IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received March 12, 2019, accepted April 6, 2019, date of publication April 16, 2019, date of current version April 29, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2911630

Medical Image Retrieval Based on Convolutional Neural Network and Supervised Hashing

YIHENG CAI, YUANYUAN LI, CHANGYAN QIU, JIE MA, AND XURONG GAO
School of Information and Communications Engineering, Beijing University of Technology, Beijing 100124, China

Corresponding author: Yiheng Cai (caiyiheng@bjut.edu.cn)

This work was supported in part by the National Key Research and Development Program of China under Grant 2017YFC1703302, and in part by the Science and Technology Project of the Beijing Municipal Education Commission under Grant KM201710005028.

ABSTRACT In recent years, with extensive application in image retrieval and other tasks, a convolutional neural network (CNN) has achieved outstanding performance. In this paper, a new content-based medical image retrieval (CBMIR) framework using CNN and hash coding is proposed. The new framework adopts a Siamese network in which pairs of images are used as inputs, and a model is learned to make images belonging to the same class have similar features by using weight sharing and a contrastive loss function. In each branch of the network, CNN is adapted to extract features, followed by hash mapping, which is used to reduce the dimensionality of feature vectors. In the training process, a new loss function is designed to make the feature vectors more distinguishable, and a regularization term is added to encourage the real value outputs to approximate the desired binary values. In the retrieval phase, the compact binary hash code of the query image is achieved from the trained network and is subsequently compared with the hash codes of the database images. We experimented on two medical image datasets: the cancer imaging archive-computed tomography (TCIA-CT) and the vision and image analysis group/international early lung cancer action program (VIA/I-ELCAP). The results indicate that our method is superior to existing hash methods and CNN methods. Compared with the traditional hashing method, feature extraction based on CNN has advantages. The proposed algorithm combining a Siamese network with the hash method is superior to the classical CNN-based methods. The application of a new loss function can effectively improve retrieval accuracy.

INDEX TERMS Binary hash coding, convolutional neural network, image retrieval, loss function.

I. INTRODUCTION

There are an increasing number of medical images, such as X-ray, magnetic resonance imaging (MRI) and computed tomography (CT), that provide necessary anatomical and functional information concerning various body parts for detection, diagnosis, treatment planning, and monitoring, as well as education and medical research [1]. The powerful retrieval of such data has become a crucial task for medical information systems. Traditional methods used for image retrieval are based on annotating images with text; however, image annotation is time consuming and tedious, and it is difficult to describe the contents of medical images with limited words. Recently, content-based image retrieval (CBIR) has received increasing emphasis in the fields of education, the military, and bioinformatics for image retrieval and classification and has also been applied for medical purposes [2]–[4].

CBIR aims to search for similar images by analyzing image content. In addition, the matching of its feature descriptors facilitates the matching of two images [5]; hence, image representations and similarity measure become critical. In the initial stage, intensity histogram-based features were employed for medical image retrieval [6]. However, their retrieval performance was frequently limited, particularly on large databases, because of the low discrimination power of such descriptors. To solve the existing problems, texture-based features were proposed for medical image retrieval. Ojala *et al.* [7] have raised the local binary pattern (LBP), which has exhibited highly promising effects in medical applications as well as less computational complexity in terms of texture classification. In addition, variants of the LBP can also be found, such as a local ternary pattern (LTP) [8], center symmetric LBP (CSLBP) [9],

The associate editor coordinating the review of this manuscript and approving it for publication was Yongping Pan.

center symmetric local ternary pattern (CSLTP) [10], and data-driven LBP (DDLBP) [11]. Other texture descriptors, such as peak valley edge pattern (PVEP) [12], local mesh pattern (LMeP) [13], local mesh peak valley edge pattern (LMePVEP) [14], and local ternary co-occurrence pattern (LTCoP) [15], which are dependent on the relationship of the center pixel with its neighbors as well as the relationship among the surrounding neighbors for a given referenced pixel in an image, were then proposed and applied to medical image retrieval. Shiv *et al.* proposed the local bit-plane decoded pattern (LBDP) [16], which is created by discovering a binary pattern based on the difference of the center pixel's intensity value.

The main drawback of the above traditional artificial methods is their relatively low performance because these visual features often fail to describe the high-level semantic information in the user's mind. Recently, with the rapid progress of deep learning, the features obtained from pretrained CNN models [21]–[28] have realized higher performance and flexibility than traditional descriptors in typical image retrieval tasks (e.g., image retrieval or object recognition). This feature information contains rich image semantic information, which is essential to increase the accuracy of image retrieval. In addition, the feature extraction using CNN models is applied to medical image retrieval [29]. However, the features extracted from CNN models are always high-dimensional, which increases the computational cost and slows the efficiency of the retrieval. To retrieve medical images from large datasets quickly, the hashing-based method, in which the high-dimensional features are projected to a lower dimensional space, and the compact binary codes are subsequently generated, began to be considered for merging into CNN models.

Lu and Member [30] proposed a deep hashing (DH) method to seek multiple hierarchical nonlinear transformations. Xia *et al.*[31] proposed a convolutional neural network hashing (CNNH) method to encode the image into a binary coding scheme. Lai *et al.*[32] proposed a deep neural network hashing (DNNH) method that uses triple-based constraints to describe more complex semantic information. Because of the benefits of the produced binary codes, fast image retrieval can be conducted based on Hamming distance measurement, significantly reducing the computational cost and further enhancing the efficiency of the retrieval.

In this paper, a new applicable image retrieval framework is proposed. In the feature extraction phase, the Siamese network is adopted, and a new loss function is designed that can make the features more distinguishable. The CNN model is combined with hash coding, which can extract image semantic information with lower dimensional feature vectors. In the retrieval phase, the compact binary hash code of the query image is achieved from the trained network and is subsequently compared with the hash codes of the database images.

