Human-Object Interaction Detection Network Proposal - Team 13

Chenhui Gou
u7194588@anu.edu.au

Yuanfei Fan
u7155106@anu.edu.au

Zicheng Duan
u7170273@anu.edu.au

1. Research Question

The project aims at coming up with a novel and robust deep network system which is capable of detecting Human-Object Interactions, this technology is vital for machine to understand the semantic space, which could be used in automatic pilot driving, behavior detection in sports, etc.

The input of the system should be images containing human and objects and the outputs are images with frames and text label indicating the relationship between the human and the object.

2. Previous Work

In this paper [3], two main constraints of current behavior recognition are proposed. One is how to correctly represent the visual relationship contained in the picture, and the other is how to deal with the scarcity of unseen visual relationship data sets. In the latter part. The article gave their own solution. First, they established a hybrid model, which included the composition of the picture and the visual relationship and expressed it in the form of <subject, predicate, and object>, which solved the first restriction. In order to solve the second limitation, the article transfers the existing triples to the unseen triples by analogy with similar visual phrase combinations. Another paper [2] proposed a hybrid model based on Faster Rcnn [4] object detection network, based on the object detection result from Rcnn, three more network branches are introduced in this model, that is three branched in total, except that the first one is pure Faster Rcnn, the other two aims at providing a gaussian probability distribution of the object possible location according to the human action detected and combining the distribution with the object coordinates from feature extraction to obtain the accurate position of the object. This paper [1] has two contributions in Visual recognition of human-object interactions (HOI) . Firstly, they constructing a new large-scale data set called HICO-DET by augmenting the current HICO with instance annotations. Secondly, they proposed Human-Object Region-based Convolutional Neural Networks (HO-Rcnn), a DNN-based framework which can detect a pair of bounding boxes. And this HO-Rcnn can detect HOI by two steps. First step is generating human-object proposal. A proposal is a pairing between a human and object box. Then they divide this proposal into three parallel steams: human and Object stream and Pairwise stream. And after that, they passed this three stream into HO-Rcnn to generate HOI classification scores. The core part of this paper is to build a ConvNet model with Multi-stream architecture.

The difference might come from changing the network structures or embedding the current state-off-the-art network design into our own implementations, one of the most probable plan to successfully make improvements is to change the feature extraction network into some other networks with higher efficiency. Besides, our project also hopes to use the structure of the mixed model for behavior analysis, and we also use the form of triples to represent the action relationship, such as <object,action,object>.

3. Proposal

The main idea is to develop several networks for object detection, human action identification and relationship recognition between human and objects.

3.1. Designed Algorithms

We plan to design a hybrid model that combines target detection, recognition of actions, and object and action matching. Specifically, we need to first obtain the object through the target detection module, and then predict the action by analyzing the relative positions of detected objects and generate multiple target action triples.Finally ,we can obtain the highest score triplet through evaluation. That triplet is used as the prediction result.This is the process we currently envision.

3.2. Extension Ideas

In general, We have two extension ideas based on the basic ideas implementation. Firstly, we want to imple-

	Aug W4	Sep W1	Sep W2	Sep W3	Sep W4	Oct W1	Oct W2	Oct W3
Proposal	■							
Literature review	■	■						
Algorithm study	■	■						
Data Processing	■	■						
Network Design		■	■					
Implementation			■	■	■			
Parameter Tuning					■	■		
Problem Shooting			■	■	■	■		
Final Retouch					■	■	■	
Final Report						■	■	■

Figure 1. Project Gant Chart.

ment a model which can identify two or more Human-Object inter- actions pairs in one picture, which means if there are two people raise the wine glass at the same time, the model will recognize that there are two people raising the wine glass. Secondly, we are also thinking using YOLO V3 to build the object detection model instead of Faster Rcnn. Because YOLO V3 usually has higher accuracy and feature processing speed than Fast Rcnn.

4. Software Datasets and Equipments

The software includes PyTorch, Anaconda, Python and LabelImg, while there might also be extra softwares. There are two dataset suitable for our tasks and they are HICO version 20150920, HICO-DET version 20160224. We will development our project with the help of laptops, Google Colaboratory, ANU student GPU server and local GPUs.

5. Timeline and Milstones

A Gant Chart is used for Timeline monitoring. Please refer to Figure 1 for detail.

6. Workload Distribution

The current decision is that we will contribute all the sections together in case of work distribution unbalance, we will use github for code version control and Microsoft Teams for project collaboration.

References

[1] Y. Chao, Y. Liu, X. Liu, H. Zeng, and J. Deng. Learning to detect human-object interactions. In 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), pages 381–389, 2018.

[2] Georgia Gkioxari, Ross Girshick, Piotr Dollár, and Kaiming He. Detecting and recognizing human-object interactions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2018.

[3] Julia Peyre, Ivan Laptev, Cordelia Schmid, and Josef Sivic. Detecting unseen visual relations using analogies. In Proceedings of the IEEE International Conference on Computer Vision, pages 1981–1990, 2019.

[4] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, Advances in Neural Information Processing Systems 28, pages 91–99. Curran Associates, Inc., 2015.