Fixed-Size Objects Encoding for Visual Relationship Detection

Hengyue Pan¹ · Xin Niu¹ · Siqi Shen² · Yixin Chen¹ · Peng Qiao¹ · Zhen Huang¹ · Dongsheng Li¹

Accepted: 4 February 2022 / Published online: 23 February 2022
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2022

Abstract
In this paper, we propose a fixed-size object encoding method called FOE-VRD to improve performance of visual relationship detection tasks. For each relationship triplet in a given image, FOE-VRD not only considers the subject and object, but also uses one fixed-size vector to encoding all background objects of the image. In this way, we introduce more background knowledge to assist the relationship detector for better performance. We firstly use a regular convolution neural network as a feature extractor to generate high-level features of input images. Then, for each relationship triplet, we apply ROI-pooling as the feature generator on the bounding boxes of subject and object to get two corresponding feature vectors. Moreover, we propose a novel method to encode all background objects in each image by using one fixed-size vector (i.e., FBE vector). By concatenating the 3 generated feature vectors, we successfully encode the relationship using one fixed-size vector. The generated feature vector is then feed into a fully connected neural network to get the predicate classification result. Experimental results on VRD and Visual Genome databases show that the proposed method works well on both predicate classification and relationship detection tasks, especially on the situation of zero-shot detection.

✉ Hengyue Pan
hengyuepan@nudt.edu.cn
Xin Niu
niuxin@nudt.edu.cn
Siqi Shen
siqishen@xmu.edu.cn
Yixin Chen
chenyixin@nudt.edu.cn
Peng Qiao
pengqiao@nudt.edu.cn
Zhen Huang
zhenhuang@nudt.edu.cn
Dongsheng Li
dsli@nudt.edu.cn

¹ National University of Defense Technology, 109 Deya Street, Changsha, Hunan, People’s Republic of China
² Xiamen University, 422 Siming South Road, Xiamen, Fujian, People’s Republic of China
1 Introduction

To understand a given image, we should learn many pieces of information from it. One natural idea is that we need to know the locations and categories of foreground objects, and the corresponding task is called object detection in computer vision. In the past few years, many successful object detection algorithms have been proposed, such as SSD [14], SPPNets [5], Faster RCNN [18], Mask R-CNN [6] and so on. The existing object detection methods achieve excellent performance on several databases like ImageNet [3] and COCO [13]. Moreover, many of them have already been applied on real-world tasks.

However, simply knowing locations and class IDs of objects in images is far from enough. In many cases, two images that include the same kinds of objects may have quite different semantic meaning. This fact requires us to classify the relationship between each pair of objects in images, which can be presented as a relationship triplet, i.e., \(<\text{subject}-\text{predicate}-\text{object}>\). This task is the so-called Visual Relationship Detection. Recently, comparing with object detection, visual relationship detection attracts more attention, not only because it is much more challengeable than object detection, but also it may help us to develop better algorithms to understand natural images automatically. [15] proposed VRD algorithm, which was a successful research at the early stage of visual relationship detection. By combining a visual model and a language model, the VRD algorithm achieved good performance on several visual relationship detection databases. Moreover, this paper proposed to use Recall@X as the performance metric instead of mAP, and Recall@X is applied in many subsequent researches. In [22], Yu et al. argued that it is unfair to simply evaluate the top detected relationship between each object pair since some correct predictions may be penalized mistakenly. Therefore, a new hyper-parameter \(k\) was introduced, which is the number of chosen predictions per object pair to calculate Recall@X. The introducing of \(k\) made the evaluation of visual relationship detection tasks more reasonable, thus most of recent works tend to apply \(k\) as a hyper-parameter in their experiments. In [11], a visual relationship detection framework called Deep Structural Ranking was proposed. This method introduced structural ranking loss into the deep neural network framework to reduce the negative impact of incomplete annotations. Deep Structured Learning [25] is another efficient relationship detection method. The deep structured model included feature-level and label-level relationship predictions, and the two prediction results should be weighted summarized to generate the final result. Moreover, this work presented to use SSVM loss as the optimization goal, which resulted in simpler and more independent optimization procedures. Zoom-Net [21] considered to apply two pooling cells, i.e., Contrastive ROI Pooling and Pyramid ROI Pooling, to improve relationship detection ability of the network. The algorithm achieved excellent performance on several widely-used databases. In [16], a Hierarchical Graph Attention Network (HGAT) was proposed to address one important problem of existing graph-based visual relationship detection methods, i.e., they ignored the dependencies between triplets. HGAT explicitly models dependencies between relationship triplets by introducing triplet-level graph attention, which provides positive influence on the final detection results. [10] proposed neural models to learn relationships between movie characters. Different from regular visual relationship detection tasks, the relationships between movie characters can be inferred via not only visual information, but also dialog cues. Therefore, a multi-modal architecture was designed to capture
those pieces of information. This method achieves good performance on the MovieGraphs database.