We propose this method especially for supervised learning. We consider that when a powerful learning model such as deep CNN is employed and the data labels are accessible, the compact binary codes could be studied by optimizing the loss function. The main purpose of our method is to enhance the discrimination ability of binary hash codes to indicate that similar images should have similar binary hash codes and vice versa. Our contributions are as follows:

1. The CNN and hash coding-based method is applied to medical image retrieval, which has not been attempted in previous studies.
2. A Siamese network is adopted in which pairs of images are used as inputs, and a model is learned to make images belonging to the same class have similar features by using weight sharing and a contrastive loss function.
3. Unlike other CNN + hashing methods, in which the processes of network learning and binary code formation are separated, some mismatch problems may occur. In our framework, thanks to the Siamese network, the CNN model can be followed by hash coding and can learn features and perform hash coding in one network.
4. In the coding process, a new loss function is designed to make the binary codes more distinguishable in which a regularization term is added to encourage the real value outputs to approximate the desired binary values.

The rest of the paper is organized as follows. Section ii mainly introduces the proposed approach as well as the loss function design and optimization. Extensive experimental results and comparison analysis are shown in Section iii. The conclusion is presented in Section iv.

II. PROPOSED METHOD

A. NETWORK STRUCTURE

As the Siamese network [36] has a significant effect on image recognition tasks, such as judging whether two similar images are ''similar'', we make improvements on the Siamese convolutional network and form a network structure, as shown in Fig. 1.

The structure consists of two identical branches that share weights and parameters. Each branch poses a deep neural net and includes a set of convolutional layers, pooling layers, and fully connected layers. Pairs of images are fed into the branches during training. The outputs of these branches are fed to a contrastive loss function. From labeled examples of matching and nonmatching image pairs, the contrastive loss function tries to minimize the distance between the features of similar image pairs and maximize it for dissimilar pairs. Therefore, the network structure could learn the optimal feature representations of the input pairs, where matched images in a pair are pulled closer and unmatched images are pushed further away.

For each branch of the network, the convolutional layers of CNN are used to learn the distinctive image feature expression. According to Fig. 1, the network framework consists of three convolution layers, three pooling layers and two fully

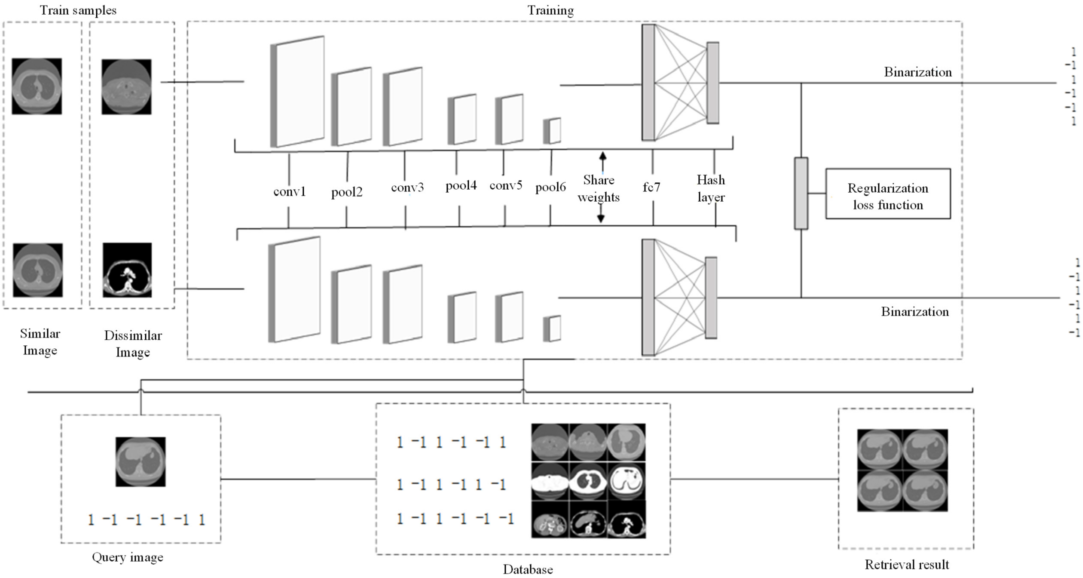

FIGURE 1. The network framework of our method. The network consists of two identical branches that share weights and parameters. Each branch consists of three convolution layers, three pooling layers and two fully connected layers. The filter size of the three convolutions is 5 * 5 with 32, 32 and 64 filters, respectively. The pooling layer has a window size of 3 * 3. The first full connection layer contains 500 nodes, and the second has k nodes that are the length of the hash codes. The outputs of these branches are fed to a new optimized loss function, which attempts to minimize the distance between the features of similar image pairs and maximize the distance between the features of dissimilar pairs. After network training, the binary code is obtained by binary processing. Finally, the compact binary hash code of the query image is compared with the hash codes of database images to retrieve the most similar image.

connected layers. The filter sizes of the three convolutions are 5 * 5 with 32, 32 and 64 filters, respectively, and the step is 1. The pooling layers have a window size of 3 * 3, and the step is 2. Max pooling is performed in the first pooling layer, while average pooling is performed in the other two layers. The purpose of the full connection layers is to learn the hash function. There are 500 nodes in the first full connection layer, and the second layer involves k nodes, where k is the length of the hash code. All convolution layers and the first full connection layer are activated by the ReLU function.

Usually, in hash coding, the binary constraint must be applied to the feature vectors to achieve binary codes. However, in our framework, the binarization of hash codes is not immediately implemented after hash mapping. Instead, binarization is carried out after network training because the binarization can cause the computation error of the loss function, which will affect the accuracy of network training. In addition, the binary constraint is discrete and nonconductive, and the error cannot be propagated backward during network training, which makes it impossible to directly optimize the loss function.

