Multimedia Tools and Applications (2019) 78:32419–32435
https://doi.org/10.1007/s11042-019-07980-9

Deep hash for latent image retrieval

Fanfeng Zeng[1] · Shengda Hu[1] · Ke Xiao[1]

Received: 7 August 2018 / Revised: 15 June 2019 / Accepted: 10 July 2019
Published online: 3 August 2019

Abstract
With the development of the era of internet, an increasing number of images flow into people's daily life, and it's really a challenge to quickly search interesting images in such a huge image database. The most advanced method is using a deep neural network to get hash code of images to achieve fast image retrieval at present. However, people usually use the single pooling method to screen the image pixels when designing the neural network, and some effective information of the image will be gradually lost due to the single pooling method with the number of network deepening. Aiming at this problem, this paper proposes a convolutional neural network combining multiple pooling methods to preserve the effective information of the image as much as possible. To verify the effectiveness of the proposed method, many experiments are carried out on the CIFRA-10, NUS-WIDE and MNIST datasets. The experimental results show that the proposed method is better than most existing hash-based image retrieval methods.

Keywords Image retrieval · Hash · Convolutional neural network · Deep learning

1 Introduction

With the continuous updating of image acquisition equipment, especially camera devices on smart phones, people are more likely to use image to record information, which also lead to a large increase in the size of image data. Massive images are constantly being replicated on the Internet or on user storage devices, and because of the redundancy that may occur during the storage of images, it is hard for users to get interesting images in short time even though these images exist. Therefore, how to quickly find valuable images in large scale images has always

✉ Ke Xiao
xiaoke@ncut.edu.cn

Fanfeng Zeng
zengfanfeng@ncut.edu.cn

Shengda Hu
20163121201O7@mail.ncut.edu.cn

[1] North China University of Technology, Beijing, China

been a very valuable research problem, and hash-based image retrieval technology has emerged.

The hash-based image retrieval technology performs binary hash encoding based on the content of the images, and it uses the Hamming distance between the hash codes to measure the similarity of different images. Compared with traditional content-based image retrieval technology, the image feature adopted by hash-based image retrieval technology is binary hash coding, which has better storage and computational advantages [1]. Thus, it is possible to perform fast image retrieval within limited storage space.

The core issue of hash-based image retrieval technology is how to obtain a compact binary hash code that fully represents the image content. In other words, this is means to design and construct a hash function, and map the image to a binary vector space, so that the obtained vector can correctly tell the difference between images. In the process of mapping, the original image is not directly hashed, but the feature vector of the image is extracted first, and then the feature vector is converted into binary hash 发 code by a hash function. The traditional feature vector extraction method is often based on the visual understanding of images. The visual information of these images of interest is generally the color, texture, shape and other characteristics of images. However, these features only describe the low-level visual information and cannot improve the "semantic gap" problem [2]. Therefore, people begin to design some higher-level image features, such as BOW (Bag of visual words, BOW) [3], VLAD (Vector of Aggregate Locally Descriptor, VLAD) [4] and FV (Fisher Vector, FV) [5, 6]. These features are based on some low-level features, which can better express image information and improve the performance of image retrieval. However, if these features are directly used for the input of a hash function, the performance of the image retrieval will be affected by the length of the binary hash code. It means that a satisfactory image retrieval performance may need a longer hash code length. And it is exactly the opposite of the target of the hash-based image retrieval algorithm that reduces the time and storage overhead of image retrieval.

At the same time, deep learning, especially the development of convolutional neural networks, has brought profound changes to the field of computer vision in recent years. Convolutional neural networks have shown phenomenal performance in more and more computer vision problems [7–10] which are hard for traditional algorithms to achieve. As convolutional neural networks have powerful feature extraction capabilities, some methods called deep hash that use convolutional neural networks to generate hash code for image retrieval have emerged recently. Compared with the previous process of extracting image features and generating the binary hash code of images, it does not have the image feature extraction process in the traditional sense, but the hash code of the image can be directly obtained after inputting the image into the neural network. It has the advantage that the obtained binary hash code is achieved directly from the original image and does not go through the image feature extraction process, so the error that may accused in the image feature extraction process can be reduced. In addition, lots of experiments [11–16] have also verified that the image hash retrieval method based on convolutional neural network has better retrieval performance than the traditional image hash retrieval method, and its retrieval performance is not very dependent on the hash code length.

In view of above situation, this paper proposes an end-to-end image retrieval method based on deep hash which uses the convolutional neural network to get hash code. The innovation of this paper is mainly focus on the design of convolutional neural network structure, and the specific contents are listed as follows:

(1) An improved bilinear network [17–19] is employed in deep-hash based image retrieval technology, which is the first time in the field to use bilinear network, and the bilinear model uses multiple pooling methods in every layer of the network to ensure that all the effective information of images can be preserved, so the image retrieval performance can get improved.
(2) Multiple pooling methods are used in the same layer of the network, and the final pooling result is fused by the convolution layer, which greatly preserves the information of the image.
(3) A new method of similarity calculation is proposed, which uses both visual and semantic information of images to better measure the similarity between images.

2 Related work

The most classic hash coding method is the LSH (Locality-Sensitive Hashing, LSH) [20] proposed in 1999. This algorithm uses a kind of random mapping to map image data to Hamming space. The construction of the hash function is simple, and the calculation speed is fast, but the retrieval performance is limited. Then there have been many improved methods [21, 22] based on LSH. SBLSH (Super-Bit Locality-Sensitive Hashing, SBLSH) [21] uses the angle as the kernel function metric to orthogonalize the random projection vector, and it theoretically proves that the variance of the Hamming distance after packet orthogonalization is smaller than the LSH. KSH (Kernel Supervised Hashing, KSH) [22] extends and generalizes the LSH so that it can adapt to any kernel function with better flexibility.

