Detecting Human-Object Interactions in Real-Time

Yong Meng
Deepconv. Technologies Co., Ltd.
lmingyin@163.com

Zhongxing Peng
Deepconv. Technologies Co., Ltd.
singlepetrol@sohu.com

Abstract

This paper proposes a novel neural network to detect Human-Object Interactions in Real-Time (HOI-RT). The problem of Human-Object Interactions (HOI) aims to discover the visual relationship between humans and objects. Prior work on HOI generates region proposals before exploring the interaction relationships, obviously, the long pipeline of which impedes the possibility of real-time inference. In a different way, we propose to utilize a single unified network to predict bounding boxes, class probabilities and interaction probabilities directly from the input image, simultaneously. The HOI-RT network consists of two parallel branches: one branch detects objects with find-level bounding boxes and class probabilities, and the other branch detects the interaction between humans and objects, then the outputs of these two branches are fused together to achieve a higher accuracy. Thanks to its single network architecture, HOI-RT can be inferred extremely fast, achieving more than 40 frames per second, with comparable accuracy to the other state-of-the-art methods, in the V-COCO dataset. Our code is available at https://github.com/lmingyin/HOI-RT

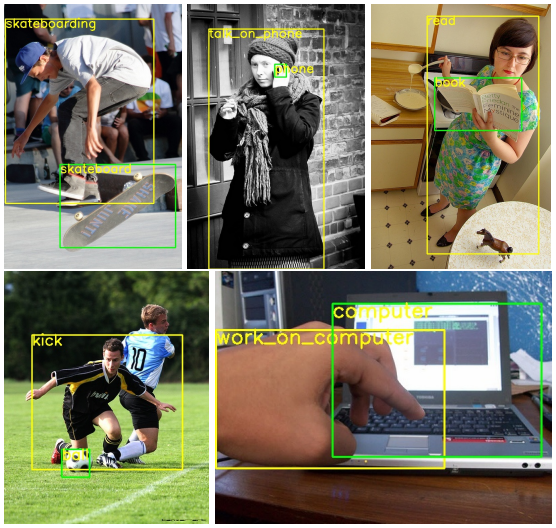

Figure 1. **Detecting human-object interactions with the proposed HOI-RT.** The result of each image highlights some capabilities of our network. (1) detect *skateboarding* with flipped skateboard, (2) detect *talk on phone* with almost occluded and tiny phone, (3) detect *read* from an unusual angle, (4) detect *kick* without messing up with the defender, (5) detect *work on computer* while only a small part of the body (i.e. a left hand) is visible.

1. Introduction

Recently, deep Convolutional Neural Networks (CNN) bring breakthroughs to object detection in terms of accuracy and speed [5, 4, 20, 18, 15, 19], which make it possible to further investigate the redundant information in images. One useful information is the interactive relationships between humans and objects. For example, if a human is on the back of a horse, we can easily recognize that a man *riding a horse*, rather than tie, groom or lead a horse. The problem of HOI aims to detect a triplet of <*human, verb, object*> to represent a human-object interaction, which can provide a deeper understanding of visual semantics.

Several works have been successfully conducted to discover HOIs [7, 6]. They are usually based on the object detection framework of Region-based CNN (RCNN), which consists of two sequential steps. The first step is to use a region proposal network to generate bounding boxes for humans and objects. Then, in the second step, another network is used to classify HOIs for all bounding boxes. Obviously, training and inferences through two sequential networks require a lot of computational time.

Can we design a single unified network to detect HOIs in real-time? Yes. In this paper, we introduce a novel neural network called Human-Object Interactions in Real-Time (HOI-RT). It consists of two parallel branches: (1) branch of interaction detection, and (2) branch of object detection. More specific, the architecture of HOI-RT is illustrated in Fig-2. Firstly, the CNN layers extract features from input image. Secondly, these features are fed into the two parallel branches of interaction detection and object detection. In the branch of interaction detection, the outputs are interaction probabilities, grained-level bounding boxes and class

probabilities. Meanwhile, in the branch of object detection, the outputs are fine-level bounding boxes and class probabilities. Finally, these fine-grained level bounding boxes and class probabilities are fused to improve the final accuracy of HOI detection. After avoiding to inference through two sequential networks, the proposed HOI-RT can do inference at 40 frames per second (fps) on a Nvidia GTX 1080 Ti GPU.

