information

Article

Image Retrieval Algorithm Based on Locality-Sensitive Hash Using Convolutional Neural Network and Attention Mechanism

Youmeng Luo, Wei Li *, Xiaoyu Ma and Kaiqiang Zhang

School of Computer and Information Engineering, Xiamen University of Technology, Xiamen 361024, China
* Correspondence: drweili@hotmail.com

Abstract: With the continuous progress of image retrieval technology, in the field of image retrieval, the speed of a search for a desired image from a great deal of image data becomes a hot issue. Convolutional Neural Networks (CNN) have been used in the field of image retrieval. However, many image retrieval systems based on CNN have a poor ability to express image features, resulting in a series of problems such as low retrieval accuracy and robustness. When the target image is retrieved from a large amount of image data, the vector dimension after image coding is high and the retrieval efficiency is low. Locality-sensitive hash is a method to find similar data from massive high latitude data. It reduces the data dimension of the original spatial data through hash coding and conversion, and can also maintain the similarity between the data. The retrieval time and space complexity are low. Therefore, this paper proposes a locality-sensitive hash image retrieval method based on CNN and the attention mechanism. The steps of the method are as follows: using the ResNet50 network as the feature extractor of the image, adding the attention module after the convolution layer of the model, and using the output of the network full connection layer to retrieve the features of the image database, then using the local-sensitive hash algorithm to hash code the image features of the database to reduce the dimension and establish the index, and finally measuring the features of the image to be retrieved and the image database to get the most similar image, completing the content-based image retrieval task. The method in this paper is compared with other image retrieval methods on corel1k and corel5k datasets. The experimental results show that this method can effectively improve the accuracy of image retrieval, and the retrieval efficiency is significantly improved. It also has higher robustness in different scenarios.

Keywords: image retrieval; convolutional neural networks; attention mechanism; locality sensitive hash

check for updates

Citation: Luo, Y.; Li, W.; Ma, X.; Zhang, K. Image Retrieval Algorithm Based on Locality-Sensitive Hash Using Convolutional Neural Network and Attention Mechanism. *Information* **2022**, *13*, 446. https://doi.org/10.3390/info13100446

Academic Editors: Miltiadis D. Lytras and Andreea Claudia Serban

Received: 1 September 2022
Accepted: 20 September 2022
Published: 24 September 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

With the advent of the information age, the number of digital images on the internet is growing exponentially. How to search for desired images from many image databases has become a hot issue that needs to be solved urgently [1], and content-based image retrieval (CBIR) can solve this problem. CBIR is a technology based on computer vision, which is used to search similar images from databases. Images are represented by high-dimensional features, which use different distance metric measures to calculate the similarity between images [2]. The retrieval performance is based on the shape, structure, color, or other features extracted from the image and the quality of matching.

Generally, content-based image retrieval includes two parts, feature extraction and similar image search. In the aspect of feature extraction, the Convolutional Neural Network can be used to extract features. However, the training samples of the actual data set are limited, which will lead to low robustness of the model, and extracting effective features is the key to improving the accuracy of image retrieval. However, most of the current image retrieval technologies only focus on the whole image feature extraction, and cannot

identify the salient parts. Therefore, researchers have added an attention mechanism to the neural network. Attention mechanism is a method of imitating the human brain. It tends to selectively acquire some important parts of the observed things according to needs, while ignoring the unimportant parts [3], so that the model can make more accurate judgments. In the aspect of similar image searches, compared with the methods of manually selecting image features to express image content, such as SIFT, HOG, etc., the high-dimensional image features learned by Convolutional Neural Network can better express image information, but its computational complexity will also increase. In the past, researchers compared the feature vectors of high-dimensional query images with those of images in the database one by one. This method consumes feature storage space and matching time. The hash-based image retrieval technology can map high dimensional feature vectors to the same space while maintaining similarity through the hash algorithm to obtain binary hash codes that are much lower than the original dimensions. At the same time, the generated hash codes can be used to build efficient indexes, which not only realize data dimension reduction, but also improve the efficiency of feature matching.

Given the above problems, this paper designs a locality-sensitive hash image retrieval method based on CNN and attention mechanism. The main contributions of this paper are:

1. Preprocessing before CNN training. Adopting standard data enhancement methods for existing training data, such as random scaling, rotation, clipping, noise addition and color contrast change, to increase the sample size, avoiding overfitting and promoting the robustness of the model.
2. By adding a simple and effective attention module (CBAM) to the Convolutional Neural Network, it can be extensively adopted to promote the representation ability of CNN and improve image retrieval accuracy to a certain extent.
3. According to the features extracted by CNN, a locality-sensitive hashing dimension reduction method is designed to build a hash index, which solves the problem of large-scale and high-dimensional images and greatly shortens the retrieval time.

