Multi-task Learning with Cartesian Product-Based Multi-objective Combination for Dangerous Object Detection

Yaran Chen[1,2] and Dongbin Zhao[1,2(✉)]

[1] The State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China
[2] The University of Chinese Academy of Sciences, Beijing, China
dongbin.zhao@ia.ac.cn

Abstract. Autonomous driving has caused extensively attention of academia and industry. Vision-based dangerous object detection is a crucial technology of autonomous driving which detects object and assesses its danger with distance to warn drivers. Previous vision-based dangerous object detections apply two independent models to deal with object detection and distance prediction, respectively. In this paper, we show that object detection and distance prediction have visual relationship, and they can be improved by exploiting the relationship. We jointly optimize object detection and distance prediction with a novel multi-task learning (MTL) model for using the relationship. In contrast to traditional MTL which uses linear multi-task combination strategy, we propose a Cartesian product-based multi-target combination strategy for MTL to consider the dependent among tasks. The proposed novel MTL method outperforms than the traditional MTL and single task methods by a series of experiments.

Keywords: Dangerous object detection · Multi-task learning and convolutional neural network

1 Introduction

Nowadays, more and more people pay attention to driving safety. Dangerous object detection is an effective measure for improving driving safety which has been widely studied for several decades by many researchers. However, it is still challenging to accurately and promptly detect dangerous object.

Dangerous object detection aims to identify the potentially dangerous vehicles and pedestrians for drivers. According to input signals, dangerous object detection methods usually are divided into: general sensor-based methods and vision-based methods. Sensor-based methods mainly apply lasers and radars to sense surroundings and detect dangerous object. They have been widely used, thanks to the

D. Zhao—This work is supported by National Natural Science Foundation of China (NSFC) under Grants 61573353 and 61533017, and the National Key Research and Development Plan under Grant No. 2016YFB0101000.

© Springer International Publishing AG 2017
F. Cong et al. (Eds.): ISNN 2017, Part I, LNCS 10261, pp. 28–35, 2017.
DOI: 10.1007/978-3-319-59072-1_4

great environmental perception capability. Autonomous cars such as Google Car and Baidu Car [1], use a rotating light detection and ranging (LIDAR) scanners to obtain the environment information and warn drivers about dangerous objects. However, these lidar sensors are too expensive to apply in a large scale. Compared with sensor-based methods, vision-based dangerous object detection is low cost and captures more traffic information, such as object distance, object categories and traffic signs [2]. In previous work, vision-based methods are usually formulated as an object detection problem and a distance prediction problem, which are dealt with using two independent models. The typical methods of object detection including faster R-CNN [5], and SSD (Single Shot MultiBox Detector) [6], can be used for object detection in autonomous driving. Object distance is generally measured by a RGB-D cameras LIDAR or radar [9].

In fact, vision-based object detection and distance prediction present prominent visual relationship. The objects far from the camera usually look small and cover few pixels of an image, while the closer ones are generally distributed in the near field of view and cover more pixels, shown as Fig. 1(a) and (b). In addition, Fig. 1(c) and (d) show that objects taken from different camera angles present different poses. Obviously, the visual relationship is very worthy to be exploited for detecting dangerous objects. However, it is much ignored in previous work which deals with object detection and distance prediction using two independent models. Therefore, simultaneously optimizing object detection and distance prediction in one model will probably improve the performance of dangerous object detection.

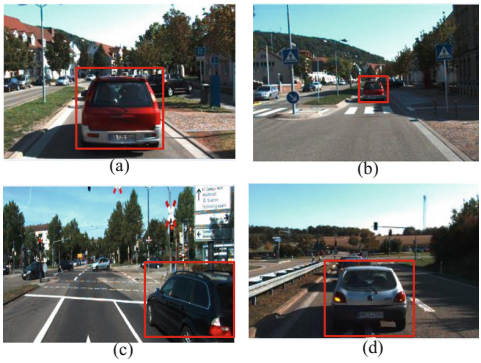

Fig. 1. The cars with different distances and poses

Multi-task learning (MTL) is a well-known method for simultaneously optimizing multiple tasks. MTL exploits shared information among multiple tasks to improve the performance of each other [3]. MTL has been widely applied in computer vision community: such as action recognition [14], pose estimation [10], face detection [11], facial landmark localization [12], and achieved great successes. MTL method generally linearly combines the objectives of multiple tasks to exploit the shared information and jointly optimizes the related tasks. However, it much ignores the correlations of multiple tasks.