The rest of the paper is organized as follows. Firstly, in Section 2, we discuss the related work. Then, the details of the proposed HOI-RT network is presented in Section 3. Furthermore, experiments are carried out in Section 4. Finally, Section 5 concludes this paper.

2. Related Work

2.1. Object Detection

In recent years, the performances of object detection progress rapidly. Generally speaking, these deep learning algorithms are based on one of two different designs.

One design is originated from R-CNN [5] with two stages: (1) generates numbers of region proposals from an input image, to isolate each individual object in one region proposal, then (2) classify each region proposal into different category. Typical algorithms belonging to this kind are Fast R-CNN [4], SPPNet [10] and Faster R-CNN [20]. The advantage of such design is to obtain accurate results of object detection. However, their computational cost of inference is high, even after sharing some part of feature extractions between region proposal network and classification network [20], because they needed to infer through two neural networks sequentially. Additionally, their training procedures are more complicate and time consuming than a single end-to-end neural network.

The other type of design builds single end-to-end neural network to boost the speed of object detection. YOLO [18] proposes to reconsider the object detection problem as a regression problem, which is able to predict bounding boxes and class probabilities simultaneously. Then, SSD [15] combines predictions from multiple feature maps with different resolutions to handle object of various sizes, which make it more precise than YOLO. More recently, YOLOv2 [19] improves the original YOLO by utilizing anchor boxes, dimension clusters, direct location prediction, and fine-grained features, to get a better, faster and stronger detector. Then, standing on the shoulder of YOLOv2, the proposed HOI-RT can reach the speed of 40 fps with the state-of-the-art accuracy in HOI problem.

2.2. Object Relationship Detection

The problem of Object Relationship Detection (ORD) aims to detect relationships between ordinary objects, without limiting to human-centered scenario. DenseCap [13] solves the dense captioning task to localize and describe salient regions in images in natural language. It depends on both CNN and RNN neural networks to do inference, with a speed of 4 fps. Vip-Net [14] employs RPN to generate triplet proposals, then feed them into the phrase recognition network inspired by Faster R-CNN. In VTransE [23], the input image is first processed by an object detection module, then every pair of detected objects are fed into the relation prediction module.

Although many of these ORD algorithms can be applied to solve HOI problems, HOI algorithm still worth to investigate because of (1) lots of images are human-centered, (2) context or human activities can imply totally different interaction relationships for the same pair of human and object.

2.3. HOI Detection

Human-object interactions had been studied before the prevalence of deep learning [8, 22, 2]. Due to the limitation of the space in this paper, we just review the deep learning based HOI detecton algorithms. HO-RCNN [1] use RPN to generate anchor box of human and object, and then put the region proposal to three stream which contain another deep network , and finally combine the streams to make finally interaction classification. Similarly, InteractNet [6] detects human and object interaction by first generate Region Proposal Network based on ResNet [11], then combines object detection and localization to conclude the interaction. This model localizes the object according to the appearance of the human. It is also a two stage neural network, which impeded its possibility to be real-time.

3. Model

The architecture of the proposed HOI-RT is depicted in figure 2, which consists of three parts: (1) feature extraction, (2) two branches of detections, (3) merge of the two branches.

The feature extraction network is inspired by YOLOv2, which is in turn originated from GoogleNet [21]. It consists of 24 convolutional layers. Except using linear activation to the last layer, we apply batch normalization [12] to all other layers, with following leaky rectified linear activations[16]:

$$f(x) = \begin{cases} x & \text{if x} > 0 \\ 0.1x & \text{otherwise} \end{cases} \quad (1)$$

The extracted features are feed into two detection branches: interaction detection and object detection. The interaction detector output the probabilities of interaction relationships, as well as a grained level bounding boxes for humans and objects. Since the objects can not be localized accurately by interaction detector, we introduce an additional object detector to provide a more accurate bounding boxes for human and objects.

