Research on 3D Object Detection Method Based on Convolutional Attention Mechanism

Zhang Yong¹, Zhang Xiaoxia¹*, Da Nana¹
School of Computer and Communication, Lanzhou University of Technology, Lanzhou, 730050, China
*Corresponding author’s e-mail: renw@lut.edu.cn

Abstract. In 3D object detection, the illumination and occlusion of the input point cloud data lead to inaccurate feature extraction, and the maximum pooling method destroys the information structure of the point cloud, leading to the problem of weak local feature expression. This paper proposes a 3D object detection method based on Convolutional Attention Mechanism (CAM). CAM first adds an attention mechanism to the first and last layers of the traditional feature extraction network structure, then fuses the feature information of different layers, and finally performs normalization operations. Experimental results show that this method has achieved better results on the KITTI and SUN-RGB data sets compared with mainstream algorithms DoBEM, MV3D, DSS, COG, 2D-driven, and FPNet. The mAP index has increased by 0.6%-12.5%. While CAM realizes the fusion of local and global information, it significantly improves the accuracy of object detection in illuminated and occluded scenes.

1. Introduction
Object detection refers to the use of computer technology to detect and recognize the category and location information of objects of interest (such as vehicles, pedestrians, obstacles) in images or videos. It is one of the important research fields in the field of computer vision. With the continuous improvement and development of deep learning technology, object detection technology based on deep learning has been widely used in many real-world fields, such as intelligent robots, autonomous driving, assisted driving, human-computer interaction, behavior recognition and other related fields.

Imran [1,2] firstly fused the RGB image with the depth image, and then used the RGB-D four-channel data stream to train a CNN target detection network. Dolson [3] designed an accelerated Gaussian interpolation algorithm to upsample camera images and radar point clouds in real time, using high-resolution depth images for computer vision applications. In [4,5], the point cloud data is expressed as a 2D image, and then a convolutional neural network is used to detect and regress the target category and location in the 2D image. MV3D [6] maps radar point cloud data to a bird's-eye view, and specially trained a region proposal network (RPN) for 3D candidate frame extraction to extract candidate targets in 3D space. [7] proposed a new end-to-end multi-view fusion (MVF) algorithm, which uses dynamic voxelization to make each point learn to fuse information from different views, effectively learning to use perspective and point cloud information. [8] proposed an effective method to learn an end-to-end probabilistic three-dimensional target detector. It was the first to capture the uncertainty of detection by simulating the distribution of the corners of the bounding box. By estimating the accuracy of detection, this method can make the downstream components in the fully automated driving system show different behaviors around objects with different uncertainties. [9] proposed a new network framework BirdNet for lidar data. First, project the lidar data into a new cell code for bird's-eye projection. Then, a
convolutional neural network originally designed for image processing is used to estimate the position and direction of the target on the plane. Finally, 3D-oriented detection tasks are calculated in the post-processing stage. [10] proposed a point cloud-based three-dimensional target detection method. The first stage is used for bottom-up 3D region proposal, and the second stage is used to refine the proposal in the standard coordinate system to obtain the final detection result.

Although the existing 3D object detection has made remarkable achievements, when the input point cloud data has illumination and occlusion and the maximum pooling method destroys the information structure of the point cloud, it leads to inaccurate feature extraction and weak expression ability. Based on this, this paper proposes a 3D detection method based on Convolutional Attention Mechanism (CAM). First, the attention mechanism is added to the first and last layers of the traditional feature extraction network structure, and then the feature information of different layers is merged, and finally the normalization operation is performed. The algorithm in this paper is compared with a variety of mainstream algorithms on the KITTI and SUN RGB-D data sets. The experiment shows that CAM is significantly better than other compared algorithms in feature expression and detection accuracy.