Our network framework is different from other CNN-based image retrieval methods. In the CNNH network, feature learning and hashing function learning to generate hash codes cannot be performed simultaneously because the two processes are separated. As our framework takes advantage of the infrastructure of the Siamese CNN, which is an end-to-end learning process, it can perform both feature learning and hash coding learning in one network. In the first part of the network, CNN is used to learn the image expression; in the second part, a hash function is learned to map the image expression into a low-dimensional vector by using full connection layers, and after training, the vector is encoded into a hash code. Moreover, the Siamese CNN network allows pairs of images as input, effectively improving the distinguishability of the extracted features.

To make binary hash coding of similar images more similar, and therefore to improve the distinction of binary codes, a new reasonable loss function is designed that uses the punishment items to make similar images have similar network outputs and adds a regularization item to make the real value outputs approximate the desired binary values. The specific process is shown in Section ii-b of this article.

After the training procedure, in the image retrieval phase, the compact binary hash code of the query image, which is achieved from the trained network, is compared with the hash codes of the database images. The images with the hash codes that are most similar to the query image will be retrieved.

B. LOSS FUNCTION DESIGN AND OPTIMIZATION

The loss function is optimized so that the network outputs of similar images are closer, and the outputs of dissimilar images are further away. To avoid the optimization of the nondifferentiable loss function, the network outputs are extended to the real value, and a regularization term is defined to inspire the real value outputs to approximate the desired discrete value. Based on this framework, it is easy to obtain a binary hash code of an image by quantifying the network output after the network training.

Suppose G represents grayscale space; the aim of the whole framework is to study a mapping from G to k-bit binary code: $f : G \to \{-1, 1\}^k$, which makes similar images be coded as similar binary codes. Given a pair of images $I_1, I_2 \in G$,, the corresponding network output is b1, b2, and in general, the loss function [36] for this pair of images is defined as follows:

$$\begin{aligned} L(\boldsymbol{b}_1, \boldsymbol{b}_2, y) = & \frac{1}{2}(1-y)D(\boldsymbol{b}_1, \boldsymbol{b}_2) \\ & + \frac{1}{2}y \max(m - D(\boldsymbol{b}_1, \boldsymbol{b}_2), 0) \end{aligned} \tag{1}$$

If two images are similar, we define $y = 0$; otherwise, $y = 1$, where $D(\cdot, \cdot)$ represents the distance between two output vectors, and $m > 0$ is a margin threshold parameter. When $D < m$, the first item in (1) penalizes similar images mapping to different output vectors, and the second item penalizes dissimilar images mapping to similar output vectors. For a training set containing N pairs of images, the following loss function is obtained:

$$\begin{aligned} L = & \sum_{i=1}^{N} L(\boldsymbol{b}_{i,1}, \boldsymbol{b}_{i,2}, y) \\ & \textit{where}, \quad i \in \{1, \ldots, N\}, \quad j \in \{1, 2\} \end{aligned} \tag{2}$$

To restrict the range of parameters and make the network outputs approximate the binary hash codes, which will be beneficial to the follow-up binarization, a regularization term is added to the loss function, and the Euclidean distance is used. The proposed loss function is shown as follows:

$$\begin{aligned} L_r(\boldsymbol{b}_1, \boldsymbol{b}_2, y) = & \frac{1}{2}(1-y)\,\|\boldsymbol{b}_1 - \boldsymbol{b}_2\|_2^2 \\ & + \frac{1}{2}y \max(m - \|\boldsymbol{b}_1 - \boldsymbol{b}_2\|_2^2, 0) \\ & + \alpha(\||\boldsymbol{b}_1| - \mathbf{1}\|_1 + \||\boldsymbol{b}_2| - \mathbf{1}\|_1) \end{aligned} \tag{3}$$

where L with subscript r represents the relaxation loss function, $\mathbf{1}$ represents all-one vector, $\|\cdot\|_1$ indicates the L_1 norm of the vector, and α is the weight parameter that controls the regularization term. The regularization term can restrict the elements of the vectors to ± 1.

We choose the L_1 norm rather than the higher-order norm because of its lower computational cost and because it is helpful in speeding up the training process. According to (3), we can write the overall loss function as follows:

$$\begin{aligned} L_r = & \sum_{i=1}^{N} \{\frac{1}{2}(1-y_i)\,\|\boldsymbol{b}_{i,1} - \boldsymbol{b}_{i,2}\|_2^2 \\ & + \frac{1}{2}y_i \max(m - \|\boldsymbol{b}_{i,1} - \boldsymbol{b}_{i,2}\|_2^2, 0) \\ & + \alpha(\left\| |\boldsymbol{b}_{i,1}| - \mathbf{1} \right\|_1 + \left\| |\boldsymbol{b}_{i,2}| - \mathbf{1} \right\|_1)\} \end{aligned} \tag{4}$$

The network training uses back propagation and gradient descent, and the gradient about $\boldsymbol{b}_{i,j}$ needs to be calculated. Because the maximum and absolute value in the loss function are nondifferentiable at some points, the subgradients are used, and the subgradients of these points are defined as $\mathbf{1}$. The subgradients of the first two parts and the third part of (4) (regularization terms) are as follows:

$$\frac{\partial Part\ 1}{\partial \boldsymbol{b}_{i,j}} = (-1)^{j+1}(1-y_i)\left(\boldsymbol{b}_{i,1} - \boldsymbol{b}_{i,2}\right)$$

$$\frac{\partial P\,\mathrm{art}\,2}{\partial \boldsymbol{b}_{i,j}} = \begin{cases} cc(-1)^j y_i \left(\boldsymbol{b}_{i,1} - \boldsymbol{b}_{i,2}\right), & \left\|\left(\boldsymbol{b}_{i,1} - \boldsymbol{b}_{i,2}\right)\right\|_2^2 < m \\ 0, & \text{others} \end{cases} \tag{5}$$

$$\frac{\partial \mathrm{Re}}{\partial \boldsymbol{b}_{i,j}} = \alpha\delta\left(\boldsymbol{b}_{i,j}\right), \tag{6}$$

where $\boldsymbol{\delta}(\mathbf{X}) = (\delta(x_1), \delta(x_2), \ldots, \delta(x_k))$, $x_k\mathbf{X}$,
$$\delta(x) = \begin{cases} 1, & -1 \le x \le 0 \text{ or } x \ge 1 \\ -1, & \text{others} \end{cases}$$

