ISSN: 2472-3703 DOI: 10.6911/WSRJ.202106_7(6).0059

Influence of Depth and Structure of Convolutional Neural Network on Loop Closure Detection

Dongcan Zhang[1, 2, a], Guoliang Zhang[1, 2, b], Junxue Li[1, 2], Yujie Chen[1, 2]

[1]School of Automation and information Engineering, Sichuan University of Science&Engneering, Zigong, China

[2]Atificial Intelligence key laboratory of Sichuan province, Zigong, China

[a]896679264@qq.com, [b]zhgl@sohu.com

Abstract

With the increasing popularity of computer vision and deep learning, the use of convolutional neural network in loop closure detection has become more and m-ore popular In this paper, the influence of the structure and depth of Convoluti-onal Neural Network on the performance of loop closure detection algorithm is verified by comparing the convolutional descriptors extracted from different laye-rs of Convolutional Neural Network. Group-based experiments were carried out on three open data sets respectively to draw the curves of precision-recall rate, compare performance parameters, and obtain the robustness of the convolutiona-l image descriptors of different layers for scenes of seasonal change, lighting ch-ange and multi-interference, so as to find out the appropriate number of netwo-rk layers to complete loop closure detection. According to the experimental res-ults on the data set, the experimental conclusion is drawn.

Keywords

VSLAM; Loop Closure; CNN.

1. INTRODUCTION

Classical SLAM algorithms such as ORB-SLAM [1] use the bag of word model for loop closure detection, but the word of bag model relies on the features of artificial design. When the appearance of images changes greatly due to illumination, season, weather and other reasons, the algorithm based on the artificial bag of word model has poor performance. With the popularity of deep learning, more and more loop closure detection algorithms begin to use descriptors of convolutional neural network, such as [2-6].

Sunderhauf et al. [7] used AlexNet [8] to study the loop closure detection algorithms of different layers of convolutional neural network, and concluded that Conv3 had a good performance in the scene of serious appearance change, while FC6 had the best loop closure detection effect in the case of medium program appearance change and viewpoint change. Compared with AlexNet, VGG16[9] is suitable for studying the influence of the depth of neural network on the performance of loop closure detection algorithm because its model has more depth and simple structure.

2. PROPERTIES

We use the VGG16 [9] model to carry out our loop closure detection algorithm. Among them, VGG16 has 16 layers of network structure, 13 layers of convolution, and 3 layers of full

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202106_7(6).0059

connection. And the convolution kernel all uses 3x3 convolution kernel, and then uses 2x2 pooling and ReLU.

The VGG model we used was trained on the dataset Palces-365 [10], which contains 365 scene categories containing more than 1.8 million images. This makes the model have good generalization ability. The network structure of VGG16 is deep enough to well verify the effect of the results of different layers on the performance of the loop closure detection algorithm.

Loop closure detection is an indispensable link in SLAM, which can eliminate the localization error of robot. Loop closure detection requires the algorithm to determine whether two images are taken in the same scene. First, we adjust the input image to 224x224x3 and load it into VGG16, make the neural network propagate forward, take out the tensor of different layers and flatten it into a vector as the convolutional descriptor of the image, and then form a descriptor calculation library for loop closure detection. During loop closure detection, The query frame was first adjusted to 224x224x3 and then input to the network to take out the descriptors of different layers to calculate the cosine similarity and compare the performance of the closed-loop algorithm under the descriptors of different layers. Since it is necessary to compare the quality of the loop closure detection algorithm for performance inspection, we can judge it through the precision-recall curve. In algorithm design, we first need to extract the picture descriptors of different layers of VGG16, and also need the reference frame image database formed by the picture descriptors of different convolution layers with reference frames. Then, image descriptors of different convolutional layers of VGG16 of query frames are extracted, and similarity calculation is carried out with reference frame database one by one to obtain the picture with the highest similarity. The pictures of the same scene were selected according to this, and then compared with the ground truth to judge whether our algorithm was correct. Finally, the precision-recall curve was drawn. The calculation formula of precision rate and recall rate is as follows:

$$\text{Presicion}=\frac{TP}{TP+FP} \tag{1}$$

$$\text{Recall}=\frac{TP}{TP+FN} \tag{2}$$

3. THE EXPERIMENTAL RESULTS

3.1. Planning

According to the precision-recall curve, we can judge the performance of our loop closure detection algorithm from the two indicators in the figure, namely, AUC(Area Under Curve) , and the maximum recall rate with 100% accuracy, represented by r, which can be seen from the decline of the curve. We validate our algorithm on three public datasets to verify the robustness of the image descriptors obtained by the Convolutional Neural Network for different environments. We use conv1, conv2, conv3, conv4, and conv5 with fc6, fc7, fc8 to represent the convolutional layer and the fully connected layer of VGG16. Where, conv1 represents conv1-2 of VGG16, namely, the second convolution layer, and conv2 represents conv2_2 of VGG16, namely, the fourth convolution layer. Cov3 represents conv3-3 of VGG16, which is the seventh convolution layer. Similarly, Conv4 is the tenth convolution layer. Cov5 is the thirteenth convolution layer. And fc6, fc7 and fc8 are fully connected layers of 14, 15 and 16.

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202106_7(6).0059

3.2. The Nordland Dateset

The Nordland Dateset contains a 729km train journey video [11], which has The same location, spring, summer, autumn and winter video, as shown in Figure 1. This test dataset contains extreme appearance changes of the picture.

Figure 1. An example of the Nordland Dateset

We conducted three sets of experiments on this data set. First, two sets of spring a-nd summer pictures were used for loop closure detection, and then two sets of sprin-g and autumn pictures were used for loop closure detection. Finally, two groups of s-pring and winter photos were used for loop closure detection.

The spring/summer group: as shown in Figure 2, the maximum AUC was obtained by the loop closure detection with the conv4 descriptor.

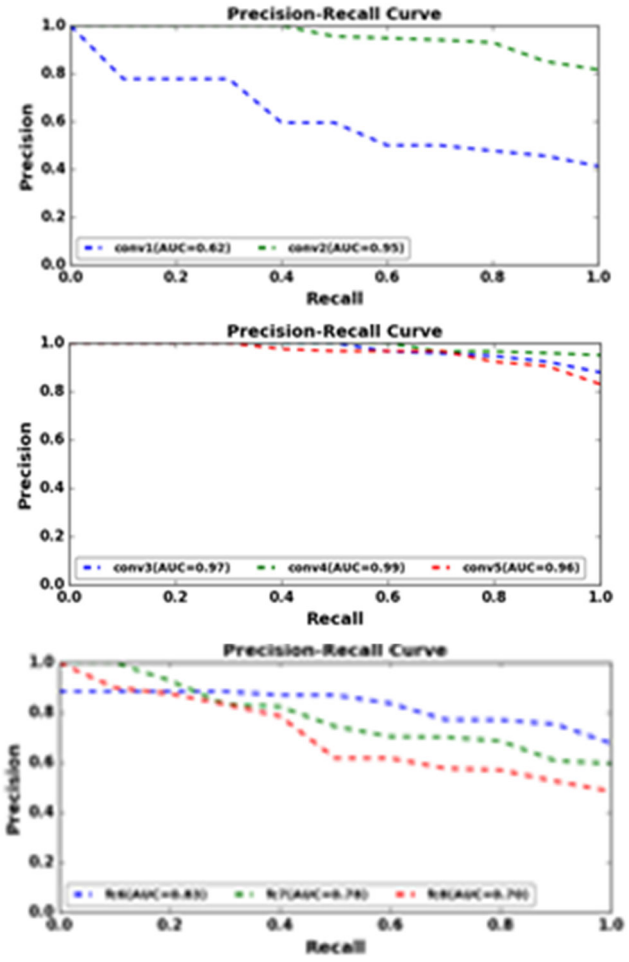

Figure 2. The precision-recall Curve

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202106_7(6).0059

The spring/autumn group: As shown in Figure 3, the maximum AUC was obtained by loop closure detection of conv3 and conv4 descriptors.