However, the above-mentioned methods have great limitations on the presentment of image content. To further improve the retrieval performance of images, researchers have proposed some methods for image retrieval using convolutional neural networks. The CNNH+(Convolution neural network hashing, CNNH+) [11] proposed by Xia et al. described whether the two images were similar by constructing a similarity matrix, that is, the value of the corresponding position in the similarity matrix represented the similarity of different images and decomposing the similarity matrix to obtain the target hash code of the image in advance. Then the resulting hash code was fitted by using a convolutional neural network. It is worth noting that in the process of training, if the category label of the image was available, the classified soft-max loss function can be added into network in the training process, and a better image retrieval performance can be got. Otherwise, the Cross-Loss function is directly used to fit the target hash code. The DSH (Deep Supervised Hashing, DSH) [14] proposed by Liu et al. considered that in the training process of hash coding, the output of the network was usually activated by the sigma or tanh function, which will make the training of the network become difficult. In addition, the hash coding obtained from the network was not a global optimization due to the difference between the Euclidean distance and the Hamming distance. Instead of using sigma or tanh function to constrain the output of the network, they added a regular term to constrain the output of the network, making the output of the network close to between −1 and 1 as well as speeding the converge of training. Different from DSH, SUBIC (Structured binary codes, SUBIC) [15] proposed by Jain et al. introduced a block-SoftMax structure in the design process of loss function and verified the validity of the structure with many experiments. Lu et al. [16] pre-calculated the hash codes of images by designed algorithm, and then fit them through deep neural networks. Compared with the method that uses category as ground truth, its hash code has more ability to tell differences with different images.

In fact, the existing image retrieval methods based on deep hash are mainly to improve the two aspects of loss function and network structure, and the method of this paper belongs to the latter. Considering the existing convolutional neural network structure usually use a single max pooling method in the same network layer, and ignore other pooling methods, this paper proposes a bilinear network struct to get hash code of images, and it uses multiple pooling methods in every layer of the network to ensure that all the effective information of images can be preserved, so the image retrieval performance can get improved.

3 Proposed method

3.1 Network structure

The pooling layer is an important struct of convolutional neural network. On the one hand, the pooling layer plays a role of subsampled which reduces image interference information and retains important information of the image with the neural network layers deepening; on the other hand, it reduces the number of pixels in each layer of the image, ensuring that the network size is not too large, and also reduces the problem of too many network parameters and excessive memory consumption.

The existing pooling methods are mainly three types: max pooling, average pooling and stochastic pooling [23]. The average pooling will blur the image, retain more background information, the max pooling will sharpen the image, retain more texture information, and the effect of stochastic pooling is between max pooling and average pooling. And max pooling and average pooling are most commonly used pooling methods. However, compared with average pooling, max pooling is more popular, because many experiments [24, 25] verify that the max pooling has better performance than average pooling in various tasks. In fact, the neural network filters the image information through the pooling layer, and finally fits the information by other structs of network. And various pooling methods have different effect depend on different strategies, if only the max pooling is used in network, the network will focus on the texture information and ignore background information. So, it's advisable to use various pooling method in networks to get more effective information of images.

In addition, in the design of convolutional neural networks, there is a useful network structure called bilinear network structure worthy of reference. This structure is essentially a combination of a single network structure and updating the parameters of all networks simultaneously during the training of the network. In fact, the advantages of bilinear network are mainly two points. On the one hand, the network has two sub-network that can perform feature extraction on images twice, which means bilinear network can extract more effective information of images compared others. On the other hand, bilinear operator achieves fusion of image features extracted from two sub-networks, and it further improved the ability of network to extract image features.

Consider the advantage of bilinear network and various pooling methods, this paper designs a network shown in Fig. 1. The network is a kind of improved bilinear network that using both max pooling and average pool in every layer in the part of feature extraction model of network, and it also employ simple convolution as bilinear operator. What's more, the details of network are shown in Table 1, and all the Batch Normalization layer used in input of every layer is skipped for simplicity of tables.

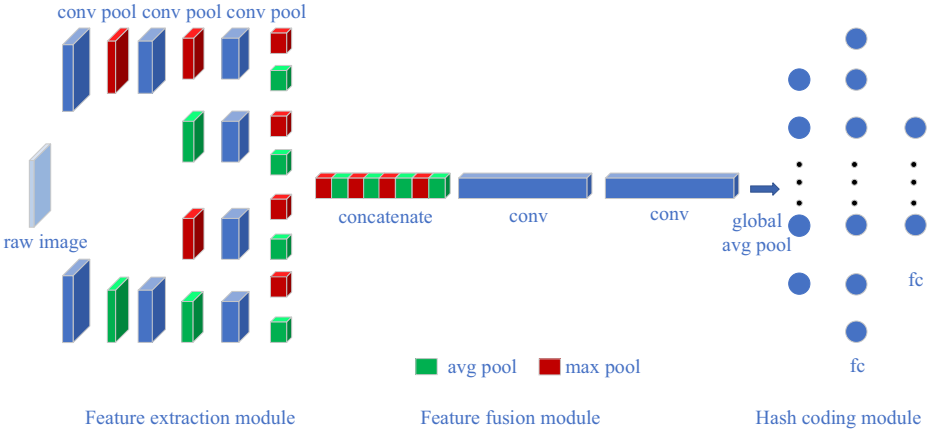


Fig. 1 The structure of network

The network used in this paper is different from conventional bilinear network, these differences are mainly as followed.