After the network training, binary hash coding is easily obtained by binary processing to the vector $\boldsymbol{b}$, which is *sign*($\boldsymbol{b}$). Unlike the current CNN-based hashing methods, which use nonlinear functions such as *sigmoid* or *tanh* to approximate quantization steps, our approach outputs the real value, and a regularization item is then defined to encourage the real value outputs to approximate the desired discrete value, which can effectively accelerate the training speed.

C. TRAINING NETWORK

The network parameters adopt the "Xavier" [33] initialization method. If each layer is randomly initialized with N (0, 0.01), the data distribution of each layer is inconsistent. Along with the increasing number of network layers, neurons will be concentrated on very large values or very small values, which is not conducive to the transmission of information. In contrast, the "Xavier" initialization method can ensure that the data distribution is the same (mean variance is consistent) and accelerate the convergence. During training, the number of data processed at one time is 200, the weight attenuation coefficient is 0.004, the initial learning rate is set to 0.001, and the number of data is reduced by 40% after 20,000 iterations. To learn network models that correspond to different code lengths, it would be wasteful to train each model from scratch because the first few layers can be shared by these models. In addition, when the length of the code increases, the model will include more parameters in the output layer, making it tend to overfit. Therefore, in the training process, a smaller hash code length is initially set; after a number of parameters are obtained, the code length is increased. Finally, the values of these parameters are fine-tuned to achieve the target model with the desired encoding length.

D. IMAGE RETRIEVAL FOR PROPOSED METHDOLOGY

The methodology for experimentation with the proposed algorithm is as follows:

Step 1: Train the proposed network using the training method presented in Section ii-c.

Step 2: Input the images from the dataset to the trained network. Obtain the k-bit feature vectors of the images and calculate the binary hash codes of the images through binary processing.

Step 3: Obtain the hash codes of the query image using a process similar to that of step 2.

Step 4: Calculate the Hamming distances between the query image and the images in the database. Retrieve the most relevant images based on correlation.

In our experiments, we obtain a set K of all computed distances. The values in set K are grouped in ascending order, with the smallest value corresponding to the most similar image retrieved. We retrieve the first k images, the number of which varies from 5 to 200.

III. EXPERIMENT AND ANALYSIS

We conduct two experiments on two public datasets to evaluate our proposed method for medical image retrieval and compare it with several other methods.

The first experiment is intended to explore the benefits of feature extraction based on CNN. To do so, we compare our proposed approach with several classical hashing methods, including locality-sensitive hashing (LSH) [17], spectral hashing (SH) [18], minimal loss hashing (MLH) [19], supervised hashing with kernels (KSH) [20], and deep hashing (DH) [30].

The second experiment evaluates our framework by comparing our approach with two CNN-based hashing methods, CNNH [31] and DNNH [32].

Because the classical hashing methods have no advantage on big datasets, the first experiment was performed on a relatively small dataset. However, deep learning methods always demand large-scale training data, so a large dataset is used for the second experiment. Based on these considerations, two different medical image databases are constructed.

A. DATASET

Vision and image analysis group/international early lung cancer action program (VIA/I-ELCAP) dataset [34]: The data used in the experiment were collected by the early lung cancer action program (ELCAP) and the vision and image analysis (VIA) research groups, which contain 12645 medical images of whole-lung computed tomography (CT) scans. The CT scans were achieved in a single breath hold with a 1.25 mm slice thickness, and the size of each image was 512*512. After eliminating the blurred and repeated images, we had 12000 images of 6 different body parts as the dataset for the first experiment.

The cancer imaging archive (TCIA) dataset [35]: The data used in the experiment were collected from The Cancer Imaging Archive (TCIA), including the collections of NSCLC-Radionics (51513 images), PANCREAS-CT (19328 images), RIDER NEURO MRI (70220 images), TCGA-BLCA (69481 images), and RIDER Lung CT (15419 images), for a total of 225961 images. In the second experiment, we used 199000 images of 4 different body parts, including 60000 lung images, 19000 pancreas images, 60000 neurological images, and 60000 urothelial bladder images. The size of each image is 512*512. Fig. 2 provides some examples of the dataset.

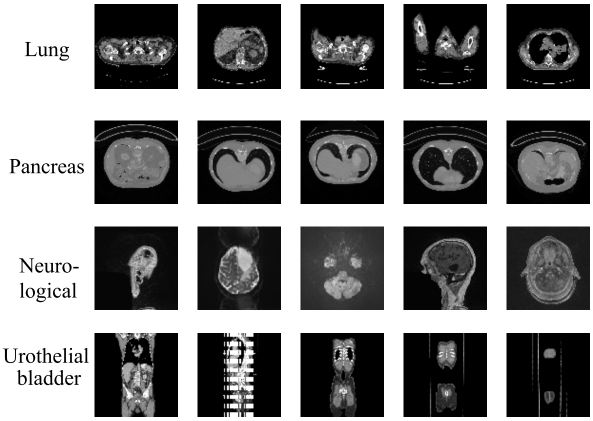

FIGURE 2. **Sample images from the TCIA dataset, including lung images, pancreas images, neurological images, and urothelial bladder images.**