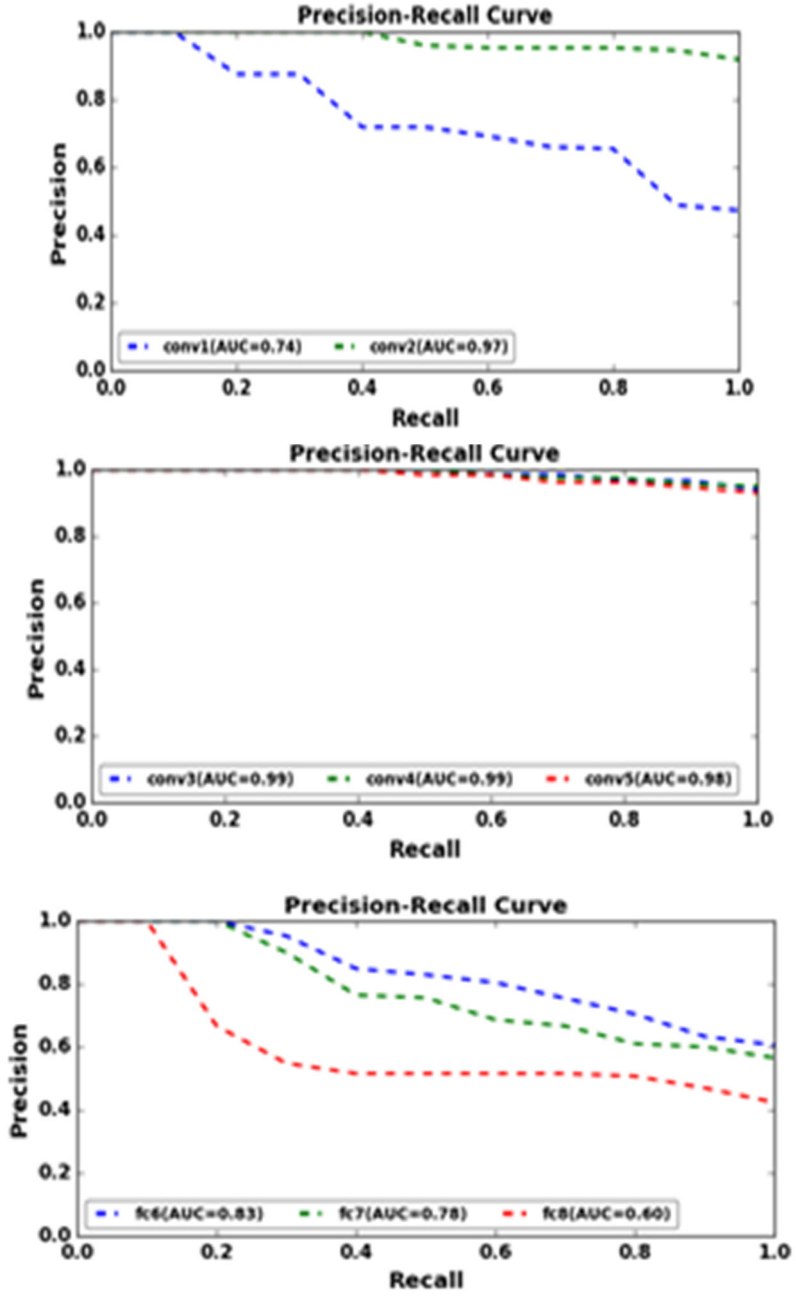

Figure 3. The precision-recall Curve

The spring/winter group: As shown in Figure 4, the maximum AUC was obtained by loop closure detection of conv4 descriptors.

