Proceeding of the IEEE
International Conference on Robotics and Biomimetics
Dali, China, December 2019

A DenseNet feature-based loop closure method for visual SLAM system*

Chao Yu
Department of Electronic Engineering
Tsinghua University
Beijing, China
yc19@mails.tsinghua.edu.cn

ZuXin Liu
School of Opto-electronics Engineering
Beihang University
Beijing, China
xinye@buaa.edu.cn

Xin-Jun Liu
Department of Mechanical Engineering
Tsinghua University
Beijing, China
xinjunliu@mail.tsinghua.edu.cn

Fei Qiao and Yu Wang
Department of Electronic Engineering
Tsinghua University
Beijing, China
qiaofei, yu-wang@mail.tsinghua.edu.cn

Fugui Xie
Department of Mechanical Engineering
Tsinghua University
Beijing, China
xiefg@mail.tsinghua.edu.cn

Qi Wei and Yi Yang
Department of Electronic Engineering
Tsinghua University
Beijing, China
weiqi, yangyy@mail.tsinghua.edu.cn

Abstract— Loop closure is a crucial part in SLAM, especially for large and long-term scenes. Utilizing off-the-shelf networks' features in loop closure becomes a hot spot. However, what kind of network is more suitable in loop closure and how to use their features have not been well-studied. In this paper, DenseNet is introduced in this field according to its own characters. The features of DenseNet preserve both semantic information and structure details and outweigh other popular networks' features significantly. Based on this, a DenseNet feature-based framework, named Dense-Loop, is proposed to address the loop closure problem. Weighted Vector of Locally Aggregated Descriptor (WVLAD) method is used to encode the local descriptors as the final global descriptor, which could resist geometry structure and viewpoint changes. Furthermore, 4 max-pooling by channel and locality-sensitive hashing (LSH) are adopted to accelerate the search process. Extensive experiments are conducted on public datasets and the results demonstrate Dense-Loop could achieve state-of-the-art performance.

Index Terms— Convolutional Neural Network, loop closure, DenseNet, SLAM

I. Introduction

Loop closure is a basic part in SLAM for mobile robots [3]. If a robot could determine whether a place has been visited before, it can use this message to correct the error and drift accumulated in the simultaneous localization and mapping (SLAM) process [4], [5].

However, this problem can be very challenging. For example, the same place may have different appearances at different time due to illumination or viewpoint changes. In addition, two different places may have the similar texture and appearance. A false positive recognition of a place may corrupt the global optimization process and cause unrecoverable localization and mapping failure [6].

*This work is partially supported by the National Natural Science Foundation of China under Grant 91648116 and 51425501.

Many effective methods have been proposed to solve loop closure problem in robotics field. One of the most prevalent methods is visual bag-of-words (BoWs) [5], [6], which treats descriptors of local features as visual words. This kind of method can achieve good performance in SLAM, and it is robust against viewpoint changes. However, the hand-crafted features can hardly deal with environment changes, such as illumination changes and similar textured regions [6], [7], [8].

In recent years, many researchers find the features extracted from Convolutional Neural Network(CNN) have better performance than hand-crafted features [9] and begin to investigate how to use these features in loop closure [10], [11], [12], [13], [14]. Even so, the research in this field is preliminary and incomplete, partially because of the weak interpretability of neural network.

When talking about these issues, We will at least think about such questions: First of all, there are numerous outstanding neural network architectures, which one is more suitable for loop closure and why? Secondly, CNN features vary from hand-crafted features in respect of the quantity and dimension. Is traditional loop closure detection framework (such as BoWs) suitable for CNN features? If not, what is the better solution?

In this paper, we try to dig into these issues further. The main contributions include:

1) We compare many off-the-shelf networks and find DenseNet outweighs other popular networks in loop closure, because this dense-connected network could preserve both semantic information and structure details of the input image.
2) A loop closure framework (Dense-Loop) using DenseNet features is proposed in this paper. Decoupling by feature-maps (DBF) and Weighted

978-1-7281-6321-5/19//$31.00 ©2019 IEEE

Vector of Locally Aggregated Descriptor (WVLAD) method is utilized to make full use of DenseNet features according to its own distinctions.

3) Extensive experimental results show Dense-Loop could achieve state-of-the-art performance on public datasets.

In the rest of the paper, the structure is as follows. Section 2 briefly introduces some current accomplishments of loop closure. Section 3 describes the proposed framework in detail. Subsequently, extensive comparative experiments and evaluation are presented in Section 4. Finally, a brief conclusion and the future work are summarized in Section 5.

II. Related works

We categorize current accomplishments on loop closure into three groups: traditional hand-crafted feature-based approaches, end-to-end training approaches, and off-the-shelf CNN features approaches.