2. Related Work

The Attention Mechanism [11] is widely used in various tasks, such as machine translation, scene classification and semantic segmentation. Non-local Neural Networks [12] uses self-attention mechanism as a sub-module of computer vision tasks, such as video classification, target detection and instance segmentation. CCNet [13] collects contextual information of all locations by stacking two consecutive cross-attention modules. DANet [14] uses a similar spatial and channel attention module to generate feature information from all pixels, but requires a lot of calculation and GPU memory. A2-Nets [15] constructs a dual attention mechanism module for collecting and distributing long-distance features, which significantly improves image recognition performance. EMANet [16] expressed the attention mechanism as a way of expectation maximization, and estimated the attention map on this basis. ACFNet [17] calculates and adaptively combines different types of centers according to each pixel. Given an intermediate feature map, CBAM [18] will infer the attention map in sequence along two independent dimensions (channel and space), and then multiply the attention map by the input feature map for adaptive feature modification.

3. Overall Network Structure

This paper generates 2D area proposals from the image and locates them in the 3D point cloud data, thereby extracting the corresponding 3D point cloud data, and then based on the segmentation network of PointNet [19] and the 3D bounding box evaluation network to achieve the goal of 2D driving 3D Detection. The original network structure of this research is shown in Figure 1. It can be clearly seen that the model is mainly composed of three modules, namely the frustum proposal module, the 3D instance segmentation module and the amodal 3D box estimation.

The method in this paper is to improve the feature extraction part of the frustum proposal module, and the improved structure is shown in Figure 2. Due to the use of two-dimensional drive for three-
dimensional object detection, the features extracted from the three-dimensional point cloud depend heavily on the features extracted from the RGB image. When there is illumination and occlusion, there will be a missed detection situation, and the feature extraction is very simple. So this article is in RGB. When extracting features from images, a CAM network is proposed, and its structure is shown in Figure 3. CAM can assign weights to different parts of the input data, extract key information from it, and suppress unimportant information, help the network learn important information in the data to make more accurate judgments, so that the characteristics of the point cloud are more accurate.

Figure 2. Improved network structure

Figure 3. CAM network structure

In order to relieve the 3D point cloud features that rely heavily on RGB image extraction features, CAM adds a channel attention mechanism and a spatial attention mechanism to the first and last layers of the CNN convolution, respectively. This paper first adds a channel attention mechanism after the first layer of CNN, as shown in Figure 4.

Figure 4. Channel attention mechanism
First, in order to summarize the spatial information, the channel attention uses both maximum pooling and average pooling to aggregate the spatial information of the feature map, denoted by $F_{\text{max}}$, $F_{\text{avg}}$, respectively. Then transfer the above information to the shared network to generate channel attention $M_c \in \mathbb{R}_c \times 1 \times 1$, the shared network consists of a single hidden layer multi-layer perceptron. In order to reduce the parameter overhead, the hidden layer activation size is set to $\mathbb{R}_{(C/r) \times 1 \times 1}$, where $r$ is the reduction ratio. Finally, element-wise addition is used to output the merged feature vector. As shown in formula 1.

$$M_c = \sigma(MLR(\text{AvgPool}(F)) + MLR(\text{MaxPool}(F))) = \sigma(W_1(F_{\text{avg}}^c) + W_0(F_{\text{max}}^c))$$  \hspace{1cm} (1)$$

Among them, MLP is multi-layer perceptron; AvgPool is average pooling; MaxPool is maximum pooling; $\sigma$ is sigmoid function; $W_0$ and $W_1$ are the learnable parameters of the multilayer perceptron, $W_0 \in \mathbb{R}_{(C/r) \times C}$, $W_1 \in \mathbb{R}_{C \times (C/r)}$.

Secondly, we add a spatial attention mechanism to the last layer of the convolutional layer. The spatial attention mechanism pays more attention to the specific position of the target object, which is a supplement to the channel attention, as shown in Figure 5.

In order to calculate the spatial attention, the average pooling and maximum pooling are sequentially performed after the channel attention, and two two-dimensional feature maps are sequentially generated through two pooling operations $F_{\text{avg}}^s$, $F_{\text{max}}^s$. Then connect the two and input the convolutional layer to generate $M_s \in \mathbb{R}_1 \times H \times W$. As shown in formula 2, where $f^{7 \times 7}$ is a $7 \times 7$ convolution kernel convolution operation.

$$M_s(F) = \sigma(f^{7 \times 7}([\text{AvgPool}(F);\text{MaxPool}(F)])) = \sigma(f^{7 \times 7}(F_{\text{avg}}^s;F_{\text{max}}^s))$$ \hspace{1cm} (2)$$

4. Experimental Results and Analysis

4.1. Dataset and Evaluation Index

All experiments in this paper are conducted on the KITTI and SUN RGB-D datasets. The experiment is based on the pytorch framework, and the experimental platform is configured as Ubuntu 18.04, Intel GPU E5-2620, GTX 1080Ti GPU.