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202106_7(6).0059

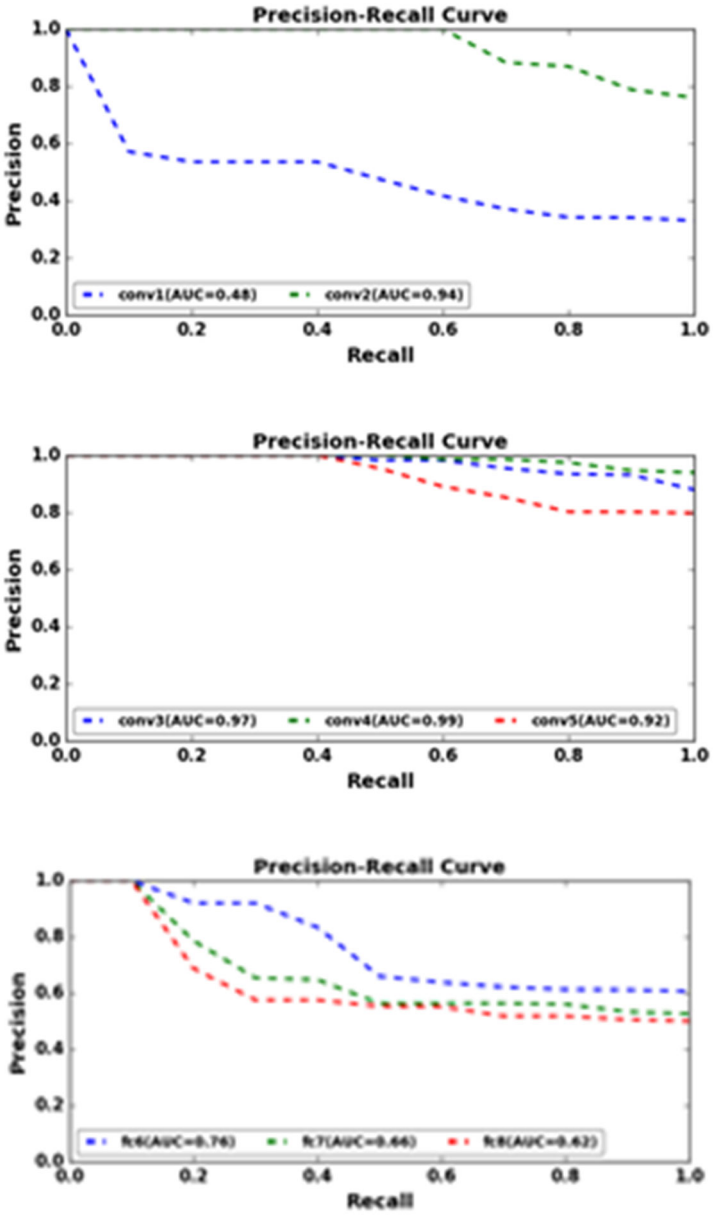

Figure 4. The precision-recall Curve

3.3. The Gardens point Dataset

The Gardens point Dataset consists of three sets of trajectories in the same environment [12]. Two of them were taken during the day and the other at night. Two sets of tracks during the day were filmed along the left and right side of the sidewalk for two days. The track at night is shot along the right side of the sidewalk, as shown in Figure 5.

Figure 5. An example of the Gardens Point Dataset

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202106_7(6).0059

We conducted two experiments on this dataset, the day left and day for a group, to test the robustness of image descriptors extracted from different layers of VGG to viewpoint changes. Day right and night right for a group. To test the robustness of image descriptors extracted from different layers of VGG to light changes.

The day left/day right group: As shown in Figure 6 and Table 1, the picture descriptors extracted by conv5, fc6 and fc7 were used in loop closure detection and have the maximum AUC which is 0.95, fc7 has the maximum r value 0.32.

