset based on VG, but visual predicates are less than 500 more than 50. For the original VG dataset, a substantial fraction of the object annotations have poor quality and overlapping bounding boxes and/or ambiguous object names (Xu et al. 2017). We further process the extracted dataset from VG to construct our dataset. For VRD-One, we use 19/10 to divide visual predicates for training/testing. For VG-One, the division ratio is 115:49.

Tasks. Following (Xu et al. 2017), we evaluate the proposed model with two tasks: predicate classification and scene graph classification. Predicate classification (PredCls) is to predict the predicates of all pairwise relationships of a set of localized objects. Scene graph classification (SGCls) is to predict the class labels for the set of objects with ground truth bounding boxes, and to predict the relationship label of each object pair.

Evaluation metrics. Unlike most previous work, we do not use the image-wise recall evaluation metrics $R@50$ and $R@100$. The $R@k$ metric measures the fraction of ground-truth relationship triplets that appear among the top k most confident triplets in an image. For our data, the relationships on each picture are more sparse, and the quantity variance of each relationship is smaller than that in the original datasets. We only focus on visual relationships in specific areas of the image. Accuracy as an evaluation metrics is more suitable for our task than recall. Following the standard one-shot learning setting, we will compare the accuracy in 5-way 1-shot, and 10-way 1-shot.

Details. Our model was trained by Adam optimizer with an initial learning rate of 1×10^{-3} and weight decay of 10^{-6}. The batch sizes for training were set to be 10 and 5 for 5-way and 10-way experiments, respectively. All our code was implemented in PyTorch (Paszke et al. 2017) and ran with 4 GPUs.

Quantitative Evaluation

To validate the effectiveness of our model, we compare it with several state-of-the-art methods on VG and VRD. The details of methods are described as follows:

Table 2: Comparison with state-of-the-art baselines on the VRD-One and VG-One datasets.

	VRD-One				VG-One			
	PredCls		SGCls		PredCls		SGCls	
	5-way 1-shot	10-way 1-shot	5-way 1-shot	10-way 1-shot	5-way 1-shot	10-way 1-shot	5-way 1-shot	10-way 1-shot
VRD	37.4%	24.7%	14.7%	12.5%	41.4%	27.2%	10.8 %	9.6 %
VTransE	37.3%	24.3%	15.8%	13.4%	39.7%	23.4%	10.1%	9.4 %
LSVRU	40.3%	27.1%	16.9%	14.0%	43.4%	27.0%	10.7%	10.1%
RelDN	40.1%	26.4%	17.2%	14.3%	43.7%	28.3%	11.3%	10.1%
RelDN w/o sem	40.6%	27.3%	17.4%	14.9%	44.1%	28.2%	11.6%	10.4%
Ours	**48.4%**	**33.5%**	**22.3%**	**20.9%**	**56.3%**	**37.5%**	**14.9%**	**13.2%**

Table 3: Predicate classification accuracy in 5-way-1-shot. We compare our final model with VTransE on the VRD-One dataset and the PredCls task.

Predicates	VTransE	Ours	Predicates	VTransE	Ours
talk	34.67%	63.51%	lean on	26.43%	39.02%
pull	29.26%	40.25%	at	42.46%	51.39%
drive	21.89%	38.23%	watch	31.94%	42.85%
attach to	35.61%	56.96%	sit on	36.36%	44.44%
fly	34.21%	44.87%	use	49.29%	51.47%

Table 4: Ablation studies on our model.

	5-way 1-shot	
	PredCls	SGCls
ours w/o Object Graph	47.3%	20.4%
ours w/o Message Passing Network	47.2%	21.8%
ours w/o Attention Network	45.7%	21.6%
ours All	48.4%	22.3%

VRD (Lu et al. 2016): This method uses language prior knowledge to help detect visual relationships.

VTransE (Zhang et al. 2017): This method is a novel visual relation learning model that incorporates translation embedding and knowledge transfer.

LSVRU (Zhang et al. 2019a): This method achieves visual relation detection by projecting visual and linguistic features into a common space.

RelDN (Zhang et al. 2019b): This method uses graphical contrastive losses which explicitly force the model to disambiguate related and unrelated instances. This method is the state-of-the-art method on VG and VRD.

For fairness, we adjusted these models to avoid the impact of sample imbalance on our datasets. The training of these models is divided into two stages: the training stage and the fine-tuning stage. In the training stage, we use T_{train} to train these models. In the fine-tuning stage, for each episode, we use support set in T_{test} to fine-tune these models. All models use ResNet101 (He et al. 2016) as a feature extractor.

For results of other models on our datasets in Table 2, the classical methods and state-of-the-art methods do not show a significant difference in accuracy rate. Good accuracy can be obtained by simple visual information and semantic information. Besides, we find that statistical information may not help the model, and removing the statistical information can improve the accuracy. This may be related to the fact that the distribution of the support set and query set is not consistent in the one-shot learning setting. Our model achieves state-of-the-art results on the two new datasets and demonstrates its efficacy in one-shot visual relation detection. For each task, our model outperforms the strong baseline by a large margin.