2. Related Work

In the early CBIR system, people used manually extracted features to represent the content of images. In recent years, with the appearance of large-scale labeled datasets such as ImageNet, CNN have been widely used for deep learning. People have found that deep convolution neural network has a strong feature extraction ability in image processing, and the use of Convolution Neural Network has become the mainstream feature extraction method [4]. As various CNN have been put forward, such as AlexNet, VGGNet, ResNet, and GoogleNet, they have achieved good results in many fields such as image classification [5], object detection, and image retrieval.

In the field of image retrieval, the attention mechanism also has relevant applications. Noh et al. [6] posed a spatial domain attention model, which was added to the convolution layer of the CNN to select the key points in the image and raise the accuracy of image retrieval. Li ZongMin et al. [7] employed an attention mechanism to the field of hand-drawn image retrieval, added an attention module to the VGG16 network, and used the extracted 4068-dimensional feature vector for hand-drawn image retrieval, achieving excellent retrieval results. Ng et al. [8] combined the second-order loss and the second-order spatial attention to re-weight the features to perform image matching. However, the feature dimension extracted by these methods is too high, which is not conducive to calculation.

In 2012, Krizhevsky et al. [9] used the full connection layer output of the AlexNet network proposed by him as the feature vector of the image for image retrieval. However, its output feature vector is up to 4096 dimensions, which has the disadvantages of too high dimensions and large computation. Babenko et al. [10] used Principal Component Analysis (PCA) to compress the dimension of the feature vector to lower the dimension of the data, which significantly accelerated the retrieval speed. However, comparing the similarity of the two matrices is still not an efficient operation. The feature representation of CNN learning is inefficient for image retrieval applications because of its high dimensional

features. To adapt the feature representation learned from CNN to the image retrieval task, hash technology has been successfully applied, because it has the characteristics of fast query processing speed and low storage cost [11]. In 2014, Xia et al. [12] proposed the CNNH method, which is divided into two steps: the first step is to learn the hash code, the second step is to train the neural network, and learn the image features and hash function simultaneously. After that, Lai et al. [13] improved the CNNH method and proposed the DNNH model, so that the features learned by the neural network can be fed back to the hash code in time. The literature [14,15] proposes to use the approximate nearest neighbor (ANN) algorithm to speed up the operation, such as local sensitive hash, map the high dimensional feature data into the low-dimensional binary space, form the binary code, and compare their Hamming distance to further improve the retrieval speed.

To sum up, the existing large-scale image retrieval methods have their advantages and disadvantages, especially in terms of retrieval accuracy and retrieval time. These problems will be effectively addressed in the methods proposed in this study.

3. Image Retrieval Framework Based on Locality-Sensitive Hash Using CNN and Attention Mechanism

This study is the image retrieval based on the descending dimension of locality-sensitive hash. The process consists of feature extraction, hash coding, index construction and similarity calculation. Firstly, we obtain the feature vector from the image database through the feature extraction model, then save the binarized feature encoding of the image after the feature vector is processed by locality-sensitive hashing, and build an index to record in the image feature library. Finally, we input the image to be retrieved into the network model to obtain the feature code of the image and the feature database in the database for similarity calculation and comparison, and the closest n images are returned. The retrieval is completed. The flow diagram is shown in Figure 1.

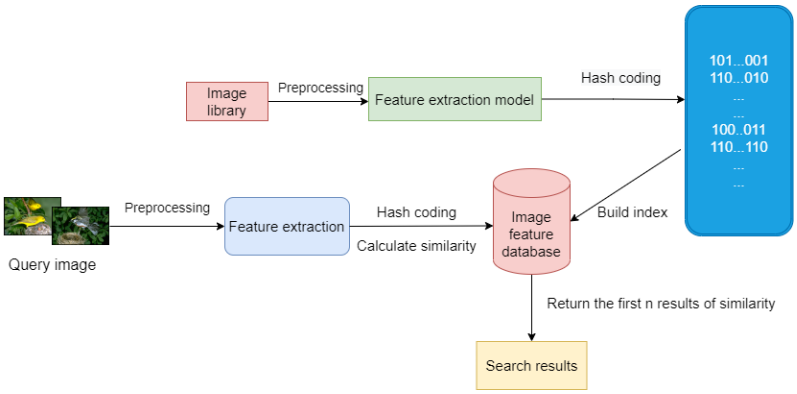


Figure 1. Flow chart of the image retrieval algorithm.