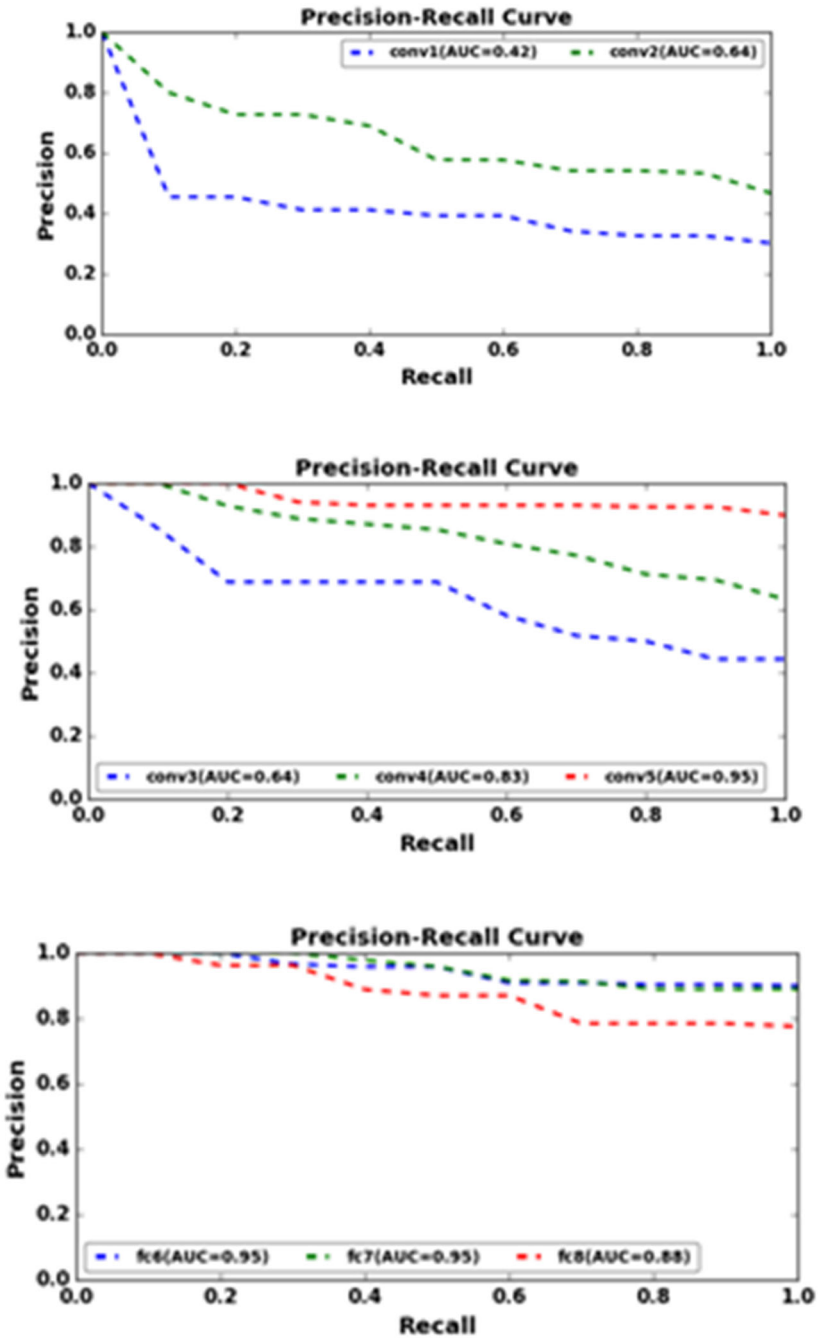

Figure 6. The precision-recall Curve

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202106_7(6).0059

Table 1. the r value

Name	CNNlayyers	r
1	Conv1	0.12
2	Conv2	0.23
3	Conv3	0.07
4	Conv4	0.17
5	Conv5	0.28
6	Fc6	0.30
7	Fc7	0.32
8	Fc8	0.13

The day right/night right group: As shown in Figure 7 and Table 2, the picture descriptors extracted from conv5 in loop closure detection have the maximum AUC and r vaule, which are 0.99 and 0.79, respectively.