In this paper, we propose a novel MTL method based on CNN to jointly optimize object detection and distance prediction. In order to facilitate distance prediction, it is formulated as a classification problem, through discretizing continuous distance. In the proposed MTL method, we propose a joint optimization objective according to the Cartesian product of object classes and distance categories. We prove that the proposed Cartesian product-based multi-task combination strategy outperforms the linear multi-task combination strategy in mathematics and experiments.

Our contributions are shown as follows. First, we use the MTL mechanism to dangerous object detection for exploiting the visual relationship between object detection and object distance prediction for the first time. Second, we propose a novel multi-task combination strategy based on the Cartesian product, and prove it outperforms the linearly combination strategy.

2 Multi-task Learning

Dangerous object detection deals with object detection and distance prediction. Object detection is usually expressed as a classification task. Namely we detect objects by classifying the proposed regions of images. It is difficult to accurately predicting continuous distance owing to the non-linear variation of the sight distance. Therefore, the distance prediction task is transformed into a classification problem. MTL is a popular technique for dealing with related multiple tasks. In this paper, we propose a novel MTL to jointly optimize the two classification problems by shared information.

2.1 Linear Multi-task Combination

Traditional MTL methods generally optimize multiple tasks by a linear multi-task combination strategy (LC-MTL). Namely the loss is a weighted linear combination of the multiple objective functions [12] shown as:

$$L_{c+d} = \alpha \cdot L_c + (1 - \alpha) \cdot L_d, \tag{1}$$

where L_c and L_d are the objective functions of the object detection task C and distance prediction task D, respectively. And α specifies the relative importance of each task and can be experimentally chosen.

Due to the powerful ability of representation learning, CNN has been widely used in multi-task learning, especially for the classification task. For dangerous object detection, through shared model parameters, CNN can jointly model the object detection C and distance prediction D. We use y_c to denote a class of objects, and $y_c \in \{c_1, c_2, ..., c_p\}_{1\times p}$ where p represents the number of object classes. Similarly, y_d denotes a category of object distance, where $y_d \in \{d_1, d_2, ..., d_p\}_{1\times q}$ and q is the number of object distance categories. For a given image $\mathbf{x} \in \Re_+^{\mathrm{m}\times\mathrm{n}}$, CNN simultaneously computes the probabilities of object recognition and distance classification: $p(y_c = c_i|\mathbf{x})$ the probability of the

image $\mathbf{x}$ belonging to the i-th class of object and $p(y_d = d_j|\mathbf{x})$ the probability of the image $\mathbf{x}$ belonging to the j-th class of object distance.

A typical objective function of the classification with multiple categories is the cross entropy loss:

$$L_c = y_c \cdot log(p(y_c = c_i|\mathbf{x})). \tag{2}$$

Similarly, we get $L_d = y_d \cdot log(p(y_d = d_j|\mathbf{x}))$. Then the loss of the MTL (Eq. (1)) can be rewritten as:

$$L_{c+d} = y_c \cdot log(p(y_d|\mathbf{x})) + y_d \cdot log(p(y_d|\mathbf{x})), \tag{3}$$

where we ignore the constant α for simplification.

Through the MTL with the linear multi-task combination strategy, CNN can exploit the shared information for the related tasks from input images. However, it much ignores the dependence among multiple targets.

2.2 Cartesian Product-Based Multi-task Combination

To exploit the dependence among related targets, we propose a Cartesian product-based multi-task combination strategy (CP-MTL) to jointly optimize object detection and distance prediction. We denote the combined task based on the Cartesian product as $M = C \otimes D$, where $\otimes$ represents the Cartesian product operator. Concretely, we use $y_{c\otimes d} = y_c \otimes y_d$ as a category of M and $y_{c\otimes d} \in \{c_1d_1, c_1d_2, ...c_1d_q, \ldots, c_id_j, \ldots, c_pd_q\}_{1\times pq}$, where pq is the number of the combined task category.