3.1. Feature Extraction

The framework feature extraction model is based on CNN and attention mechanism. The feature extraction model is composed of image preprocessing, convolution layer, attention module and full connection layer, as shown in Figure 2. ResNet50 is selected as the feature extraction network. The user first enhances the training set in the image library with standard data to increase the sample size, and then performs preprocessing operations. The preprocessed image is input into the model for training, and an attention module is added after the convolution layer of the network. By training the attention module, the important features in the retrieval task are given higher weights, so that the model can extract more effective features. Lastly, the output feature map is inputted to the full connection layer.

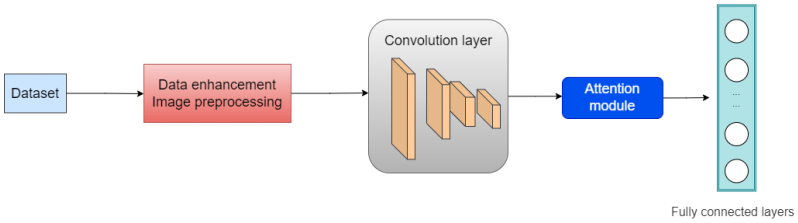


Figure 2. Feature extraction model.

Image Preprocessing

Deep convolution neural networks need a large number of training data to obtain good results, prevent overfitting, and improve robustness. However, it is often difficult to obtain enough training samples, so this paper adopts standard image enhancement methods before convolution training, image enhancement is a process of creating new training data from existing training data. Figure 3 shows the original image conversion examples of several image enhancement methods adopted in this paper. This paper applies these transformations to the original image data set and adds some training data. Image enhancement greatly improves the quality and performance of the deep neural network.



Figure 3. Several data enhancement examples.

After image enhancement of the training set, the training set is preprocessed to make the network model in this paper converge faster during training and feature extraction, and increase the accuracy of the retrieval results.

(1) Image size processing. The image dimension of the image database used in this study is 192 × 128 and 128 × 192, and the size is inconsistent. The size is processed as 192 × 192 to maintain the original characteristics.

(2) Mean and normalization. In the mean removal process, the mean value is subtracted from the RGB 3D, and the image data is centered to 0 to prevent overfitting, see Formula (1). For normalization processing, calculate the RGB maximum value, and compress the image data between 0–1. After normalization processing, the data can better respond to the activation function and improve its expressiveness of the data. The conversion function of the data is shown in Formula (2).

$$Y = Y_{\mathrm{n}} - \frac{1}{m}\sum_{n=1}^{m} Y_n \tag{1}$$

$$Y = \frac{Y_n - Y_{\min}}{Y_{\max} - Y_{\min}} \tag{2}$$

3.2. Using Attention Mechanism (CBAM) to Improve Retrieval Accuracy

The attention mechanism used in this paper is the CBAM module, and CBAM can be added to any position of the convolution layer. Spatial attention and channel attention have different functions. CBAM combines the two to better weigh the prominent parts in the image. In this study, the attention mechanism is applied to the locality-sensitive hashing image retrieval method. This study inputs an image of size 192 × 192 into the network model. A 7 × 7 × 2048 feature map is obtained after the image passes through the

convolution layer of the model. The attention module will perform weighting operations on the channel domain and the spatial domain of the feature map, respectively, to generate a new feature map. The size of the feature map does not change before and after being input into the attention module. Lastly, the new feature map is flattened and input to the full connection layer of the model.

CBAM [16] is composed of a one-dimensional channel attention module $F_c \in R^{C\times1\times1}$ and a two-dimensional spatial attention module $F_s \in R^{1\times H\times W}$, which are arranged in series, as shown in Figure 4. The image features are first input into the channel attention module, and the channel features are generated using the channel relations of the features β'. Then, the generated features β' go through the spatial attention module to obtain the final features β''. The channel attention module retains more image texture information and detailed semantic information, while the spatial attention module extracts effective spatial features by focusing on the contour and spatial structure of the image. The calculation formula of β'' is as follows:

$$\beta' = F_c(\beta) \otimes \beta \tag{3}$$

$$\beta'' = F_s(\beta') \otimes \beta' \tag{4}$$

where: β' is the resulting output of the channel attention module and $\otimes$ is the array element multiplied in sequence.

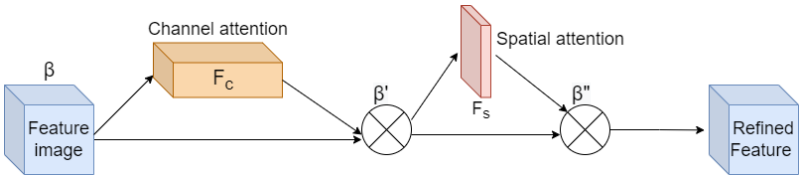


Figure 4. CBAM module (channel and spatial).