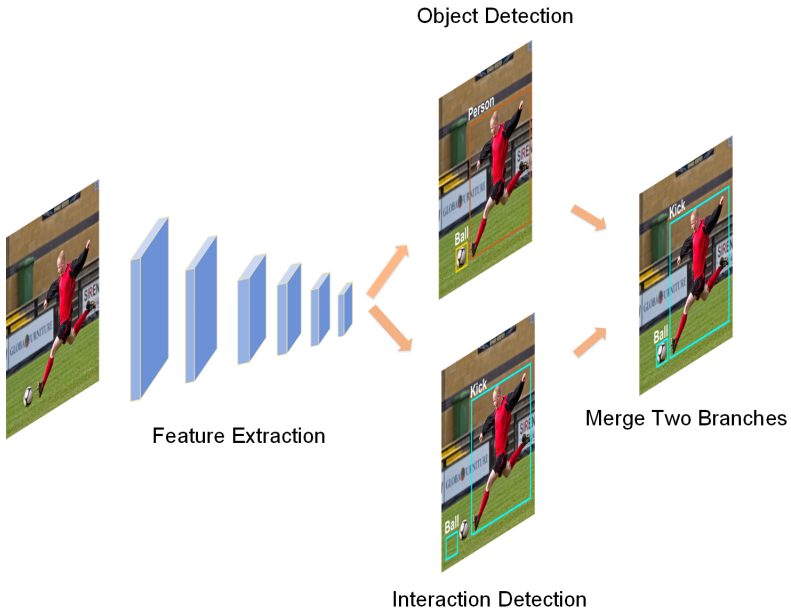

Figure 2. **Architecture of the proposed HOI-RT:** Input an image with size of $N \times N$. The features are extracted by a CNN. Then, we proposed two parallel branches to do object detection and interaction detection, simultaneously. Finally, our model merge the output of the previous two branches to achieve a accurated interaction detection.

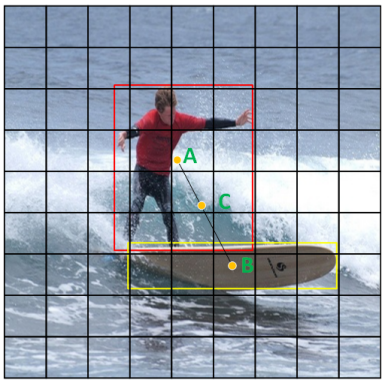

Figure 3. **Three possible choices of grid for detection**. The input image is divided into $N \times N$ grids. Either of the grids containing A, B or C can be responsible for predicting the interaction. A is in the center of the human bounding box. B is the center of the object. C is the center of a cover of both human and object.

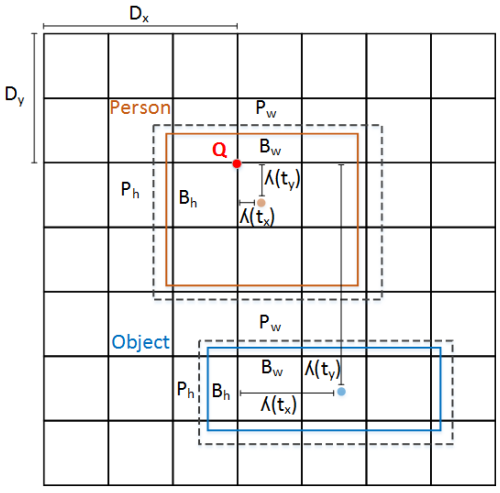

Figure 4. **Illustration of a pair of bounding boxes for one interaction.** Choose the center of the human as the grid for detection. The coordinates of the human bounding box is related to top-left corner of the image, which the bounding box of the object is related to the human bounding box.

The third part of our network merges the results from two previous branches, to make a more accurate result.

3.1. Interaction Detection

In the interaction detection, we divide the input image into $N \times N$ grids. Each grid is in charge of detecting an interaction relationship if its center falls into this grid. Due to the limited number of $N \times N$, our detection can be run very fast. Naturally, how to choose the correct grid to detect the relationship becomes a problem. Since our detection is based on the presences of human and object, we have three reasonable choices as ABC, which are depicted in Figure 3. More specifically, (A) chooses the grid containing the center of bounding box for human, and (B) chooses the grid

containing the center of bounding box for object, and (C) chooses the grid containing the center of a bigger bounding box covering both human and object. The experimental comparisons among different choices of grids will be showed in latter section.

