Dense-Loop: A Loop Closure Detection Method for Visual SLAM using DenseNet Features

Chao Yu
Tsinghua University, Beijing, China 100084
yc19@mails.tsinghua.edu.cn

ZuXin Liu
Beihang University, Beijing, China, 100191
xinye@buaa.edu.cn

Xin-Jun Liu*, Fei Qiao*, Yu Wang, Fugui Xie, Qi Wei,Yi Yang
Tsinghua University, Beijing, China 100084
{xinjunliu, qiaofei, yu-wang, xiefg, weiqi, yangyy}@mail.tsinghua.edu.cn

Abstract

Loop closure detection (LCD) is an important part in SLAM for the autonomous mobile robot. A recent trend is to employ off-the-shelf networks' features to address LCD problem, which outperform traditional hand-crafted features. However, what kind of network is more suitable in LCD and how to use their CNN features have not been well-studied. In this paper, we compare many popular networks and introduce DenseNet in this field. The features extracted by DenseNet, which preserve both semantic information and structure details, outweigh other popular CNN features signicantly. Then a DenseNet feature-based framework (Dense-Loop) is proposed to address the LCD problem. We use the Weighted Vector of Locally Aggregated Descriptor (WVLAD) method 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 ensure the real-time search. Extensive experiments are conducted on public datasets using Precision-Recall Curve evaluation method. The results demonstrate Dense-Loop could achieve state-of-the-art performance.

1 Introduction

In recent years, the combination of semantics and SLAM has become a research hotspot, and many related works have appeared, such as DS-SLAM[YLL+18], DA-RNN[XF17] and so on. Most of these SLAM systems utilize semantics in Visual Odometry (VO) and Mapping, while introducing semantic information into loop closure detection (LCD) is indispensable and requires further research.

Visual place recognition is a basic part in re-localization and loop closure detection for mobile robots[LSN+16]. If the robot could determine whether an image of a place has been visited before, then this information could help the robot re-localize itself, or correct the error and drift accumulated in the simultaneous localization and mapping (SLAM) process[LM13, MAT17].

However, this problem is very challenging. On the one hand, the same place may have different appearances at different time due to the illumination or viewpoint changes. On the other hand, 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 severer unrecoverable localization and mapping failure[Cum08].

Many effective methods have been proposed to solve loop closure detection problem in robotics field. One of the most prevalent methods is visual bag-of-words (BoWs)[MAT17, Cum08], which treats descriptors of local features as visual words. This kind of method can achieve good performance on place recognition, and it is robust against viewpoint changes. However, the hand-crafted features can hardly deal with environment changes, such as the illumination changes and similar textured regions[Cum08, UMCM14, GSM18].

Recently, many researchers have found the features extracted from off-the-shelf convolutional neural networks (CNN) have better performance than hand-crafted features[KSH12] and began to investigate how to use CNN features in LCD[CLJM14, SSD+15, AGT+18, SSJ+15, BWZ+16]. Even so, the research in this field is preliminary and incomplete, partially because of the weak interpretability of neural networks.

Before delving into the paper, we first see some frequently asked questions when people want to employ CNN in LCD. First of all, there are numerous outstanding neural network architectures, which one is more suitable for LCD and what is the reason? 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, do we have better solutions?

In this paper, we try to explore these problems in depth and give corresponding explanations. The main contributions include:

1. We compare many off-the-shelf networks and find DenseNet outweighs other popular networks in loop closure detection, because this dense-connected network could preserve both semantic information and structure details of the input image.

2. A loop closure detection framework (Dense-Loop) using DenseNet features is proposed in this paper. 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.

3. Extensive experimental results show Dense-Loop approach 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 detection. Section 3 presents 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.

2 Related works

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

Many well-designed local features are widely used in place recognition and loop closure detection tasks because their ability to resist scale changes or orientation changes. One of the most successful use is FAB-MAP, which employs SUFT[BETG08] and BoWs for place recognition and demonstrates robust performance against viewpoint changes[Cum08].[MAT17] integrate ORB[RRKB11] 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 detection is to train a CNN network in an end-to-end manner. NetVLAD[AGT+18] 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.[LAGOPGJ17] adopt 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. [CLJM14] employ CNN features based on OverFeat for place recognition. The performance of feature-maps