3.2.1. Channel Attention Module

The channel attention module can be expressed as:

$$F_c(\beta) = \sigma(MLP(Avg\,\mathrm{Pool}(\beta)) + MLP(\mathrm{MaxPool}(\beta))) \tag{5}$$

Firstly, the input feature map is entered into the channel attention, and then the max pooling and average pooling are carried out. Then, the output of the features by the MLP are added through the shared network, and then the RELU function is used to activate to obtain the final channel attention feature map. The flow is shown in Figure 5. Finally, we multiply this channel attention feature map by the input feature map to generate the input features required by the spatial attention module. Among β expressed as input characteristic map, $F_c(\beta)$ shows an output characteristic diagram. In this paper, the size of the input feature map of the channel attention module is $7 \times 7 \times 2048$, and the weight of each channel can be obtained after the operation of the pooling and perceptron. The weight is $1 \times 1 \times 2048$.

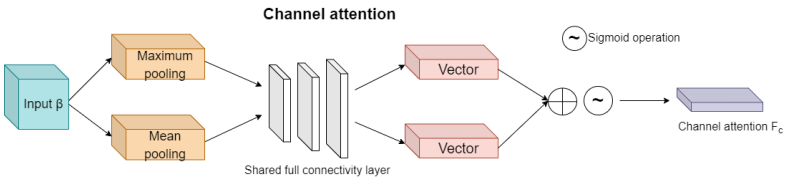


Figure 5. Channel attention structure.

3.2.2. Spatial Attention Module

The spatial attention module can be expressed as:

$$F_s(\beta') = \sigma(\mathrm{Conv}([\mathrm{mean}(\beta'); \max(\beta')])) \tag{6}$$

The input of this module is the feature map β' output by the channel attention module, We subject the feature map to maximum pooling along the channel dimension and mean pooling for dimensionality reduction. Then, the two feature maps with the number of channels of 1 are synthesized into one, and the convoluted attention feature map is obtained after Conv operation. Finally, the final spatial attention feature map $F_s(\beta)$ is obtained after RELU function activation, and the flow is shown in Figure 6. Finally, we multiply the spatial attention feature map with the input feature map β' to obtain the ultimate features.

From the above, it can be seen that the feature map output by the convolution module is weighted by the attention module to gain a new feature map of the same size. It focuses on the target area of the input image, filters out some invalid background information, and effectively represents the characteristics of the input image.

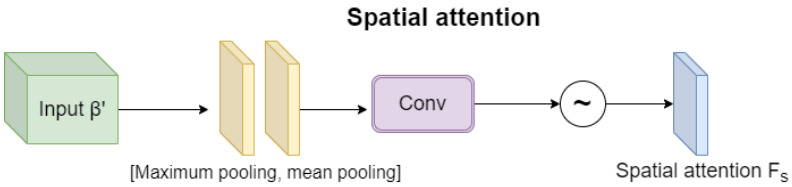


Figure 6. Spatial attention structure.

3.3. Using Local-Sensitive Hash Algorithm to Improve Retrieval Speed

There are many common similarity search algorithms in the image search part. The first thought is linear search. Firstly, the similarity measure is defined, and then the similarity is calculated in pairs, and the top-n is calculated for filtering. However, under the background demand of large-scale image retrieval, this time complexity is too large. At the same time, for high-dimensional sparse data, the calculation of similarity is very time-consuming. At this time, we need some approximation algorithms to improve the retrieval speed. In this study, the LSH algorithm is used to represent the original image with a low latitude binary hash code to avoid directly storing the high latitude image features. Moreover, the generated hash code is used to build an index to search for the nearest neighbor of the image. By calculating the distance between the image to be retrieved and each image in the database, the most similar result to the retrieved image is obtained. In the actual calculation, since the hash code is composed of 0 and 1, their hash distance can be directly performed by the bit operation in the computer, so that the operation speed is greatly improved. Compared with using image features in the original feature space for retrieval, using hash coding for nearest neighbor search reduces the cost of feature storage, improves the efficiency of feature matching, and significantly speeds up the retrieval speed.

LSH constructs a hash function so that the closer the feature points in the original feature space are to be mapped by this hash function, the more likely they will fall into the same hash bucket, and vice versa. Taking Figure 7 as an example, there are three images in total. The last two images have high similarities. From the perspective of a semantic level, the latter two images look more similar. Then, after mapping into the hash code, the distance should be smaller than that of the first image. Then, when querying the image, it is only necessary to obtain the bucket number, and then take out all the data in the bucket from the corresponding bucket, perform linear matching on them, and finally find the similar images that meet the query requirements. In other words, a super large original data set is divided into several subsets after the mapping transformation of the hash function, so as to reduce the query scope to improve retrieval performance.