In this paper, we propose a novel visual relationship detection algorithm called Fixed-size Objects Encoding for Visual Relationship Detection (FOE-VRD). Comparing with the previous counterparts, the proposed method pays more attention to one problem: how to encode all objects in given images to improve the performance of visual relationship detection? The basis of this question is the consideration that the kind of relationship between two selected objects may not simply related to the two objects themselves, but all background objects in the image. Therefore, encoding all objects to a fixed-size vector may bring about conveniences to further computations. The main obstacle of this idea is fact that the number of objects in images are not fixed, thus it is not a trivial task to use a fixed-size vector to do object encoding. The main contribution of this paper is that we propose a novel method to generate a fixed-size feature vector to model all objects in one image, no matter how many objects are included in the image. The fixed-size feature vector discards most visual information, but keeps two important things: the class IDs and the relative distribution of background objects. By introducing the fixed-size background encoding (FBE) method, the proposed FOE-VRD algorithm achieves state-of-the-art performance on both Stanford Visual Relationship Detection [15] (VRD) and Visual Genome [9] (VG) databases.

2 Related Works

Generally speaking, the proposed FOE-VRD algorithm has some important bases, i.e., ROI pooling method, VRD algorithm, and FOFE algorithm. In this section, we review the core ideas of the three above mentioned methods.

2.1 ROI Pooling

In [4], the ROI Pooling layer is proposed for object detection tasks. A ROI (Region Of Interest) is a rectangular region on feature maps of one convolution layer, which is corresponding to a foreground object of the input image. By introducing max-pooling operation, we can pool each ROI rectangle to a group of smaller fixed-size feature maps. By doing ROI Pooling, we can convert all ROIs to fixed-size features, despite the different size of their corresponding objects. ROI Pooling brings about convenience for further detection and classification operations. In this work, we apply ROI pooling to encode the subject and object of every given relationship triplet.

2.2 VRD Algorithm

The VRD algorithm proposed in [15] is an important visual relationship detection algorithm, which serves as significant basis for many researches. The VRD algorithm introduced two models, i.e., visual model and language model, for relationship detection tasks. Specifically, for each pair of objects in input images, a convolutional neural network serves as the visual model to calculate the possibility of each kind of predicate. On the other hand, a projection function is trained to project all possible relationships into an embedding space, where the relationships with similar semantic meaning should close to each other. By proposing an objective function combines the loss of visual model and language model, the VRD model can be trained end-to-end efficiently. This work reveals the importance of semantic information
Fixed-size Ordinally-Forgetting Encoding (FOFE) [23] is an encoding method of natural language processing. The basic idea of FOFE is shown in Fig. 1: assuming that we have a dictionary that includes \( K \) words (we use \( K = 4 \) as an example), it is obvious that we can use 1-of-K code to represent each word. For any word sequences \( S = w_1, w_2, ..., w_T \), we can always use corresponding \( K \)-length vectors \( V \) to represent them by introducing a forgetting factor \( \rho \):

\[
V_t = \rho V_{t-1} + c_t
\]

where \( c_t \) is the 1-of-K code of the word \( w_t \), and \( V_t \) is the FOFE vector of the sequence \( w_1, w_2, ..., w_T \).

Moreover, [23] proved the uniqueness of FOFE vector, which means that FOFE can serve as a good encoding method for language models.

FOFE is an important basis of our work, since it provides a possible way to encode variable-length word sequences to fixed-size feature vectors. Based on FOFE, we propose Fixed-size Background Encoding (FBE) method to encode all objects in a given image to a fixed-size feature vector.

