Localize, Assemble, and Predicate: Contextual Object Proposal Embedding for Visual Relation Detection

Ruihai Wu, Kehan Xu, Chenchen Liu, Nan Zhuang, Yadong Mu*
School of Electronics Engineering and Computer Science, Peking University
{wuruihai, yurina, liuchenchen, zhuangn53, myd}@pku.edu.cn

Abstract

Visual relation detection (VRD) aims to describe all interacting objects in an image using subject-predicate-object triplets. Critically, valid relations combinatorially grow in $\mathcal{O}(C^2R)$ for C object categories and R relationships. The frequencies of relation triplets exhibit a long-tailed distribution, which inevitably leads to bias towards popular visual relations in the learned VRD model. To address this problem, we propose localize-assemble-predicate network (LAP-Net), which decomposes VRD into three sub-tasks: localizing individual objects, assembling and predicting the subject-object pairs. In the first stage of LAP-Net, Region Proposal Network (RPN) is used to generate a few class-agnostic object proposals. Next, these proposals are assembled to form subject-object pairs via a second Pair Proposal Network (PPN), in which we propose a novel contextual embedding scheme. The inner product between embedded representations faithfully reflects the compatibility between a pair of proposals, without estimating object and subject class. Top-ranked pairs from stage two are fed into a third sub-network, which precisely estimates the relationship. The whole pipeline except for the last stage is object-category-agnostic in localizing relationships in an image, alleviating the bias in popular relations induced by training data. Our LAP-Net can be trained in an end-to-end fashion. We demonstrate that LAP-Net achieves state-of-the-art performance on the VRD benchmark while maintaining high speed in inference.

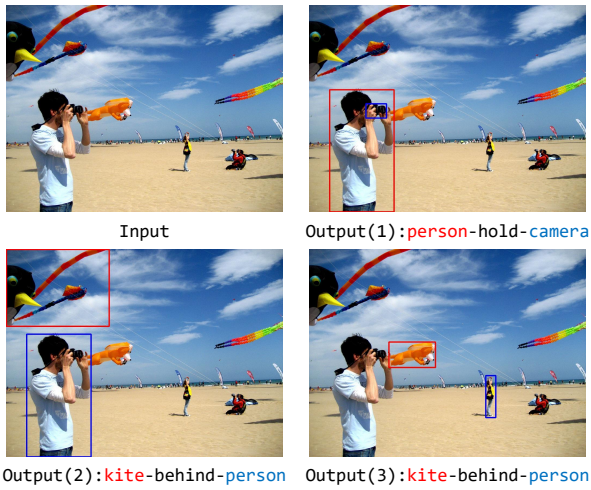

Figure 1: This figure shows some examples of visual relations from VRD dataset. Local information is insufficient for accurate relation detection. For example, in the this image, only the content in object bounding boxes is not enough to decide the relative position between kite and person. Therefore, we propose a novel contextual embedding to prompt global-local information interaction for relation detection.

Introduction

To interpret an image, it is insufficient to locate and recognize objects in the scene. The interactions between objects also need to be carefully estimated. While research on object detection is rapidly progressing, understanding visual relationships is still a hard task and the results of existing pipelines are far from being satisfactory. Visual relationships are defined as <subject-predicate-object> tuples, where subject and object are related by the predicate. There can be many types of predicates, for example, verb (cat-sit on-chair), spatial (book-on-shelf), preposition (person-near-dog), comparative (person1-taller-person2). The goal of visual relationship detection is to both localize objects in the image and predict the relationship between object pairs.

Comparing to object detection, visual relationship detection is difficult in that the distribution of relation triplets is long-tailed. Given C object classes and R relationships, the total number of possible relationships will be $\mathcal{O}(C^2R)$, and learning so many relationships with limited labeling data is a challenging task. One common solution to this problem is to learn separate models for detecting objects and relations, reducing the complexity of training detectors to $\mathcal{O}(C + R)$. In such pipelines, an input image is first fed into an object detection module to get a set of detected objects, then all pairs are considered as potential <subject, object> pairs and will go through the relation detection module sequentially, as proposed in (Lu et al. 2016;

*Corresponding author.

