Detecting Visual Relationships with Networks

Hongzhi Liu

May 31, 2018

Abstract

In daily life, we often see kinds of complicated pictures and photos taken by other people, which we may not understand meanings even relationships behind them. Today, I read a thesis written by Bo Dai, who studys at Department of Information Engineering in the Chinese University of Hong Kong. His team propose an integrated framework to reason about the relationships among individual objects which achieves substantial improvement over state-of-the-art on two large data sets.

1. Overview of Deep Relational Network

Images in the real world often involve multiple objects that interact with each other. Relationships among objects play a crucial role in image understanding. To understand such images, being able to recognize individual objects is generally not sufficient. Thanks to the great success and advances of deep learning techniques in recognizing individual objects, the past several years witness remarkable progress in several key tasks in computer vision such as attribute detection [5]. However, visual relationship detection remains a very difficult task.

In the paper, Bo Dai and his team develop a new framework to tackle the problem of *visual relationship detection* [1]. This framework formulates the prediction output as a triplet in the form of *(subject, predicate, object)*, and jointly infers their class labels by exploiting two kinds of relations among them, namely spatial configuration and statistical dependency. Such relations are ubiquitous, informative, and more importantly they are often more reliable than visual appearance.

This new way of formulation also allows the model parameters to be learned in a discriminative fashion, using the latest techniques in deep learning. On two large datasets, the proposed framework outperforms not only the classification-based methods but also the CRFs based on deep potentials.

2. Deep Relational Network

Visual relationships play a crucial role in image understanding. Whereas a relationship may involve multiple parties in general, many important relationships occur between exactly two objects. In this thesis, Dai focuses on such relationships. In particular, the team follow a widely adopted convention [4] and characterize each visual relationship by a triplet in the form of *(s, r, o)*. Here, *s*, *r*, and *o* respectively denote the subject category, the relationship predicate, and the object category. The task is to locate all visual relationships from a given image, and infer the triplets.

2.1. Visual Relationship Detection

From the paper, I can know that the team adopt the paradigm that is to recognize each component individually for relationship detection and aim to take its performance to a next level. Particularly, the team focus on developing a new method that can effectively capture the rich relations among the three components in a triplet and exploit them to improve the prediction accuracy. As shown in Fig. 1, the overall pipeline of their framework comprises three stages, as described below.

(1) **Object detection**. Given an image, we use an object detector to locate a set of candidate objects. In this work, the team use Faster RCNN [3] for this purpose. Each candidate object comes with a bounding box and an appearance feature, which will be used in the joint recognition stage for predicting the object category.

(2) **Pair filtering**. The next step is to produce a set of *object pairs* from the detected objects. With n detected objects, they can form $n(n-1)$ pairs. They found that a considerable portion of these pairs are obviously meaningless and it is unlikely to recognize important relationships therefrom. Hence, Dai introduces a low-cost neural network to filter out such pairs, so as to reduce the computational cost of the next stage. This filter takes into account both the spatial configurations and object categories.

(3) **Joint recognition**. Each retained pair of objects will be fed to the *joint recognition* module. Taking into account multiple factors and their relations, this module will produce a triplet as the output.

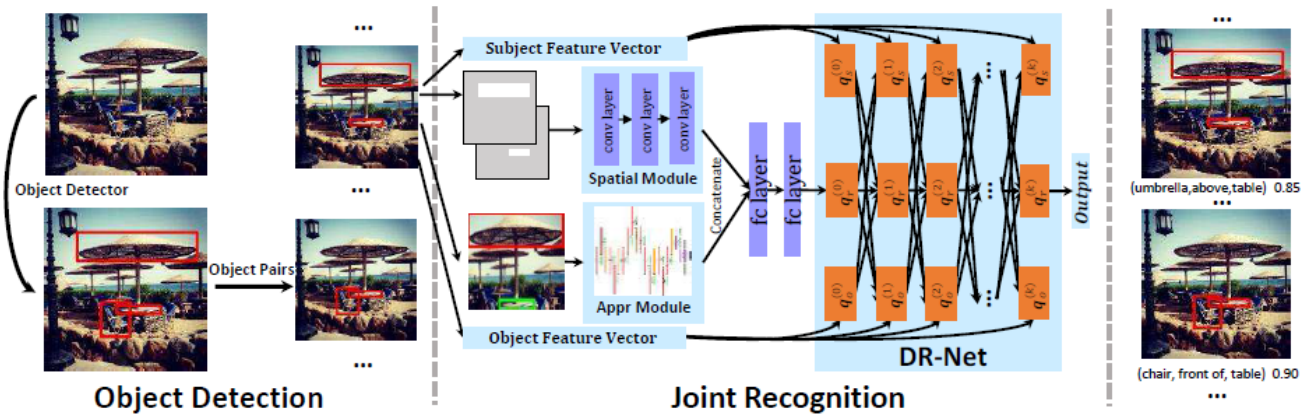

Figure 1. The proposed framework for visual relationship detection. Given an image, it first employs an object detector to locate individual objects. Each object also comes with an appearance feature. For each pair of objects, the corresponding local regions and the spatial masks will be extracted, which, together with the appearance features of individual objects, will be fed to the DR-Net. The DR-Net will jointly analyze all aspects and output q_s, q_r, and q_o, the predicted category probabilities for each component of the triplet. Finally, the triplet *(s, r, o)* will be derived by choosing the most probable categories for each component.