We use regression to predict the bounding boxes. Since the interaction triplet is $\langle human, verb, object\rangle$, for every *verb*, the proposed HOI-RT need to predict two bounding boxes B_{per} and B_{obj}, as showed in figure 4, for human and object, respectively. For example, if we use the grid in the center of the human bounding box to predict the interaction, we have: (1) to regress the bounding boxes for human, we compute its coordinates relative to the grid cell which is in charge of detecting that person. Meanwhile, the coordinates of the grid cell, denoted as D_x and D_y, are relative to the left and top corner of the image, which means the ground truth bounding boxes will fall between 0 and 1. To guarantee the output is within $[0, 1]$, we apply logistic activation to make the output prediction. (2) To regress the bounding box for object, we also compute its coordinates relative to the grid cell which is in charge of detecting that person. Obviously, the coordinates of B_{obj} should not within 0 and 1, so they can be located at any part of the image. As a result, either of the bounding box for human or object can be calculated by (2).

$$\begin{aligned} b_x &= \lambda(t_x) + D_x \\ b_y &= \lambda(t_y) + D_y \\ b_w &= p_w e^{t_w} \\ b_h &= p_h e^{t_h} \end{aligned} \tag{2}$$

where $\lambda(t_x)$, and $\lambda(t_x)$ is relative to the same point Q. D_x and D_y are width and height distance from Q to left up corner of image respectively, more detail are show in figure 4.

The output tensor of the interaction detection is $N \times N \times (R+1+8)$, where $N \times N$ is the number grid. R is the number of interaction relationships we want to detect. 1 represents for C_r, which is the confidence of the relationship. 8 is the number of coordinates for the bounding boxes of human and object, i.e. x_h, y_h, w_h, h_h, and x_o, y_o, w_o, h_o, respectively.

3.2. Loss Function

Since there are two detection branches, the loss function L of the proposed HOI-RT is consists of two parts: one for interaction detection L_r, and the other is for the object detection L_o. Thus, we have

$$L = L_r + \lambda L_o \tag{3}$$

For interaction detection, its loss L_r is defined as follows

$$\begin{aligned} L_r =& \lambda_{\boldsymbol{coord}} \sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbb{1}_{ij}^{per} \Big[(x_{per} - \hat{x_{per}})^2 + (y_{per} - \hat{y_{per}})^2 \\ &+ (x_{obj} - \hat{x_{obj}})^2 + (y_{obj} - \hat{y_{obj}})^2 \Big] \\ &+ \lambda_{\boldsymbol{coord}} \sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbb{1}_{ij}^{per} \Big[(\sqrt{w_{per}} - \sqrt{\hat{w_{per}}})^2 \\ &+ (\sqrt{h_{per}} - \sqrt{\hat{h_{per}}})^2 + (\sqrt{w_{obj}} - \sqrt{\hat{w_{obj}}})^2 \\ &+ (\sqrt{h_{obj}} - \sqrt{\hat{h_{obj}}})^2 \Big] \\ &+ \sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbb{1}_{ij}^{per} \left(C_i - \hat{C}_i \right)^2 \\ &+ \lambda_{noper} \sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbb{1}_{ij}^{noper} \left(C_i - \hat{C}_i \right)^2 \\ &+ \sum_{i=0}^{S^2} \mathbb{1}_{i}^{per} \sum_{c \in cls} \left(p_i(c) - \hat{p_i}(c) \right)^2 \end{aligned} \tag{4}$$

where $\mathbb{1}_{ij}^{per}$ denotes that in the i-th cell the j-th bounding box predictor is used for prediction. $\mathbb{1}_{j}^{per}$ denotes if the i-th cell is responsible for detecting the interaction between this human and the object, and for object detection branch loss L_o, we use the exact same loss function as YOLOv2.

Usually, in many images, most of the grid cells do not contain any object we interested, which leads to overpower the gradient from cells that has no object, causing model instability. As a result, we set $\lambda_{coord} = 5$ and $\lambda_{noobj} = 0.5$, to increase the loss from bounding box coordinate predictions and decrease the loss from confidence predictions for boxes that don't contain objects.