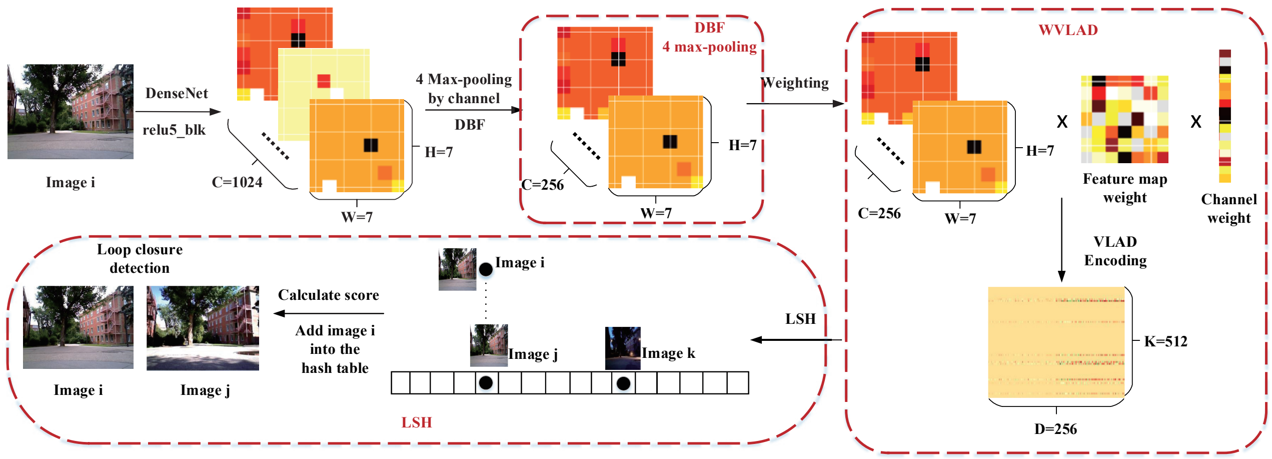


Figure 1: The pipeline of Dense-Loop

of different layers is explored. [HZZ15] focus on using AlexNet to generate an image representation appropriate for visual loop closure detection in SLAM. They find CNN features outperform hand-crafted features when illumination changes signicantly.[SSD$^+$15] deploy 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 LCD and how to use them to achieve better performance without specific supervised training.

3 Framework of Dense-Loop

In the proposed framework, 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)[RPH05] is employed according to the characteristic of Dense-Loop descriptors. 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.

3.1 Image 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.

3.1.1 DenseNet features

DenseNet 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. The architecture of a 5-layer block in DenseNet is shown in Figure 2(a). DenseNet adopted in Dense-Loop is made up of 5 dense blocks. 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 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.

3.1.2 DBF and 4 max-pooling by channel

Here are two ways to map these features to 2-dimension, as shown in Figure 2(b). 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.

In order to ensure the real-time search, 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.

3.2 WVLAD method

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 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.

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)[JDSP10] method and inspired by Cross-dimensional Weighting for Aggregated Deep Convolutional Features (CROW)[KMO16] method.

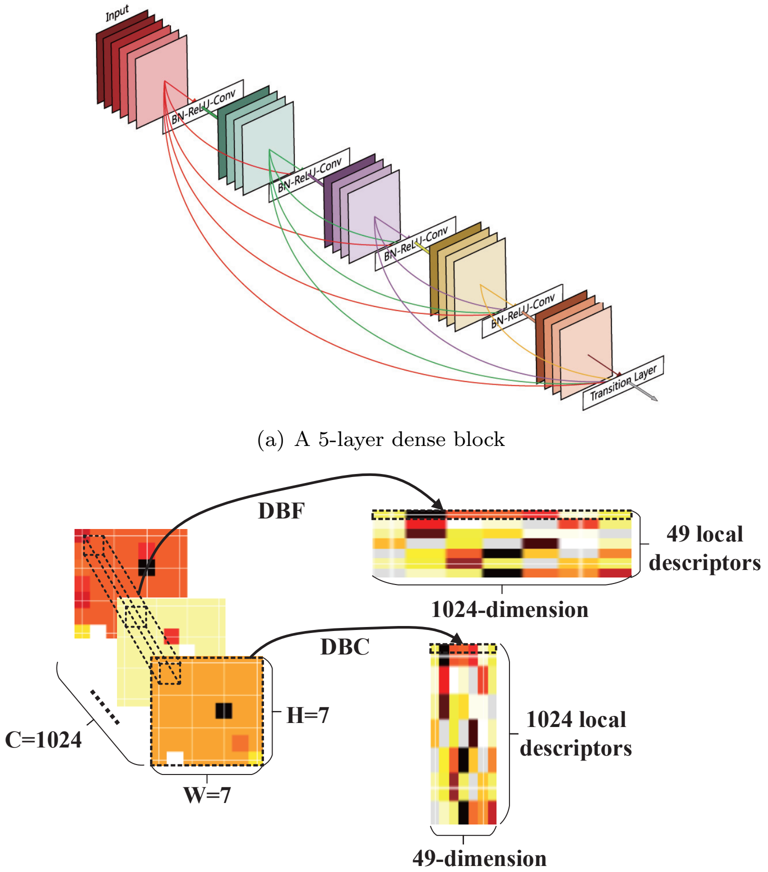
Usually we want the descriptors care more about the distinctions of the image and reduce the importance of the plain areas (e.g. sky). It's similar to the human perception system, which is conducive to improving resistance to environment changes. One way is to use region proposal methods and compute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