Then, the loss function of M is formulated as:

$$L_{c\otimes d} = y_{c\otimes d} \cdot log(p(y_{c\otimes d} = c_id_j|\mathbf{x})). \tag{4}$$

Through taking the Cartesian product operator into Eq. (4), we can obtain:

$$\begin{aligned} l_{c\otimes d} = {} & c_1d_1 \cdot log(p(y_{c\otimes d} = c_1d_1|\mathbf{x})) + c_1d_2 \cdot log(p(y_{c\otimes d} = c_1d_2|\mathbf{x}))+ \\ & \cdots + c_pd_q \cdot log(p(y_{c\otimes d} = c_pd_q|\mathbf{x})). \end{aligned} \tag{5}$$

Equation (5) is the sum of pq entries, and each one contains a probability $p(y_{c\otimes d} = c_id_j|\mathbf{x})$. It means that the image $\mathbf{x}$ belongs to c_i of task C and d_j of the task D. If the task C and D are completely independent, we can obtain:

$$p(y_{c\otimes d} = c_id_j|\mathbf{x}) = p(y_c = c_i|\mathbf{x}) \cdot p(y_d = d_j|\mathbf{x}). \tag{6}$$

Then we take Eq. (6) into Eq. (5) and deduce that:

$$L_{c\otimes d} = L_c + L_d = L_{c+d}. \tag{7}$$

Compared Eqs. (7) and (3), we prove that if the two tasks are independent, the loss function of the traditional LC-MTL method is equal to the loss function of the proposed CP-MTL method. Otherwise, the CP-MTL method can exploit the dependency between two tasks, which is ignored by LC-MTL method. For dangerous object detection, the object detection task and object distance classification task are probably not independent, which may be more suitable for being modeled by the proposed CP-MTL model.

3 CP-MTL SSD Method

Dangerous object detection consists of object detection and distance prediction. Owing to the strong capability of learning representation, CNN-based object detection methods have achieved satisfactory performance. SSD is one of the art-of-the-state CNN-based object detection methods. It directly predicts object bounding boxes and object classes by sharing convolutional features, resulting a short detection time and high accurate. In this paper, we incorporate the proposed CP-MTL (Cartesian product-based combination multi-target) into the optimization objective of SSD to simultaneously optimize the object detection and distance classification tasks.

3.1 Model Architecture

Figure 2 shows the structure of the proposed CP-MTL SSD Method. It consists of multiple hierarchical convolutional layers, some default bounding boxes with different aspect ratios, and a number of detections. By the convolution operation, the hierarchical convolution layers can produce a lot of feature maps of different scales and resolutions for an input image. There are some default bounding boxes on these feature maps. For one default bounding box, the following detection consists of a full-connected classification layer and a regression layer, to regress the bounding box and classify the object category simultaneously. Due to the larger number of default bounding boxes, the model can produce a lot of detections of boxes. Through non-maximum suppression [8], the model will predicts the final boxes.

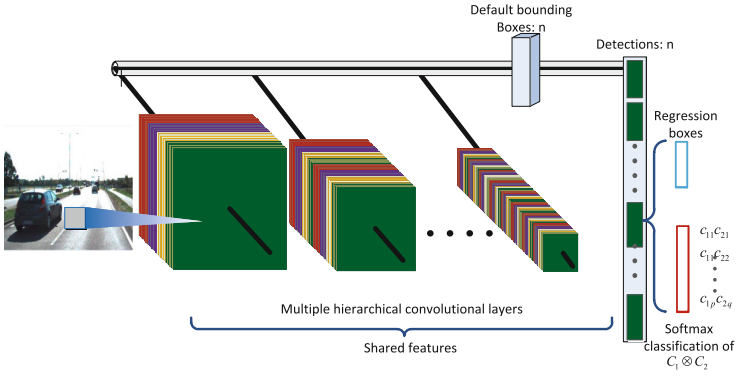

Fig. 2. The architecture of the proposed CNNVA.

CP-MTL is a variant of SSD. Although they seem similar, there is an essential difference between them. Namely CP-MTL optimizes the Cartesian product-based combination targets of object recognition and object distance classification, while SSD just only optimizes the target of object recognition.