Many well-designed local features are widely used in place recognition and loop closure tasks because their ability to resist scale changes or orientation changes. One of the most successful use is FAB-MAP, which employs SUFT [15] and BoWs for place recognition and demonstrates robust performance against viewpoint changes [6]. Reference [5] integrates ORB [16] and BoWs in SLAM. This kind of method becomes the most popular framework to detect loop closure in real-time visual SLAM systems. However, these hand-crafted features only care about low-level information of the image and can hardly deal with environment changes, such as illumination changes. Furthermore, these statistics based methods' performance depends heavily on the quality of the features and may be easily deceived by the textured dynamic objects in the environment.

Considering the shortcomings of the hand-crafted features, a recent trend in loop closure is to train a CNN network in an end-to-end manner. NetVLAD [12] is a novel architecture which aims to minimize the distance of two image representations of the same place. The training images are categorized into many tuples, where each training query image has corresponding potential positive samples and definite negative samples. Reference [17] adopts the similar triplet training scheme and could produce a 128 dimension descriptor vector for each image. However, all of these supervised learning approaches require a large amount of labeled datasets to train. It is also a bottleneck for others to use the network for their own needs.

Another trend is to exploit the learned features of the off-the-shelf networks with pre-trained weights. Reference [10] employs CNN features based on OverFeat for place recognition. The performance of feature-maps of different layers is explored. Reference [18] focuses on using AlexNet to generate an image representation appropriate for visual loop closure in SLAM. They find CNN features outperform hand-crafted features when illumination changes significantly.Reference [11] deploys pre-trained AlexNet as CNN features and using locality-sensitive hashing and semantic search space partitioning optimization techniques to ensure real-time search. These kind of methods do not require specific end-to-end training and thus are more convenient. The feature could be extracted without interference to the pre-trained networks that designed for other tasks. However, since there are numerous outstanding network architectures in recent years, which one is better and how to make good use of its inner features have not been fully explored.

In this paper, we will explore what kind of network is more suitable in loop closure and how to use them to achieve better performance without specific supervised training.

III. Framework of Dense-Loop

The pipeline of Dense-Loop is shown in Figure 1, where C, H, W represent the dimension of the channel, the height and weight of feature-maps. K is the number of cluster centers and D represents the dimension of one cluster center.

A. Descriptors Extraction

In the traditional BoWs, a lot of disordered local descriptors with low dimensions are extracted and they are designed to resist scale or viewpoint changes. However, CNN features are ordered and 3-dimension. Therefore, the first thing is to exact good features from CNN and map them to 2-dimension.

1) DenseNet Features: DenseNet [22] is a compact network and made up of dense blocks. All layers in one dense block are directly connected to ensure maximum information flow between feature-maps. The input of each layer is all the preceding layers' output, and thus, the block's final classifier could obtain all the information of the previous feature-maps. This kind of compact internal representation could reduce feature redundancy and help to solve vanishing-gradient problem. DenseNet adopted in Dense-Loop is made up of 5 dense blocks [22]. The output of ReLu layer in the last dense block is used as the raw features of the input image, where 7×7 is the size of feature-maps and 1024 is the number of channels. The reason for choosing the ReLu layer is that it is cleaner and contains less noise.

The reason of using DenseNet is its reuse of feature-maps. The features of low layers contain more structural information and measure fine-grained similarity, which is similar to hand-crafted features. While the features of higher layers care more about semantic information and