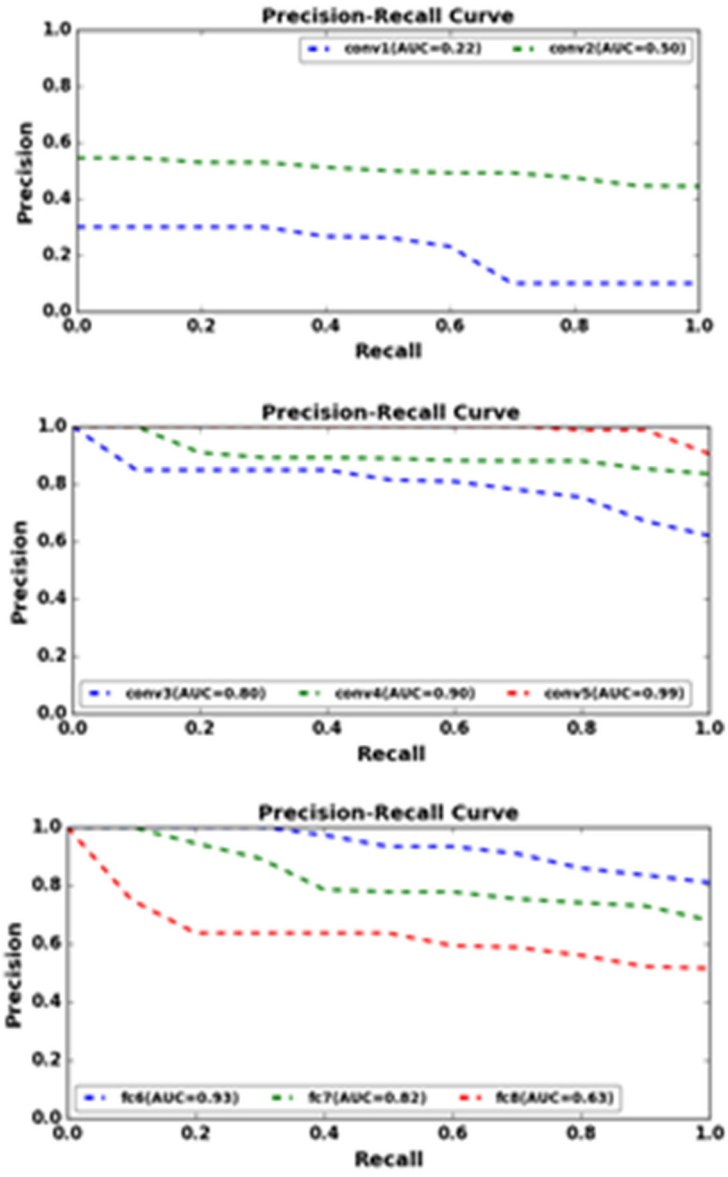

Figure 7. The precision-recall Curve

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202106_7(6).0059

Table 2. The r value

Name	CNNlayyers	r
1	Conv1	/
2	Conv2	/
3	Conv3	0.01
4	Conv4	0.14
5	Conv5	0.79
6	Fc6	0.35
7	Fc7	0.16
8	Fc8	0.03

3.4. The Campus Loop Dataset

The Campus Loop Dataset [6] contains more disturbances than the previous two datasets, including seasonal changes, weather changes, illumination changes, viewpoint changes and dynamic objects, as shown in Figure 8.

Figure 8. The Campus Loop Dataset

We performed a set of experiments on this dataset to test the robustness of image descriptors extracted from different layers of VGG in a multi-interference environment.

As shown in Figure 9, the maximum AUC of fc6 is 0.72, the highest r value of fc7 is 0.21, and the second highest value of fc6 is 0.21.