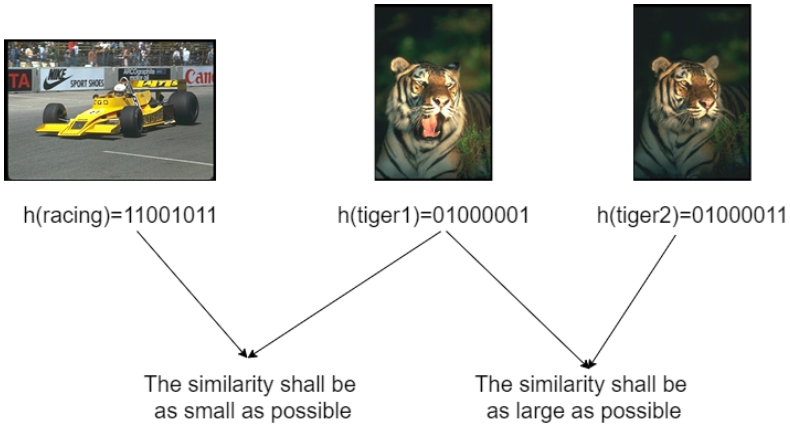


Figure 7. Hash code and image similarity.

(1) Similarity measure and LSH hash function

(a) LSH hash function family:

It is known that $d(x, y)$ is the distance between two points x, y, h is the hash function, $h(x)$ and $h(y)$ are the hash transformation of points x, y, and $P_1 > P_2$. For any two points x, y in the high dimensional space:

If $d(x, y) \leq h_1$, there must be $Pr[h(x) = h(y)] \geq P_1$.

If $d(x.y) \geq h_2$, there must be $Pr[h(x) = h(y)] \leq P_2$.

In similarity calculation, different methods can be adopted to measure the similarity between two points. For dissimilar measurement methods, the hash function used in the local hash is different. There are many methods to measure the similarity of two vectors. This study uses the LSH function based on cosine distance.

(b) LSH hash function of the cosine of the included angle of vector:

Almost similar eigenvectors can be found by representing them in k-dimensional space based on cosine distance. We use the angle between two vectors to measure whether two vectors are similar. The smaller the angle between two vectors, the more similar they are. The calculation formula of vector included angle is as follows:

$$\cos(\theta) = \frac{x \cdot y}{\|x\| \cdot \|y\|} \tag{7}$$

$$\cos(\theta) = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2} \sqrt{\sum_{i=1}^{n} y_i^2}} \tag{8}$$

where x and y represent two sample data points, $\theta \in [0, \pi]$.

(2) Building Index

(a) The step to building index:

- Calculate a hash function $h(x)$ to store similar points in the same bucket.
- For a new query point $x_n \rightarrow h(x_n) \rightarrow \text{Bucket}_n$, calculate that x_n should belong to a certain slot.

Since the image features extracted in this paper are a large number of high latitude features, this paper projects the high latitude features into the multi-dimensional space, and each dimension represents a basic feature. However, it is very complicated to project K feature vectors in the high dimensional space with n features. Therefore, the feature vectors are projected into the m-dimensional space using the random projection method, where $m \ll n$, while maintaining the same cosine distance. Firstly, this paper divides the original data space with a random hyperplane. After each data is projected, it will fall into one

side of the hyperplane. After multiple random hyperplane partitions, the original space is divided into many cells, and the data in each cell is considered likely to be adjacent, then, hashes the data in each cell into the corresponding slot through the hash function $h(x)$ to form a hash table. In a hash table, different numbers of buckets and their corresponding feature vectors may be created. Figure 8 shows LSH using the random projection method to create a set of hash tables to build an index. For example, x_1, x_5, x_8, x_{13} are similar feature vectors that are divided into the same slot.

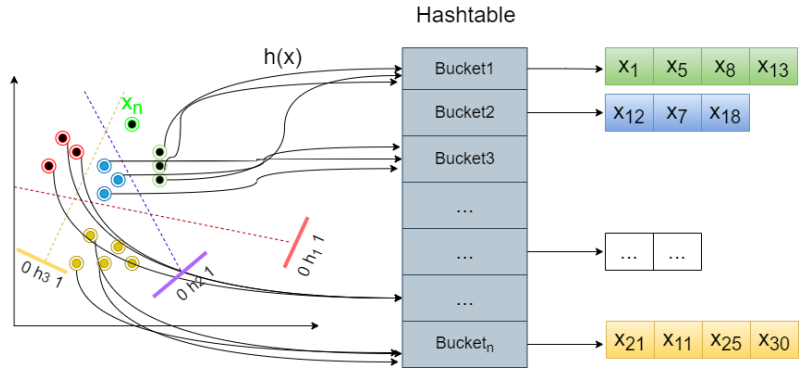


Figure 8. LSH builds index.