### 3 Method

The most important motivation of our work is the consideration that the category of a predicate not only related to the corresponding subject and object, but also the other background objects in the whole image. Figure 2 shows a good example: in both images the subjects are human beings and the objects are horses, but they have different predicates. In left image the predicate is riding, while in right image the predicate is feeding. The two images have quite different background objects. For instance, in the right image we can find barriers and fodders, which may not appear around the relationship triplet of \(<\text{human}-\text{riding-\text{horse}}\>\). Therefore, finding a suitable way to model the background objects may benefit the performance of visual relationship detection. However, due to the uncertain number of background objects, it is hard to generate a fixed-size feature vector for all background objects. This may disturb the further processing that applies fully-connected neural network layers. Moreover, because of the large number of background objects, directly modeling all objects may result in very
huge time and space complexity. Thus in practice, we hope to find a better way to model them.

3.1 Fixed-Size Background Encoding

Based on FOFE, we propose Fixed-size Background Encoding (FBE) method to encode all objects in a given image to a fixed-size feature vector. According to the basic definition of FOFE, we introduce an important feature to FBE: the algorithm outputs a fixed-size background vector (FBE vector) no matter how many background objects the input image contains. We use $\mathbf{B}$ to denote the FBE vector, and the length of $\mathbf{B}$ (denote by $L$) should equal to the number of object classes of the corresponding dataset (for instance, in case of VRD dataset we set $L = 100$, since this dataset contains 100 categories of objects). We generate FBE vectors following the rules below:

1. For the subject and object of one relationship triplet, we set the corresponding two elements (based on the class IDs of the subject ($c_s$) and object ($c_o$)) in $\mathbf{B}$ to 1;
2. For every background object, we calculate the distance between the center of its bounding box and the center of the smallest bounding box that covers both of the subject and object. Assuming that we have $n$ background objects $b_1, b_2, ..., b_n$ with class IDs $c_1, c_2, ..., c_n$, and the corresponding distances to the subject and object are $d_1, d_2, ..., d_n$. Without loss of generality, we set $d_1 < d_2 < ... < d_n$.
3. We use $\rho$ to denote the forgetting factor ($0 < \rho < 1$). For $b_1$, we add $\rho$ to the $c_1$th element of $\mathbf{B}$, then for $b_2$, we add $\rho^2$ to the $c_2$th element. We repeat the operation above until all background objects have been processed.

In Fig. 3, we show an example of the generation of $\mathbf{B}$. Even though $\mathbf{B}$ discard most visual information, such as color and texture, of background objects, it still keeps two important pieces of information: the class IDs and the relative distribution of the background objects in the image. Here still leaves one question: is the FBE vector $\mathbf{B}$ unique? This question can be converted into Theorem 1 below:

**Theorem 1 (The uniqueness of FBE vectors)** For two object lists with class ID sequences $C^1 = [c^1_s, c^1_o, c^1_1, c^1_2, ..., c^1_n]$ and $C^2 = [c^2_s, c^2_o, c^2_1, c^2_2, ..., c^2_n]$, we have $\mathbf{B}^1 \neq \mathbf{B}^2$ if $C^1 \neq C^2.

**Proof** Considering the $c^1_m$th element in $\mathbf{B}^1$. Assuming that we have a forgetting factor $\rho$, then we should add $\rho^m$ to $\mathbf{B}^1_{c^1_m}$ (if $m = s$ or $m = o$, we should add 1 to $\mathbf{B}^1_{c^1_m}$). If $c^2_m \neq c^1_m$, we need to consider the inequality below to prove the uniqueness of FBE vectors:

$$\rho^m \neq \rho^n$$
Fig. 3 The process to generate the FBE vector $B$ of a given relationship triplet: (1) feed 1 to corresponding elements of the subject and object (person and motorbike respectively in this image); (2) for the nearest background object to the subject and object, feed the forgetting factor $\rho$ to the corresponding element (helmet in this case); (3) The second nearest background object is person again, so we add $\rho^2$ to the corresponding element of person; (4-6) we add all background elements one-by-one based on the distance from the center of the smallest bounding box that cover both subject and object, and finally we get $B$ of the relationship <person-riding-motorbike>

$$\sum_{i=m+1}^{\infty} \rho^i \geq \rho^m$$  \hspace{1cm} (2)