$$S = \sum_{c} X_c \tag{1}$$

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

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



(a) A 5-layer dense block

(b) DBF and DBC

Figure 2: (a) A 5-layer dense block with a growth rate of k = 4. The figure is reproduced from[HLvMW17]. (b) The description of DBF and DBC.

$CW \in \mathbb{R}^{(1\times C)}$ 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.

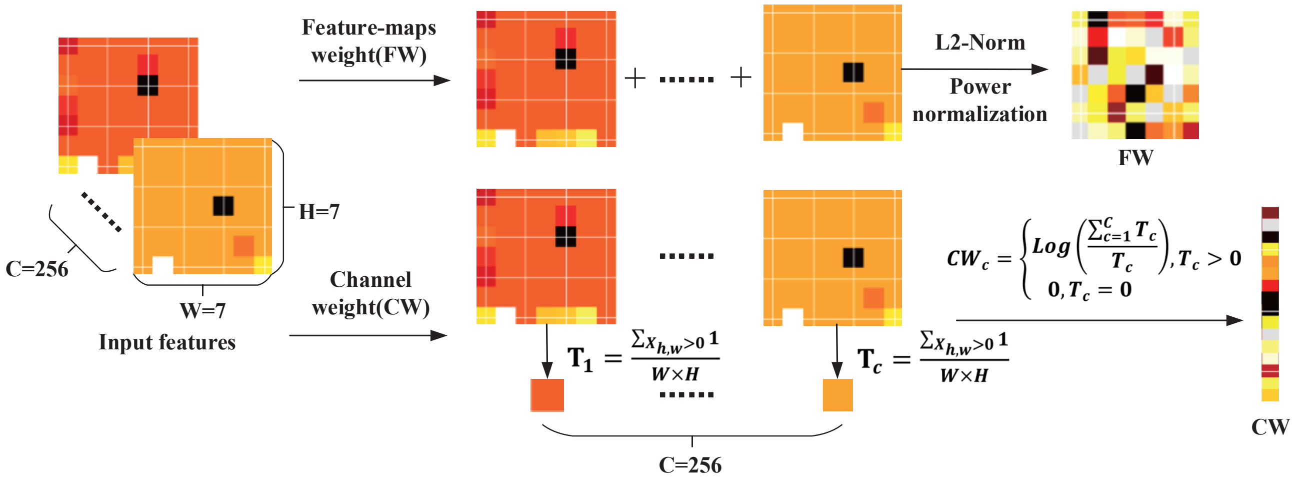

Feature-maps weight(FW)
L2-Norm
Power normalization
FW
H=7
C=256
W=7
Input features
Channel weight(CW)
$CW_c = \begin{cases} Log\left(\frac{\sum_{c=1}^{C} T_c}{T_c}\right), T_c > 0 \\ 0, T_c = 0 \end{cases}$
$T_1 = \frac{\sum_{x_{h,w}>0} 1}{W \times H}$
$T_c = \frac{\sum_{x_{h,w}>0} 1}{W \times H}$
C=256
CW

Figure 3: The detailed process of calculating FW and CW.