In joint recognition, multiple factors are taken into consideration. These factors are presented in detail below. First, each detected object comes with an appearance feature, which can be used to infer its category. In addition, the type of the relationship may also be reflected in an image visually. To utilize this information, the team extract an appearance feature for each candidate pair of objects, by applying a CNN [2] to an *enclosing box*, *i.e.* a bounding box that encompasses both objects with a small margin. Besides, the relationship between two objects is also reflected by the spatial configurations between them, *e.g.* their relative positions and relative sizes. Such cues are complementary to the appearance of individual objects, and resilient to photometric variations.

Furthermore, In a triplet *(s, r, o)*, there exist strong statistical dependencies between the relationship predicate *r* and the object categories *s* and *o*. Finally, we describe how these factors are actually combined as integrated prediction. The framework produces the prediction by choosing the most probable classes for each of these components. In the training, all stages in their framework, namely object detection, pair filtering and joint recognition are trained respectively. As for joint recognition, different factors will be integrated into a single network and jointly fine-tuned to maximize the joint probability of the ground-truth triplets.

2.2. Deep Relational Network

The *Conditional Random Field* (CRF) is a classical formulation to incorporate statistical relations into a discriminative task. Specifically, for the task of recognizing visual relationships, the CRF can be formulated as Equation 1:

$$p(r, s, o|x_r, x_s, x_o) = \frac{1}{Z} \exp(\Phi(r, s, o|x_r, x_s, x_o; W)). \quad (1)$$

Here, x_r is the *compressed pair feature* that combines both the appearance of the enclosing box and the spatial configurations; x_s and x_o are the appearance features respectively for the subject and the object; W denotes the model parameters; and Z is the normalizing constant, whose value depends on the parametersW. The joint potential Φ can be expressed as a sum of individual potentials as Equation 2:

$$\begin{aligned}\Phi = &\psi_a(s|x_s; W_a) + \psi_a(o|x_o; W_a) + \psi_r(r|x_r; W_r) \\ &+ \phi_{rs}(r, s|W_{rs}) + \phi_{ro}(r, o|W_{ro}) + \phi_{so}(s|W_{so})\end{aligned} \quad (2)$$

Here, the unary potential ψ_a associates individual objects with their appearance; ψ_r associates the relationship predicate with the feature x_r; while the binary potentials ϕ_{rs}, ϕ_{ro} and ϕ_{so} capture the statistical relations among the relationship predicate *r*, the subject category *s*, and the object category *o*.

3. Experimental Results of Framework

The team demonstrates the effectiveness of the framework in experiments.We tested our model on two datasets: (1) VRD: the dataset contains 5, 000 images and 37, 993 visual relationship instances that belong to 6, 672 triplet types. (2) sVG: a substantially larger subset constructed from Visual Genome. sVG contains 108K images and 998K relationship instances that belong to 74, 361 triplet types. All instances are randomly partitioned into disjoint training and

Table 1. Comparison of different variants of the proposed method, using *Recall*@50 as the metric.

		A_1	A_2	S	A_1S	A_1SC	A_1SD	A_2SD	A_2SDF
VRD	Predicate Recognition	63.39	65.93	64.72	71.81	72.77	80.66	80.78	-
	Union Box Detection	12.01	12.56	13.76	16.04	16.37	18.15	19.02	19.93
	Two Boxes Detection	10.71	11.22	12.16	14.38	14.66	16.12	16.94	17.73
sVG	Predicate Recognition	72.13	72.54	75.18	79.10	79.18	88.00	88.26	-
	Union Box Detection	13.24	13.84	14.01	16.04	16.08	20.21	20.28	23.95
	Two Boxes Detection	11.35	11.98	12.07	13.77	13.81	17.42	17.51	20.79

testing sets, which respectively contain 799K and 199K instances.

In Table 1, Dai compares A_1, A_2, S, A_1S, A_1SC, A_1SD, A_2SD and A_2SDF. These results show the benefit of exploiting statistical dependencies for joint recognition.

References

[1] B. Dai, Y. Zhang, and D. Lin. Detecting visual relationships with deep relational networks. In *CVPR*, 2017. 1

[2] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 2

[3] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: towards real-time object detection with region proposal networks. In *NIPS*, 2015. 1

[4] M. A. Sadeghi and A. Farhadi. Recognition using visual phrases. In *CVPR*, 2011. 1

[5] N. Zhang, M. Paluri, M. Ranzato, T. Darrell, and L. Bourdev. Panda: Pose aligned networks for deep attribute modeling. In *CVPR*, 2014. 1