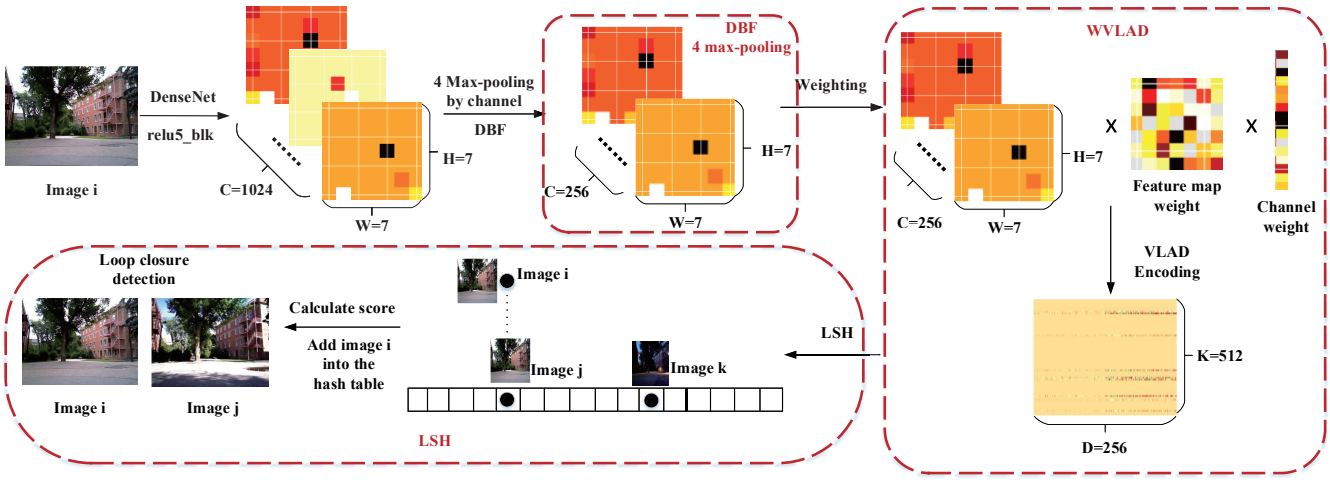


Fig. 1. The pipeline of Dense-Loop

measure semantic similarity. A natural idea is to utilize the complementary of high-layer and low-layer features. The outputs of last few layers preserve all extracted features of preceding layers, which means, the low-level features and high-level features are merged together in an efficient way. It is helpful for more fine-grained features expression of an image. The superiority of DenseNet will be illustrated in the experiment section in detail.

2) Decoupling By Feature-maps (DBF): Here are two ways to map these features to 2-dimension, as shown in Figure 2. One is decomposing the global feature into 49 local descriptors with 1024 dimensions, called decoupling by feature-maps (DBF). Anther way is to decompose 1024 local descriptors with 49 dimensions, called decoupling by channel (DBC). The former plan is chosen because it is of physical meaning, and it has better performance than DBC. Each pixel in the feature-map is corresponding to a receptive field in the input image, and all the channels of the pixel could describe the distinctions of the corresponding receptive field. As for DBC, it's more like using many global descriptors to describe an image. But image's viewpoint change may cause a shift in the feature-maps and thus the ability to resist geometry structure or viewpoint changes will be weaken.

B. Weighted Vector of Locally Aggregated Descriptor (WVLAD)

In the traditional BoWs, BoW encoding method is used to measure the similarity of two images. BoWs is a statistical method and usually needs a large number of visual words (e.g. 10^6) in the dictionary. A lot of local descriptors with low dimensions are more suitable in this situation, while the CNN descriptors, which are

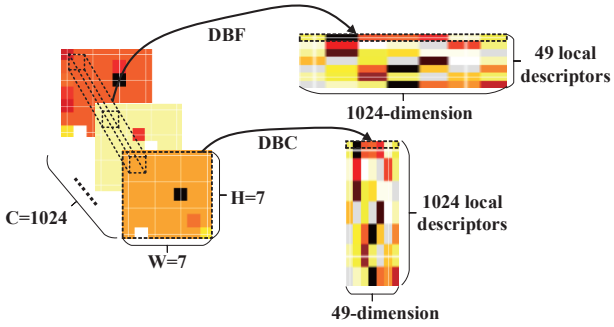


Fig. 2. The description of DBF and DBC.

decoupled by feature-maps, often have small quantity but large dimensions. Besides, it is hard to train such a huge BoW dictionary. Instead, Weighted Vector of Locally Aggregated Descriptor (WVLAD) is proposed in this paper to encode the 49×256 local descriptors of an image (256 can be get after 4 max-pooling by channel, which will be described in next part).

WVLAD could ignore the geometric structure of the image via clustering and care more about the distinctions via weight. Therefore, it's more resistant to viewpoint and scale changes than calculating euclidean distance of CNN features. It is an improved method of famous Vector of Locally Aggregated Descriptor (VLAD) [20] method and inspired by Cross-dimensional Weighting for Aggregated Deep Convolutional Features (CROW) [21] method.

Usually we want the descriptors care more about the distinctions of the image and reduce the importance of the plain areas (e.g. sky). It's similar to the human perception system, which is conducive to improving resistance to environment changes. One way is to use

region proposal methods and compute regions' descriptors respectively. Another way is to adopt the self-adaptive weight methods to adjust the importance of the textured regions and ordinary areas. The first way is computational expensive. Considering the need for real time, the second way is integrated in Dense-Loop. Figure 3 shows the detailed process of calculating the feature-maps weight (FW) and the channel weight (CW).

The strong response of convolution is usually corrsponding to the region of objects. FW can force features to focus on the textured regions and help solving scale changes. Let $F \in \mathbb{R}^{(C\times H\times W)}$ denotes the 3-dimension features of the inner layer. $X \in \mathbb{R}^{(H\times W)}$ represents one feature-map. c, h, w is the location of the feature vector. $FW \in \mathbb{R}^{(H\times W)}$ can be calculated by summing feature-maps of all channels. Then L2-norm and a power normalization with power 0.5 are utilized to get aggregated feature-maps weight.

$$S = \sum_c X_c \tag{1}$$

$$S' = \sqrt{\sum_{h,w} S_{h,w}^2} \tag{2}$$

$$FW = \sqrt{S/S'} \tag{3}$$

$CW \in \mathbb{R}^{(1\times C)}$ is similar to the idea of inverse documentary frequency (IDF) in BoWs, that is, reducing the importance of high-frequent features.