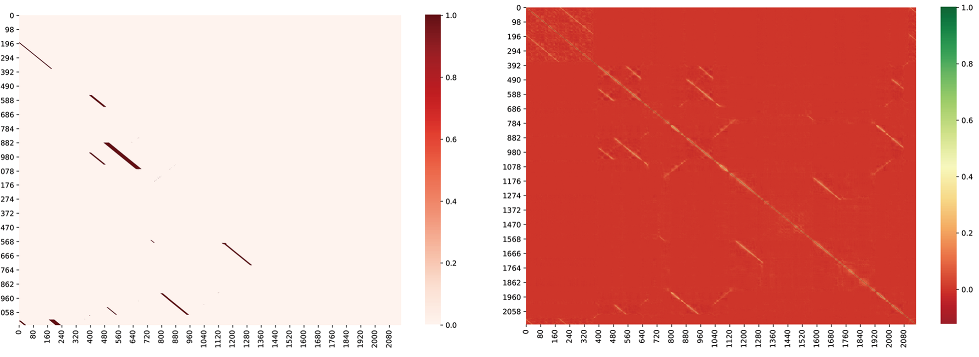
3.3 Locality-Sensitive Hashing

An important feature of loop closure detection for robotic application (e.g. SLAM) is real-time. In the traditional BoWs, K-D tree is adopted as the nearest neighbor search. 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 in such case. 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, which is a cheap operation, is used to evaluate the similarity. According to our test, using 1024 bits retains approximately 99% performance but much more quick than brute search.

4 Experimental Results and Explanations

4.1 Datasets and evaluation method

City Center dataset[Cum08] and New College dataset[Cum08] 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 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

Figure 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.

4.2 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.

4.2.1 Why DenseNet

In recent years, there are many prevalent and excellent convolutional networks showing up, such as ResNet50[HZRS16], VGG[SZ14], DPN[CLX+17], SENet[HSS17], ResNeXt[XGD+17], NasNet[ZVSL17], SqueezeNet[IMA+16], Xception[Cho17], Inceptionv3[SVI+15], Inceptionv4 and Inception-ResNet[SIV17]. 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.

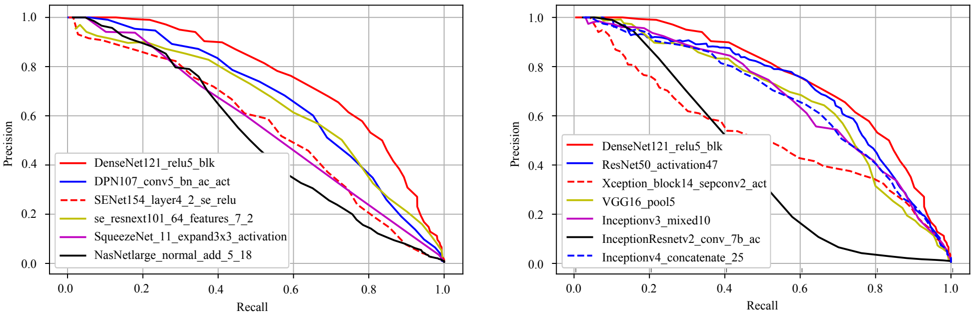
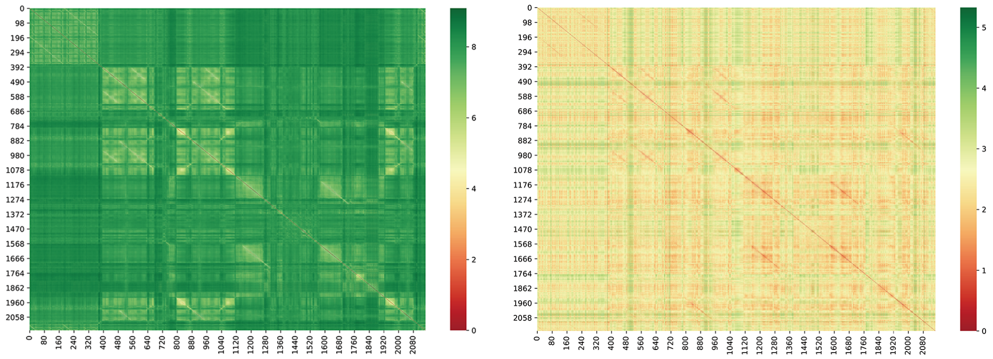


Figure 5: The PR-Curves of different networks on New College dataset.

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.

4.2.2 Why DBF and 4 max-pooling

Figure 7(a) 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. Figure 7(b) 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.



(a) DenseNet (b) Xception

Figure 6: The euclidean distance of images on New College Dataset.