Firstly, a basic requirement for a kind of feature fusion method is that different image features must be extracted. If the differences between these different features are large, in the other words, these features can complement each other's limitations, a better result can be obtained. In previous bilinear network models, it focused more on the bilinear operate while ignore the basic feature extraction in two sub-networks, and these bilinear network models usually choose typical networks like VGGs as two sub-networks, so the choice of sub-networks lacks theoretical guidance. And this paper ensures the differences of image features from feature extraction module is rather huge by simply applying max pooling method and average pooling in every sub-network.

Secondly, compared with the conventional bilinear network, the network in this paper offer more images feature to feature fusion module. In Fig. 2, it can find that in a traditional net, there is only one image information stream that used to extract image feature, and in a typical bilinear network, there are two image information streams to offer two image features to latter network layers, so the image features offered by sub-network is rather limited. While in this paper, the feature extraction module uses limited parameter layers, more specifically, three convolution layers, offer eight image features to feature fusion module. Therefore, the network in this paper is more thorough for the extraction of image features.

Finally, the network in this paper in every layer uses two pooling methods, and the number of pooling is increasing with the network going deeper. This is a useful method to make up the loss of image information. Although the pooling layer will cause the loss of partial information of images, the network in this paper increase the number of pooling to keep this image information from pervious layer which may be lost in current layer. This method can preserve image information as much as possible to ensure the robots of image features entreated by network.

3.2 Loss function and Binarization

There are three main loss functions commonly used to train hash codes, which are SoftMax-loss, Contrast-Loss [26] and Triplet-Loss [27]. SoftMax-Loss is the most common used for

Table 1 The details of network

Layer								ksize	stride	num
Conv				Conv				3 × 3	1 × 1	32
Avg pool				Max pool				3 × 3	2 × 2	32
Relu				Relu				–	–	32
Conv		Conv		Conv		Conv		3 × 3	1 × 1	64
Avg pool		Max pool		Avg pool		Max pool		3 × 3	2 × 2	64
Relu		Relu		Relu		Relu		–	–	64
Conv	Conv	Conv	Conv	Conv	Conv	Conv	Conv	3 × 3	1 × 1	64
Avg pool	Max pool	Avg pool	Max pool	Avg pool	Max pool	Avg pool	Max pool	3 × 3	2 × 2	64
Relu	Relu	Relu	Relu	Relu	Relu	Relu	Relu	–	–	64
Concatenate								–	–	–
Conv								3 × 3	1 × 1	512
Relu								–	–	512
Conv								3 × 3	1 × 1	512
Relu								–	–	512
Golobal average pool								–	–	512
Relu								–	–	512
Fc								–	–	1000
Fc								–	–	bits

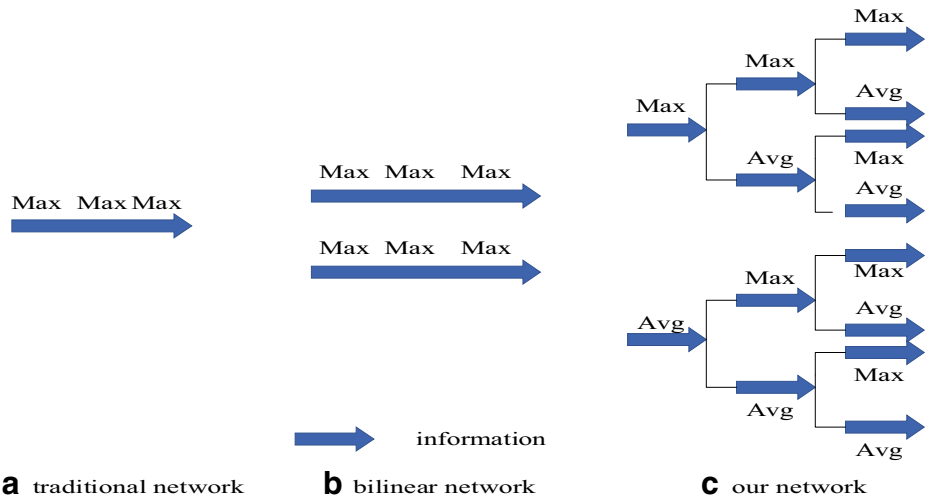


Fig. 2 Flow direction of image information in feature extraction part of network

classification problems. It limits the output of the neural network from 0 to 1, so researchers can threshold it to a binary hash code with a median of 0.5. For Contrast-Loss and Triplet-Loss, the former displays the similarity of the two images, while the latter constrains that the distance between two different images should be farther than the distance between the same images, thus displays more information of images. However, it is more difficult to train the network using the Triple-Loss function, because for all the trained images, the proportion of similar images in a batch will be much lower than the dissimilarity, which likely leads to a bad training result. And an ideal training model needs a well-designed image batch which feed to network. Therefore, this paper adopts improved Contrast-Loss which proposed by Ref. 14 to simplify experiments.