$$T_c = \frac{\sum_{X_{h,w}>0} 1}{H \times W} \tag{4}$$

$$CW_c = \begin{cases} \log(\frac{\sum_{c=1}^{C} T_c}{T_c}), T_c > 0 \\ 0, T_c = 0 \end{cases} \tag{5}$$

Then, we can calculate the weighted feature-maps $F_{weight} \in \mathbb{R}^{(C\times H\times W)}$. And decompose it into weighted local descriptors L, which means 49 local features with 256 dimensions.

$$F'_c = F_c \times FW \tag{6}$$

$$F_{weight} = F'_{c,h,w} \times CW_c \tag{7}$$

In order to improve the ability of resisting geometry structure or viewpoint changes, VLAD is used to encode weighted local descriptors as a global descriptor. Firstly K-means is used to cluster all the weighted local descriptors of the datasets and get the codebook $\{u_1, ..., u_K\}$, where K is the number of cluster centers. Each local descriptor L_i has its corresponding cluster center u_j: $NN(L_i) = argmin_j \|L_i - u_j\|$, where NN represents nearest neighbor. VLAD is denoted as a set of vector $V = [v_1^T, ..., v_K^T]$, where each v_i is associated with a cluster center u_i and has the same size. Then V is calculated by the concatenation of the residual of each L_i and $NN(L_i)$:

$$v_i = \sum_{L_t : NN(L_t)=i} L_t - u_i \tag{8}$$

Finally, a power normalization with power 0.5 and L2-norm is utilized to normalize V.

C. Fast Search

In order to ensure accelerate the search process, a method called 4 max-pooling by channel is proposed to reduce the descriptors' dimensions with minimal accuracy reduction. 1024-dimension descriptors are divided into 256 groups and the maximum value of each group is used as the final descriptor. Compared with PCA, which is widely used to reduce dimensions, 4 max-pooling by channel has less computational complexity but similar performance. More results can be found in the experimental part.

In the traditional BoWs, K-D tree is adopted to accelerate the search process. However, the spatial dimension of Dense-Loop descriptors is far more than the number of words in the codebook, K-D tree will be unsuitable anymore. Instead, locality-sensitive hashing (LSH) is employed to speed up the search with minimal accuracy degradation.The detailed process is shown in Figure 1. The Hamming distance between the respective hashed bit vectors is used to evaluate the similarity. According to our test, using 1024 bits retains approximately 99% performance but much more quick than brute search.

In summary, Dense-Loop can be presented as follows: First of all, the output of ReLu layer in the last dense block of DenseNet is adopted as the initial features and decoupling by feature-maps (DBF) is utilized to decompose the global feature into local descriptors. Then, 4 max-pooling by channel is adopted to reduce the computational complexity. Finally, Weighted Vector of Locally Aggregated Descriptor (WVLAD) method is proposed to improve the ability of resisting scale or viewpoint changes. To accelerate the searching process, locality-sensitive hashing (LSH) [19] is employed according to the characteristic of Dense-Loop descriptors.

IV. Experimental Results and Explanations

A. Datasets and Evaluation

City Center dataset [6] and New College dataset [6] are widely used in visual SLAM research and loop closure detection evaluation in particular. The former dataset has many dynamic objects like pedestrians and vehicles. Besides, the sunlight, wind and viewpoint change may cause the features like shadow unstable. The latter New College dataset has many dynamic elements and repeated elements, such as similar walls and bushes. Ground truth