Compared with the VRD-One dataset, the VG-One dataset has more visual relations. More data can improve the generalization ability and performance in our model. For other models, more data can not improve the performance because the number of support set is constant. For the SGCls task, the results of our model on the VRD-One dataset are better than that on the VG-One dataset, because a substantial fraction of the object annotations have poor quality and overlapping bounding boxes and/or ambiguous object names on original VG dataset. Our VG-One dataset is selected from VG, the accuracy of object recognition declines on the VG-One dataset.

We compare the accuracy of every predicate on our VRD-One dataset with VTransE, as shown in Table 3. The accuracy of all predicate in our model is higher than that in VTransE. For some predicates, our model outperforms VTransE by a large margin, such as "`talk`" and "`attach to`". It indicates that our model has the great ability for one-shot learning in different predicates.

Ablation Study

To evaluate the effectiveness of our model, we consider several ablations in Table 4. We validate the effect of attention mechanism, object graph and message passing from different graphs on the model. We remove each module to verify the effectiveness of utilizing all the proposed modules. As shown in Table 4, we can see the performance improvement when we use all modules jointly. This indicates that each module plays a critical role in the recognition of visual relationships.

In this experiment, the attention mechanism can help the model improve accuracy. On the one hand, features that are obtained by the feature extraction network from the support set have the problem of data sparsity. Several dimensions of the representation of predicate are more discriminative

Figure 4: Sample results from our model trained with different numbers of layers of dual graph network. Green indicates the correct result of prediction, red indicates the wrong result.

for classifying in the embedding space. The attention mechanism can enhance the representation of predicate. On the other hand, the attention mechanism also brings more visual information to predicate embedding.

For object graph and message passing from different graphs, the accuracy of the scene graph classification (SGCls) task is greatly improved. There is no object label in SGCls task, but annotation information is important for inferring the visual relationship. When the model uses wrong object recognition information, it will not only lead to the wrong recognition of visual relations, but also affect the recognition of visual predicates. Although objects in visual relations are not small sample data, intra-cluster similarity information and inter-cluster dissimilarity information can be added to the embedding of objects by object graph. The accuracy of object recognition can be improved. Message passing mechanism can also bring more accurate object information for predicate recognition. The accuracy of visual relationship is greatly improved through these modules.

Qualitative Results

Figure 4 illustrates the qualitative results. we compare our final model trained with different numbers of layers in our dual graph network and show several wrong cases. Although the prediction of visual predicates in the query set relies on the visual predicates in support set, the results show visual predicates are easy to make wrong predictions when the object in the same visual relation is incorrectly identified, such as <person-at-table>. For objects, obscured or confusing objects are easily misidentified, such as <person-at-desk> and <person-fly-kite>. The dual graph network plays a key role in one-shot visual relation detection, and it enables our model to outperform previous state-of-the-art methods. For the effect of the number of layers on the network, our final model trained with one layer is also able to recognize the visual relationships in images, such as <person-pull-luggage>, but several predicates or objects which are easy to make confusion can cause wrong predictions like <person-use-phone>, <tower-attach to-building. The final model trained with three layers can make semantically correct predictions.

Conclusion

In this paper, we address the problem of one-shot visual relation detection by a novel model. Our model first extracts visual features and obtains the embedding of predicates by visual feature extraction network. Then, a dual graph network is employed to achieve message passing and detect the visual relation. We construct two new datasets, i.e., VRD-One and VG-One, for the one-shot experiment. The experimental results show that our model achieves state-of-the-art results. Moreover, through extensive ablation experiments, we demonstrate the efficacy of our approach. In the future, a possible improvement direction would be able to explore its capability in other prediction problems in visual relation detection. We hope our model becomes a generic framework

for the one-shot visual relation detection problem.

Acknowledgment

This work was supported by the National Science Foundation of China with Grant Nos. 61906037 and U1736204; National Key Research and Development Program of China with Grant Nos. 2018YFC0830201 and 2017YFB1002801; the Judicial Big Data Research Center, School of Law at Southeast University with Grant No.4313059291.

References

Chen, W.; Liu, Y.; Kira, Z.; Wang, Y. F.; and Huang, J. 2019. A closer look at few-shot classification. In *Proceedings of ICLR*.

Dai, B.; Zhang, Y.; and Lin, D. 2017. Detecting visual relationships with deep relational networks. In *Proceedings of the CVPR*, 3076–3086.

Dornadula, A.; Narcomey, A.; Krishna, R.; Bernstein, M.; and Fei-Fei, L. 2019. Visual relationships as functions: Enabling few-shot scene graph prediction. *arXiv preprint arXiv:1906.04876*.

