Human-centric Visual Relation Segmentation Using Mask R-CNN and VTransE

Fan Yu, Xin Tan, Tongwei Ren, and Gangshan Wu

State Key Laboratory for Novel Software Technology, Nanjing University, China
{yf, tx}@smail.nju.edu.cn, {rentw, gswu}@nju.edu.cn

Abstract. In this paper, we propose a novel human-centric visual relation segmentation method based on Mask R-CNN model and VTransE model. We first retain the Mask R-CNN model, and segment both human and object instances. Because Mask R-CNN may omit some human instances in instance segmentation, we further detect the omitted faces and extend them to localize the corresponding human instances. Finally, we retrain the last layer of VTransE model, and detect the visual relations between each pair of human instance and human/object instance. The experimental results show that our method obtains 0.4799, 0.4069, and 0.2681 on the criteria of R@100 with the m-IoU of 0.25, 0.50 and 0.75, respectively, which outperforms other methods in Person in Context Challenge.

Keywords: Human-centric, visual relation segmentation, Mask R-CNN, VTransE

1 Introduction

Visual relation detection attracts much attention from both academic and industry recently because it provides more comprehensive visual content understanding beyond objects. A visual relation involves a relation triplet represented by <subject, predicate, object> together with two bounding boxes to localize the subject and the object. Researchers have made efforts to analyse visual relations in both images [7, 4, 8] and videos [9]. It can effectively support various computer vision applications, such as image/video captioning [10], visual question answering [11] and visual search [12].

Current research aims to extract all the visual relations between each pair of subject and object. Though it may benefit to some applications, such as visual content indexing, it has obvious limitation in emphasizing the main content of images and videos. Many visual relations between an arbitrary pair of subject and object are usually meaningless to represent visual content. For example, to two distant objects, they only have spatial relation and the relation is omitted in captioning or visual question answering. Instead, a viewer concerns the human-centric visual relations, *i.e.*, the subjects in the triplets of these relations are *human instances*. Moreover, human-centric relations involves more predicates as

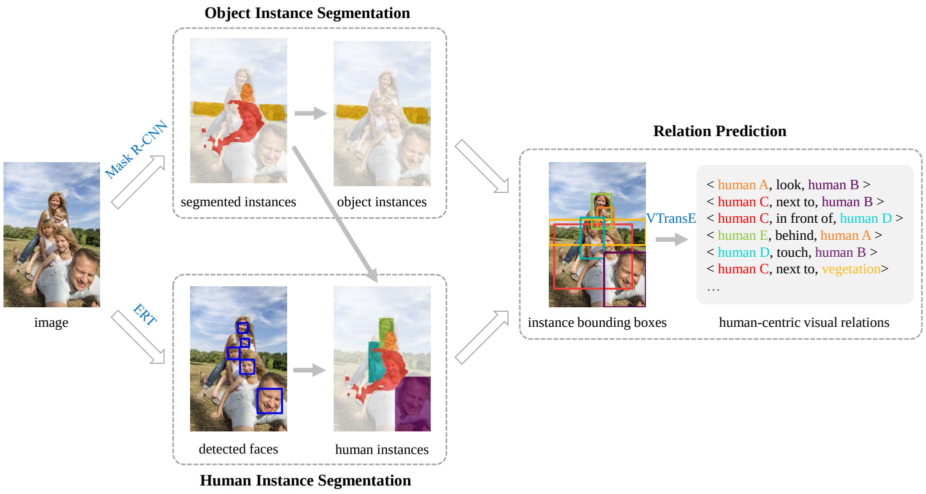
Object Instance Segmentation
segmented instances
object instances
Mask R-CNN
image
ERT
detected faces
human instances
Human Instance Segmentation
Relation Prediction
VTransE
instance bounding boxes
< human A, look, human B >
< human C, next to, human B >
< human C, in front of, human D >
< human E, behind, human A >
< human D, touch, human B >
< human C, next to, vegetation>
…
human-centric visual relations

Fig. 1. An overview of the proposed method.

compared to the visual relations between two object, because a person may perform different actions in the interactions with objects. It means human-centric visual relation detection is more meaningful and challenging than traditional visual object detection. Based the above observation, Person in Context (PIC) Challenge aims to promote the progress of human-centric visual relation detection. Beyond human-centric visual relation detection, PIC Challenge focuses on *human-centric visual relation segmentation*, *i.e.*, the subject and the object of each visual relation should be localized with masks instead of bounding boxes.