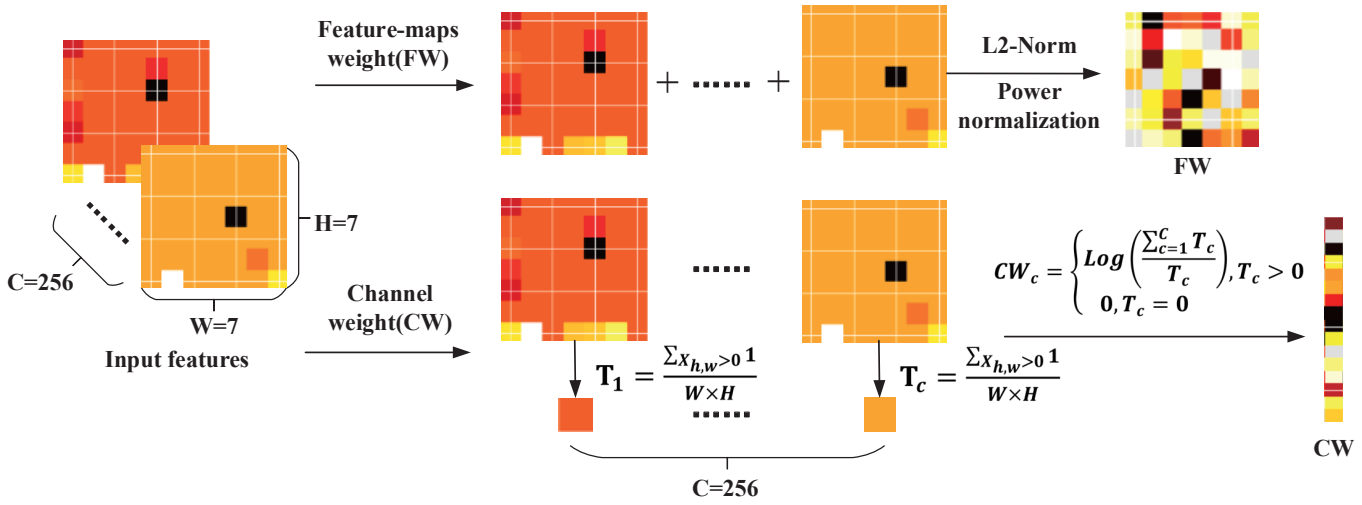
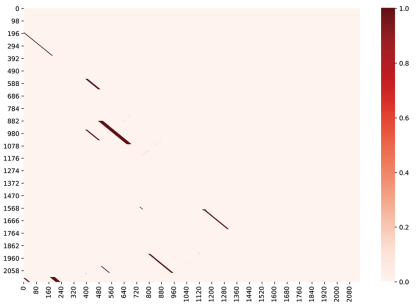
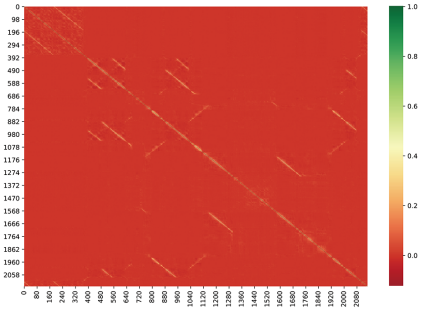


Fig. 3. The detailed process of calculating FW and CW.

are given in two datasets. Figure 4 shows the ground truth and the results of Dense-Loop.



(a) Ground truth



(b) Results of Dense-Loop

Fig. 4. The ground truth and the results of Dense-Loop on New College Dataset. Pixel (i, j) represents the relationships of image i and image j.

However, the provided ground truth can't be used directly. It's inconsistent with the goal of loop closure detection because we only need to identify one loop in the same place. Therefore, new definition of the true loop are made based on the original ground truth. The images in one dataset are divided into two groups, named left and right, and so is the ground truth. If a loop is detected, we will stop searching loop in 10 images (according to GPS) to avoid getting the same loop. When we vary the threshold if a loop closure is accepted, the precision and recall value will change and the PR-Curve can be gained.

B. Experiments and evaluation

Some comparative experiments are conducted to explore the validity of Dense-Loop. Dense-Loop could achieve state-of-the-art performance on public datasets, The reason can be summarized as two points. One is excellent features from DenseNet, which take high-level semantic information and fine-grained information into account. Another is WVLAD method, which could ignore the geometric structure of the image via clustering and care more about the distinctions via weight.

1) Why DenseNet?: In recent years, there are many prevalent and excellent convolutional networks showing up, such as ResNet50 [23], VGG [24], DPN [25], SENet [26], ResNeXt [27], NasNet [28], SqueezeNet [29], Xception [30], Inceptionv3 [31], Inceptionv4 and Inception-ResNet [32]. To verify the excellent features of DenseNet, extensive comparative experiments were conducted. Figure 5 exhibits the PR-Curves of different networks on New College dataset. Curves are named by the following formats: network name_layer name. For example, DenseNet_relu5_blk represents the features extracted from relu5_blk layer of DenseNet. All the networks are pre-trained on the ImageNet2012 dataset and euclidean distance is adopted as the similarity score. The layer with best performance in each network is chosen to draw in the figure and it is apparent that DenseNet outweighs other popular network

architectures.

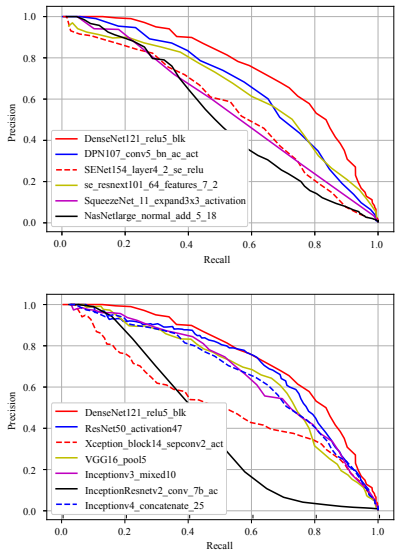


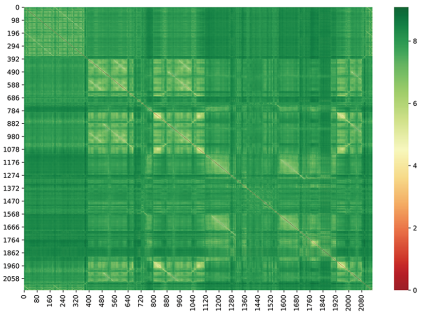
Fig. 5. The PR-Curves of different networks on New College dataset.