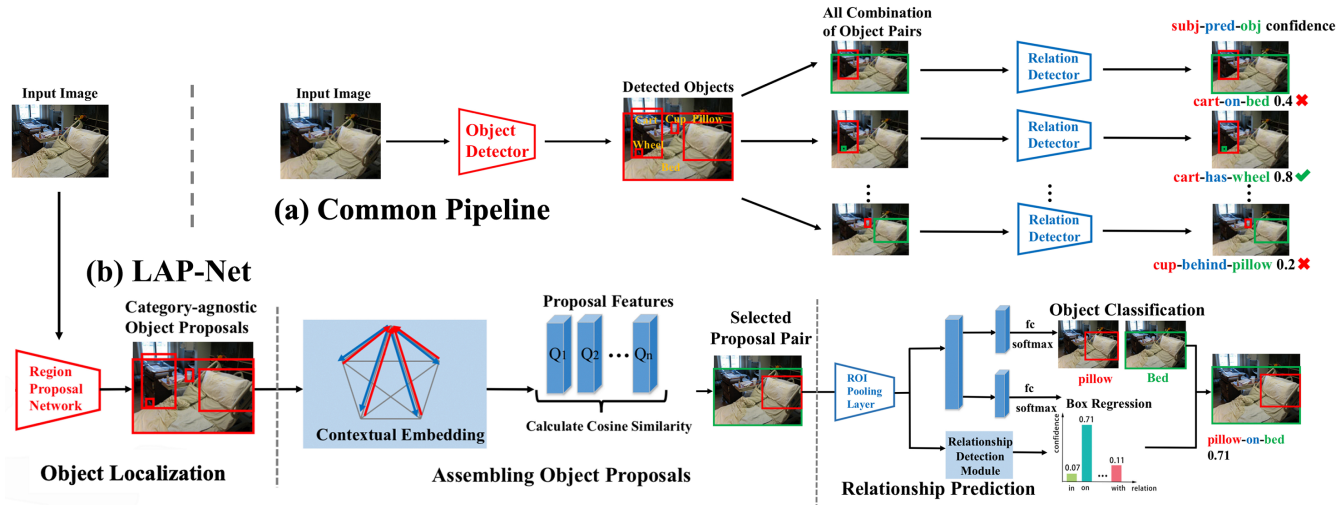

Figure 2: Network comparison between common pipelines and our work. (a) shows the architecture of most VRD pipelines, in which an object detector first outputs objects, then relationship prediction is performed on each pair of them. These pipelines contain much redundant computation, as most object pairs are unrelated. (b) demonstrates our three-stage LAP-Net. In the first stage, an RPN outputs category-agnostic proposals. Stage two assembles compatible subject-object proposals by utilizing a novel contextual embedding. Stage three takes in selected object pairs and classify objects and their relationship in two parallel networks. Contextual embedding and relationship detection module will be illustrated in detail in Fig. 3 and 4.

Zhang et al. 2017a). Usually, the first module generates hundreds of proposals, leading to tens of thousands of possible pairs. To detect the relationship between all the pairs would be too time-consuming and redundant, as in most images there only exist a handful of relationships.

To this end, we introduce our localize-assemble-predicate network (LAP-Net) to address both the problem of redundant proposal pairs and that of biased relationships. After generating object proposals, we intend to select only highly compatible pairs, which possibly contain valid relationships, for relationship prediction. The relation prediction module takes in visual and spatial information of object pairs but is agnostic to their class information, thus the relation prediction is independent of object and subject category, alleviating the bias towards popular subject-predicate-object relationships induced by labeling data.

To be more specific, we decompose VRD into three simpler sub-tasks: localizing individual objects, assembling subject-object pairs, and predicting the categories and relationship between pairs, and let each task solved by a corresponding stage in our network. In the first stage, we generate class-agnostic object proposals using the Relation Proposal Network (RPN) in FasterRCNN (Ren et al. 2015). Our main contribution is in the second stage, in which we propose a novel **Pair Proposal Network** (PPN) that considers visual cues to assemble subject-object pairs from previous proposals. In this module, we construct a complete graph in which each object proposal represents a node and the edge between every two nodes stands for a relation, making proposing valid relationships among objects equal to choosing edges from the graph. A novel contextual embedding is employed to formulate the global and local information exchange between nodes, ensuring that good subject-object pairs admit large inter-vector similarity. Our PPN is light and fast in selecting related pairs from all possible proposal combinations. As the first two stages of our network are respectively class- and predicate-agnostic, the process of relationship pair proposal can be applied to any input image regardless of specific training data or object classes. The selected object-subject pairs are then fed into the third stage, in which two parallel networks classify object proposals and their relation into correct categories simultaneously. A three-branch stacked interacting hourglass network takes in the object, subject and union bounding box features and precisely estimates the relationship. While other methods (Zhang et al. 2017a; Yin et al. 2018; Zhang et al. 2017c; Dai, Zhang, and Lin 2017) claiming to be end-to-end trainable use pre-trained object detection model and only train the rest of their framework, our network can be trained end-to-end as a whole, including the RPN in stage one. In contrast to many existing models that utilize language priors, such as exploring statistical dependency between labels (Dai, Zhang, and Lin 2017), mining external linguistic knowledge (Yu et al. 2017) or constructing word-embedding-related classifier (Zhuang et al. 2017), our LAP-Net is a purely visual model, surpassing existing visual models by a large margin on VRD (Lu et al. 2016) dataset. In summary, our main contributions are as follows:

- We propose LAP-Net, an end-to-end trainable three-stage visual relation detection network aiming to alleviate the relation bias in the training data.
- We introduce a novel Pair Proposal Network (PPN), which utilizes contextual embedding and a complete graph model to assemble class-agnostic object pairs.

- LAP-Net achieves state-of-the-art performance on the VRD benchmark as a purely visual model, surpassing most existing models by a large margin, while maintaining high speed in inference.