If the inequality in Eq. 2 is true, then it is possible that the FBE vectors are not unique. We use $S_\infty$ to denote the summation of the geometric series $\rho, \rho^2, \rho^3, \ldots$ ($0 < \rho < 1$), then we have:

$$\sum_{i=m+1}^{\infty} \rho^i = S_\infty - S_m = \frac{\rho}{1 - \rho} - \frac{\rho (1 - \rho^m)}{1 - \rho} = \frac{\rho}{1 - \rho} \cdot \rho^m$$  \hspace{1cm} (3)

(1) If $0 < \rho < 0.5$, then $\rho$ always less then $1 - \rho$, thus we have:

$$\sum_{i=m+1}^{\infty} \rho^i < \rho^m$$  \hspace{1cm} (4)

therefore, in this case, FBE vectors are unique.

(2) If $0.5 \leq \rho < 1$, we can consider the situation that the number of terms (denote by $N$) of the left side in Eq. 2 is finite, and it is easy to show that $\sum_{i=m+1}^{N} \rho^i < \rho^m$ under the condition that $m > N + 1$. Thus FBE vectors are unique in this situation.

Even though we cannot strictly prove the uniqueness of FBE vectors when $0.5 \leq \rho < 1$ and $m < N + 1$, the situations that violate the uniqueness are extremely hard to happen.
Based on the analysis above, the FBE vectors have potential to assist the predicate classification task.

3.2 The FOE-VRD Algorithm

The proposed FOE-VRD model can be divided into 3 parts: 1. several convolution layers and corresponding pooling layers, which serve as feature extractor to generate abstract feature maps from input images; 2. relationship feature extractor, which is a fixed-size encoder to encode all given relationship triplets; 3. several fully-connected layers to do the final predicate classification.

By combining ROI pooling [4] and FBE vectors, we find a suitable way to encode all objects in a given image. Specifically, for the subject and object in a relationship triplet, we use ROI pooling and a fully connected layer to generate their feature vectors (denote by $S$ and $O$ respectively). $S$ and $O$ are then concatenated, and feed into one fully connected layer to generate a feature vector $M$. For the FBE vector $B$, we use the method described in Sect. 3.1 to do the feature generation. Then $M$ and $B$ are concatenated as the feature vector (denote by $R$) of the given relationship triplet. By using fully-connected layers to further extract features from $R$, we finally get the results of predicate classification.

Figure 4 describes the overall structure of the proposed FOE-VRD algorithm.

4 Experimental Results

In this section, we firstly present some important information of our experiments, which include database selection, experimental configurations and implementation details. Then we show the performance of FOE-VRD algorithm on predicate classification and relationship detection tasks. Moreover, we select many state-of-the-art relationship detection algorithms as baselines to compare with the proposed method.

4.1 Database

We perform our experiments on Standford Visual Relationship Detection [15] (VRD) and Visual Genome [9] (VG) databases.
Fig. 5 The detailed network structure of FOE-VRD algorithm we used in our experiments (for VRD database)

VRD is an important database for visual relationship detection, which applied by many related works [2, 7, 15, 17, 24]. This database has 4000 training images and 1000 test images respectively. VRD database has 100 categories of objects and 70 categories of predicates, and the total number of the relationship triplets is 37993 (7701 types of triplets). Moreover,
there are 1169 relationships (1029 types of triplets) only appear in the test set, thus we can use them to perform zero-shot tests.

Comparing with VRD, VG is a more changeable database. The original VG databases contains 108,077 images, 18,136 categories of objects and 13,894 unique relationships [9]. However, there are many kinds of objects and predicates that only appear few times, which may bring about negative impact on network training. Therefore, following [20], we only keep the most frequent 150 object categories and 50 predicates in our experiments. We randomly select 70% images as training set and 30% as test set. Also, we use the 7376 relationships (7220 types of triplets) that only appear in our test set as the zero-shot test set.

4.2 Performance Measurements

We consider two tasks to evaluate the proposed FOE-VRD algorithm, i.e., predicate classification and relationship detection.

4.2.1 Predicate Classification

In case of predicate classification, the inputs of FOE-VRD include raw images, bounding boxes of objects in all images, and class IDs of all objects. Our algorithm predicts predicates between every pair of objects. This task shows the performance of FOE-VRD on the basis that the class IDs and locations of all objects are known.