Figure 6 shows the euclidean distance of images on New College dataset when employing DenseNet and Xception respectively. The high-level features of Xception, which care more about semantic information, have a poorer discrimination on images than those of DenseNet. A common method to combine various levels' features is to concatenate them directly, but DenseNet aleady did this during the forward processing. The output of the last few layers integrate both low-level and high-level features naturally.

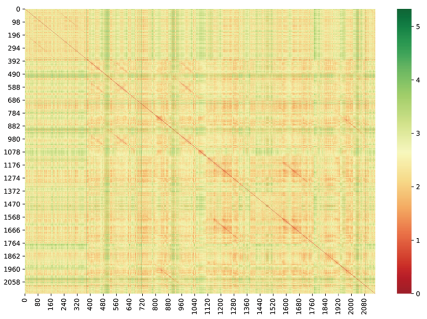
2) Why DBF?: Figure 7 shows the PR-Curves of DBF and DBC on City Center dataset. In order to make a quick comparison, euclidean distance is adopted as the similarity score. It is obvious that DBF far outweighs DBC and similar results can be gained on New College dataset.

3) Why 4 max-pooling by channel?: Figure 8 illustrates the PR-Curves of different dimensionality reduction methods on City Center dataset. The label named relu5_blk means the original features without dimensionality reduction. The label named 4 max-pooling by channel represents applying 4 max-pooling to the feature's channel dimension. The label named 256 PCA means reducing the channel dimension to 256 through PCA method. We can observe that utilizing 4 max-pooling by channel can maintain 99% accuracy and have alomost the same performance as PCA. Considering the processing time, 4 max-pooling by channel is adopted finally.

4) Why WVLAD?: In order to compare the performance with traditional methods, two hand-crafted



(a) DenseNet



(b) Xception

Fig. 6. The euclidean distance of images on New College Dataset.

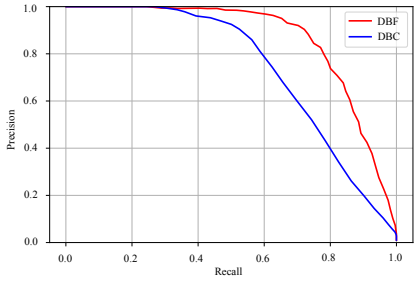


Fig. 7. The PR-Curves of DBF V.S. DBC on City Center dataset.

features (ORB and SIFT) and two encoding methods (BoW and VLAD) are adopted. The VLAD codebooks have 512 cluster centers, just the same as Dense-Loop, while BoW codebooks have 10000 visual words. The results on two datasets are shown in Figure 9.

It's clear that WVLAD could achieve better performance than BoW and VLAD encoding method based on DenseNet. And we can notice Dense-Loop far outweighs hand-crafted features. Here are two typical examples. In Figure 10(a) and 10(b), high similarity score is obtained based on hand-crafted features because of similar textured regions on the trees and sky, while score of Dense-Loop is close to zero in this case. This is because Dense-Loop could utilize high-level semantic and global information to judge the similarity. In Figure 10(c) and 10(d), Dense-Loop can recognize the two images as the

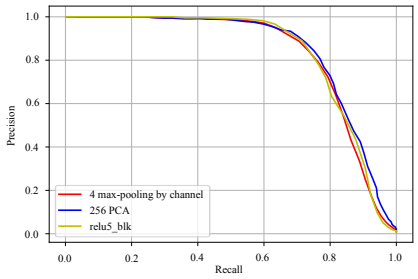
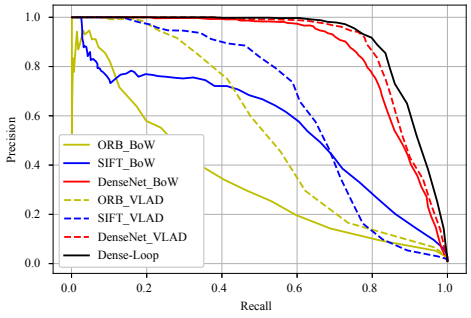
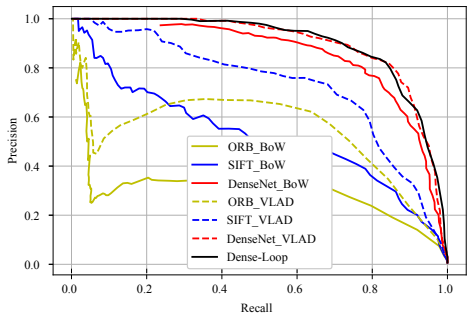


Fig. 8. The PR-Curves of different dimensionality reduction methods on City Center dataset.

same place with high score but hand-crafted features can't achieve that due to illumination changes. Besides, in this case, we can also find Dense-Loop can resist the viewpoint changes. As for WVLAD and VLAD, WVLAD can reduce channel redundancy by CW and focus on the distinguished and unique parts of the image by FW. Therefore, better performance can be obtained in some cases by solving the problem of scale and viewpoint changes.