Related Work

Object Proposals: Before deep CNNs become popular, object proposal methods can generally be classified into two categories: grouping proposal methods and window scoring proposal methods. The first method is based on merging super-pixels (Uijlings et al. 2013; Carreira and Sminchisescu 2011; Arbeláez et al. 2014), while the other is based on scoring candidate windows and filtering out those with low scores (Alexe, Deselaers, and Ferrari 2012; Zitnick and Dollár 2014). Then CNN-based methods (Szegedy et al. 2014; Cai et al. 2016) gradually emerge, predicting regions based on features extracted by deep neural networks. These methods are adopted in object detection networks (Girshick 2015; Ren et al. 2015) to generate bounding boxes before detectors, forming an end-to-end trainable pipeline.

Visual Relationships: Earlier works exploring visual relationships often focus on a particular type of relationship, including positional (Choi et al. 2013; Johnson et al. 2015) or interactive (Rohrbach et al. 2013; Gkioxari, Girshick, and Malik 2015) relations. As these relationships are from a given category, they are usually extracted using handcrafted features. These relations are utilized to improve the performance on other tasks, such as scene understanding (Kumar and Koller 2010) or image retrieval (Gong et al. 2014).

The problem of detecting generic visual relation is first proposed in (Sadeghi and Farhadi 2011), in which relations are formulated as visual phrases (*e.g.*, "A person riding a horse."). Following the definition in natural language processing (Zhou et al. 2007), (Lu et al. 2016) introduced the task of visual relation detection as both predicting the <subject, predicate, object> triplet and localizing object and subject. The VRD dataset is also proposed in (Lu et al. 2016), which is one of the most common benchmarks on this task. Most recent methods on visual relation detection can be divided into two categories: (i) train one classifier to output the <subject, predicate, object> tuple; (Sadeghi and Farhadi 2011; Divvala, Farhadi, and Guestrin 2014) (ii) predict subject, object and relationship from separate classifiers (Lu et al. 2016; Dai, Zhang, and Lin 2017; Yin et al. 2018; Yu et al. 2017). As the number of relation tuples can be extremely large and the labeled data is long-tailed, training a classifier to handle all possible tuples can not generate satisfying results. Therefore, more methods are now focused on predicting objects and relationships with separately and trying to form information exchange between different classifiers. Almost all these methods first perform object detection and then predict relationships for each pair of detected objects. This is inefficient in that the number of pairs is quadratic to that of proposals, but usually, only a handful of relationships exist in an image, so sending most of the pairs into the relation classifier is wasteful.

Down this line, our network aims to filter out unrelated object pairs before passing them into the relationship detection module to improve efficiency. To alleviate the bias induced by the long-tailed distribution of <subject, predicate, object> triplets in the dataset, we generate category-agnostic proposals for selection. The method in Relationship Proposal Network (Zhang et al. 2017c) is the most related to ours, in which proposal pairs are also selected before relationship prediction. Our pipeline differs in two aspects: (i) We utilize only one object proposal module to generate potential pairs, while Relationship Proposal Network uses three modules to predict subject, object, union box respectively; (ii) We propose a novel contextual embedding to effectively formulate the compatibility between object proposals. Experiments demonstrate that our framework surpasses Relationship Proposal Network by a considerable margin when evaluating both proposal selection and final relationship detection results on the VRD benchmark.

Localize-Assemble-Predicate Network

LAP-Net is an end-to-end trainable network with three stages. Category-agnostic object proposals are generated in stage one. Stage two assembles compatible subject-object pairs and feeds them into stage three to predict the category of object, subject, and relation in parallel. Our framework is highly efficient in that stage two accurately selects related object proposal pairs from all possible combinations, reducing the computational waste of sending irrelevant pairs into stage three. Relationship prediction in stage three is object-class-agnostic to alleviate the bias in popular relations and is still able to achieve state-of-the-art performance in relation detection without category information. Each stage is explained in detail in a subsection below. Fig. 2 illustrates the overview of our pipeline.

Stage One: Object Localization

In stage one, we use the Region Proposal Network (RPN) in FasterRCNN (Ren et al. 2015) to generate object proposals. Each proposal is represented as a bounding box in the image, from which visual and spatial features but category information is extracted. We choose ResNet-50 (He et al. 2016) network as our backbone architecture. Generally, RPN outputs thousands of proposals. Non-maximum-suppression (NMS) with IoU > 0.7 is performed on the results of RPN and prunes them to no more than 512 object proposals before feeding them into the next stage.

Stage Two: Assembling Object Proposals

Stage two aims at filtering out incompatible proposal pairs and feeding remaining pairs into the next stage for relationship prediction. We construct a complete graph with object proposals as its nodes and relationships as edges between every two nodes. A novel contextual embedding is proposed to select edges from the graph model, which is equal to assembling relationships from all possible proposal pairs. To select edges that represent valid relationships, our contextual embedding utilizes the Conditional Random Field (CRF) model (Lafferty, Mccallum, and Pereira 2001) to formulate global and local information exchange. In our contextual embedding, we design our message passing procedure in CRF, then fully integrate CRF modeling with CNNs,