(3) Online Searching

The query image is first input into the network, and the feature vector is produced after the feature extraction model, and then the feature vector is encoded and hashed into the homologous hash bucket of the hash table employing the LSH algorithm random projection, and the corresponding bucket number is obtained. Then, the corresponding data in the bucket number is taken out, the distance between the query image to be retrieved and these data is calculated, and the most similar n data are returned. The search process is shown in Figure 9.

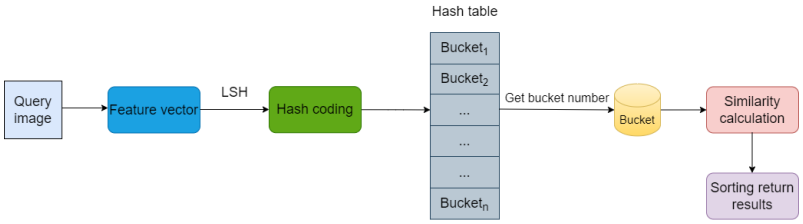


Figure 9. LSH lookup process.

4. Experimental Results and Analysis

4.1. Experimental Environment

The operating system of this experiment is ubuntu 20.04, the CPU is 2.5 GHz 6-core Intel Core i5, the GPU is NVIDIA GeForce RTX 3080, the memory is 12 GB, the programming language of the program is Python 3.6, and the framework based on deep learning is PyTorch.

4.2. Data Sets

This paper selects two data sets that are extensively adopted in the field of image retrieval: corel1k and corel5k. The corel1k dataset contains 10 categories of images such as buses and dinosaurs, with 100 images in each category, totaling 1000 images. Corel5k contains 50 categories of images, including racing cars, beaches, cats, airplanes, etc. Similar to the corel1k dataset, each category contains 100 images and a total of 5000 images. In the experiment, we will randomly extract 10 images from each type of image from corel1k and corel5k as query images, and the rest of corel5k as training sets.

4.3. Evaluating Indicator

The performance of content-based image retrieval is generally evaluated using recall, average retrieval accuracy (MAP) and retrieval time. Therefore, this paper uses these three evaluation indicators. The calculation formula of recall is:

$$B_{Recall} = \frac{1}{m}\sum_{i=1}^{n} f(i) \tag{9}$$

where m is the total number of images in the image library that are similar to the image to be retrieved, $f(i)$ is the similarity relationship between the ith image of the top n images in the search results and the image to be retrieved. If the image to be retrieved is similar to the retrieved image, the value of $f(i)$ is 1, otherwise it is 0.

For k searches, the average accuracy is calculated as follows:

$$B_{MAP} = \frac{1}{x}\sum_{b=1}^{x}\left[\frac{1}{n}\sum_{i=1}^{n} f(i)\right] \tag{10}$$

4.4. Performance Evaluation

4.4.1. Performance Comparison

We compare the proposed method with four different existing image retrieval methods to demonstrate the effectiveness of our method. The five comparison algorithms are SIFT, SVM active learning [17], LeNet-L [18], ref. [19] and ResNet50+CBAM+LSH (ours). Among the five models compared in this paper, the SIFT algorithm combined with the k-means clustering algorithm can achieve fast retrieval of small-scale simple images, which belong to the traditional image retrieval algorithm. The SVM active learning algorithm uses a classifier to find the optimal retrieval target by maximizing the interval. LeNet-L is an improvement of the LeNet-5 model in ref. [18]. Its structure is relatively simple and clear, and its parameters are medium. It is the best choice to compare other networks. Ref. [19] uses CNN to extract features, and further uses the LSH algorithm to build an index for searching. For fairness, we reproduce the method of ref. [19] on the same dataset. ResNet50+CBAM+LSH is the method suggedted in this study. ResNet50, as the feature extractor, adds the CBAM module after the convolution layer to increase the representation ability of CNN. Then, the LSH algorithm is used for hash coding to drop the dimension of high latitude features in the image library before searching. Table 1 shows the changing trend of the MAP of the five models on the dataset corel1k with the number of returned images and the comparison results of the recall rate.

Table 1. MAP, recall and Top-n of different methods on the dataset corel1k.

Method	MAP	Recall	Top-5	Top-10	Top-20	Top-30
Traditional SIFT	0.7748	0.7024	0.8543	0.7933	0.7893	0.7693
SVM Active Learning [17]	0.7241	0.6512	0.8403	0.7806	0.7425	0.6916
VGG-N [18]	0.8951	0.7858	0.9304	0.9125	0.8927	0.8841
Literature [19]	0.9186	0.8192	1	0.9932	0.9842	0.9646
Ours	0.9578	0.842	1	1	1	0.9994

According to Table 1, for the same Top-n, the method in this paper is superior to SIFT and other CNN models. At the same time, when the number of returned images grows, the method in this paper can still maintain a high MAP. In the dataset corel1k, the MAP of this paper is about 18.3% higher than that of the traditional SIFT method, about 23.4% higher than that of the SVM active learning method, and about 6.3% higher than that of the VGG-N method, which shows that the model in this study is better than other models in MAP. From Table 1 we know that compared with the literature [19], which does not use an attention

mechanism, the MAP of the model in this paper using an attention mechanism is increased by about 3.9% on average. This result proves that under the same other conditions, the neural network with the CBAM module can enhance the representation ability of CNN, thus improving the accuracy of image retrieval. In addition, the recall rate of this method is also significantly better than the other four existing image retrieval methods, which proves the utility of the algorithm model proposed in this study.