For an image pairI_1,I_2, the corresponding binary outputs areb_1,b_2.$y$is the value which define the similarity of images. If two images are from same class,$y$is set to 0, and if not, y is 1. So, the loss of image pair$L_{pair}(b_1, b_2, y)$is as follows.

$$L_{\text{pair}}(b_1, b_2, y) = \frac{1}{2}(1-y)\|b_1-b_2\|_2^2 + \frac{1}{2}y\cdot\max\left(m-\|b_1-b_2\|_2^2, 0\right) + \alpha\left(\||b_1|-1\|_1 + \||b_2|-1\|_1\right) \tag{1}$$

Where the **1** is a vector of all ones, $\|\cdot\|_1$is L1-norm of vector, $|\cdot|$ is the element-wise absolute value operation, and$\alpha$is a weighting parameter that controls the strength of the regularizes.

Therefore, for N training pairs randomly selected from training images $\{(I_{i,1}, I_{i,2}, y_i) | i = 1, \cdots, N\}$, the total loss L is as follows.

$$L = \sum_{i=1}^{N} L_{pair}\left(b_{i,1}, b_{i,2}, y_i\right) \tag{2}$$

For the final hash code is generated by sign of every element in b. if the sign is positive, the corresponding bit of hash code is 1, and if not, the value of this bit is 0.

3.3 Similarity

After the hash code of images is generated, there is an import operator that return the relevant images from high to low according to similarity in database. In fact, the similarity of two images calculated by Hamming distance is not consequent, which means that two images that share the same hash code may have the same semantic information while they can have different visual information, such as different color and texture.

To solve this problem, this paper proposes a new similarity calculation method. The core idea is that the best similarity description of images is both adapting the visual information and semantic information, and the semantic similarity is more import. Consider that the hash code got by traditional hash method is usually visual information, this paper chooses to get a kind of traditional hash method to improve the calculation of similarity between images.

For every image pairI_1,I_2, their hash code generated by network areH_1andH_2, while hash code generated by traditional method are$H_1^{'}$and$H_2^{'}$, and$h$is a bit of$H$. Thesim(I_1, I_2) is as follows:

$$sim(I_1, I_2) = \begin{cases} s(H_1, H_2) + \frac{1}{n} s\left(H_1^{'}, H_2^{'}\right) \; if \; s(H_1, H_2)! = 1 \\ \frac{n-1}{n} + \frac{1}{n} s\left(H_1^{'}, H_2^{'}\right) \; else \end{cases} \tag{3}$$

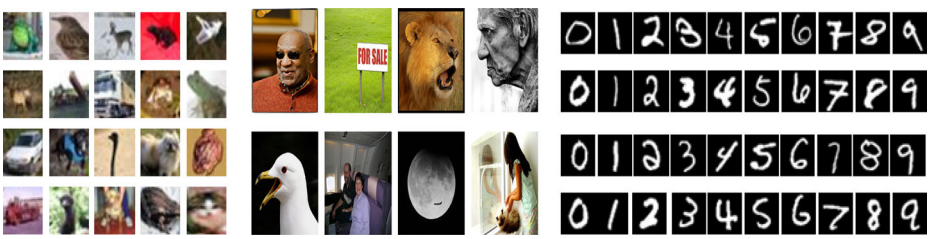
In the above formula,$n$is the bit of hash code got from network, and$s(\cdot)$is the similarity of two hash code. And the definition of$s(\cdot)$is as follows:

$$s(H_1, H_2) = \frac{1}{n} \sum_{k=1}^{n} |h_{1,k} - h_{2,k}| \tag{4}$$

4 Experiment

4.1 Datasets

The datasets used in this paper are several famous benchmark datasets which are widely applied in various computer vision tasks. And some images from these datasets are shown in Fig. 3.



a CIFRA-10 **b** NUS-WIDE **c** MNIST

Fig. 3 From (**a**) to (**c**) are some images from CIFRA-10, MNIST and NUS-WIDE respectively

CIFRA-10 [28]: The CIFRA-10 dataset consists of 60,00032 × 32color images in 10 classes, with 6,000 images per class. There are 50,000 training images and 10,000 test images.

NUS-WIDE [29]: The NUS-WIDE dataset consists of 269,648 images in 81 classes collected from Flickr. The size of images is about240 × 180and usually different from each other, and the number of images in every class is also not always same.

MNIST [30]: The MNIST contains 10 classes of handwriting digits from 0 to 9, and all the images are normalized to28 × 28.

For the CIFRA-10, this paper uses the default 50,000 training images for training and divides the default 10,000 test images into image databases and test query sets randomly, which have 9,000 images and 1,000 images respectively. And all pixels of each image are scaled from 0.0 to 1.0.

For the NUS_WIDE, this paper follows Ref. 14 that chose 21 most frequent classes, including 19,5834 images, and each class have at least 5,000 images. This paper selects 10,000 images from these images as test images and the remains are as training sets. The division of image databases and query sets is the same as the CIFRA-10.

For the MNIST, this paper uses the default 60,000 training images for training and the rest 10,000 images for testing like CIFRA-10.

