On Exploring Undetermined Relationships for Visual Relationship Detection

Yibing Zhan , Jun Yu , Ting Yu , and Dacheng Tao

6. Supplement Materials

6.1. MF-URLN-IM

To tackle the problem of zero-shot learning, we propose a multi-modal feature based undetermined relationship learning network with inferring model (MF-URLN-IM). The inferring model is inspired by humans' natural gift for inference wherein a person is able to predict the relationship between two objects from partial information obtained from learned object pairs. This process is illuminated in Fig. 1. Therefore, when encountering unseen relationships, MF-URLN-IM still performs robustly, according to the information obtained from the individual subjects and objects.

Specifically, MF-URLN-IM has three separate types of relationship learning networks: a union relationship learning network, a subject relationship learning network, and a object relationship learning network. All relationship learning networks share the same architecture of MF-URLN, except they have different input features. The union relationship learning network includes all of the features. The subject relationship learning network includes the subjects' visual features, subjects' external linguistic features, and spatial features. The object relationship learning network includes the objects' visual features, objects' external linguistic features, and spatial features. The visual features of union boxes and internal linguistic features are not used in the subject and object relationship learning network because these two features contain both subject and object information. A joint loss function is used to simultaneously train the three relationship learning networks. The joint loss function is defined as:

$$L = L_{sub+obj} + L_{sub} + L_{obj}, \qquad (1)$$

where $L_{sub+obj}$, L_{sub}, and L_{obj} represents the loss functions for the union, subject, and object relationship learning network, respectively.

By using this joint loss function, the three relationship learning networks can share the same parameters as in previous modules. The final relationship is predicted by calculating the geometric average of the predictions from the three networks. This is calculated as:

Table 1. Performance comparison on the zero-shot set of the VRD dataset.

	Pre.	Phr.		Rel.	
	$R_{50/100}$	R_{50}	R_{100}	R_{50}	R_{100}
MF-URLN	26.9	5.9	7.9	4.3	5.5
MF-URLN-IM	**27.2**	**6.2**	**9.2**	**4.5**	**6.4**

$$P(R) = P(R|u) \cdot P(R|s) \cdot P(R|o). \qquad (2)$$

where $P(R|u)$, $P(R|s)$, and $P(R|o)$ represents the relationship probabilities of the union, subject, and object relationship learning network, respectively.

Table 1 compares performances of MF-URLN and MF-URLN-IM. As shown, MF-URLN-IM outperforms MF-URLN in all tasks. This results reveal the potential usefulness of the inferring model for visual relationship detection.

6.2. More Discussion of Undetermined Relationships

In this subsection, more qualitative results are provided. Fig. 2 provides two examples of top-5 object pairs detected by Faster R-CNN and Faster R-CNN-DC. Faster R-CNN-DC refers to the method, which uses Faster R-CNN to detect objects and uses determinate confidence subnetwork to produce determinate confidence scores for object pairs. In Faster R-CNN, object pairs are ranked by the product of subject boxes' and object boxes' probabilities. In Faster R-CNN-DC, object pairs are ranked by the product of subject boxes', object boxes', and determinate confidence probabilities. As shown in Fig. 2, Faster R-CNN-DC outperforms Faster R-CNN in both examples. By adding determinate confidence subnetwrok, the object pairs with determinate relationships are highlighted. These highlighted determinate relationships results in better performance of visual relationship detection. Since determinate confidence subnetwork is trained based on undetermined relationships, the advantage of determinate confidence subnetwork again

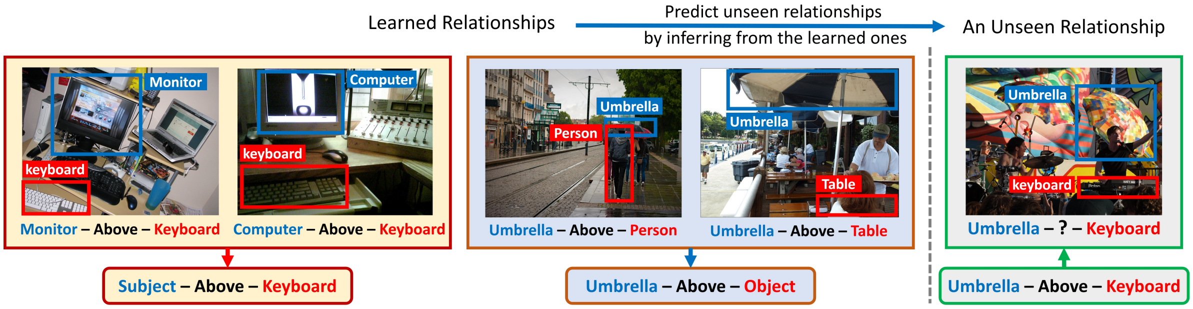

Learned Relationships
Predict unseen relationships by inferring from the learned ones
An Unseen Relationship
Monitor
keyboard
Computer
keyboard
Monitor – Above – Keyboard
Computer – Above – Keyboard
Umbrella
Person
Umbrella
Table
Umbrella – Above – Person
Umbrella – Above – Table
Umbrella
keyboard
Umbrella – ? – Keyboard
Subject – Above – Keyboard
Umbrella – Above – Object
Umbrella – Above – Keyboard