3.2 Cartesian Product-Based Combination Targets

We propose a Cartesian product-based combination of object detection task and distance classification task to simultaneously optimize object detection and object distance prediction. Based on the sizes and shapes of objects, we classify objects into three categories: cars, vans and pedestrians, denoted as $\{c_1, c_2, c_3\}$. Due to the relationship between the distance and the object distance, we consider the distance category task from two dimensions: the vertical distance and the horizontal distance, shown as Fig. 3.

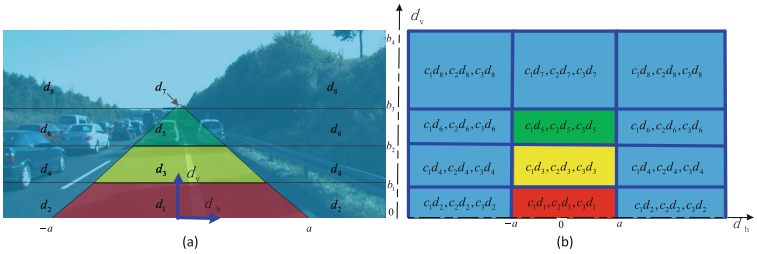

Fig. 3. (a) The image geographic division according to the distance and the visual angle. (b) The categories of the Cartesian product-based combination target, where (a) is mapped to the two dimensional plane (b) (Color figure online)

Figure 3(a) shows that the space is parted into 12 regions and 8 categories denoted as $\{d_1, d_2, \cdots, d_8\}$, due to the symmetry of vehicles. And the red one denotes the shortest vertical distance and the most dangerous category, followed by the yellow one, the green one, and the blue one. In Fig. 3(b), each region is a distance category and contains all the categories $\{c_1, c_2, c_3\}$ of C. Recognizing objects during a given distance category is much easier than recognizing them at all the range of distance.

4 Experiment

In this section, we comprehensively evaluate the proposed CP-MTL model on dangerous object detection task by comparing the proposed CP-MTL with the single task learning model (SSD) and the LC-MTL method with the linear multi-task combination strategy.

Dataset: KITTI dataset [4] contains more than 40,000 images which are collected by a car driving in European cities. About 16,000 images contain information of object positions. In the experiments, we randomly divide the 16,000 images into 3 parts: training set, testing set and validating set. Among them, the training set contains 12,000 images, the testing set contains 3000 images and the validating set contains 1000 images. All experimental configures are experimentally chosen according to the performances on the validating set.

Evaluation Metrics: In object detection, a common evaluation metrics is the average precision (AP). AP measures the comprehensive performance, including the recall rate and precision rate of object detection. The mAP is the mean value of the APs of different object categories.

Experimental Setup: In this study, we take SSD as the baseline model. It has 18 convolutional layers and 5 detectors. The early 13 layers are initialized by the Oxford VGG [7,13], and others are randomly initialized. There are five bounding boxes with different aspect ratios ($\{1, 2, 3, \frac{1}{2}, \frac{1}{3}\}$) at each position of 10-th, 15-th, 16-th, 17-th, and 18-th convolutional feature layers. The detection with multiple shapes, resolutions and scales, can deal with various objects with different shapes and sizes. The proposed models, whether the CP-MTL or the LC-MTL, are the variants of SSD. They have the same network architecture and configures with SSD. But the key difference is the output target of detector. We divide the whole image into 12 regions as shown in Fig. 3, and set $a = 3\,\text{m}, b_1 = 10\,\text{m}, b_2 = 20\,\text{m}$, and $b_3 = 40\,\text{m}$. In addition to the object detection, the proposed CP-MTL and the LC-MTL also take the object distance prediction into account.

Comparison Experiments: Table 1 reports the detection performance of SSD, LC-MTL and CP-MTL. The proposed MTL methods (LC-MTL and CP-MTL) consistently outperform the SSD on mAP and APs of each object category. It mainly owes to MTL methods capturing the visual relationship between object detection and object distance.