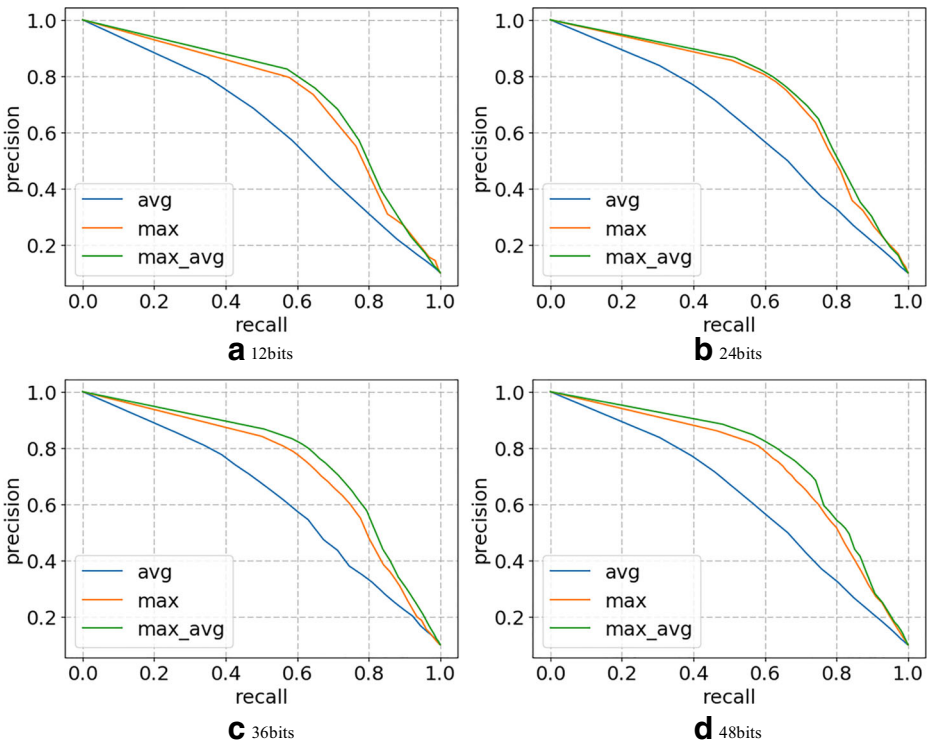


Fig. 4 *Precision-recall* cures in cifra-10

4.2 Training details

This paper has performed necessary data enhancement of all images. Specifically, all images are addressed by random cutting, flipping, brightness enhancement, contrast enhancement and normalization before feed to network.

For the parameters of loss function,$\alpha$is 0.01,$m$is twice the length of hash code. In addition, this paper randomly selects 100 images feed to network and uses Adam gradient descent method for training. And initial learning rate is 0.0001, which decrease to 50% every 20,000 iterations both in CIFRA-10 and MNIST, and 80% every 18,000 iterations in NUS-WIDE respectively. And all datasets will finish training after 50 epochs.

4.3 Metrics

This paper adopts standard recall and precision to measure the performance of image retrieval. The definitions of precision and recall are as respectively as follow:

$$precision = \frac{|\{relevant\ images\} \cap \{retrieved\ images\}|}{|\{retrieved\ images\}|} \tag{5}$$

$$recall = \frac{|\{relevant\ images\} \cap \{retrieved\ images\}|}{|\{relevant\ images\}|} \tag{6}$$

4.4 Benefits of multiple pooling methods

To verify the effectiveness of multiple pooling methods in every layer of network on the image retrieval problem, this paper changes the pooling method of each layer to a single max pool or average pool as two control groups which means it will perform the same pooling method twice in every layer. The experiment is performed on CIFRA-10, and the ***precision-recall*** curve and ***map*** curve at different binary hash code lengths are shown in Fig. 3 and Fig. 4(a) respectively.

It can be found from Fig. 4 and Fig. 5(a) that the method of using the max pooling and the average pooling method has better image retrieval performance under different hash code lengths and has different degrees of retrieval performance improvement