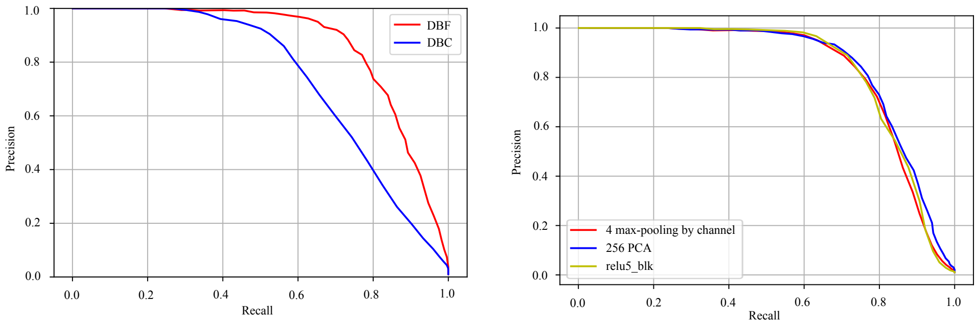


Figure 7: The PR-Curves of DBF V.S. DBC and different dimensionality reduction methods on City Center dataset.

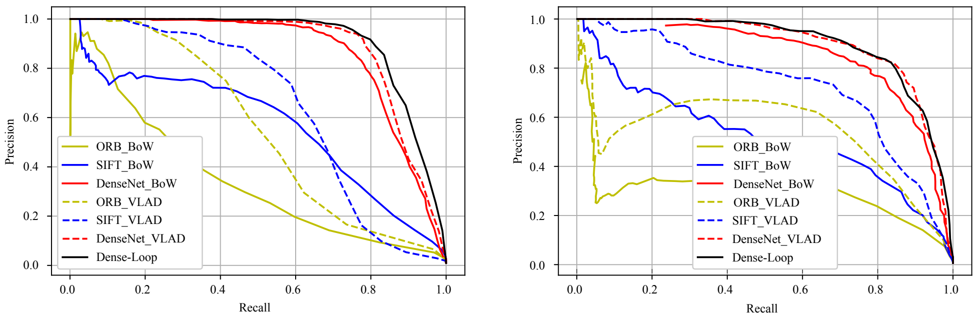
4.2.3 Why WVLAD

In order to compare the performance with traditional methods, two hand-crafted 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 8.

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 9(a) and 9(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 9(c) and 9(d), Dense-Loop can recognize the two images as the 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.

5 Conclusion

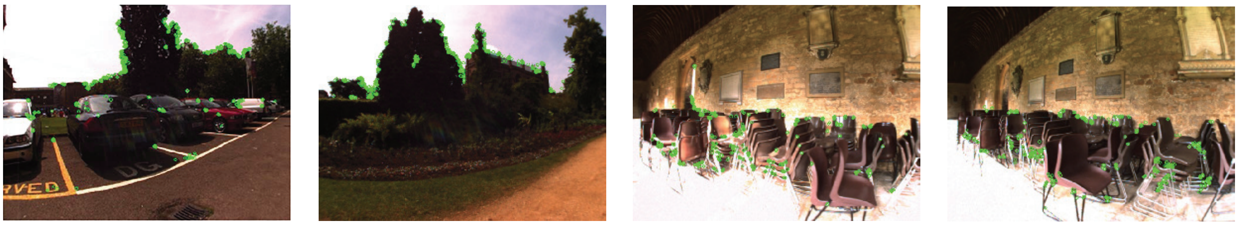
Loop closure detection is used to detect if the robot has passed through the same place. It's crucial for the robot to establish a globally consistent map, especially for large and long-term scenes. A framework of loop closure detection based on CNN features is proposed in this paper. We find that features 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



(a) City Center dataset

(b) New College dataset

Figure 8: The PR-Curves of DenseNet V.S. (ORB, SIFT) and Dense-Loop V.S. (BoW, VLAD)



(a)

(b)

(c)

(d)

Figure 9: 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.

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.

Acknowledgement

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

References

[AGT+18] 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*, 40(6):1437–1451, June 2018.

[BETG08] Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool. Speeded-up robust features (surf). *Computer Vision and Image Understanding*, 110(3):346–359, 2008.

[BWZ+16] Dongdong Bai, Chaoqun Wang, Bo Zhang, Xiaodong Yi, and Yuhua Tang. Matching-range-constrained real-time loop closure detection with cnns features. *Robotics and Biomimetics*, 3(1):15, Sep 2016.

[Cho17] F. Chollet. Xception: Deep learning with depthwise separable convolutions. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 00, pages 1800–1807, July 2017.

[CLJM14] Zetao Chen, Obadiah Lam, Adam Jacobson, and Michael Milford. Convolutional neural network-based place recognition. *CoRR*, abs/1411.1509, 2014.