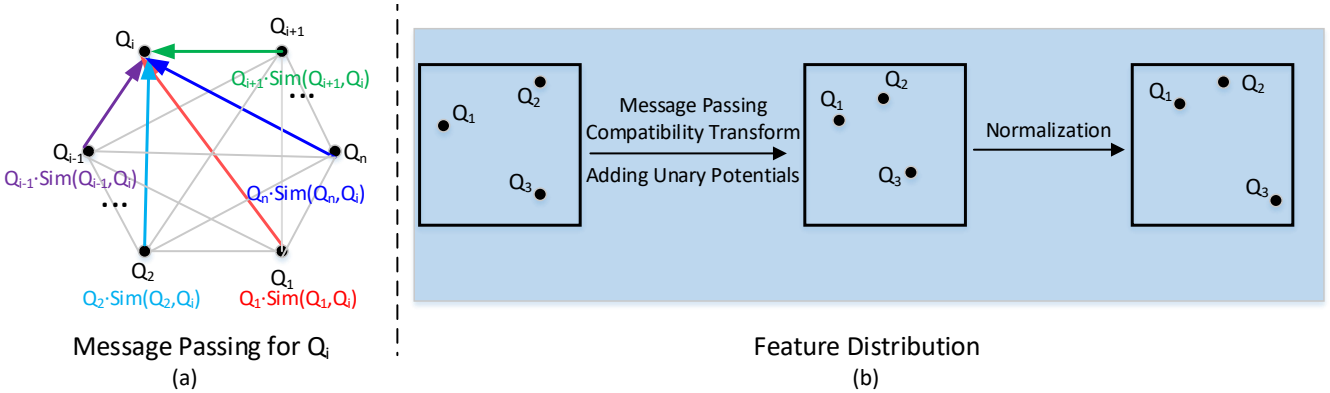

Figure 3: (a) Illustration of message passing in contextual embedding. Here shows the feature interaction procedure for the ith proposal's embedding Q_i. Q_i receives information from all other $Q_j (j \neq i)$, and the information intensity is determined by the cosine similarity between Q_i and Q_j. Message passing can be done for all Q_is in parallel. (b) A simplified demonstration on the distribution of proposal features in each iteration. After the information exchange is calculated in message passing, embeddings originally having cosine similarity> 0 will get closer, while those with similarity< 0 will become further. Embeddings admitting larger cosine similarity will contribute more to the feature interaction between each other, thus making close features closer. The feature distribution scale is aligned back to standard in the following normalization step. These steps will help to form clearer boundaries between clusters of embeddings. In this case, the boundary is between $\{Q_1, Q_2\}$ and $\{Q_3\}$.

making the whole network end-to-end trainable while maintaining the desirable interpretability of CRF. The optimization target of our contextual embedding is to ensure good subject-object pairs admit large inter-vector similarity, and we prove that our self-designed message passing procedure in CRF is following this objective.

Contextual Embedding with Conditional Random Field Following (Krähenbühl and Koltun 2011), we first provide a brief overview of CRF modeling for assembling object-subject pairs from all proposals. A CRF in this context models features of each proposal as random variables that form a Markov Random Field (MRF) conditioning on a global observation. Here, the global observation is the input image and all object proposals. We let the random variable related to the ith proposal be denoted by X_i, and the total number of proposals be N. We construct a complete undirected graph $G = (V, E)$, in which $V = \{X_1, ..., X_N\}$ and $E = \{e_{ij} | 1 \leq i < j \leq N\}$. Consider one random field X defined over $\{X_1, ..., X_N\}$ and another random field I defined over the whole image as the global observation, a fully connected pairwise conditional random field (I, X) is characterized by a Gibbs distribution:

$$P(X|I) = \frac{1}{Z(I)} exp(-\sum_i \phi_i(X_i|I)) \tag{1}$$

We denote $\sum_{i \in \{1...N\}} \phi_i(x_i|I)$ as $E(x)$, the energy of configuration x. $Z(I)$ is the partition function (Lafferty, Mccallum, and Pereira 2001). For convenience, we drop the conditioning on I in the rest. The energy function of x is:

$$E(x) = \sum_i \psi_u(x_i) + \sum_{i<j} \psi_p(x_i, x_j), \tag{2}$$

where i and j range from 1 to N. The unary potential $\psi_u(x_i)$ is computed independently for each proposal, representing the original object feature without information exchange. In our implementation, the unary potential considers visual cues to generate initial features that fully represents each region's context. The pairwise potential is formulated as:

$$\psi_p(x_i, x_j) = \mu(x_i, x_j) \sum_{m=1}^{K} \omega^{(m)} k^{(m)}(f_i, f_j). \tag{3}$$

Here, μ is the category compatibility function, capturing the compatibility of labels between overlapping or nearby proposals. For example, the categories of two near proposals being "floor" and "desk" should be penalized lesser than "sky" and "desk". $k^{(m)}(f_i, f_j)$ represents the information exchange between the ith and jth proposal. Different m stands for different ways of feature interaction, and ω are learnable weights for them.