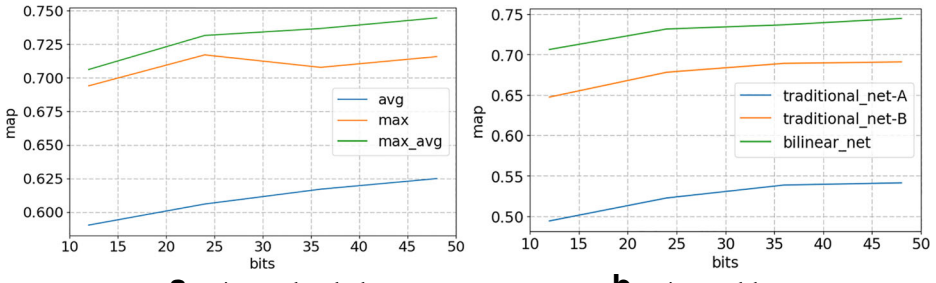


Fig. 5 *Map* cures in cifra-10

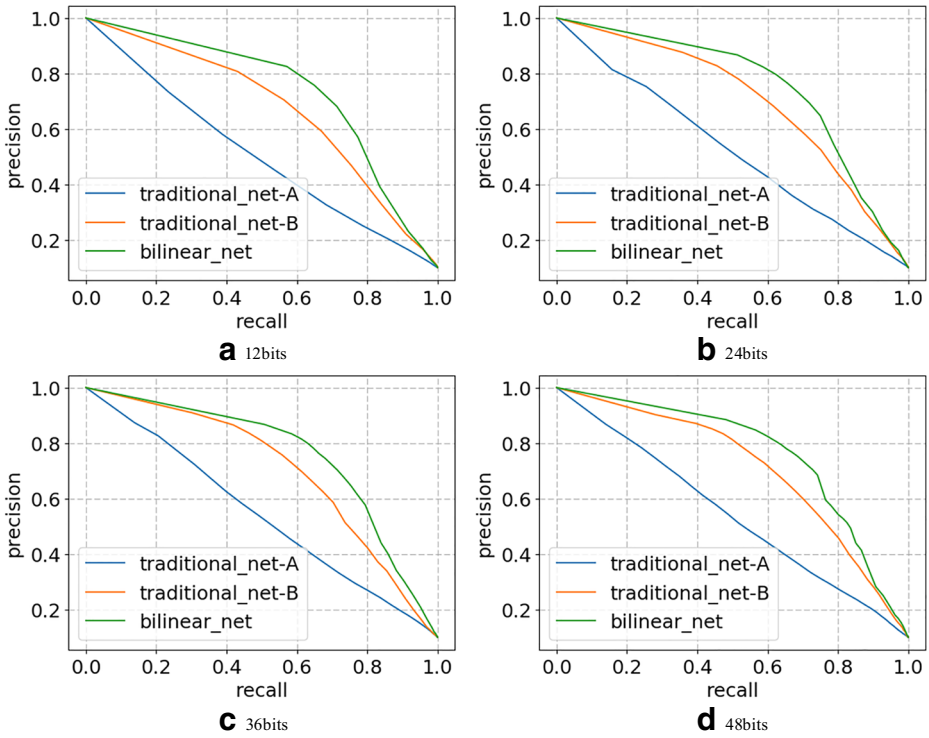


Fig. 6 Precision-recall cures in cifra-10

compared with other methods. As mentioned above, the image retrieval results obtained using a single average pooling method is the worst. In fact, the network in this paper both uses max pooling and average pooling in every layer can enrich the pixel filtering mechanism, which can enlarge the differences around eight image features offered to later layers. In addition, bilinear network is a fusion model, and the features offered by our feature extraction module is a complementary feature, so our network can get a better image retrieval performance.

4.5 Benefits of bilinear models

To verify the effectiveness of the bilinear model on the image retrieval problem, instead of adopting the bilinear model in first half of the network, the traditional single information flow is applied. Then single max pooling and average pooling method are used as control method, and the method which only use average pooling and max pooling are called traditional_net-A and traditional_net-B respectively. And the comparative experiment is performed on CIFRA-10. The ***precision-recall*** curve and the ***map*** curve at different binary hash code lengths are shown in Figs. 4a and 5 respectively.

It can be seen from Figs. 5b and 6 that the bilinear_net has the best image retrieval performance under different hash code lengths, and the image retrieval of traditional_net-A is the worst. The network structure using the bilinear model extracts the features of the

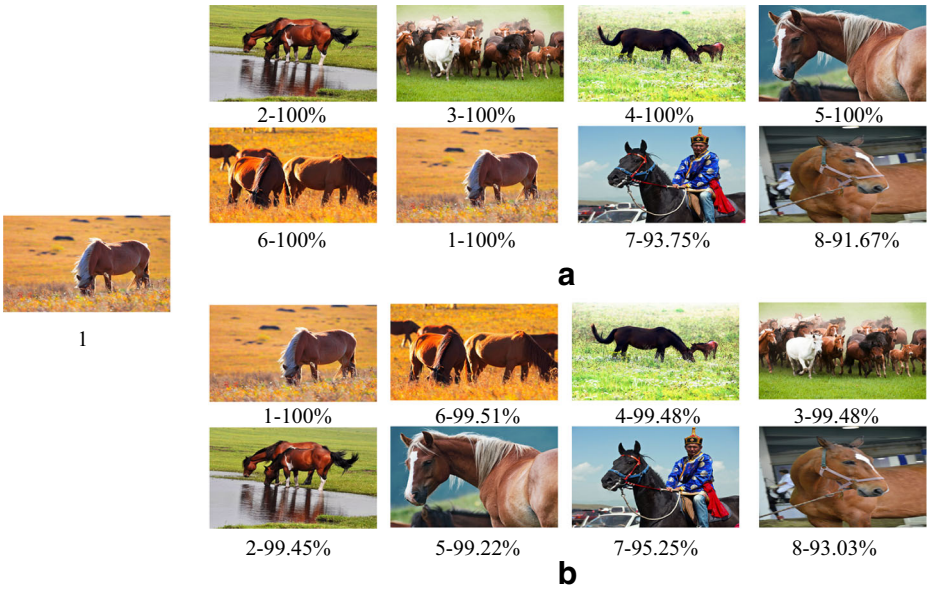


Fig. 7 A retrieval example, (**a**) is retrieval result of original similarity and (**b**) is proposed in this paper

image twice in each layer of the network, so the effective features of the image can be greatly preserved, and the performance of image retrieval is improved. In addition, a very noteworthy problem is that every time a pooling operation is performed, some of the effective image information may be missed, and the grade of information loss will increase as the depth of the network deepens. This is because that the pooling in current layer is based in previous pooling result. If the information is lost in previous layer, the next pooling will preserve less effective information. And in the network proposed in this paper, as the network deepens, the number of pooling of in every layer increase, so that the effective information of the image which from previous layer can further retained. In other words, it is the multiple pooling that improves the performance of image retrieval.

4.6 Benefits of new similarity calculation method

To verify the effectiveness of the proposed similarity calculation between images, this paper lists a convincing retrieval example in Fig. 6. The retrieval result is got by 48-bits hash code from network and 64-bits hash code from traditional hash method called PSH (PSH, Perceptual hashing) [31].