(a) City Center dataset



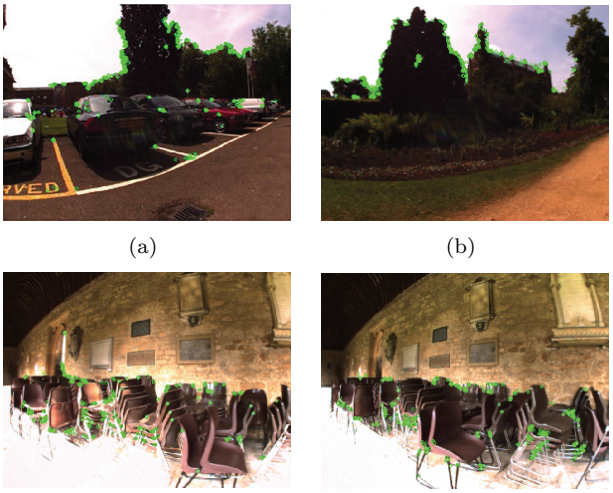
(b) New College dataset

Fig. 9. The PR-Curves of DenseNet V.S. (ORB, SIFT) and Dense-Loop V.S. (BoW, VLAD)

V. Conclusion

Loop closure detection is used to detect if the robot has passed through the same place. It's crucial for the robot to establish a globally consistent map, especially for large and long-term scenes. A framework of loop closure detection based on CNN features is proposed in this paper. We find that features extracted from DenseNet outweigh hand-crafted features and other popular networks' features. The reason is DenseNet can preserve both semantic information and structure details of the input image via dense connection. In order to improve the ability of resisting scale or viewpoint changes, decoupling by feature-maps (DBF) and Weighted Vector of Locally Aggregated Descriptor (WVLAD) method is utilized to make full use of DenseNet features according to its own distinctions. Locality-sensitive hashing (LSH) and 4 max-pooling by channel are adopted to ensure the real-time search for robotic application. Extensive experiments illustrate Dense-Loop approach could achieve state-of-the-art performance on public datasets.

However, the impact of the training datasets on the network's performance has not been investigated. In the future, we will conduct more extensive experiments to explore the generalization ability of Dense-Loop, which is important in real-world robot applications. Besides, we would consider to utilize semantic information of the network's prediction results and establish a multi-level semantic knowledge base to speed up the search and improve the loop closure detection performance.



(a) (b) (c) (d)

Fig. 10. Picture (a) and (b) with ORB features come from different scenes, but they share similar textured regions (e.g. trees and sky). Picture (c) and (d) with ORB features come from the same place, but they have different apparences, such as illumination changes.

Acknowledgment

This work was supported in part by the National Natural Science Foundation of China under Grant 91648116 and 51425501.

References

[1] C. Yu, Z. Liu, X. Liu, F. Xie, Y. Yang, Q. Wei, and Q. Fei, "Ds-slam: A semantic visual slam towards dynamic environments," in 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Oct 2018, pp. 1168–1174.

[2] Y. Xiang and D. Fox, "DA-RNN: semantic mapping with data associated recurrent neural networks," CoRR, vol. abs/1703.03098, 2017. [Online]. Available: http://arxiv.org/abs/1703.03098

[3] S. Lowry, N. Sünderhauf, P. Newman, J. J. Leonard, D. Cox, P. Corke, and M. J. Milford, "Visual place recognition: A survey," IEEE Transactions on Robotics, vol. 32, no. 1, pp. 1–19, 2016.

[4] M. Labbe and F. Michaud, "Appearance-based loop closure detection for online large-scale and long-term operation," IEEE Transactions on Robotics, vol. 29, no. 3, pp. 734–745, June 2013.

[5] R. Mur-Artal and J. D. Tardós, "Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras," IEEE Transactions on Robotics, vol. 33, no. 5, pp. 1255–1262, 2017.

[6] M. Cummins, "Fab-map : Probabilistic localization and mapping in the space of appearance," The International Journal of Robotics Research, vol. 27, no. 6, pp. 647–665, 2008.