Inspired by the highly efficient mean-field approximation to the CRF distribution in (Krähenbühl and Koltun 2011), we imitate them to generate an approximation for inference in our fully connected CRF model. The whole inference algorithm is summarized in Algorithm 1 and will be explained in detail in the following paragraph. We follow (Zheng et al. 2015) to formulate our mean-field iteration algorithm as a Recurrent Neural Network (RNN) to enable the end-to-end trainable property of the whole pipeline. After obtaining contextual embeddings for proposals as output of the algorithm, we calculate cosine similarity between the embeddings and select those with high similarity to be subject-object pairs. A pair-assembling loss is imposed on the output pairs of stage two. It is defined as:

$$\begin{aligned} L_{pair} = & \sum_{(i,j)related} \|Sim(f_i, f_j) - 1\|_2 \\ & + \sum_{(i,j)notrelated} \|Sim(f_i, f_j) + 1\|_2 \end{aligned} \tag{4}$$

Algorithm 1 Mean field approximation in fully connected CRFs for pair proposal selection

$Q_i(l) \leftarrow \frac{1}{\sqrt{\sum_l Q_i(l)^2}} Q_i(l)$ for all i — Initialization

while not converged **do**

$\widetilde{Q}_i(l) \leftarrow \alpha \sum_{j \neq i} k(f_i, f_j) Q_j(l)$ — Message Passing

$\widehat{Q}_i(l) \leftarrow \sum_{l'} \mu(l, l') \widetilde{Q}_i(l')$ — Compatibility Transform

$\overline{Q}_i(l) \leftarrow U_i(l) + \widehat{Q}_i(l)$ — Adding Unary Potentials

$Q_i(l) \leftarrow \frac{1}{\sqrt{\sum_l \overline{Q}_i(l)^2}} \overline{Q}_i(l))$ — Normalization

end while

$Sim(f_i, f_j) = \frac{f_i \cdot f_j}{\|f_i\| \times \|f_j\|}$ is the cosine similarity. This pair-assembling loss enforces pairs labeled as related in ground truth to have top ranked feature similarity, making it the optimization goal of the CRF model.

Here, Q_i is the contextual embedding for proposal i that is gradually updated in the algorithm. l stands for each dimension in the contextual embedding vector. To improve efficiency, we propose only one simple but effective way of feature interaction in the message passing step ($k = 1$). Our $k(f_i, f_j)$ is formulated as the cosine similarity between contextual embeddings of object proposal i and j. If the feature of proposal j is the closest to that of proposal i among all proposals, feature j will contribute most in affecting feature i towards changing in its direction. α here is a hyperparameter adjusting the extent to which other features affects feature i. The message passing step is efficient in that feature calculation for each proposal can be executed in parallel. So the time spent in this step is equal to that of feature calculation on one proposal, which is $O(N)$ instead of $O(N^2)$. In compatibility transform, information is exchanged between different dimensions in each contextual embedding vector, while μ contains learnable parameters trained by the pair-assembling loss mentioned above. We believe the learnable interaction inside feature vectors can prompt contextual embedding to adjust itself towards better representation and more effective relationship detection. In the next step, updated features are added to the unary potential to generate new proposal embeddings. U_i is the unary potential for proposal i, and we assign its value to be that of Q_i after the initialization step in the algorithm. The information from other features gathered in message passing step is now transmitted to feature i. According to the claim above, embeddings that are originally close to each other would contribute more than further embeddings to one another's feature and thus become relatively closer after each iteration. See Fig. 3 for an illustration of the whole feature interacting process. Though the scale of distance distribution between all proposal features might fluctuate after updating, a normalization step followed immediately to stabilize the scale, thus enabling proposal features to form more obvious clusters without altering the scope of feature distribution. This normalization also enables contextual embedding to align in the afterward cosine similarity calculation procedure.

After several iterations, the CRF model outputs the updated features for all proposals, each of them is an effective contextual embedding containing the interactive information with other boxes. One advantage of our contextual embedding is that the global knowledge each proposal receives is not the same, general information, but the knowledge is proposal-specific, in that every object sees the whole image by gathering information from interacting with other proposals. These interactions include information from the object itself, implementing object-specific knowledge into global information, making the contextual embedding unique for the object and effective for relationship proposal.

The optimization goal of the CRF model is to generate contextual embeddings that ensure good subject-object pairs admit large inter-vector similarity. As message passing in CRF intuitively enables similar proposal features to congregate, boundaries will naturally form between groups of related proposals in the embedding space. This together with the pairwise objective function makes the training of our pair proposal network highly efficient and effective. Thus, the optimization procedure in CRF follows its goal.

Stage Three: Relationship Prediction

In stage three, two parallel modules, object classification module and relation detection module, classify object proposals and relations simultaneously, and their outputs are assembled to generate final results of visual relation detection.

Object Classification Module The object classification module takes in all proposals from RPN and generates two branches of outputs: the object categories and the precise positions of their bounding boxes. Following the multi-task loss in Fast RCNN (Girshick 2015), the objective function we minimize at training is a combination of classification loss and bounding box regression loss, which is defined as:

$$L(\{p_i\}, \{t_i\}) = \frac{1}{N_{cls}} \sum_{i=1} L_{cls}(p_i, p_i^*) + \lambda \frac{1}{N_{reg}} \sum_{i=1} p_i^* L_{reg}(t_i, t_i^*), \tag{5}$$

where i is the index of a proposal. p_i is the predicted probability of object i belonging to each category, and t_i is a vector representing the four coordinates of the bounding box. p_i^* and t_i^* are the corresponding ground-truth label and box of the proposal. $p_i^* L_{reg}$ ensures regression loss calculate only on the right label. Details can be referred to (Girshick 2015).