[CLX$^+$17] Yunpeng Chen, Jianan Li, Huaxin Xiao, Xiaojie Jin, Shuicheng Yan, and Jiashi Feng. Dual path networks. *CoRR*, abs/1707.01629, 2017.

[Cum08] M Cummins. Fab-map : Probabilistic localization and mapping in the space of appearance. *The International Journal of Robotics Research*, 27(6):647–665, 2008.

[GSM18] 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)*, pages 3645–3652, May 2018.

[HLvMW17] 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)*, pages 2261–2269, July 2017.

[HSS17] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. *CoRR*, abs/1709.01507, 2017.

[HZRS16] 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)*, pages 770–778, June 2016.

[HZZ15] 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)*, pages 2238–2245, Aug 2015.

[IMA$^+$16] Forrest N. Iandola, Matthew W. Moskewicz, Khalid Ashraf, Song Han, William J. Dally, and Kurt Keutzer. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and <1mb model size. *CoRR*, abs/1602.07360, 2016.

[JDSP10] 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)*, pages 3304–3311, June 2010.

[KMO16] Yannis Kalantidis, Clayton Mellina, and Simon Osindero. Cross-dimensional weighting for aggregated deep convolutional features. In Gang Hua and Hervé Jégou, editors, *Computer Vision – ECCV 2016 Workshops*, pages 685–701, Cham, 2016. Springer International Publishing.

[KSH12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems (NIPS)*, NIPS'12, pages 1097–1105, USA, 2012. Curran Associates Inc.

[LAGOPGJ17] Manuel Lopez-Antequera, Ruben Gomez-Ojeda, Nicolai Petkov, and Javier Gonzalez-Jimenez. Appearance-invariant place recognition by discriminatively training a convolutional neural network. *Pattern Recognition Letters*, 92:89–95, 2017.

[LM13] Mathieu Labbe and Francois Michaud. Appearance-based loop closure detection for online large-scale and long-term operation. *IEEE Transactions on Robotics*, 29(3):734–745, June 2013.

[LSN$^+$16] Stephanie Lowry, Niko Sünderhauf, Paul Newman, John J. Leonard, David Cox, Peter Corke, and Michael J. Milford. Visual place recognition: A survey. *IEEE Transactions on Robotics*, 32(1):1–19, 2016.

[MAT17] Raúl Mur-Artal and Juan D. Tardós. Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras. *IEEE Transactions on Robotics*, 33(5):1255–1262, 2017.

[RPH05] Deepak Ravichandran, Patrick Pantel, and Eduard 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*, ACL '05, pages 622–629, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics.

[RRKB11] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. Orb: An efficient alternative to sift or surf. In *Proceedings of the 2011 International Conference on Computer Vision (ICCV)*, ICCV '11, pages 2564–2571, Washington, DC, USA, 2011. IEEE Computer Society.

[SIV17] Christian Szegedy, Sergey Ioffe, and Vincent Vanhoucke. Multi-scale orderless pooling of deep convolutional activation features. In *Proceeding of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI)*, pages 4278–4284, 2017.

[SSD$^+$15] 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)*, pages 4297–4304, Sept 2015.

[SSJ$^+$15] Niko Sünderhauf, Sareh Shirazi, Adam Jacobson, Feras Dayoub, Edward Pepperell, Ben Upcroft, and Michael Milford. Place recognition with convnet landmarks: Viewpoint-robust, condition-robust, training-free. In *Robotics: Science and Systems (RSS)*, Auditorium Antonianum, Rome, July 2015.

[SVI$^+$15] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. *CoRR*, abs/1512.00567, 2015.

[SZ14] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[UMCM14] B Upcroft, C Mcmanus, W Churchill, and W Maddern. Lighting invariant urban street classification. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 1712–1718, Hong Kong, China, 2014. IEEE.

[XF17] Yu Xiang and Dieter Fox. DA-RNN: semantic mapping with data associated recurrent neural networks. *CoRR*, abs/1703.03098, 2017.

[XGD$^+$17] 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)*, volume 00, pages 5987–5995, July 2017.

[YLL$^+$18] 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)*, pages 1168–1174, Oct 2018.

[ZVSL17] Barret Zoph, Vijay Vasudevan, Jonathon Shlens, and Quoc V. Le. Learning transferable architectures for scalable image recognition. *CoRR*, abs/1707.07012, 2017.