3.3. Merge the two Detection Branches

We add detection branch to improve the localization accuracy for humans and objects.

Suppression of Object Detection. For bounding boxes with probabilities higher than a given threshold, we apply non-maximum suppression (NMS) to them to suppress the ones with high IoU. The remaining bouding box b_{obj} with probability of S_{prob} will be used to rectify the object detected in the interaction detection.

Suppression of Interaction Detection. The branch of interaction detection outputs interaction probabilities for pairs of humans and objects. Each pair contains the probability of the interaction, as well as a bounding box b_p for human, and b_o for the object interacted by this person. For pairs with interaction probability higher than a given threshold, we apply NMS to them to suppress pairs with high IoU.

Merge the two branches. After the above suppressions, for every pair of $\langle b_p, b_o \rangle$, we find a b_{obj} from object detection to rectify b_o, i.e. to maximize a following probability

$$S_{obj} = S_{dis} * S_{prob} \tag{5}$$

where, $S_{obj} \in [0, 1]$ is the new confident score, and S_{prob} is the score of object detection, and the distance score S_{dis} is defined as follows:

$$S_{dis} = \min\left(\frac{\min(w, h)}{\alpha * d_{xy} + \beta * d_{wh}}, 1\right) \tag{6}$$

where d_{xy} is defined as the distance of the centers of two bounding boxes, and d_{wh} describes the shape differences between two bounding boxes, which is defined as follows

$$d_{wh} = \sqrt{(w_1 - w_2)^2 + (h_1 - h_2)^2} \tag{7}$$

where, w_1 , h_1 and w_2, h_2 are the heights and widths for the two bounding boxes, respectively. α and β are the weights for d_{xy} and d_{wh} repectively, which we set to 1 for simplicity. Thus, we attempt to find the bounding box from object detection branch, which results in the highest S_{obj}, to replace the bounding boxes from interaction detection. Of course, more dedicated strategy can be applied to such merge to make more accurate results.

Obviously, according to (5), our merge strategy can suppress false positive in interaction detection. Because if no corresponding object detected in the object detection branch, S_{prob} will be zero, which further result in $S_{obj} = 0$. Then, this interaction detection result will be discarded.

3.4. Limitation of the Proposed HOI-RT

In the proposed HOI-RT, we need to combine two bounding boxes to discover one interaction relationship, which means the same person can have multiple interaction relationships to different objects at the same time. However, the proposed HOI-RT can only detect one interaction relationship for one person. In future, we will extend HOI-RT to predict multiple interactions for one person at the same time.

4. Experiments

4.1. Dataset

We train and evaluate our model on V-COCO dataset [9], which consists of images spreading across 26 action classes. We choose 8 action classes from V-COCO, since they only contain one object except human. A list of these 8 actions can be found in table 1. Thus, we have around 4k images in V-COCO for these 8 interaction relationships. We split them into two parts, one for trainval set, and the other for test set.

Because the dataset based on V-COCO is too small for a deep learning model, we collect and label an extra 10000 images of the same 8 action classed from Internet. It will be used to train our model alongside the V-COCO data. However, in testing, we only use data from the testing set of V-COCO.

4.2. Metrics

According to [9], we employ two metrics based on Average Precision, i.e. AP_{role} and AP_{agent}, to evaluate the proposed HOI-RT.

AP_{role} is an AP to describe the HOI for the triplet of $\langle human, verb, object \rangle$. It is the central interest in HOI task. A triplet is considered as true positive satisfying: 1) the predicted bounding box b_h for human has IoU greater than 0.5 comparing to a ground-truth human bounding box, 2) the predicted bounding box b_o for object has IoU greater than 0.5 comparing to a ground-truth object bounding box, 3) the predicted and the ground-truth actions are match.

AP_{agent} is an AP to describe a pair of $\langle human, verb \rangle$. Its true positive should satisfy the above 1) and 3) standards. Obviously, AP_{agent} don't care about the target object.