In order to further examine the effectiveness of the algorithm model suggested in this study, we tested the MAP under this method and other methods on the corel5K dataset. The experimental results can be seen from Table 2. For the MAP on the dataset corel5K, this paper is about 17.6% higher than SIFT, 22% higher than SVM active learning, 5.5% higher than VGG-N, and 4.5% higher than literature [19]. This shows that the model suggested in this study has an effective role in improving MAP on different data sets, and further verifies that the CBAM module added to the neural network in this study can effectively heighten the accuracy of image retrieval.

Table 2. The MAP of different methods on data set corel5k.

Method	Traditional SIFT	SVM Active Learning [17]	VGG-N [18]	Literature [19]	Ours
MAP	0.7528	0.7091	0.8733	0.8841	0.9286

4.4.2. Comparison of Retrieval Time of Different Methods

This paper compares the retrieval time of this method with that of other algorithms on corel5k and corel1k through experiments, as shown in Table 3.

Table 3. Retrieval time of different methods on data sets corel1k and corel5k (time/s).

Method	corel1k	corel5k
Traditional SIFT	18.81	76.91
SVM Active Learning [17]	19.56	78.57
VGG-N [18]	12.42	43.81
Literature [20]	2.33	11.56
Ours	0.13	0.64

Table 3 illustrates the retrieval time of five different methods on the corel1k and corel5k datasets. Based on the traditional SIFT and SVM, the retrieval time is the longest, and the retrieval time of the method in this paper is significantly lower than that of VGG-N and literature [20]. Compared with VGG-N and literature [20] that do not use the LSH algorithm, the retrieval speed of this method is significantly improved. Although the ResNet50 network model is used for feature extraction in this paper, with high feature dimensions, the local-sensitive hash algorithm is adopted to reduce the feature dimensions, and there is no need to compare and calculate the image features in the image library one by one, which greatly reduces the retrieval time. Thus, the feasibility of the algorithm model in this study to improve the efficiency of retrieval time is verified.

4.4.3. Model Robustness

In this experiment, the target retrieval image is preprocessed by various methods such as rotation, clipping and brightness change, and compare it with the primitive image input to test the robustness of this model. The experimental results are displayed in Figure 10. It is manifest from Figure 10, the input image can still maintain good search results after being rotated, cropped, and brightness changed. It shows that the model in this paper uses the standard data enhancement method to augment the training amount of samples in the training. In the process of image retrieval, under the influence of factors such as rotation, clipping, scaling and brightness change in the image, the model has certain robustness.



Figure 10. For example, the top 8 return results of image retrieval under different processing.

5. Conclusions

This study puts forward a local sensitive hash image retrieval algorithm based on CNN and attention mechanism. The feature is extracted by Convolutional Neural Network, and the channel and spatial attention module are embedded in the network so that the attention module can effectively identify the importance of different regions in the feature map, and give more weight to the key regions in a weighted way, to raise the feature expression ability of hash coding. The experiments were carried out on corel1k and corel5k datasets, respectively. It can be known from the experimental results that the proposed method has an advantage to improve the retrieval accuracy and recall rate after adding the attention mechanism module, and can still maintain a good accuracy rate when top-n increases. The search part uses a locally sensitive hash algorithm to hash code the high-dimensional features of the image and map them to a low latitude hash space to establish an image index and retrieve the image. The algorithm not only ensures a certain accuracy, but also greatly reduces the retrieval time, which fully demonstrates the effectiveness of the algorithm and good retrieval performance. At last, the experiment fully shows that the algorithm has certain robustness under the influence of rotation, cropping, brightness change, and other factors by using data enhancement methods. Because this paper uses the features of the full connection layer of the network to retrieve, high-level features are easy to lose a lot of detailed information; therefore, it will influence the capability of image retrieval in part. So, in future research, we will not only use the semantic information of high-level features, but also consider the details of texture information of low-level features, and improve the attention module to integrate the high-level semantic features and low-level features of images, to make image retrieval more accurate.