In Fig. 7, all images are horses, and the sixth image and the first image are visually closest among all the images retrieved. In Fig (a), all search results have a similarity of 100% and there is no intuitive visual distinction. In the method of this paper, the original image and the most similar image are at the forefront. In fact, if only the hash code obtained by the deep hash method is used to calculate the similarity of the image, the visual similarity of the image cannot be distinguished, and the visual similarity in the example given herein mainly refers to the difference in color. However, if non-hash image features are introduced on the similarity calculation problem, it will affect the

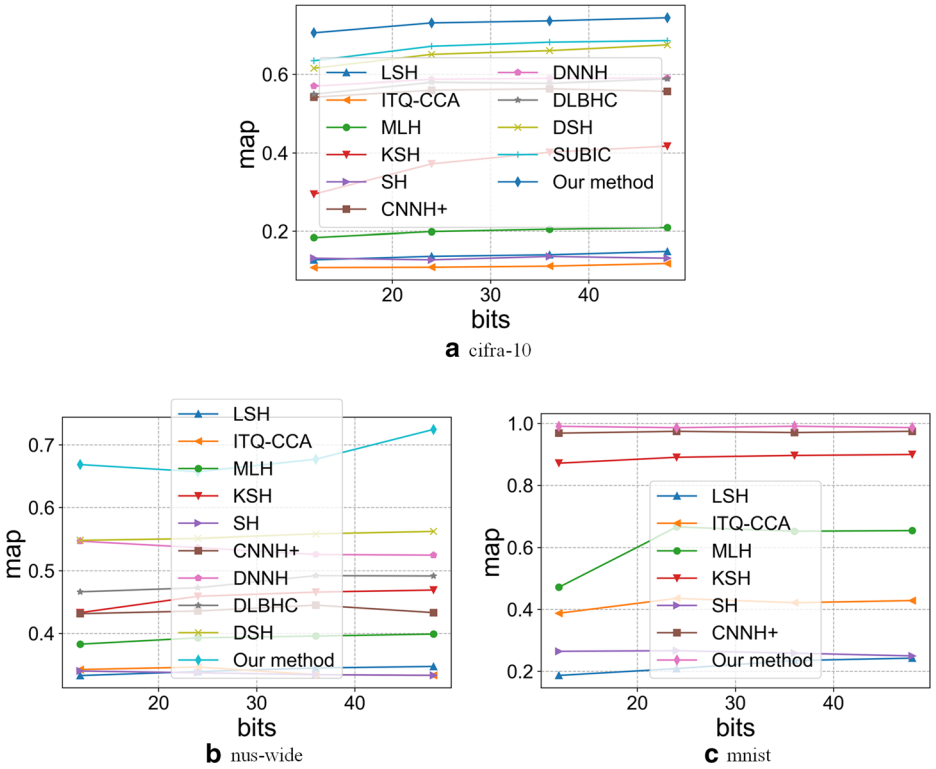


Fig. 8 *Map* cures in CIFRA-10, NUS-WIDE and MNIST respectively

speed of image retrieval, so it is better to use traditional hash methods that represent visual information.

4.7 Compared with other methods

To verify the effectiveness of the proposed method, this paper compares our method with two image retrieval methods. One is traditional method which don't use neural network to get hash code, and these methods are LSH [20], ITQ [32], MLH [31], KSH [22] and SH [33]. The other is some deep hash method like us, and these methods are CNNH+ [11], DNNH [12], DLBHC [13], DSH [14], and SUBIC [15]. The experiment results of other method are quoted from authors' paper except DSH. The experiment result that the ***map*** under different hash code length in CIFRA-10, NUS-WIDE and MNIST is shown in Fig. 8.

In these three datasets, the method proposed in this paper has highest ***map*** at all the lengths of hash code. And in CIFRA-10, NUS-WIDE and MNIST, the proposed method has about an improvement of 6%, 12% and 2% respectively than the second method. And the method of deep hash generally has better image retrieval performance than traditional hash method which means the neural network is very useful to get hash code of images. In fact, the most advantage of our method is that we have extract more different image features by using both max pooling and average pooling in every layer.

It's also these effective image features have been further refined in the subsequent feature fusion module to improve the image retrieval effect.

5 Conclusion

This paper proposes a new deep hash method for image retrieval. It improves the feature extraction module of bilinear network that using both max pooling and average pooling in every layer of network. And it also increases the number of pooling methods with the network going deeper, which is a good way to make up the loss information of image caused by pooling method. In addition, a new method of similarity calculation between images is also proposed to appropriately measure the similarity of images both in vision and semantic. Experimental results show that the proposed method has better image retrieval performance than the most of existing hash-based image retrieval methods. However, due to time and energy constraints, there are also some limitations in our method. Specifically, our network is a more complicated model and has more parameters to learn in training. Therefore, this paper will make some corresponding improvements in this aspect in the future work.

Funding This study was funded by Science and Technology Innovation Service Capacity Building Project of china (PXM2017–014212-000002).