Fang, H.; Gupta, S.; Iandola, F.; Srivastava, R. K.; Deng, L.; Dollár, P.; Gao, J.; He, X.; Mitchell, M.; Platt, J. C.; et al. 2015. From captions to visual concepts and back. In *Proceedings of CVPR*, 1473–1482.

Finn, C.; Abbeel, P.; and Levine, S. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of ICML*, 1126–1135.

Gao, T.; Han, X.; Liu, Z.; and Sun, M. 2019. Hybrid attention-based prototypical networks for noisy few-shot relation classification. In *Proceedings of AAAI*, 6407–6414.

Garcia, V., and Bruna, J. 2018. Few-shot learning with graph neural networks. In *Proceedings of ICLR*.

Hamilton, W.; Ying, Z.; and Leskovec, J. 2017. Inductive representation learning on large graphs. In *Proceedings of NIPS*, 1024–1034.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of CVPR*, 770–778.

Johnson, J.; Krishna, R.; Stark, M.; Li, L.-J.; Shamma, D.; Bernstein, M.; and Fei-Fei, L. 2015. Image retrieval using scene graphs. In *Proceedings of CVPR*, 3668–3678.

Kipf, T. N., and Welling, M. 2016. Semi-supervised classification with graph convolutional networks. In *Proceedings of the ICLR*.

Koch, G.; Zemel, R.; and Salakhutdinov, R. 2015. Siamese neural networks for one-shot image recognition. In *Proceedings of ICML Workshop*.

Li, Y.; Tarlow, D.; Brockschmidt, M.; and Zemel, R. 2016. Gated graph sequence neural networks. In *Proceedings of ICLR*.

Liu, Y.; Lee, J.; Park, M.; Kim, S.; Yang, E.; Hwang, S. J.; and Yang, Y. 2019. Learning to propagate labels: Transductive propagation network for few-shot learning. In *Proceedings of ICLR*.

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

Paszke, A.; Gross, S.; Chintala, S.; Chanan, G.; Yang, E.; DeVito, Z.; Lin, Z.; Desmaison, A.; Antiga, L.; and Lerer, A. 2017. workshop. In *Proceedings of NIPS Workshop*.

Peng, H.; Li, J.; He, Y.; Liu, Y.; Bao, M.; Wang, L.; Song, Y.; and Yang, Q. 2018. Large-scale hierarchical text classification with recursively regularized deep graph-cnn. In *Proceedings of WWW*, 1063–1072.

Ravi, S., and Larochelle, H. 2017. Optimization as a model for few-shot learning. In *Proceedings of ICLR*.

Shi, J.; Zhang, H.; and Li, J. 2019. Explainable and explicit visual reasoning over scene graphs. In *Proceedings of the CVPR*, 8376–8384.

Snell, J.; Swersky, K.; and Zemel, R. 2017. Prototypical networks for few-shot learning. In *Proceedings of NIPS*, 4077–4087.

Sung, F.; Yang, Y.; Zhang, L.; Xiang, T.; Torr, P. H.; and Hospedales, T. M. 2018. Learning to compare: Relation network for few-shot learning. In *Proceedings of CVPR*, 1199–1208.

Vinyals, O.; Blundell, C.; Lillicrap, T.; Wierstra, D.; et al. 2016. Matching networks for one shot learning. In *Proceedings of NIPS*, 3630–3638.

Wang, M.; Chen, W.; Wang, S.; Liu, J.; Li, X.; and Stantic, B. 2018. Answering why-not questions on semantic multimedia queries. *Multimedia Tools and Applications* 77(3):3405–3429.

Xie, M.; Yin, H.; Wang, H.; Xu, F.; Chen, W.; and Wang, S. 2016. Learning graph-based poi embedding for location-based recommendation. In *Proceedings of the CIKM*, 15–24. ACM.

Xiong, C.; Merity, S.; and Socher, R. 2016. Dynamic memory networks for visual and textual question answering. In *Proceedings of ICML*, 2397–2406.

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

Ying, Z.; You, J.; Morris, C.; Ren, X.; Hamilton, W.; and Leskovec, J. 2018. Hierarchical graph representation learning with differentiable pooling. In *Proceedings of NIPS*, 4800–4810.

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

Zhang, J.; Kalantidis, Y.; Rohrbach, M.; Paluri, M.; and Elhoseiny, M. 2019a. Large-scale visual relationship understanding. In *Proceedings of AAAI*, 9185–9194.

Zhang, J.; Shih, K. J.; Elgammal, A.; Tao, A.; and Catanzaro, B. 2019b. Graphical contrastive losses for scene graph parsing. In *Proceedings of CVPR*, 11535–11543.

Zhou, J.; Cui, G.; Zhang, Z.; Yang, C.; Liu, Z.; and Sun, M. 2018. Graph neural networks: A review of methods and applications. *arXiv preprint arXiv:1812.08434*.