Relationship Detection Module Pairs of related object proposals are assembled from stage two. For each pair of proposals, we define the predicate bounding box as the union box of object and subject bounding boxes. The relationship detection module is a three-branch stacked interacting hourglass network, as shown in Fig. 4, taking the visual features of the object, subject, and predicate bounding box as input. LAP-Net is efficient as it only requires each image to pass through the feature extractor once. Here, the predicate bounding box feature can be quickly inferred from the image feature map, instead of being calculated from scratch.

Object, subject, predicate features first all go through an ROI Align layer (He et al. 2017), then three features are concatenated and passed into a fully-connected (FC) layer with

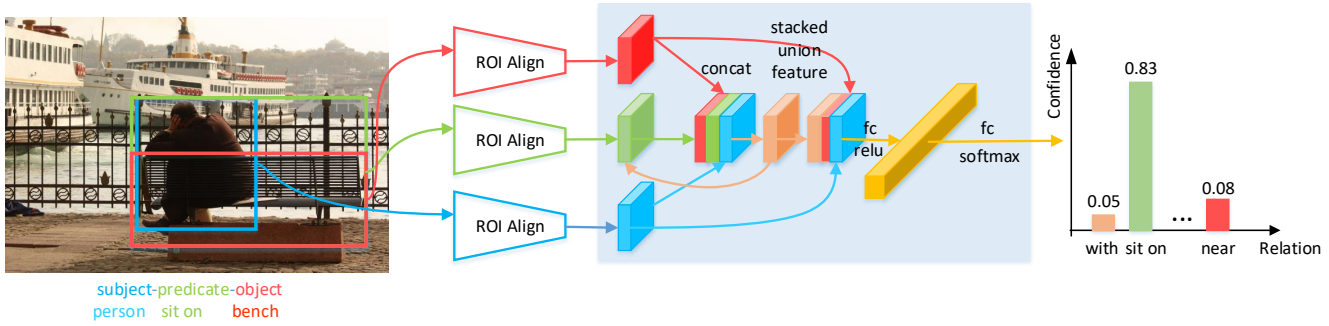

Figure 4: The relationship detection module is a three-branch stacked interacting hourglass network. It takes in features from subject, object, and union bounding boxes and first performs ROI Align on these features. Predicate features are updated by concatenating object and subject features continuously, and the update can be performed for several iterations to prompt interaction between global and local information. Then three features are congregated to predict the confidence in each kind of relationship between subject and object boxes.

a ReLU layer to generate an updated predicate feature. The motivation to update the predicate feature is that, though the union bounding box itself contains global information that would complement for the local information in object and subject features, this global feature only sees the whole region without knowing the exact position of subject and object in the area. Therefore, to extract more effective global features for relation detection, we seek to incorporate object and subject information into the union box feature. This global-local interaction to improve predicate feature can be performed multiple times. After obtaining effective predicate feature, object, subject, predicate features are concatenated and sent into two FC layers, then passed into a softmax layer to predict the confidence for each type of relationship between object and subject. After getting the probability distribution for all relationships in all pairs, we select those relationships with the highest confidence, pairing them with object and subject categories obtained from the object detection module as the final output of our LAP-Net.

Experiments

We evaluate LAP-Net on two kinds of outputs: (i)<subject, predicate, object> triplets, which is the final output of our framework and the most commonly evaluated target in visual relationship detection; (2)relationship proposals, the output of our pair proposal network (PPN), following the setup in (Zhang et al. 2017c), as it shares similarity with our work in pair proposal selection. We demonstrate the effectiveness of our pipeline by surpassing all existing visual models and most of the language-based models.

Evaluation on Visual Relationship Detection

Dataset and Metrics We conduct experiments on VRD dataset (Lu et al. 2016). This dataset contains 5,000 images and 37,993 visual relations with 6,672 relation types. There are 100 object categories and 70 predicates. We follow the splitting of train/test set in (Lu et al. 2016), using 4000 images in training and test on the remaining 1000 images.

Here, visual relation detection is evaluated under two measurements: relationship detection and phrase detection. In relationship detection, the framework is considered to generate a correct relationship if <subject, predicate, object> triplets are categorized correctly, and the subject and object bounding boxes must each have at least 0.5 overlap with their ground truth boxes. Phrase detection is similar to relationship detection, but different in that it requires the union bounding box's overlap to be at least 0.5. The evaluation metrics are recall@50 and recall@100. Recall@x stands for the portion of ground truth relationship predicted in the top x confident relationship predictions. Unlike object detection, the annotation of relationship may be incomplete due to its diversity and ambiguity. Thus, precision is not the proper metric because predicted relationships might be correct but not labeled in the image, while recall is effective in that it only evaluates on labeled ground truth. As multiple predicates might exist to describe the relationship between two objects, we use k to represent the maximum number of relationships chosen per object pair. We evaluate on $k = 1$ and 70. For $k = 1$, only one relationship can be predicted for each pair. For $k = 70$, as the total number of predicates is 70, this is equivalent to evaluating all possible relationships.