Author Contributions: Conceptualization, W.L.; Data curation, Y.L., X.M. and K.Z.; Formal analysis, Y.L.; Funding acquisition, W.L.; Investigation, K.Z.; Methodology, W.L.; Project administration, W.L.; Resources, Y.L., X.M. and K.Z.; Software, Y.L.; Supervision, W.L.; Validation, X.M.; Visualization, X.M.; Writing—original draft, Y.L.; Writing—review & editing, W.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Fujian Natural Science Foundation of China (2022J011233) and Xiamen University of Technology (XPDKT20027).

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Wang, H.; Qu, H.; Xu, J.; Wang, J. Texture image retrieval based on fusion of local and global features. *Multimed. Tools Appl.* **2022**, *81*, 14081–14104. [CrossRef]
2. Mohite, N.B.; Gonde, A.B. Deep features based medical image retrieval. *Multimed. Tools Appl.* **2022**, *81*, 11379–11392. [CrossRef]
3. Xiong, B.; Lou, L.; Meng, X.; Ma, H.; Wang, Z. Short-term wind power forecasting based on Attention Mechanism and Deep Learning. *Electr. Power Syst. Res.* **2022**, *206*, 107776. [CrossRef]
4. Zhu, Y.; Liu, R.; Huang, Q. Weakly supervised information fine-grained image recognition based on deep neural network. *J. Electron. Meas. Instrum.* **2020**, *32*, 8.
5. Qin, J.; Pan, W.; Xiang, X.; Tan, Y.; Hou, G. A biological image classification method based on improved CNN. *Ecol. Inform.* **2020**, *58*, 101093. [CrossRef]

6. Noh, H.; Araujo, A.; Sim, J.; Weyand, T.; Han, B. Large-Scale Image Retrieval with Attentive Deep Local Features. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017.
7. Li, Z.; Li, S.; Liu, Y.; Li, H. Hand-drawn image retrieval method based on attention model. *Comput. Sci.* **2020**, *47*, 6.
8. Ng, T.; Balntas, V.; Tian, Y.; Mikolajczyk, K. SOLAR: Second-Order Loss and Attention for Image Retrieval. *arXiv* **2020**, arXiv:2001.08972.
9. Sutskever, I.; Hinton, G.E.; Krizhevsky, A. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25*; Neural Information Processing Systems Foundation, Inc.: San Diego, CA, USA, 2012; pp. 1097–1105.
10. Babenko, A.; Slesarev, A.; Chigorin, A.; Lempitsky, V. *Neural Codes for Image Retrieval*; Springer International Publishing: Berlin, Germany, 2014.
11. Wang, W.; Jiao, P.; Liu, H.; Ma, X.; Shang, Z. Two-stage content based image retrieval using sparse representation and feature fusion. *Multimed. Tools Appl.* **2022**, *81*, 16621–16644. [CrossRef]
12. Xia, R.; Pan, Y.; Lai, H.; Liu, C.; Yan, S. Supervised Hashing for Image Retrieval via Image Representation Learning. In Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence, Québec City, QC, Canada, 27–31 July 2014.
13. Lai, H.; Pan, Y.; Ye, L.; Yan, S. Simultaneous Feature Learning and Hash Coding with Deep Neural Networks. In Proceedings of the IEEE International Conference on Pattern Recognition and Computer Vision, Boston, MA, USA, 7–12 June 2015; pp. 3270–3278.
14. Shan, T. Research on the Nearest Neighbor Search Algorithm Based on Image Features. Master's Thesis, University of Science and Technology of China, Hefei, China, 2017.
15. Gao, X. Research on Large-scale Image Nearest Neighbor Retrieval Algorithm Based on Hash Algorithm. Master's Thesis, University of Electronic Science and Technology of China, Chengdu, China, 2018.
16. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. *CBAM: Convolutional Block Attention Module*; Springer: Cham, Switzerland, 2018.
17. Wang, X.; Luo, G.; Qin, K.; Zhang, Y. An Image Retrieval Method Based on SVM and Active Learning. *Comput. Appl. Res.* **2016**, *33*, 3836–3846.
18. Wei, Y.; Yan, Z. Research on Image Retrieval Technology Combined with Attention Convolutional Neural Network. *Small Microcomput. Syst.* **2021**, *42*, 2368–2374.
19. Balasundaram, P.; Muralidharan, S.; Bijoy, S. An mproved Content Based Image Retrieval System using Unsupervised Deep Neural Network and Locality Sensitive Hashing. In Proceedings of the 2021 5th International Conference on Computer, Communication, and Signal Processing, ICCCSP 2021, Chennai, India, 24–25 May 2021; pp. 65–71.
20. Qin, J.; Huang, J.; Xiang, X.; Tan, Y. Image retrieval based on convolutional neural network and attention mechanism. *Telecommun. Technol.* **2021**, *61*, 304–310.