4.3. Implementation Details

Our feature extraction layers are based on YOLOv2[19], which are pre-trained on the ImageNet dataset[3]. Then, we fine tune our model based on the pre-trained model. During training, we choose the center of the human bounding box to predict the pairs of interaction relationships.

Darknet[17] is used for all training and inference. We fine-tune our model for 40k iterations, with a learning rate of 0.001 in the first 20k iterations, then reduce the learning rate to 0.0001 for the last 20k iterations. We use a momentum of 0.9 and a weight decay of 0.0005. Batch size is 64. We use random scaling for data augmentation. We also randomly adjust the exposure and saturation of the image by up to a factor of 1.5 in HSV color space.

4.4. Result Comparison

The proposed HOI-RT can infer very fast. According to Table 3, HOI-RT is more than $10\times$ times faster than the InteractNet which can only ran in 7 FPS, and still has comparable precision. Because our network is a simple end-to-end one to produce different predictions simultaneously, while the InteractNet is a two stage network and need to generate RPN before carrying out detection.

Table 1 compares accuracies between InteractNet and the proposed HOI-RT for different cases: HOI-RT without object detetion branch, standard HOI-RT and HOI-RT trained on our extra data. According to the table, we have following observations

- The branch of object detection is important. Without object detection, the branch of interaction detec-

	InteractNet [6]		**Ours no loc**		**Ours with loc**		**Ours with loc (more data)**	
	AP_{agent}	$\textbf{AP}_{\textbf{role}}$	AP_{agent}	$\textbf{AP}_{\textbf{role}}$	AP_{agent}	$\textbf{AP}_{\textbf{role}}$	AP_{agent}	$\textbf{AP}_{\textbf{role}}$
kick	77.5	69.4	75.1	34.2	70.8	64.9	**89.5**	**84.5**
read	41.6	23.9	27.0	5.9	21.5	10.1	**59.3**	**40.9**
skateboard	90.0	75.5	88.6	51.1	83.5	66.7	**93.0**	**78.7**
ski	**84.7**	36.5	76.6	28.4	73.8	31.8	77.1	**45.7**
snowboard	81.1	**63.9**	85.4	56.9	81.7	58.0	**86.7**	58.4
surf	93.5	65.7	**93.8**	56.0	88.9	62.3	93.2	**68.3**
talk-on-phone	82.0	31.8	84.0	13.2	80.2	19.4	**84.0**	**36.5**
work-on-computer	**75.7**	**57.3**	61.0	39.1	56.4	48.6	71.1	57.2
mean AP	78.2	53.0	73.9	35.6	69.6	45.2	**81.7**	**58.7**

Table 1. Comparison of Accuracy for InteractNet and various versions of HOI-RT

tion can not localize the object precisely, since the bounding box of human has great influence in the prediction of detected object. Fortunately, once add the branch of object detection, AP_{role} is improved significantly. For example, considering *kick* and *talk-on-phone*, their AP_{role} are increased by more than 89.77% and 46.97%, respectively. Worth mentioning that, AP_{agent} is a little bit worsen after adding object detection, because the merge of the two branches will affect the AP_{role}. However, such drawback can be remedied after trained on a bigger dataset, which will be showed below.

- Big data is important. Trained on 2k V-COCO data, our model is over-fitting, both AP_{agent} and AP_{role} are worse than InteractNet. Once add extra training data, the AP_{agent} and AP_{role} of the proposed HOI-RT improve significantly to top the other cases in most interactions.

Choice of detected grid	AP_{agent}	AP_{role}
center in human	**81.7**	58.7
center in human and object	79.1	**59.6**
center in object	78.8	57.4

Table 2. Results for different choices of detected grid

As being mentioned in previous section and Figure 3, there are three potential choices of the predicted grid: in the center of the human, in the center covering human and object, in the center of the object. The results are listed in Table 2. Obviously, if we want to obtain high AP_{agent}, we need to use the center of human as the predicted grid, because it puts its attention on human detection. Meanwhile, we can place the predicted grid at the center of human and object to get a high AP_{role}, because it pays its attention to both of them. However, the accuracy of AP_{agent} will be worse than the previous choice. Without surprise, the worst accuracies for both AP_{agent} and AP_{role} exist when we focus our attention on the center of the object. Because in HOI problem, human will always be the center of interest to motivate different kinds of actions, so set the center on the object will distract the attention of the model.