B. EVALUATION METRICS

To assess performance in our experiments, mean average precision (mAP), which has been shown to have particularly good discrimination and stability, is used to measure the ranking quality of retrieval database images:

$$mAP = \frac{1}{Q}\sum_{q=1}^{Q} AP(q) \tag{7}$$

where Q is the number of relevant images. For a query image, AP is the average of the precision values achieved for the set of top k images after each relevant image is retrieved. This value is then averaged over the information needs.

$$AP = \frac{\sum_{p=1}^{M} \prod(r_p) prep@p}{M_r} \tag{8}$$

where $\prod(.) \in \{0, 1\}$ is an indicator function, $r_p > 0$, r_p is the similarity level of the images on the $p - th$ position of a ranking with the query image, and $M_r > 0$ is the number of relevant images. $prep@p$ is calculated by taking the average of the levels of similarity of the top-p images to the query image:

$$prep@p = \frac{1}{p}\sum_{i=1}^{p} r_i, \tag{9}$$

$prep@p$ is the precision weighted by the similarity level of each image.

C. RESULTS ON THE VIA/I-ELCAP DATASET

In this experiment, we randomly selected 1000 samples as query images and adopted the remaining 11000 images as the training set. In the training sample set, two images from different categories (different body parts) form dissimilar image pairs. The images from the same category are judged

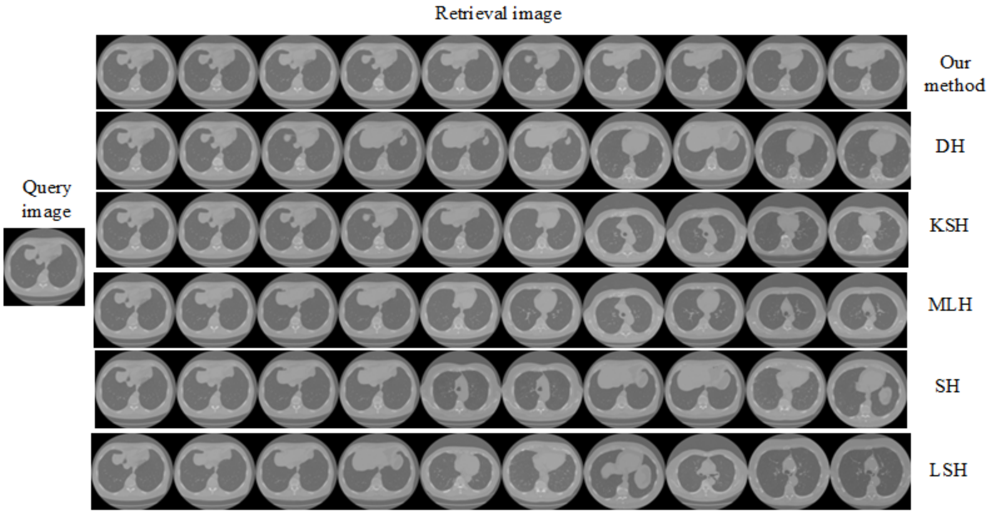

FIGURE 3. Top 10 retrieval images from the VIA/I-ELCAP dataset by varying methods for the same query image; the closest images are shown at the top.

manually, and the similar images are selected to form the similar image pairs. The ratio of the number of dissimilar image pairs to similar image pairs is 3:2. To assess the retrieval performance, this method is compared with several other methods (LSH, SH, KSH, MLH and DH) by applying them to the VIA/I-ELCAP dataset.

Fig. 3 presents the retrieval results of the top 10 images using different methods. In general, image retrieval based on deep learning methods performs better than traditional hash-based methods on image representation. When the datasets have a relatively small number of images, all of the above methods can effectively retrieve images that are similar to the query images. The top 10 images retrieved by our method have greater similarity to the query image, which shows that our method is more capable of distinguishing between similar and dissimilar images by improving the loss function.

We further analyze the mAP values of the retrieval results with different hash bits. The mAP values of the different tested methods with different hash bits regarding the top 100 retrieval images are presented in Table 1. Obviously, the proposed method achieves substantially better retrieval accuracies (mAP) than other methods. For example, in Table 1, in comparison with the best competitor, DH, the mAP values of the proposed method suggest a relative increase of 0.0846, 0.0822, 0.0767, and 0.0811 in 16 bits, 32 bits, 48 bits, and 64 bits, respectively. This suggests that our CNN method has a better performance than the hashing and DH methods. Furthermore, we found that in 48 hash bits, the image retrieval performance is optimal, so in the follow-up experiment, the hash bit is set to 48 to generate a binary hash code.

The methodology is tested by varying the number of retrieval images from 5 to 100. Fig. 4 presents the mAP values of different methods for 5, 10, 15, 25, 35, 50, 65, 80, and 100 retrieval images. We can see from the figure that when the number of top k is small, the mAP values of all tested methods are relatively high; when k increases, the mAP values decrease continually. Among all the tested methods, the mAP values of our method are highest, regardless of the number of images retrieved, indicating that our method outperforms the other tested methods in retrieval accuracy.

TABLE 1. Map of hamming ranking in different numbers of bits on the via/i-elcap dataset.