4.2.2 Relationship Detection

In case of relationship detection, the inputs of FOE-VRD only include raw images. We firstly use Faster RCNN algorithm [18] to do object detection on input images, then predict predicates between all pairs of detected objects by using FOE-VRD algorithm. The outputs are several relationship triplets <subject-predicate-object>. Comparing with predicate classification, relationship detection is more closed to real world applications, since in practice we cannot always have labeled images.

For both tasks, we use Recall@100 as the performance measurement. Recall@X is widely used in visual relationship detection researches [11, 15, 21, 25]. To calculate Recall@X, we need firstly introduce a hyperparameter $k$ [22]. The meaning of $k$ is that for each pair of subject and object, we take top $k$ predicted predicates into account for the calculation Recall@X. Assuming that a given image has $n$ pairs of subject and object, we rank $n \cdot k$ relationship predictions based on probability and take the top $X$ to compute the average fraction of correct relationships among them [15], which is the Recall@X of the image. In our experiments, we set $k = 70$ for VRD database and $k = 50$ for VG database, which equals to predicate categories of two databases respectively. This setting implies that we consider all relationship predictions to compute the Recall@X.

4.3 Implementation Details

We implement our FOE-VRD algorithm on TensorFlow 1.15 [1] platform. The network structure of FOE-VRD bases on VGG-16 network [19] (see Fig. 5) for more details). At the ROI pooling layer, we convert the regions of the subject and object to two vectors $S$ and $O$ with length 4096. $S$ and $O$ are then concatenated, and feed into one fully connected layer. The
Table 1: The comparison of performances of predicate classification task with different settings of $\rho$

| $\rho$ | VRD   | VG    |
|-------|-------|-------|
| 0.3   | 85.48 | 93.95 |
| 0.5   | 87.12 | 93.78 |
| 0.7   | 90.46 | 96.64 |
| 0.9   | 93.10 | 97.87 |

output of this fully connected layer is a feature vector $\mathbf{M}$ with length 1024. For each image, we model all background objects to one FBE vector $\mathbf{B}$ with length 100 for VRD database and 150 for VG database, and use 0.9 as the forgetting factor. $\mathbf{M}$ and $\mathbf{B}$ are then combined to one 1124 dimension feature vector $\mathbf{R}$. $\mathbf{R}$ is then fed into three fully connected layers to calculate the final predicate classification results.

In training phase, we use the ImageNet [3] pre-trained model to initialize most of model parameters, and fine-tune the model using the training set of the VRD dataset. The training data include raw images and ground truth bounding boxes along with their class IDs. During the training process, we use 0.005 as the learning rate, 0.9 momentum, and 0.0005 weight decay. We train the network for 50000 iterations, and in every iteration we only input 1 image into the model. In test phase, we input test images and the corresponding ground truth bounding boxes and their class IDs to the FOE-VRD for predicate classification. For relationship detection, the bounding boxes and the corresponding class IDs are generated by Faster RCNN algorithm [4]. Our computation platform includes Intel Core i7 9700k CPU, 64 GB memory, and Nvidia Titan RTX GPU.

4.4 The Selection of Forgetting Factor

In Theorem. 1, we show how to get FBE vectors to achieve the uniqueness requirement. Even though a less than 0.5 forgetting factor $\rho$ corresponding to a unique FBE vector, one problem is that a too small $\rho$ may result in a fast decay sequence of background encoding, which lead to a sparse FBE vector that may harm the performance. To prove this argument, we perform experiments of predicate classification on VRD and VG dataset, and the results are presented in Table 1. Via Table 1 we can learn that in practice a larger $\rho$ may result in better performance.

4.5 Experimental Results

In this part, we present experimental results of the proposed FOE-VRD algorithm on the test set of VRD and VG dataset. We compare FOE-VRD with several state-of-the-art baselines in case of predicate classification and relationship detection. Table 2 and Table 3 show the experimental results of predicate classification and relationship detection (include entire test set and zero-shot subset) respectively. Notice that to show the importance of the proposed FBE feature, we remove FBE from the FOE-VRD model and use it as another baseline. Moreover, to prove the indispensable of background information in FBE vectors $\mathbf{B}$, we add another group experimental results, in which we only keep the two elements that corresponding to the subject and object of the relationship in $\mathbf{B}$, and all elements of background objects are set to 0.