**KITTI dataset** Which contains real image data collected from scenes such as urban areas, villages and highways. Each image contains up to 15 cars and 30 pedestrians, with various degrees of occlusion and truncation. The entire data set is composed of 389 pairs of stereo images and optical flow diagrams, 39.2 km visual ranging sequence and more than 200k 3D annotated object images, which are sampled and synchronized at a frequency of 10 Hz.

**SUN-RGB dataset** Which contains 10335 indoor pictures of different scenes, 146617 2D polygon annotations, and 58659 3D borders.
mAP Which commonly used measurement indicators in target detection, the output of the model (considered as a rank list containing classification results, confidence, and target frame coordinates) is judged as the accuracy (AP) of positive examples, and multi-target Average, namely mAP.

4.2. Experimental Results and Analysis

4.2.1. KITTI dataset experimental results

On the KITTI dataset, the method proposed in this paper is compared and analyzed with several other representative algorithms, including DoBEM, MV3D and FPNet based on LiDAR images with a bird’s eye view. The experimental results are shown in Table 1.

| Method   | Cars    | Pedestrians | Cyclists |
|----------|---------|-------------|----------|
|          | Easy    | Moderate    | Hard     | Easy | Moderate | Hard | Easy | Moderate | Hard |
| DoBEM    | 7.42    | 6.95        | 13.45    | -    | -        | -    | -    | -        | -    |
| MV3D     | 71.09   | 62.35       | 55.12    | -    | -        | -    | -    | -        | -    |
| FPNet(v1)| 80.62   | 64.70       | 56.07    | 50.88| 41.55    | 38.04| 69.36| 53.50    | 52.88|
| CBM(v1)  | 80.71   | 64.81       | 56.63    | 51.13| 41.73    | 38.29| 69.70| 53.68    | 52.98|

It can be seen from Table 1 that the method proposed in this paper has achieved the best results on Cars, Pedestrians, and Cyclists on the KITTI dataset. This is mainly due to the integration of multiple feature map spaces through the channel attention mechanism. In the case of suppressing light and occlusion, there is a missed detection situation, which leads to the problem of single feature extraction.

4.2.2. SUN-RGBD dataset experimental results

In order to further verify the feature expression ability of the proposed method and the accuracy of object detection, the SUN-RGBD dataset was compared and analyzed with popular algorithms such as DSS, COG, 2D-driven and FPNet. The experimental results are shown in Table 2.

| Method       | Bathhub | Bed | Bookshelf | Chair | Desk | Dresser | Nightstand | Sofa | Table | Toilet | Runtime | mAP |
|--------------|---------|-----|-----------|-------|------|---------|------------|------|-------|--------|---------|-----|
| DSS          | 44.2    | 78.8| 11.9      | 61.2  | 20.5 | 6.4     | 15.4       | 53.5 | 50.3  | 78.9   | 19.55s  | 42.1|
| COG          | 58.3    | 63.7| 31.8      | 62.2  | 45.2 | 15.5    | 27.4       | 51.0 | 51.3  | 70.1   | 10-30min| 47.6|
| 2D-driven    | 43.5    | 64.5| 31.4      | 48.3  | 27.9 | 25.9    | 41.9       | 50.4 | 37.0  | 80.4   | 4.15s   | 45.1|
| FPNet(v1)    | 43.3    | 81.1| 33.3      | 64.2  | 24.7 | 32.0    | 58.1       | 61.1 | 51.1  | 90.9   | 0.12s   | 54.0|
| CBM(v1)      | 43.98   | 81.7| 33.65     | 64.5  | 25.15| 32.2    | 58.19      | 61.35| 52.3  | 91.12  | 0.10s   | 54.6|

It can be seen from Table 2 that the method proposed in this paper has achieved better results on the SUN-RGBD data set. Compared with DSS, COG, 2D-driven, and FPNet, mAP indicators have increased by 12.5%, 7%, 9.5% and 0.6% respectively. The main reason is that this paper uses the attention mechanism to achieve the fusion of local and global information, and at the same time, it significantly improves the detection accuracy under illumination and occlusion scenes.