Method	mAP			
	16 bit	32 bit	48 bit	64 bit
Ours	0.7932	0.8057	0.8163	0.8092
DH	0.7086	0.7235	0.7396	0.7281
KSH	0.5914	0.6053	0.6198	0.6085
MLH	0.5039	0.5228	0.5534	0.5261
SH	0.4531	0.4958	0.4871	0.4798
LSH	0.3271	0.3447	0.3653	0.3579

D. RESULTS ON THE TCIA DATASET

In the current experiment, we randomly choose 9,000 samples as the query images and the remaining 190000 as the training set. The collection of dissimilar image pairs and similar image pairs is the same as that in the experiment on the VIA/I-ELCAP dataset. In this experiment, we no longer compared the proposed approach with the traditional hashing methods because the computational complexities of hashing methods are relatively great, which makes the experiment extremely time-consuming when these methods are applied to the TCIA dataset with such great data volume. In addition, the retrieval accuracies of hashing methods are not as high as those of deep learning methods. Instead, our method is compared with the

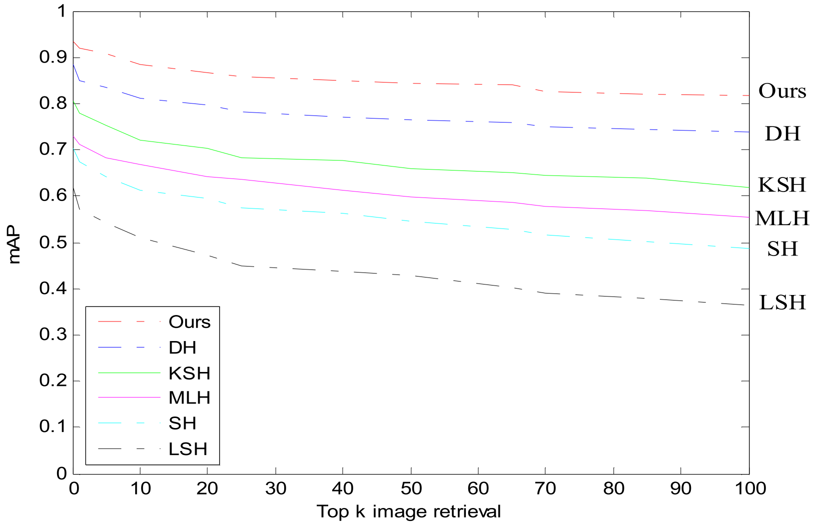

FIGURE 4. The mAP values of image retrieval with 48 bits of the VIA/I-ELCAP dataset.

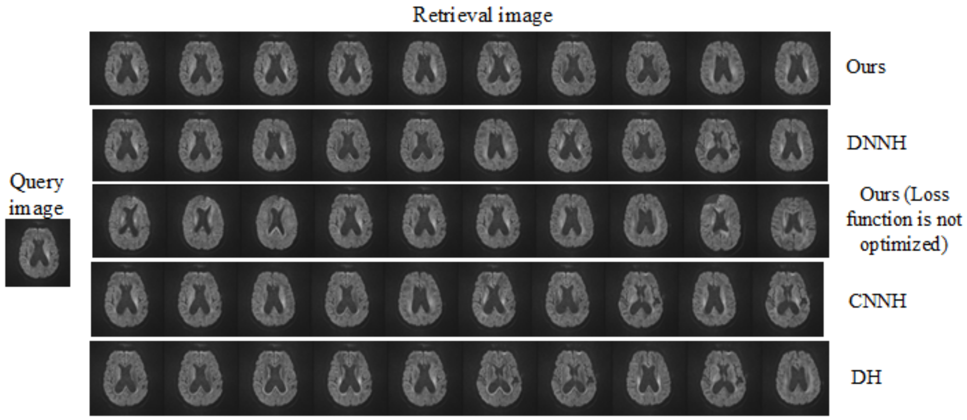

FIGURE 5. Top 10 retrieval images from the TCIA dataset using various methods for the same query image. The closest images are shown at the top.

deep learning methods DH, CNNH, and DNNH, which have all been shown to be expert at image retrieving. To test the effect of the optimization of the loss function, the proposed method with unoptimized loss functions is also compared.

Fig. 5 exhibits the retrieval results of the top 10 images under different methods with 48-bit hash. The figure clearly shows that among all of the methods, the retrieval result of our method is the closest to the query image. In addition, our method uses low-dimensional feature vectors and maintains high discriminative performance.

Fig. 6 presents the mAP values of the different top k images retrieved on the TCIA dataset. As k increases, the mAP values of DH decrease rapidly. The performances of CNNH, our framework without loss function optimization, and DNNH are relatively close. Although the performance of our method without loss function optimization is the same as that of CNNH, which is also based on CNN and hashing, it is better than that of CNNH, which indicates that the introduction of the Siamese network can improve the retrieval accuracy. Among all the tested methods, the proposed method has the best performance because it maintains high mAP values as k changes.

We also present the classification accuracy during the training phase and retrieval time consumption of all tested methods after 50000 iterations on the TCIA dataset in Table 2. Obviously, our proposed method has a stable performance compared with other methods, which have a 3.95%∼6.74% improvement. In terms of efficiency, regarding the time to retrieve an image from the TCIA database, other methods have a 34.07%∼157.92% time delay relative to our method.

In summary, we can see that the methods based on deep learning have significantly improved compared with traditional hash methods, while our method performs better than other typical deep learning methods, indicating that the proposed end-to-end learning framework is effective. This is mainly because the processes of network learning and binary encoding in DH and CNNH are separated, and there is a certain mismatch between the two phases. In DNNH, the usage