Compared with LC-MTL, the CP-MTL yields significant performance improvements in the mAP and APs of all object categories. In a sense, it is verified that the proposed the Cartesian product-based multi-task combination strategy outperforms the linear multi-task combination strategy. At the same time, we also note that the Cartesian product-based multi-task combination strategy increases the difficulty of multi-task learning due to the more detailed classification categories. Therefore, the proposed CP-MTL may require more data to be trained. Finally, we exhibit an example of real-time dangerous object detection on a video. Figure 4 shows four snapshots of the video at $t = 1\,\text{s}$, $t = 10\,\text{s}$, $t = 20\,\text{s}$ and $t = 30\,\text{s}$, respectively. Compared with other object detection systems, the proposed CP-MTL not only bounds the object in an image but also gives its danger level according to the predicted object distance, shown in Fig. 4. Moreover, the proposed CP-MTL based on a fast detection algorithm SSD can meet the real-time requirements of practical applications.

Table 1. The detection results with CP-MTL, LC-MTL and SSD

Method	mAP	AP (Cars)	AP (Pedestrians)	AP (Vans)
SSD	0.8104	0.8779	0.6741	0.8790
LC-MTL	0.8331	0.8933	0.8945	0.7113
CP-MTL	0.8405	0.8945	0.8980	0.7292

Fig. 4. Snapshots from video detection with CP-MTL model

5 Conclusion

We propose the CP-MTL algorithm for dangerous object detection in autonomous driving. Through Cartesian product-based multiple objectives combination, CP-MTL can simultaneously optimize object detection and object distance prediction to exploit the relationship between them. We mathematically prove that the proposed CP-MTL outperforms LC-MTL, when the two tasks are not independent. Also, we carry out systematic experiments to verify that the proposed method outperforms the state-of-art SSD object detection method and the traditional MTL method.

References

1. Bruch, M.: Velodyne HDL-64E lidar for unmanned surface vehicle obstacle detection. In: Proceedings of SPIE - The International Society for Optical Engineering, Florida, 05 April 2010
2. Chen, X., Kundu, K., Zhang, Z., Ma, H., Fidler, S., Urtasun, R.: Monocular 3D object detection for autonomous driving. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2016
3. Evgeniou, A., Pontil, M.: Multi-task feature learning. Adv. Neural Inf. Process. Syst. **19**, 41 (2007)
4. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? The KITTI vision benchmark suite. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2012)
5. Girshick, R.: Fast R-CNN. In: Proceedings of the IEEE Conference on Computer Vision, pp. 1440–1448 (2015)
6. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., Berg, A.C.: SSD: single shot multibox detector. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9905, pp. 21–37. Springer, Cham (2016). doi:10.1007/978-3-319-46448-0_2
7. Lv, L., Zhao, D., Deng, Q.: A semi-supervised predictive sparse decomposition based on the task-driven dictionary learning. Cogn. Comput. (2016). doi:10.1007/s12559-016-9438-0
8. Neubeck, A., Gool, L.V.: Efficient non-maximum suppression. In: International Conference on Pattern Recognition, pp. 850–855 (2006)
9. Xia, Y., Wang, C., Shi, X., Zhang, L.: Vehicles overtaking detection using RGB-D data. Sign. Proces. **112**, 98–109 (2015)
10. Yim, J., Jung, H., Yoo, B.I., Choi, C.: Rotating your face using multi-task deep neural network. In: Computer Vision and Pattern Recognition, pp. 676–684 (2015)
11. Zhang, C., Zhang, Z.: Improving multiview face detection with multi-task deep convolutional neural networks. In: IEEE Winter Conference on Applications of Computer Vision, pp. 1036–1041 (2014)
12. Zhang, Z., Luo, P., Chen, C.L., Tang, X.: Facial landmark detection by deep multi-task learning. In: European Conference on Computer Vision, pp. 94–108 (2014)
13. Zhao, D., Chen, Y., Lv, L.: Deep reinforcement learning with visual attention for vehicle classification. IEEE Trans. Cogn. Dev. Syst. (2016). doi:10.1109/TCDS.2016.2614675
14. Zhou, Q., Wang, G., Jia, K., Zhao, Q.: Learning to share latent tasks for action recognition. In: IEEE International Conference on Computer Vision, pp. 2264–2271 (2013)