Models	AP_{agent}	AP_{role}	FPS
InteractNet[6]	78.2	53.0	7
HOI-RT 416×416	74.6	54.7	90
HOI-RT 544×544	81.7	57.6	43
HOI-RT 608×608	**81.7**	**58.7**	35

Table 3. Speed and accuracy comparison for different models

To visualize the detected results of HOI-RT, please refer to Figure 5, where each sub-figure show all detected $\langle human, verb, object \rangle$ triplets. Just looking the middle figure, the person is kicking ball, the detection result can find the ball the person who is kicking, and for other person who not kicking the ball will not be selected, and looking the first line, fourth figure, there are two person reading one book together, and our method can detect the two interaction very well.

Some wrong detections are showed in Figure 6.In first line, figure from left to right, 2, 3, 4 cant locate the the object which in interaction very well, this to say, we need strength the ability of object detection branch. And in second line, figure from left to right 1, 2, 3, cant recognize the object very well, we need to improve our interaction branch ability.

5. Conclusion

In this paper, we present a real-time neural network for detecting the human-object interaction. Its fast speed thanks to the unified and end-to-end architecture. Since an only branch of interaction detection can not guarantee high accuracy, we introduce a parallel branch of object detection to result in much better predictions. Meanwhile, beyond the

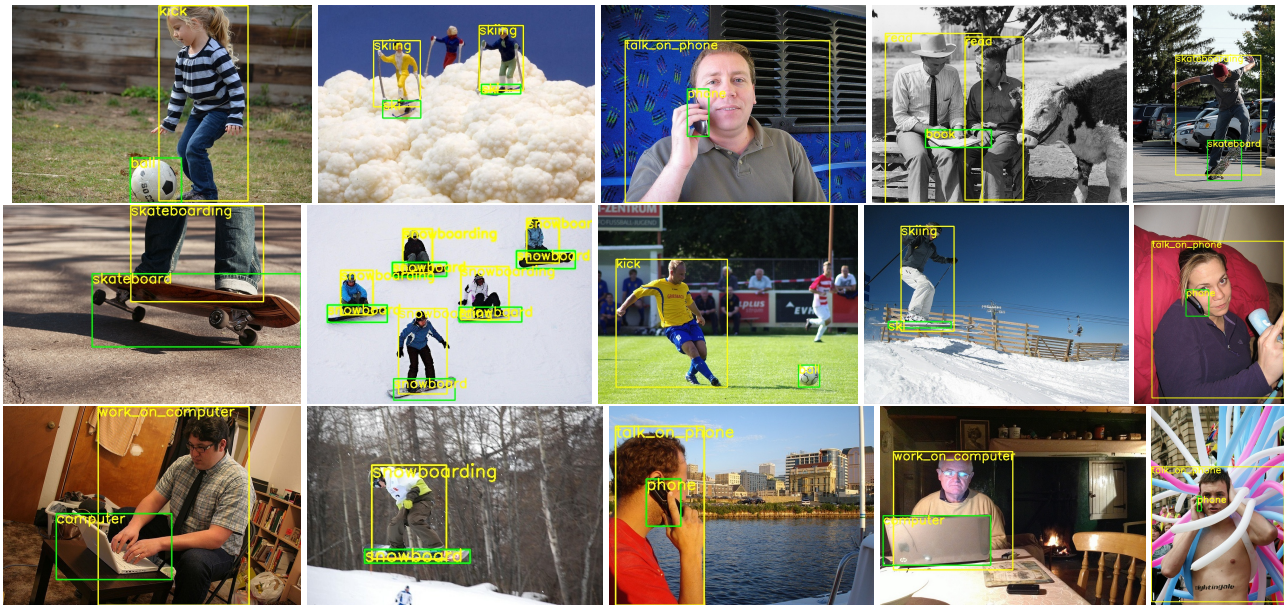

Figure 5. Selected examples of good resulted from HOI-RT. Each image shows all detected $\langle human, verb, object \rangle$ triplets.