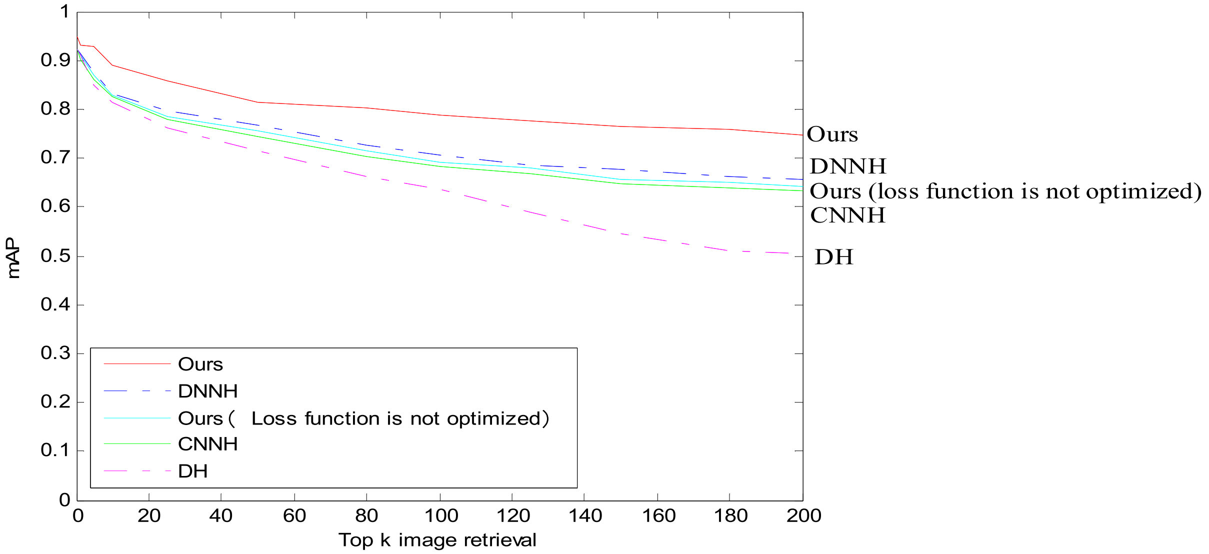

FIGURE 6. The mAP values of image retrieval with 48 bits of the TCIA dataset.

TABLE 2. Map of hamming ranking in different numbers of bits on the via/i-elcap dataset.

Method	Accuracy (%)	Time consumption of retrieving an image (ms)
(1) DH	80.23	987.724
(2) CNNH	81.53	683.766
(3) Ours (loss function is not optimized)	82.15	605.479
(4) DNNH	83.02	513.455
(5) Ours	86.97	382.952

of sigmoid function, a nonlinear and parametric threshold function, makes network training more difficult. In addition, the performance comparison between the approaches with and without the optimized loss function proves that the improvement of the loss function and the introduction of the regularization term can obviously improve the discrimination ability of binary code.

IV. CONCLUSION

The proposed medical image retrieval method based on CNN and supervised hash mainly contributes as follows: First, the network framework uses a Siamese network in which pairs of images (similar/dissimilar) are used as input; second, nonlinear feature learning and hash coding are combined to obtain the image representation; and finally, the reconstruction of the loss function, for which we propose a regularization term to reduce the difference between real-valued network outputs and binary codes, increases the ability of the network to distinguish images. Regarding efficiency, experiments show that the proposed method can retrieve similar images faster than traditional hash methods and certain typical deep learning methods. At the same time, the mAP values of the proposed method are higher than those of other methods, showing that the proposed method is effective in further improving the accuracy of image retrieval.

REFERENCES

[1] H. Müller, N. Michoux, D. Bandon, and A. Geissbuhler, "A review of content-based image retrieval systems in medical applications—Clinical benefits and future directions," *Int. J. Med. Inform.*, vol. 73, no. 1, pp. 1–23, 2004.

[2] H. Müller, A. Rosset, J.-P. Vallee, and A. Geissbuhler, "Comparing features sets for content-based image retrieval in a medical-case database," *Proc. SPIE*, vol. 5371, pp. 99–110, Apr. 2004.

[3] J. C. Felipe, A. J. M. Traina, and C. Traina, "Retrieval by content of medical images using texture for tissue identification," in *Proc. 16th IEEE Symp. Comput.-Based Med. Syst.*, Jun. 2003, pp. 175–180.

[4] C. B. Akgül, D. L. Rubin, S. Napel, C. F. Beaulieu, H. Greenspan, and B. Acar, "Content-based image retrieval in radiology: Current status and future directions," *J. Digit. Imag.*, vol. 24, no. 2, pp. 208–222, 2011.

[5] K. N. Manjunath, A. Renuka, and U. C. Niranjan, "Linear models of cumulative distribution function for content-based medical image retrieval," *J. Med. Syst.*, vol. 31, pp. 433–443, Dec. 2007.

[6] T. Ojala, M. Pietikäinen, and D. Harwood, "A comparative study of texture measures with classification based on featured distributions," *Pattern Recognit.*, vol. 29, no. 1, pp. 51–59, 1996.

[7] X. Tan and B. Triggs, "Enhanced local texture feature sets for face recognition under difficult lighting conditions," *IEEE Trans. Image Process.*, vol. 19, no. 6, pp. 1635–1650, Jun. 2010.

[8] M. Heikkilä, M. Pietikäinen, and C. Schmid, "Description of interest regions with local binary patterns," *Pattern Recognit.*, vol. 42, no. 3, pp. 425–436, 2009.

[9] R. Gupta, H. Patil, and A. Mittal, "Robust order-based methods for feature description," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 334–341.

[10] J. Ren, X. Jiang, J. Yuan, and G. Wang, "Optimizing LBP structure for visual recognition using binary quadratic programming," *IEEE Signal Process. Lett.*, vol. 21, no. 11, pp. 1346–1350, Nov. 2014.

[11] S. Murala and Q. M. J. Wu, "Peak valley edge patterns: A new descriptor for biomedical image indexing and retrieval," in *Proc. IEEE CVPR Workshops*, Jun. 2013, pp. 444–449.

[12] S. Murala and Q. M. J. Wu, "Local mesh patterns versus local binary patterns: Biomedical image indexing and retrieval," *IEEE J. Biomed. Health Inform.*, vol. 18, no. 3, pp. 929–938, May 2014.