In this paper, we propose a novel human-centric visual relation segmentation method using Mask R-CNN model and VTransE model. These two models are the state-of-the-art in instance segmentation and visual relation detection, respectively. Figure 1 shows an overview of our method. We first retain the last layer of Mask R-CNN model, fine-tune its parameters and utilize it for instance segmentation on an input image, which can generate the object instances. Then, we detect face using ensemble of regression trees method (ERT), and fuze the human instance segmented by Mask R-CNN model and the ones localized based on face detection. After this, we retrain the last layer of VTransE model and apply it on the bounding boxes of human instances and object instances to detect human-centric visual relations. Finally, we score the visual relations according to their confidences and relation prior, and return the top 100 visual relations as our result.

2 Preliminary

2.1 Mask R-CNN

Mask R-CNN is a two-stage framework [1]. The first stage, which is called Region Proposal Network(RPN), proposes candidate object bounding boxes. And the second stage performs classification and bounding-box regression, a binary mask for each RoI is generated as well. Mask R-CNN extends Faster R-CNN [2] by adding a branch that takes the positive RoIs and generates masks for them in parallel with the existing branch. Mask R-CNN only adds an overhead to Faster R-CNN and reaches the speed of 5 fps. Also, it is easily to be applied to other tasks like human poses estimation. Mask R-CNN.

Mask R-CNN has three technical essentials. Firstly, Mask R-CNN uses ResNeXt-101 and FPN as a feature extraction network and shows better result than other models. Secondly, Mask R-CNN improves the pooling net by using ROIAlign instead of ROIPooling, and solves the misalignment resulting from direct sampling by pooling net. Experiment shows that the bigger the stride, the more obvious the improvement. Thirdly, the special loss function designed by Mask R-CNN has a better result than softmax.

2.2 Face detection by ERT

Many methods have been proposed to handle face detection and face alignment. As one of the state-of-the-arts, ERT method uses the ensemble of regression trees to estimate the faces landmark positions and proposed a framework based on gradient boosting [3]. The method uses a cascade of regression functions and each regression function efficiently estimates the shape. During the learning period, ERT save the value of shape into leaf node. After learning the tree, the initial landmark positions will be gradually improved by adding all the updated shape value. Each of the regressors is composed of many trees, and the parameters of each tree are generated after training the model using the coordinate differences of current shape and ground truth and the pixel pairs which are randomly chose.

The boosting algorithm applied in this method has a good performance in classification and regression by reducing the sum of square error of initial shape and ground truth. The method performs face alignment in one millisecond on average. Also, the missing and uncertain labels are handled. The dlib library has implemented this algorithm.

2.3 VTransE

Many research works are afforded to visual relation detection in images, and most of them focus on predicting the huge number of relations by learning from few training data. Lots of methods have been proposed to reduce the complexity from O(N^2R) to O(N+R) where N and R are the numbers of objects and predicates respectively.

A Visual Translational Embedding network (VTransE) [4] has been proposed for visual relation detection. This model predicts relations from an image in an end-to-end fashion and refers to a visual relation as a subject-predicate-object triplet. VTransE proposes to model visual relations by mapping the features of objects and predicates in a low-dimensional space, which greatly reduces the volume of data to process. Simultaneously, VTransE incorporates knowledge transfer between objects and predicates. VTransE has been demonstrate its effectiveness on two datasets: Visual Relationship and Visual Genome and performs great.

3 Our method

3.1 Object instance segmentation

We first conduct instance segmentation on an input image using Mask R-CNN model implemented with ResNet-101 network, which was pre-trained on MS COCO dataset. Considering the detected object categories in our task is different to that on MS COCO dataset, we retrain the last layer of Mask R-CNN model with the following loss function:

$$L = L_{class} + L_{box} + L_{mask}, \tag{1}$$

where the classification loss L_{class} and bounding-box loss L_{box} are defined as those in [5]; the mask loss L_{mask} is defined as that in [1].

We use a mini-batch size of 2 images on 1 GPU, starting from a learning rate of 0.001 to train the network. A weight decay of 0.0001 is used as well as a momentum of 0.9. All layers are fine-tuned using stochastic gradient descent for 500K iterations with a mini-batch size of 2, and a learning rate of 0.0001. The total training time is approximately one day on a 1080Ti GPU.

To the instance segmentation result, we retain the segmented instances belong to object but not human as our object instance segmentation result.

3.2 Human instance segmentation

Though the instance segmentation result generated by Mask R-CNN model includes human instance, some human instances may be omitted in instance segmentation. It leads to serious decrease on our performance because a visual relation cannot be detected if the corresponding human instance to its subject is omitted. In order to address this problem, we use ERT method to detect the faces on a given image. To each detected face, if its bounding box has not been covered by the bounding box of a segmented human instance largely, we treat it belonging to an omitted human instance. In our experiments, the coverage threshold for omitted face filtering is 0.8.