5. Conclusions

The illumination and occlusion of the input point cloud data lead to inaccurate feature extraction, and the maximum pooling method destroys the information structure of the point cloud, leading to the problem of weak local feature expression. This paper proposes a 3D object detection method based on Convolutional Attention Mechanism (CAM). CAM first adds an attention mechanism to the first and last layers of the traditional feature extraction network structure, then merges the feature information of different layers, and finally performs normalization operations to achieve the fusion of local and global information, while significantly improving accuracy of object detection under lighting and occlusion.
scenes. In the next step, we will improve the accuracy of detection in more complex scenes and scenes containing objects with smoother surfaces, and consider the detection of multiple objects.

References

[1] J. Imran, P. Kumar, “Human action recognition using rgb-d sensor and deep convolutional neural networks”, in International Conference on Advances in Computing, Communications and Informatics. IEEE, 2016, pp. 144-148.

[2] C. Lv, H. Wang, & D. Cao, “High-Precision Hydraulic Pressure Control Based on Linear Pressure-Drop Modulation in Valve Critical Equilibrium State,” IEEE Transactions on Industrial Electronics, 2017.

[3] J. Dolson, J. Baek, C. Plagemann, S. Thrun, “Upsampling range data in dynamic environments,” in IEEE Conference on Computer Vision and Pattern Recognition, IEEE Computer Society, 2010, pp. 1141-1148.

[4] Z. Deng and L. J. Latecki, Amodal detection of 3d objects: Inferring 3d bounding boxes from 2d ones in rgb-depth images. In Conference on Computer Vision and Pattern Recognition (CVPR), volume 2, 2017.

[5] B. Li, T. Zhang, and T. Xia. Vehicle detection from 3d lidar using fully convolutional network. arXiv preprint arXiv:1608.07916, 2016.

[6] X. Chen, H. Ma, J. Wan, B. Li, and T. Xia. Multi-view 3d object detection network for autonomous driving. In IEEE CVPR, 2017.

[7] Wu, Fanzi, et al. "Mvf-net: Multi-view 3d face morphable model regression." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019.

[8] Yang B, Luo W, Urtasun R. Pixor: Real-time 3d object detection from point clouds[C]//Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. 2018: 7652-7660.

[9] Beltrán J, Guindel C, Moreno F M, et al. Birdnet: a 3d object detection framework from lidar information[C]//2018 21st International Conference on Intelligent Transportation Systems (ITSC). IEEE, 2018: 3517-3523.

[10] Shi, Shaoshuai, Xiaogang Wang, and Hongsheng Li. "Pointrcnn: 3d object proposal generation and detection from point cloud." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019.

[11] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” in Advances in neural information processing systems, 2017, pp. 5998–6008.

[12] X. Wang, R. Girshick, A. Gupta, and K. He, “Non-local neural networks” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7794–7803.

[13] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, “Ccnet: Criss-cross attention for semantic segmentation,” in Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 603–612.

[14] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, “Dual attention network for scene segmentation,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 3146–3154.

[15] Y. Chen, Y. Kalantidis, J. Li, S. Yan, and J. Feng, “A2-nets: Double attention networks,” in Advances in Neural Information Processing Systems, 2018, pp. 352–361.

[16] X. Li, Z. Zhong, J. Wu, Y. Yang, Z. Lin, and H. Liu, “Expectation-maximization attention networks for semantic segmentation,” in Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 9167–9176.

[17] F. Zhang, Y. Chen, Z. Li, Z. Hong, J. Liu, F. Ma, J. Han, and E. Ding, “Acnet: Attentional class feature network for semantic segmentation,” in Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 6798–6807.

[18] Woo, Sanghyun, et al. "Cbam: Convolutional block attention module." Proceedings of the
European conference on computer vision (ECCV). 2018.

[19] Qi, Charles R., et al. "Pointnet: Deep learning on point sets for 3d classification and segmentation." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.