[7] B. Upcroft, C. Mcmanus, W. Churchill, and W. Maddern, "Lighting invariant urban street classification," in IEEE International Conference on Robotics and Automation (ICRA). Hong Kong, China: IEEE, 2014, pp. 1712–1718.

[8] S. Garg, N. Suenderhauf, and M. Milford, "Don't look back: Robustifying place categorization for viewpoint- and condition-invariant place recognition," in 2018 IEEE International Conference on Robotics and Automation (ICRA), May 2018, pp. 3645–3652.

[9] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in Proceedings of the 25th International Conference on Neural Information Processing Systems (NIPS), ser. NIPS'12. USA: Curran Associates Inc., 2012, pp. 1097–1105.

[10] Z. Chen, O. Lam, A. Jacobson, and M. Milford, "Convolutional neural network-based place recognition," CoRR, vol. abs/1411.1509, 2014.

[11] N. Sünderhauf, S. Shirazi, F. Dayoub, B. Upcroft, and M. Milford, "On the performance of convnet features for place recognition," in 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Sept 2015, pp. 4297–4304.

[12] R. Arandjelović, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "Netvlad: Cnn architecture for weakly supervised place recognition," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 40, no. 6, pp. 1437–1451, June 2018.

[13] N. Sünderhauf, S. Shirazi, A. Jacobson, F. Dayoub, E. Pepperell, B. Upcroft, and M. Milford, "Place recognition with convnet landmarks: Viewpoint-robust, condition-robust, training-free," in Robotics: Science and Systems (RSS), Auditorium Antonianum, Rome, July 2015.

[14] D. Bai, C. Wang, B. Zhang, X. Yi, and Y. Tang, "Matching-range-constrained real-time loop closure detection with cnns features," Robotics and Biomimetics, vol. 3, no. 1, p. 15, Sep 2016.

[15] H. Bay, A. Ess, T. Tuytelaars, and L. V. Gool, "Speeded-up robust features (surf)," Computer Vision and Image Understanding, vol. 110, no. 3, pp. 346–359, 2008.

[16] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "Orb: An efficient alternative to sift or surf," in Proceedings of the 2011 International Conference on Computer Vision (ICCV), ser. ICCV '11. Washington, DC, USA: IEEE Computer Society, 2011, pp. 2564–2571.

[17] M. Lopez-Antequera, R. Gomez-Ojeda, N. Petkov, and J. Gonzalez-Jimenez, "Appearance-invariant place recognition by discriminatively training a convolutional neural network," Pattern Recognition Letters, vol. 92, pp. 89–95, 2017.

[18] Y. Hou, H. Zhang, and S. Zhou, "Convolutional neural network-based image representation for visual loop closure detection," in 2015 IEEE International Conference on Information and Automation (ICInfA), Aug 2015, pp. 2238–2245.

[19] D. Ravichandran, P. Pantel, and E. Hovy, "Randomized algorithms and nlp: Using locality sensitive hash function for high speed noun clustering," in Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, ser. ACL '05. Stroudsburg, PA, USA: Association for Computational Linguistics, 2005, pp. 622–629.

[20] H. Jégou, M. Douze, C. Schmid, and P. Pérez, "Aggregating local descriptors into a compact image representation," in 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), June 2010, pp. 3304–3311.

[21] Y. Kalantidis, C. Mellina, and S. Osindero, "Cross-dimensional weighting for aggregated deep convolutional features," in Computer Vision – ECCV 2016 Workshops, G. Hua and H. Jégou, Eds. Cham: Springer International Publishing, 2016, pp. 685–701.

[22] G. Huang, Z. Liu, L. v. Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 2017, pp. 2261–2269.

[23] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2016, pp. 770–778.

[24] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," CoRR, vol. abs/1409.1556, 2014.

[25] Y. Chen, J. Li, H. Xiao, X. Jin, S. Yan, and J. Feng, "Dual path networks," CoRR, vol. abs/1707.01629, 2017.

[26] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," CoRR, vol. abs/1709.01507, 2017.

[27] S. Xie, R. Girshick, P. Dollar, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), vol. 00, July 2017, pp. 5987–5995.

[28] B. Zoph, V. Vasudevan, J. Shlens, and Q. V. Le, "Learning transferable architectures for scalable image recognition," CoRR, vol. abs/1707.07012, 2017.

[29] F. N. Iandola, M. W. Moskewicz, K. Ashraf, S. Han, W. J. Dally, and K. Keutzer, "Squeezenet: Alexnet-level accuracy with 50x fewer parameters and <1mb model size," CoRR, vol. abs/1602.07360, 2016.

[30] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), vol. 00, July 2017, pp. 1800–1807.

[31] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," CoRR, vol. abs/1512.00567, 2015.

[32] C. Szegedy, S. Ioffe, and V. Vanhoucke, "Multi-scale orderless pooling of deep convolutional activation features," in Proceeding of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI), 2017, pp. 4278–4284.