According to the location of human faces and the common sense about the proportion of human body, we estimate the location of the whole human roughly. On average, the height of persons' bounding boxes in image equals three times the height of their heads. Considering that the mask exporting from Mask R-CNN is more accurate and should have a higher priority, we add the expanding of face detection as a supplementary result. The final result of mask will be the result given by Mask R-CNN and the area covered by the bounding boxes of the estimated additional human bodies.

3.3 Relation prediction

We use VTransE model to predict the relations between the pairs of human-human and human-object. The inputs of VTransE are the original images and bounding boxes exporting from the result of human and object segmentation.

According to the visual relations appearing in the training dataset, we retain the last layer of VTransE model with the following loss function:

$$L = \sum_{(s,p,o)\in\mathbf{R}} - \quad log\ softmax(\mathbf{t}_p^T(\mathbf{W}_o\mathbf{x}_o - \mathbf{W}_s\mathbf{x}_s)), \tag{2}$$

where s, p and o represent subject, predicate and object, respectively; $\mathbf{x}_o, \mathbf{x}_s \in \mathbb{R}^M$ denote the M-dimensional features of subject and object, and $\mathbf{R}$ is the set of valid relations; $\mathbf{t}_p \in \mathbb{R}^r$ $(r \ll M)$ is relation translation vector as the one in [6]. $\mathbf{W}_o, \mathbf{W}_s \in \mathbb{R}^{r\times M}$ are two projection matrices from the feature space to the relation space. We predict the candidate relation of every pair of human-human and human-object and compute the probability score. Instead of keeping the relation of each pair with the higher score, we mix all of the predicted relations and filter out the triples with little probability and keep the triples with the same subject and object but higher score. To make the result more accurate, we remove some the result according to language prior.

4 Experiments

4.1 Dataset and experimental settings

We validated the proposed method on the dataset provided by PIC Challenge in ECCV 2018. The dataset has three parts: training dataset, validation dataset, and test dataset. The number of subject/object categories and relaiton categories on the three parts are all 85 and 17, respectively. Specifically, the training dataset contains 10,000 images, in which there are 106,959 segments and 167,916 relation instances; the validation dataset contains 1,135 images, in which there are 12,061 segments and 18,729 relation instances; the test dataset contains 2,998 images, and its details are not released.

We used Recall@100 (R@100) under different mean Intersection of Union (m-IoU) as the evaluation criteria in our experiments. Here, R@K denotes the fraction of the correct relation instances in the top K predicated relation instances in an image. Three m-IoUs were used in our experiments, namely 0.25, 0.5 and 0.75. We also evaluate the mean score of R@100 under different m-IoUs.

All the experiments were conducted on a computer with i7 3.5GHz CPU, 32GB memory, and one 1080Ti GPU. The average time cost in processing each image is 1.9 seconds.

4.2 Component analysis

Our proposed method contains three key modules: object instance segmentation, human instance segmentation and visual relation prediction. To validate the effectiveness of each key module, we generate three baselines: 1) Mask+VTransE:

using the retrained Mask R-CNN model for object instance segmentation and human instance segmentation, and using the retrained VTransE model for visual relation prediction; 2) Mask*+VTransE: using the fine-tuned Mask R-CNN model to replace the retrained one in Mask+VTransE; 3) Mask*+RelPrior+VTransE: using relation prior to filter the infrequent <predicate, object> pairs on the object instance and human instance segmentation results in Mask*+VTransE, and further using the retrained VTransE model for visual relation prediction. As compared to Mask*+relation prior+VTransE, our method extends the human instance segmentation result by fuzing the results of human instance segmentation by the fine-tuned Mask R-CNN and face detection based person localization.

Because the groundtruths in the test dataset is not available, we carried out the component analysis on the validation dataset. Table 1 shows the performance of our method and these three baselines. We can see that: 1) The fine-tuned Mask R-CNN model only improves the performance by 0.0003 in mean score. It shows that object instance segmentation cannot be easily improved by global parameter adjustment and it requires further studies for performance improvement. 2) Face detection based person localization improves the performance by 0.016 in R@100 under m-IoU 0.25 but only by 0.003 in R@100 under m-IoU 0.75. It means that face detection based person localization can find some persons omitted in human instance segmentation, but it cannot accurately localize the persons by simply extending the face locations. 3) <predicate, object> pair filtering improves the performance by 0.055 in mean score, which is the obvious improvement in our component analysis. It shows that relation prior is effective to visual relation prediction.