Comparative Results The comparison between our proposed LAP-Net and several existing methods on the VRD dataset is listed in Table 1. The results show that our framework achieves state-of-the-art on most of the evaluation metrics. LAP-Net improves greatly when k=1, surpassing previous state-of-the-art methods by 2.30% Recall@50 and 2.90% Recall@100 in relation detection. In k=70, our pipeline only demonstrates comparable performance with some prior arts. Our Recall@50 results in relation and phrase prediction outperform previous methods by 1.32% and 0.26% respectively, while the performance on Recall@100 is slightly lower than state-of-the-art. Still, LAP-Net is superior in that it only utilizes visual information for visual relation prediction, while other methods like CAI (Zhuang et al. 2017) and LK (Liang, Lee, and Xing 2017) adopt internal and external language knowledge to

Table 1: Comparison with existing visual relationship detection methods on VRD dataset. Best performances are shown in bold.

k	Methods	Relationship		Phrase	
		Recall@50	Recall@100	Recall@50	Recall@100
k=1	LP (Lu et al. 2016)	13.86	14.70	16.17	17.03
	VTransE (Zhang et al. 2017a)	14.07	15.02	19.42	22.42
	PPR-FCN (Zhang et al. 2017b)	14.41	15.72	19.62	23.15
	TFR (Jae Hwang et al. 2018)	15.20	16.80	17.40	19.10
	CAI (Zhuang et al. 2017)	15.63	17.39	17.60	19.24
	Weak-Su (Peyre et al. 2017)	15.80	17.10	17.90	19.50
	Deep-Str (Zhu and Jiang 2018)	17.27	18.26	22.61	23.92
	Vip-cnn (Li et al. 2017)	17.32	20.01	22.78	27.91
	VRL (Liang, Lee, and Xing 2017)	18.19	20.79	21.37	22.60
	LK (Yu et al. 2017)	19.17	21.34	23.14	24.03
	Zoom-Net (Yin et al. 2018)	18.92	21.41	24.82	28.09
	CAI + SCA-M (Zhuang et al. 2017)	19.54	22.39	25.21	28.89
	LAP-Net	**21.84**	**25.29**	**28.07**	**33.05**
k=70	DR-Net (Dai, Zhang, and Lin 2017)	17.73	20.88	19.93	23.45
	LK (Yu et al. 2017)	22.68	**31.89**	26.32	29.43
	Zoom-Net (Yin et al. 2018)	21.37	27.30	29.05	37.34
	CAI + SCA-M (Zhuang et al. 2017)	22.34	28.52	29.64	**38.39**
	LAP-Net	**24.00**	29.67	**29.90**	37.23

Table 2: Relationship proposal recall rates on VRD dataset with **IoU**≥ 0.5 and different proposal numbers.

IoU≥ 0.5	2000	5000	8000	10000
SS, pairwise	22.1	28.0	31.4	33.0
EB, pairwise	15.1	20.6	24.2	25.2
RPN, pairwise	28.9	36.2	41.0	43.0
Rel-PN	38.3	44.3	46.4	47.3
LAP-Net	**57.9**	**66.6**	**71.0**	**72.6**

achieve similar or slightly higher performance.

Methods in Table 1 mostly rely on pre-trained object detectors to generate proposals with class information. We observe that object detectors sometimes fail to detect all objects in an image. Though those lost objects can be localized in RPN, they are not assigned with high confidence scores in the detection stage and thus are filtered out. These objects might be involved in some relationships, which will also go lost in this case. Our pipeline prevents this from happening, for we directly take the output of RPN to find related pairs. Our contextual embedding enables the network to precisely assemble compatible proposal pairs, forcing the following object classification module to identify the proposal's category. Such relationships can thus be accurately detected.

Evaluation on Relationship Proposals

Evaluation Settings We follow Relationship Proposal Network (Rel-PN) (Zhang et al. 2017c) to evaluate our model by localizing relationships in images, aiming to demonstrate the effectiveness of our second stage in assembling subject-object relationships. Recall rate is calculated on varying numbers of relationship proposals (2000, 5000, 8000, 10000) with IoU≥ 0.5. Here, "IoU≥ 0.5" suggests subject and object proposals overlap with their ground truth bounding boxes by at least 0.5 respectively. (Zhang et al. 2017c) also proposes three baselines to compare recall rate with, each of them containing a different object detection module and outputs relationship proposals as combinations between every two detected objects. These three detection methods are: Selective Search (Uijlings et al. 2013), Edge-Boxes (Zitnick and Dollár 2014), and Region Proposal Network (Ren et al. 2015). We compare with Relationship Proposal Network and other baselines using the output from stage two. Experiments are also conducted on VRD dataset.

Comparative Results The results are reported in Table 2. The recall rate of stage two in LAP-Net is extremely high, almost reaching a 50% increase over Relationship Proposal Network in all settings. This experiment demonstrated the efficiency of our Pair Proposal Network (PPN), proving that the proposed contextual embedding is effective in modeling the compatibility between object proposals.

Concluding Remarks

We proposed a three-stage end-to-end trainable framework LAP-Net for visual relation detection. The core is a novel Pair Proposal Network (PPN) that employs Conditional Random Field (CRF) for message passing on a complete graph to select class-agnostic object proposal pairs. By assembling compatible object pairs from all proposal combinations, our method greatly reduces computational cost. Experiments demonstrated the superior accuracy and efficiency of our model in learning complex relationships.