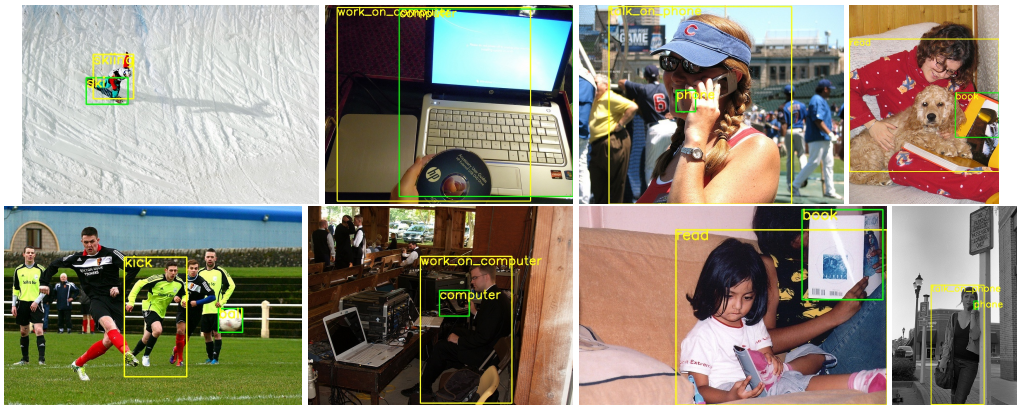

Figure 6. Examples of incorrect detections

small dataset of V-COCO, we gather and label additional data to train a stronger model for HOI problem.

References

[1] Y.-W. Chao, Y. Liu, X. Liu, H. Zeng, and J. Deng. Learning to detect human-object interactions. *arXiv preprint arXiv:1702.05448*, 2017.

[2] V. Delaitre, I. Laptev, and J. Sivic. Recognizing human actions in still images: a study of bag-of-features and part-based representations. In *BMVC 2010-21st British Machine Vision Conference*, 2010.

[3] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255, 2009.

[4] R. Girshick. Fast r-cnn. *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.

[5] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.

[6] G. Gkioxari, R. Girshick, P. Dollár, and K. He. Detecting and recognizing human-object interactions. *arXiv preprint arXiv:1704.07333*, 2017.

[7] G. Gkioxari, R. Girshick, and J. Malik. Contextual action recognition with r* cnn. *Proceedings of the IEEE international conference on computer vision*, pages 1080–1088, 2015.

[8] A. Gupta, A. Kembhavi, and L. S. Davis. Observing human-object interactions: Using spatial and functional compatibility for recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(10):1775–1789, 2009.

[9] S. Gupta and J. Malik. Visual semantic role labeling. *arXiv preprint arXiv:1505.04474*, 2015.

[10] K. He, X. Zhang, S. Ren, and J. Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. In *European Conference on Computer Vision*, pages 346–361. Springer, 2014.

[11] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[12] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, pages 448–456, 2015.

[13] J. Johnson, A. Karpathy, and L. Fei-Fei. Densecap: Fully convolutional localization networks for dense captioning. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4565–4574, 2016.

[14] Y. Li, W. Ouyang, and X. Wang. Vip-cnn: A visual phrase reasoning convolutional neural network for visual relationship detection. *arXiv preprint arXiv:1702.07191*, 2017.

[15] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg. Ssd: Single shot multibox detector. *European conference on computer vision*, pages 21–37, 2016.

[16] V. Nair and G. E. Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814, 2010.

[17] J. Redmon. Darknet: Open source neural networks in c. http://pjreddie.com/darknet/, 2013–2016.

[18] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 779–788, 2016.

[19] J. Redmon and A. Farhadi. Yolo9000: better, faster, stronger. *arXiv preprint arXiv:1612.08242*, 2016.

[20] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, pages 91–99, 2015.

[21] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[22] B. Yao and L. Fei-Fei. Modeling mutual context of object and human pose in human-object interaction activities. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 17–24. IEEE, 2010.

[23] H. Zhang, Z. Kyaw, S.-F. Chang, and T.-S. Chua. Visual translation embedding network for visual relation detection. *arXiv preprint arXiv:1702.08319*, 2017.