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202106_7(6).0059

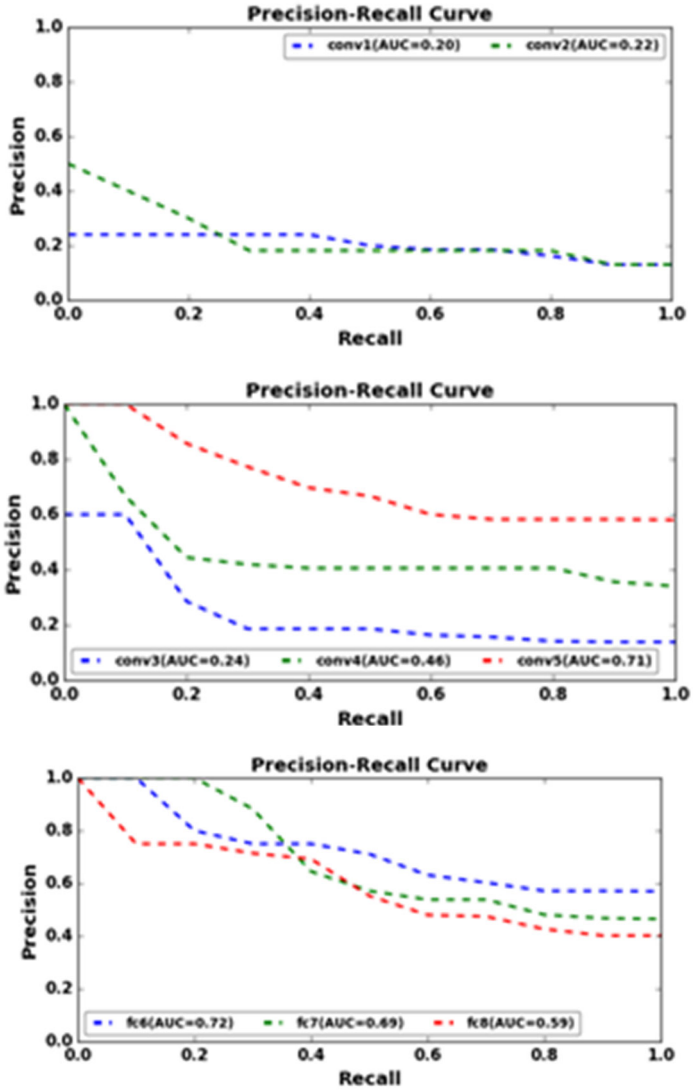

Figure 9. The precision-recall Curve

4. CONCLUSION

According to the above experiments, image descriptors extracted by conv4 performed best for scenes with severe changes in the appearance of the Nordland Dateset images. For The Gardenpoint DataSet, the image descriptors extracted from the viewpoint change group fc7 performed best. In the day and night group, picture descriptors extracted from conv5 performed best. For the Campus Loop Dataset, the image descriptors extracted from The multi-interference group, fc6 and fc7 performed best. Therefore, it can be concluded that the middle-level Conv4 picture descriptor of Convolutional Neural Network is only applicable to the scene of simple seasonal change, and conv5 picture descriptor is applicable to the scene of simple severe light change. As for the deep layer of convolutional neural network, because it contains more semantic information, fc6 and fc7 picture descriptors are suitable for scenes with viewpoint changes and multiple interference.

ACKNOWLEDGMENTS

This paper was supported by Applied Basic Research Project of Sichuan Province.

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202106_7(6).0059

REFERENCES

[1] Mur-Artal R, Tardós J D. Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras[J]. IEEE Transactions on Robotics, 2017, 33(5): 1255-1262

[2] Chen Z, Lam O, Jacobson A, et al. Convolutional neural network-based place recognition[J]. arXiv preprint arXiv:1411.1509, 2014

[3] Gomez-Ojeda R, Lopez-Antequera M, Petkov N, et al. Training a convolutional neural network for appearance-invariant place recognition[J]. arXiv preprint arXiv:1505.07428, 2015.

[4] Sünderhauf N, Shirazi S, Jacobson A, et al. Place recognition with convnet landmarks: Viewpoint-robust, condition-robust, training-free[J]. Robotics: Science and Systems XI:, 2015: 1-10.

[5] Camara L G, Přeučil L. Spatio-semantic conv-net-based visual place recognition[C]//2019 European Conference on Mobile Robots (ECMR). IEEE,2019: 1-8

[6] Merrill N, Huang G. Lightweight unsupervised deep loop closure[J]. arXiv preprint arXiv:1805.07703, 2018

[7] Sünderhauf N, Shirazi S, Dayoub F, et al. Onthe performance of convnet features for place recognition[C]//2015 IEEE/RSJ international conferen-ce on intelligent robots and systems (IROS). IEE-E, 2015: 4297-4304.

[8] Krizhevsky A , Sutskever I , Hinton G . ImageNet Classification with Deep Convolutional Neural Networks[C]// NIPS. Curran Associates Inc. 2012.

[9] Simonyan K , Zisserman A . Very Deep Convolutional Networks for Large-Scale Image Recognition[J]. Computer Science, 2014.

[10] Zhou B, Lapedriza A, Khosla A, et al. Places: A 10 million image database for scene recognition[J]. IEEE transactions on pattern analysis and machine intelligence, 2017, 40(6): 1452-146

[11] Olid D, Fácil J M, Civera J. Single-view place recognition under seasonal changes[J]. arXiv preprint arXiv:1808.06516, 2018.

[12]]Sünderhauf N, Shirazi S, Dayoub F, et al. Onthe performance of convnet features for place recognition[C]//2015 IEEE/RSJ international conferen-ce on intelligent robots and systems (IROS). IEE-E, 2015: 4297-4304.