Table 1. Evaluation of our method with different components on the validation dataset.

Method	R@100 (m-IoU: 0.25)	R@100 (m-IoU: 0.5)	R@100 (m-IoU: 0.75)	Mean score
Mask+VTransE	0.3828	0.3330	0.2203	0.3120
Mask*+VTransE	0.3831	0.3334	0.2204	0.3123
Mask*+RelPrior+VTransE	0.4534	0.3915	0.2545	0.3673
Our	0.4693	0.3933	0.2571	0.3724

4.3 Comparison with the state-of-the-arts

Because the implementation of other methods are not released in PIC Challenge, we use the comparison results provided by the leaderboard of PIC Challenge. We select the top 3 methods in PIC Challenge exclude our method in comparison: Cluster, Depth and Greedy (CDG), iCAN, and A context-aware top-down model (CATD).

Table 2 shows the performance of our method and these three methods. We can see that: 1) Our method outperforms other methods in all the criteria, which shows the effectiveness of our method. 2) Comparing the last rows in Table 1 and 2, it shows that our method obtains similar performance on the validation dataset and the test dataset. It shows that our method has good generalization ability. 3) Though our method obtains the first place in PIC Challenge, its performance is far from the requirements in real applications. Human-centric visual relation segmentation is still a challenging task which requires much research attention.

Table 2. Evaluation of different methods on the test dataset.

Method	R@100 (m-IoU: 0.25)	R@100 (m-IoU: 0.5)	R@100 (m-IoU: 0.75)	Mean score
CDG	0.3140	0.2515	0.1313	0.2323
iCAN	0.2499	0.1641	0.0939	0.1693
CATD	0.1493	0.1277	0.0879	0.1216
Our	0.4799	0.4069	0.2681	0.3850

5 Conclusion

We proposed a method handling human-centric relation segmentation, which is based on Mask R-CNN and VTransE model. We use fine-tuned Mask R-CNN model to detect and segment humans and objects in images. To remedy the defect that Mask R-CNN omits some persons in the instance segmentation results, we use face detection in addition and estimate the location of human body. After exporting the result of human and object segmentation, we fine-tuned VTransE model and get the prediction of relations. The results on the test dataset are fairly good, but we can still improve the method.

References

[1] K. He, G. Gkioxari, P. Dollr, R. Girshick. 2017. Mask R-CNN. In IEEE International Conference on Computer Vision. IEEE.

[2] S. Ren, K. He, R. Girshick, and J. Sun. 2015. Faster R-CNN: To-wards real-time object detection with region proposal net-works. In NIPS.

[3] Hanwang Zhang, Zawlin Kyaw, Shih-Fu Chang, and Tat-Seng Chua. 2017. Visual Translation Embedding Network for Visual Relation Detection. In IEEE Conference on Computer Vision and Pattern Recognition. IEEE.

[4] Kazemi V, Sullivan J. 2014. One millisecond face alignment with an ensemble of regression trees. In IEEE Conference on Computer Vision and Pattern Recognition. IEEE.

[5] R. Girshick. Fast R-CNN. In IEEE International Conference on Computer Vision. IEEE, 2015.

[6] A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, and O. Yakhnenko. Translating embeddings for modeling multi- relational data. In NIPS, 2013.

[7] Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. 2016. Visual relationship detection with language priors. In European Conference on Computer Vision. Springer, 852C869.

[8] Xiaodan Liang, Lisa Lee, and Eric P. Xing. 2017. Deep Variation-Structured Reinforcement Learning for Visual Relationship and Attribute Detection. In IEEE Conference on Computer Vision and Pattern Recognition. IEEE.

[9] Xindi Shang, Tongwei Ren, Jingfan Guo, Hanwang Zhang, and Tat-Seng Chua. Video visual relation detection. Proceedings of ACM International Conference on Multimedia. 2017.

[10] Xiangyang Li, Xinhang Song, Luis Herranz, Yaohui Zhu, and Jiang Shuqiang. 2016. Image Captioning with both Object and Scene Information. In ACM International Conference on Multimedia. ACM, 1107C1110.

[11] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In IEEE International Conference on Computer Vision. IEEE, 2425C2433.

[12] Haiyun Guo, Jinqiao Wang, Min Xu, Zheng-Jun Zha, and Hanqing Lu. 2015. Learning multi-view deep features for small object retrieval in surveillance scenarios. In ACM International Conference on Multimedia. ACM, 859C862.