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Lu X, Song L, Xie R et al (2017) Deep binary representation for efficient image retrieval[J]. Advances in Multimedia 2017:1–10
2. Wan J, Wang D, Hoi SCH, Wu P, Zhu J, Zhang Y et al. (2014). Deep learning for content-based image retrieval: a comprehensive study. (FullPaper), 157–166
3. Csurka G (2004) Visual categorization with bags of keypoints. Workshop on Statistical Learning in Computer Vision Eccv 44(247):1–22
4. J'egou H, Douze M, Schmid C, and Perez P (2010) Aggregating local descriptors into a compact image representation. In CVPR, pages 3304–3311, 1, 14
5. Shi C, Wang Y, Jia F et al (2017) Fisher vector for scene character recognition: a comprehensive evaluation[J]. Pattern Recogn 72:1–14
6. Perronnin F, Liu Y, Sanchez J, and Poirier H (2010) Large-scale image retrieval with compressed fisher vectors. In CVPR, pages 3384–3391, 14
7. Rueckauer B, Lungu IA, Hu Y, Pfeiffer M, Liu SC (2017) Conversion of continuous-valued deep networks to efficient event-driven networks for image classification. Front Neurosci 11:682
8. Xie S, Tu Z (2015). Holistically-Nested Edge Detection. IEEE International Conference on Computer Vision (pp.1395–1403). IEEE Computer Society
9. Zheng Z, Zheng L, Yang Y (2017). A discriminatively learned cnn embedding for person re-identification. Acm Transactions on Multimedia Computing Communications & Applications, 14(1)

10. Yao Y, Shi Y, Weng S, Guan B (2017) Deep learning for detection of object-based forgery in advanced video. Symmetry 10(1):3
11. Xia R, Pan Y, Lai H, Liu C, Yan S (2014). Supervised hashing for image retrieval via image representation learning. AAAI Conference on Artificial Intelligence
12. Lai H, Pan Y, Liu Y, and Yan S(2015) Simultaneous feature learning and hash coding with deep neural networks. In CVPR, pages 3270–3278, 6, 13
13. Lin K, Yang HF, Hsiao JH et al. (2015) Deep learning of binary hash codes for fast image retrieval[C]// 2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). IEEE
14. Liu H, Wang R, Shan S, and Chen X (2016) Deep supervised hashing for fast image retrieval. In CVPR, pages 2064–2072, 6
15. Jain H, Zepeda J, Perez P, Gribonval R (2018). Subic: a supervised, structured binary code for image search. 833–842
16. Lu X, Song L, Xie R, Yang X, Zhang W (2017) Deep Binary Representation for Efficient Image Retrieval, Advances in Multimedia, vol. 2017, Article ID 8961091, 10 pages
17. Alzu'Bi A, Amira A, Ramzan N. (2017) Content-based image retrieval with compact deep convolutional features[J]. Neurocomputing, S0925231217306185
18. Tenenbaum JB, Freeman WT (2000) Separating style and content with bilinear models. Neural Comput 12(6):1247
19. Lin TY, Roychowdhury A, Maji S (2016). Bilinear CNN Models for Fine-Grained Visual Recognition. IEEE International Conference on Computer Vision (pp.1449–1457). IEEE
20. Gionis A, Indyk P, Motwani R (1999). Similarity search in high dimensions via hashing., 8(2), 518–529
21. Ji J, Li J, Yan S, Zhang B, Tian Q (2012). Super-bit locality-sensitive hashing. International Conference on Neural Information Processing Systems (Vol. 1, pp.108–116). Curran Associates Inc
22. Kulis B, Grauman K (2009) Kernelized locality-sensitive hashing for scalable image search[C]// 2009 IEEE 12th International Conference on Computer Vision
23. Zhai S, Wu H, Kumar A, Cheng Y, Lu Y, Zhang Z et al. (2017). S3Pool: Pooling with Stochastic Spatial Sampling. Computer Vision and Pattern Recognition (pp.4003–4011). IEEE
24. Boureau YL, Bach F, Lecun Y et al. (2010) Learning mid-level features for recognition[C]// 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE
25. Jarrett K, Kavukcuoglu K, Ranzato M, Lecun Y (2010). What is the best multi-stage architecture for object recognition?. IEEE, International Conference on Computer Vision (Vol.30, pp.2146–2153). IEEE
26. Hadsell R, Chopra S, Lecun Y (2006). Dimensionality Reduction by Learning an Invariant Mapping. Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on (Vol.2, pp.1735–1742). IEEE
27. Schroff F, Kalenichenko D, Philbin J (2015). Facenet: a unified embedding for face recognition and clustering. 815–823
28. Krizhevsky A (2009). Learning multiple layers of features from tiny images
29. Chua TS, Tang J, Hong R, Li H, Luo Z, Zheng Y (2009). NUS-WIDE: a real-world web image database from National University of Singapore. ACM International Conference on Image and Video Retrieval (pp.48). ACM
30. Lecun, Y., & Cortes, C. (2010). The mnist database of handwritten digits
31. Norouzi M, Fleet DJ (2011). Minimal loss hashing for compact binary codes. International Conference on International Conference on Machine Learning (pp.353–360). Omnipress
32. Gong Y and Lazebnik S. Iterative quantization: A procrustean approach to learning binary codes. In CVPR, pages 817–824, 2011.6, 10, 17
33. Weiss Y, Torralba A, Fergus R (2008). Spectral hashing. International Conference on Neural Information Processing Systems (Vol.282, pp.1753–1760). Curran Associates Inc

Fanfeng Zeng is an associate professor at North China University of Technology, china, since 1999. He has long been engaged in the research and development and teaching work of information security, image processing and intelligent control.

Shengda Hu is a master candidate at North China University of Technology, China. His research interests are image search, pattern recognition and deep learning.

Ke Xiao received his Ph.D. degree in circuit and system from Beijing University of Posts and Telecommunications, Beijing, China, in 2008. He has been a professor at North China University of Technology, China, since 2018. He has long been engaged in the research and development and teaching work of image processing, wireless communications, the Internet of Things, and embedded systems.