Acknowledgement: This work is supported by Beijing Municipal Commission of Science and Technology under Grant Z181100008918005, National Natural Science Foundation of China (NSFC) under Grant 61772037.

References

Alexe, B.; Deselaers, T.; and Ferrari, V. 2012. Measuring the objectness of image windows. *IEEE transactions on pattern analysis and machine intelligence* 34(11):2189–2202.

Arbeláez, P.; Pont-Tuset, J.; Barron, J. T.; Marques, F.; and Malik, J. 2014. Multiscale combinatorial grouping. In *CVPR*.

Cai, Z.; Fan, Q.; Feris, R. S.; and Vasconcelos, N. 2016. A unified multi-scale deep convolutional neural network for fast object detection. In *ECCV*.

Carreira, J., and Sminchisescu, C. 2011. Cpmc: Automatic object segmentation using constrained parametric min-cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 34(7):1312–1328.

Choi, W.; Chao, Y.-W.; Pantofaru, C.; and Savarese, S. 2013. Understanding indoor scenes using 3d geometric phrases. In *CVPR*.

Dai, B.; Zhang, Y.; and Lin, D. 2017. Detecting visual relationships with deep relational networks. In *CVPR*.

Divvala, S. K.; Farhadi, A.; and Guestrin, C. 2014. Learning everything about anything: Webly-supervised visual concept learning. In *CVPR*.

Girshick, R. 2015. Fast r-cnn. In *ICCV*.

Gkioxari, G.; Girshick, R.; and Malik, J. 2015. Contextual action recognition with r* cnn. In *ICCV*.

Gong, Y.; Ke, Q.; Isard, M.; and Lazebnik, S. 2014. A multi-view embedding space for modeling internet images, tags, and their semantics. *International journal of computer vision* 106(2):210–233.

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

He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. 2017. Mask r-cnn. In *ICCV*.

Jae Hwang, S.; Ravi, S. N.; Tao, Z.; Kim, H. J.; Collins, M. D.; and Singh, V. 2018. Tensorize, factorize and regularize: Robust visual relationship learning. In *CVPR*.

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

Krähenbühl, P., and Koltun, V. 2011. Efficient inference in fully connected crfs with gaussian edge potentials. In *NIPS*.

Kumar, M. P., and Koller, D. 2010. Efficiently selecting regions for scene understanding. In *CVPR*.

Lafferty, J.; Mccallum, A.; and Pereira, F. C. N. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML*.

Li, Y.; Ouyang, W.; Wang, X.; and Tang, X. 2017. Vip-cnn: Visual phrase guided convolutional neural network. In *CVPR*.

Liang, X.; Lee, L.; and Xing, E. P. 2017. Deep variation-structured reinforcement learning for visual relationship and attribute detection. In *CVPR*.

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

Peyre, J.; Sivic, J.; Laptev, I.; and Schmid, C. 2017. Weakly-supervised learning of visual relations. In *ICCV*.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*.

Rohrbach, M.; Qiu, W.; Titov, I.; Thater, S.; Pinkal, M.; and Schiele, B. 2013. Translating video content to natural language descriptions. In *ICCV*.

Sadeghi, M. A., and Farhadi, A. 2011. Recognition using visual phrases. In *CVPR*.

Szegedy, C.; Reed, S.; Erhan, D.; Anguelov, D.; and Ioffe, S. 2014. Scalable, high-quality object detection. *arXiv preprint arXiv:1412.1441*.

Uijlings, J. R.; Van De Sande, K. E.; Gevers, T.; and Smeulders, A. W. 2013. Selective search for object recognition. *International journal of computer vision* 104(2):154–171.

Yin, G.; Sheng, L.; Liu, B.; Yu, N.; Wang, X.; Shao, J.; and Change Loy, C. 2018. Zoom-net: Mining deep feature interactions for visual relationship recognition. In *ECCV*.

Yu, R.; Li, A.; Morariu, V. I.; and Davis, L. S. 2017. Visual relationship detection with internal and external linguistic knowledge distillation. In *ICCV*.

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

Zhang, H.; Kyaw, Z.; Yu, J.; and Chang, S.-F. 2017b. Ppr-fcn: weakly supervised visual relation detection via parallel pairwise r-fcn. In *ICCV*.

Zhang, J.; Elhoseiny, M.; Cohen, S.; Chang, W.; and Elgammal, A. 2017c. Relationship proposal networks. In *CVPR*.

Zheng, S.; Jayasumana, S.; Romera-Paredes, B.; Vineet, V.; Su, Z.; Du, D.; Huang, C.; and Torr, P. H. 2015. Conditional random fields as recurrent neural networks. In *ICCV*.

Zhou, G.; Zhang, M.; Ji, D.; and Zhu, Q. 2007. Tree kernel-based relation extraction with context-sensitive structured parse tree information. In *EMNLP-CoNLL*.

Zhu, Y., and Jiang, S. 2018. Deep structured learning for visual relationship detection. In *AAAI*.

Zhuang, B.; Liu, L.; Shen, C.; and Reid, I. 2017. Towards context-aware interaction recognition for visual relationship detection. In *ICCV*.

Zitnick, C. L., and Dollár, P. 2014. Edge boxes: Locating object proposals from edges. In *ECCV*.