From the experimental results we can learn that the introducing of FBE feature vector obviously improves the performance, and the background information also has irreplaceable
value. Also, the proposed FOE-VRD algorithm has comparable or even better performance comparing with the state-of-the-art methods on experiments of entire test set. For zero-shot subset, FOE-VRD achieves state-of-the-art performance, which is obviously better than previous counterparts. One potential reason of this fact is that the introducing of the FBE vector helps the algorithm to classify unseen relationships, since the same predicates may corresponding to similar FBE vectors.

**5 Conclusion**

In this paper, we propose a novel visual relationship detection algorithm named FOE-VRD. Motivated by the consideration that the category of predicate of a given relationship triplet

---

**Table 2** Experimental results of predicate classification task on the VRD and VG test set, which include the results on entire test set and the zero-shot test set

| Algorithm          | VRD  | VRD Zero-Shot | VG   | VG Zero-Shot |
|--------------------|------|---------------|------|--------------|
| VRD [15]           | 84.34| 50.04         | N/A  | N/A          |
| LKD [22]           | 86.97| 74.65         | 95.68| 88.23        |
| DSR [11]           | 93.18| 79.61         | N/A  | N/A          |
| Zoom-Net [21]      | 90.59| N/A           | 77.51| N/A          |
| NLG [12]           | 92.73| 90.52         | 91.77| N/A          |
| LSV [8]            | 95.18| 83.49         | **99.37**| 95.72        |
| HGT [16]           | 97.02| 89.79         | 96.65| N/A          |
| FOE-VRD (no FBE)   | 87.04| 82.83         | 94.21| 81.70        |
| FOE-VRD (no background) | 87.64| 90.28         | 93.99| 92.32        |
| FOE-VRD            | 93.10| **93.29**     | 97.87| **96.82**    |

The measurement metric is Recall@100 ($k = 70$ for VRD, and $k = 50$ for VG)

The bold values indicate the best performance over the corresponding columns

**Table 3** Experimental results of relationship detection task on the VRD and VG test set, which include the results on entire test set and the zero-shot subset

| Algorithm          | VRD  | VRD Zero-Shot | VG   | VG Zero-Shot |
|--------------------|------|---------------|------|--------------|
| VRD [15]           | 21.51| 11.70         | N/A  | N/A          |
| LKD [22]           | **31.89**| 15.89     | N/A  | N/A          |
| DSR [11]           | 23.29| 9.20          | N/A  | N/A          |
| DSL [25]           | 18.26| N/A           | 7.50 | N/A          |
| Zoom-Net [21]      | 27.30| N/A           | 25.07| N/A          |
| NLG [12]           | 21.97| 22.03         | 22.17| N/A          |
| LSV [8]            | 20.54| 12.14         | N/A  | N/A          |
| HGT [16]           | 27.73| N/A           | N/A  | N/A          |
| FOE-VRD (no FBE)   | 26.46| 23.54         | 24.57| 19.82        |
| FOE-VRD (no background) | 25.37| 23.55         | 24.41| 20.07        |
| FOE-VRD            | 28.19| **25.91**     | **28.09**| **23.98**    |

The measurement metric is Recall@100 ($k = 70$ for VRD, and $k = 50$ for VG)

The bold values indicate the best performance over the corresponding columns.
may not only related to the subject and object, but also background objects in the image, we introduce Fixed-size Background Encoding (FBE) method, which successfully use a fixed-size vector (FBE vector) to encode all background objects in a image, and the length of the FBE vector only related to the number of object categories. Moreover, ROI Pooling method is also applied to model subjects and objects fixed-sizely. The fixed-size feature vectors make it possible to use fully-connected neural networks to do predicate classification. The experimental results on VRD and VG dataset show that FOE-VRD works well on both predicate classification and relationship detection tasks. Moreover, FOE-VRD has good performance on zero-shot cases, which implies that it has good ability to deal with wide variety of real-world visual relationship detection tasks.

Acknowledgements This paper is supported by the National Key Research and Development Program of China (Grant No. 2018YFB0204300).