[13] S. Murala and Q. M. J. Wu, "MRI and CT image indexing and retrieval using local mesh peak valley edge patterns," *Signal Process., Image Commun.*, vol. 29, pp. 400–409, Mar. 2014.
[14] S. Murala and Q. J. Wu, "Local ternary co-occurrence patterns: A new feature descriptor for MRI and CT image retrieval," *Neurocomputing*, vol. 119, no. 7, pp. 399–412, 2013.
[15] S. R. Dubey, S. K. Singh, and R. K. Singh, "Local bit-plane decoded pattern: A novel feature descriptor for biomedical image retrieval," *IEEE J. Biomed. Health Inform.*, vol. 20, no. 4, pp. 1139–1147, Jul. 2016.
[16] A. Gionis, P. Indyk, and R. Motwani, "Similarity search in high dimensions via hashing," in *Proc. VLDB*, 1999, pp. 518–529.
[17] Y. Weiss, A. Torralba, and R. Fergus, "Spectral hashing," in *Proc. Adv. Neural Inf. Process. Syst.*, 2008, pp. 1753–1760.
[18] M. Norouzi and D. J. Fleet, "Minimal loss hashing for compact binary codes," in *Proc. ICML*, vol. 11, 2011, pp. 353–360.
[19] W. Liu, J. Wang, R. Ji, Y.-G. Jiang, and S.-F. Chang, "Supervised hashing with kernels," in *Proc. CVPR*, 2012, pp. 2074–2081.
[20] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2012, pp. 1097–1105.
[21] Y. Jia *et al.*, "Caffe: Convolutional architecture for fast feature embedding," in *Proc. ACM Int. Conf. Multimedia*, 2014, pp. 675–678.
[22] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. ICLR*, 2015, pp. 1–14.
[23] P. Liu, J.-M. Guo, C.-Y. Wu, and D. Cai, "Fusion of deep learning and compressed domain features for content-based image retrieval," *IEEE Trans. Image Process.*, vol. 26, no. 12, pp. 5706–5717, Dec. 2017.
[24] J. Li, N. Wang, Z.-H. Wang, H. Li, C.-C. Chang, and H. Wang, "New secret sharing scheme based on faster R-CNNs image retrieval," *IEEE Access*, vol. 6, pp. 49348–49357, Apr. 2018.
[25] F. Xie, H. Fan, Y. Li, Z. Jiang, R. Meng, and A. Bovik, "Melanoma classification on dermoscopy images using a neural network ensemble model," *IEEE Trans. Med. Imag.*, vol. 36, no. 3, pp. 849–858, Mar. 2017.
[26] N. Liu, L. Wan, Y. Zhang, T. Zhou, H. Huo, and T. Fang, "Exploiting convolutional neural networks with deeply local description for remote sensing image classification," *IEEE Access*, vol. 6, pp. 11215–11228, 2018.
[27] S. U. Rehman, S. Tu, Y. Huang, and O. U. Rehman, "A benchmark dataset and learning high-level semantic embeddings of multimedia for cross-media retrieval," *IEEE Access*, vol. 6, pp. 67176–67188, 2018.
[28] A. Khatami, M. Babaie, A. Khosravi, H. R. Tizhoosh, S. M. Salaken, and S. Nahavandi, "A deep-structural medical image classification for a Radon-based image retrieval," in *Proc. IEEE 30th Can. Conf. Elect. Comput. Eng. (CCECE)*, Windsor, ON, USA, Apr./May 2017, pp. 1–4.
[29] J. Lu, V. E. Liong, and J. Zhou, "Deep hashing for scalable image search," *IEEE Trans. Image Process.*, vol. 26, no. 5, pp. 2352–2367, May 2017.
[30] R. Xia, Y. Pan, H. Lai, C. Liu, and S. Yan, "Supervised hashing for image retrieval via image representation learning," in *Proc. Nat. Conf. Artif. Intell.*, vol. 3, 2014, PP. 2156–2163.
[31] H. Lai, Y. Pan, Y. Liu, and S. Yan, "Simultaneous feature learning and hash coding with deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, vol. 7, no. 12, Jun. 2015, pp. 3270–3278.
[32] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," *J. Mach. Learn. Res.*, vol. 9, pp. 249–256, Jan. 2010.
[33] *VIA/I-ELCAP Database*. Accessed: Mar. 2018. [Online]. Available: http://www.via.cornell.edu/databases/lungdb.html
[34] *TCIA Database*. Accessed: Mar. 2018. [Online]. Available: https://wiki.cancerimagingarchive.net/display/Public/Wiki
[35] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *Proc. CVPR*, 2005, pp. 539–546.

YIHENG CAI received the Ph.D. degree in pattern recognition and intelligent system from the Beijing University of Technology, China, in 2006, where she is currently an Associate Professor with the School of Information and Communications Engineering. She has published over 60 papers in journals, book chapters, and conferences. Her research interests include image processing and pattern recognition. She is a member of the China Computer Federation.

YUANYUAN LI received the B.S. degree in communication engineering from the School of Information and Communications Engineering, Beijing University of Technology, Beijing, China, in 2017, where she is currently pursuing the M.S. degree in information and communication engineering. Her research interests include video analysis and image retrieval.

CHANGYAN QIU received the B.S. degree in electronic science and technology from Huanghuai University, Henan, China, in 2014, and the M.S. degree in electronic science and technology from the Beijing University of Technology, Beijing, China, in 2018. His research interests include image processing and image retrieval.

JIE MA received the B.S. degree in electronic information engineering from the School of Information and Communications Engineering, Beijing University of Technology, Beijing, China, in 2016, where he is currently pursuing the M.S. degree in information and communication engineering. His research interests include image segmentation and image retrieval.

XURONG GAO received the B.S. degree in biomedical engineering from the North University of China, Shanxi, China, in 2015, and the M.S. degree in electronic science and technology from the Beijing University of Technology, Beijing, China, in 2018. Her research interests include medical image processing and image segmentation.