Figure 1. The process of predicting an unseen relationship by inferring from the learned ones. The unseen relationship can be predicted by combining information from the relationships that only contain one of the subject and the object.

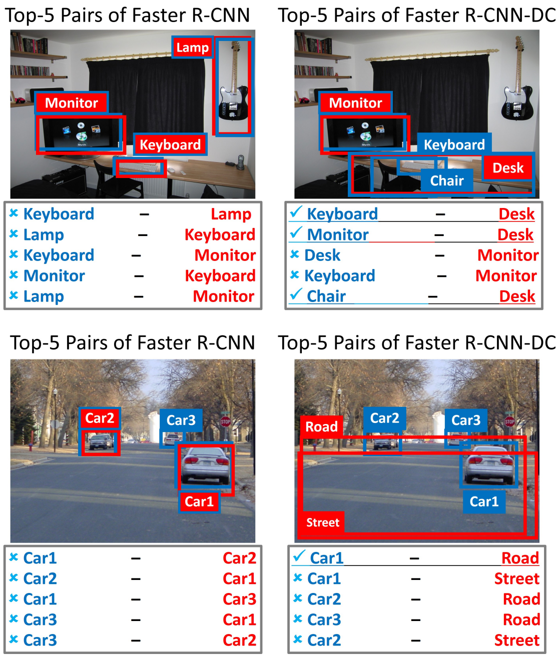
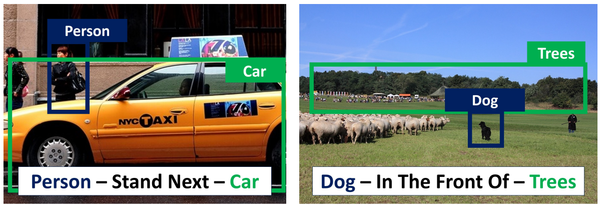


Figure 2. The top-5 detected object pairs of Faster R-CNN and Faster R-CNN with determinate confidence scores (Faster R-CNN-DC). In Faster R-CNN, object pairs are ranked by the product of subject boxes' and object boxes' probabilities. In Faster R-CNN-DC, object pairs are ranked by the product of subject boxes', object boxes', and determinate confidence probabilities. The $\surd$ represents the manual-labeled object pairs.



(a) The failed case of predicate detection.

(b) The failed case of relation detection.

Figure 3. Two failed cases of MF-URLN. (a) The failed case of predicate detection. The predicates of both MF-URLN and MFLN are Person-Sit On-Car. (b) The failed case of relation detection. Both of MF-URLN and MFLN correctly predict the predicates. In MFLN, the relationship dog-in the front of-trees is the No.11 recall of relation detection. In MF-URLN, the relationship is No.233 recall of relation detection.

confirms the necessity and usefulness of undetermined relationships in visual relationship detection. In addition, in the upper example of Faster R-CNN of Fig. 2, the guitar is falsely detected as a lamp by the Faster R-CNN. Such mistake negatively influences the performance of a visual relationship detection method. Contrarily, in the example of Faster R-CNN-DC, we observe that the object pairs that contain falsely detected objects are ignored. This is because the object pairs with falsely detected objects are labeled as undetermined relationships. Using undetermined relationships in visual relationship detection alleviate the problem of falsely detected objects to some extent.

Fig. 3 presents two failed cases of MF-URLN. Fig. 3 (a) is a failed case of predicate detection. The detected predicate of both MF-URLN and MFLN for the given person and car is "sit on". This failure is caused because legs of the person are obscured by the car and it is difficult for MF-URLN and MFLN to identify the posture of the person. Fig. 3 (b) provides a failed case of relation detection. Both of MF-URLN and MFLN predict correct predicates between the dog and the trees. In MFLN, the relationship dog-in the front of-trees is in the top-50 recall of relation detection, whereas in MF-URLN, the relationship is not in the top-50. This is because the relationship dog-in the front of-trees has low probability score of determinate confidence. The failure of relation detection indicate that better strategies to generate and utilize undetermined relationships are still necessary.