References

1. Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, Corrado GS, Davis A, Dean J, Devin M, Ghemawat S, Goodfellow I, Harp A, Irving G, Isard M, Jia Y, Jozefowicz R, Kaiser L, Kudlur M, Levenberg J, Mané D, Monga R, Moore S, Murray D, Olah C, Schuster M, Shlens J, Steiner B, Sutskever I, Talwar K, Tucker P, Vanhoucke V, Vasudevan V, Viégas F, Vinyals O, Warden P, Wattenberg M, Wicke M, Yu Y, Zheng X (2015) TensorFlow: Large-scale machine learning on heterogeneous systems. https://www.tensorflow.org/, software available from tensorflow.org
2. Bin Y, Yang Y, Tao C, Huang Z, Li J, Shen HT (2019) Mr-net: Exploiting mutual relation for visual relationship detection. Proceed AAAI Conf Artificial Intell 33:8110–8117
3. Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L (2009) Imagenet: a large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition, Ieee, pp 248–255
4. Girshick R (2015) Fast r-cnn. In: Proceedings of the IEEE international conference on computer vision, pp 1440–1448
5. He K, Zhang X, Ren S, Sun J (2015) Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Trans Pattern Anal Mach Intell 37(9):1904–1916
6. He K, Gkioxari G, Dollár P, Girshick R (2017) Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision, pp 2961–2969
7. Jung J, Park J (2018) Visual relationship detection with language prior and softmax. In: 2018 IEEE international conference on image processing, applications and systems (IPAS), IEEE, pp 143–148
8. Jung J, Park J (2019) Improving visual relationship detection using linguistic and spatial cues. ETRI J
9. Krishna R, Zhu Y, Groth O, Johnson J, Hata K, Kravitz J, Chen S, Kalantidis Y, Li LJ, Shamma DA, Bernstein M, Fei-Fei L (2016) Visual genome: Connecting language and vision using crowdsourced dense image annotations. arXiv:1602.07332
10. Kukleva A, Tapaswi M, Laptev I (2020) Learning interactions and relationships between movie characters. In: IEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)
11. Li, K, Guo Y, Chang H, Chen X (2018) Visual relationship detection with deep structural ranking. In: Thirty-Second AAAI Conference on Artificial Intelligence
12. Liao W, Rosenhahn B, Shuai L, Ying Yang M (2019) Natural language guided visual relationship detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops
13. Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P, Zitnick CL (2014) Microsoft coco: Common objects in context. In: European conference on computer vision, Springer, pp 740–755
14. Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu CY, Berg AC (2016) Ssd: Single shot multibox detector. In: European conference on computer vision, Springer, pp 21–37
15. Lu C, Krishna R, Bernstein M, Fei-Fei L (2016) Visual relationship detection with language priors. In: European Conference on Computer Vision
16. Mi L, Chen Z (2020) Hierarchical graph attention network for visual relationship detection. In: Proceedings of the IEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)
17. Mizzi N, Muscat A (2019) Optimising the input image to improve visual relationship detection. arXiv preprint arXiv:1903.11029
18. Ren S, He K, Girshick R, Sun J (2015) Faster r-cnn: Towards real-time object detection with region proposal networks. In: Advances in neural information processing systems, pp 91–99
19. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556

20. Xu D, Zhu Y, Choy CB, Fei-Fei L (2017) Scene graph generation by iterative message passing. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 3097–3106, https://doi.org/10.1109/CVPR.2017.330

21. Yin G, Sheng L, Liu B, Yu N, Wang X, Shao J, Change Loy C (2018) Zoom-net: mining deep feature interactions for visual relationship recognition. In: Proceedings of the European Conference on Computer Vision (ECCV), pp 322–338

22. Yu R, Li A, Morariu VI, Davis LS (2017) Visual relationship detection with internal and external linguistic knowledge distillation. In: Proceedings of the IEEE international conference on computer vision, pp 1974–1982

23. Zhang S, Jiang H, Xu M, Hou J, Dai LR (2015) The fixed-size ordinally-forgetting encoding method for neural network language models. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), pp 495–500

24. Zhou H, Zhang C, Hu C (2019) Visual relationship detection with relative location mining. In: Proceedings of the 27th ACM International Conference on Multimedia, pp 30–38

25. Zhu Y, Jiang S (2018) Deep structured learning for visual relationship detection. In: Thirty-Second AAAI Conference on Artificial Intelligence

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.