Multi label image retrieval based on depth hash

Zihui Qin

1 Faculty of Information Technology, Beijing University of Technology, Beijing, 100020, China

*Corresponding author's e-mail: qinzhi@emails.bjut.edu.cn

Abstract. In recent years, hash is a popular method of image retrieval. Convolutional neural network is used to generate image features and hash codes, so as to achieve fast and effective image retrieval. In order to solve the problem of poor image retrieval effect caused by complex background of Multi-target Image, this paper proposes a hash generation method based on the combination of target region recommendation network and convolutional neural network. The RPN network is selected as the target region recommendation algorithm. After the image passes through the RPN network, multiple target regions and the four-dimensional coordinates of each region will be generated. According to the four-dimensional coordinates, the main target region will be filtered and extracted. Then, the feature of the largest target area is extracted by the convolutional neural network (GoogleNet), and the corresponding hash code is generated, which is finally retrieved in the database. Experiments are carried out on voc2012 data set and self collected data set to verify the algorithm. When the number of test images is 1000, the experimental results show that the total correct rate of the retrieval results of this method is 95.5%, which is about 5 percentage points higher than the existing methods.

1. Introduction

With the rapid development of mobile Internet technology, today's image, text, video and other data show exponential growth. In the large-scale image data, how to retrieve the user's demand information quickly and effectively becomes the main problem in the field of image retrieval. Image retrieval is one of the important problems in image processing. In the early stage of image retrieval, the low-level feature extraction (color, texture, shape, etc.) or directional gradient histogram was carried out by hand. This method requires a lot of time to adjust parameters, most of which rely on experience and have uncertainty. How to effectively extract the depth semantic information contained in the image is a big problem in the academia.

With the rapid development of deep learning in recent years, there has been a major breakthrough in the field of image. Deep learning has obvious advantages in image processing. Convolutional neural network is widely used in image classification, object detection, semantic segmentation and other fields. In the fast r-cnn algorithm which performs well in the field of target detection, researchers join the region proposal network (RPN) to directly get the position of the object, which greatly improves the speed and accuracy. Since the convolution layer or full connection layer features extracted by CNN can well represent the semantic features of images, the image retrieval technology with images as input has made great progress. Compared with the early manual feature extraction, deep learning can accurately and quickly learn the super high-level semantic features of images through network training, which is widely used in the field of image retrieval. The current image retrieval methods include optimizing the loss function and proposing a new hash function to optimize the network, and
preprocessing the input image. Each method has its own advantages and disadvantages in terms of retrieval efficiency and accuracy.

The fast retrieval framework proposed by Kevin Lin et al. [1] innovatively adds a late layer between the last fully connected layer (fc8) and the penultimate fully connected layer (fc7), which can not only learn deep CNN features, but also get binary hash codes. At the same time, this paper proposes a coarse-grained to fine-grained retrieval strategy. This hash method maps image data features from high dimension to low dimension through deep learning, and produces compressed binary code. Finally, similar images are retrieved in the database according to the extracted image features and hash code. By calculating the Hamming distance of two hash codes to measure the similarity between two images, the closer the Hamming distance is, the higher the similarity of two images is. This hash method has significantly improved the retrieval speed and accuracy.

The above algorithms are mainly applied to the whole image retrieval, for the image with complex background, the background information features are more significant, and the retrieval results are poor. Aiming at this defect, this paper proposes a multi-target image retrieval network. Firstly, the target region recommendation algorithm is used to generate a series of regions that may contain targets, and the main target region is determined after filtering. Then, the convolutional neural network is used to extract the features and hash codes of the main target region, and finally similar images are retrieved in the database according to the hash code features. The main contributions of this paper are as follows

- This paper proposes a method of using the main target region feature to represent the whole image to retrieve in the database, which can eliminate the completely unrelated images and reduce the interference of image background.
- Hash layer is added to the network. By adding a hash layer in the middle of the last two fully connected layers, the image features are mapped from high dimension to low dimension, and binary hash code is generated, which reduces the retrieval cost.

Figure 1 shows a basic implementation framework. Firstly, a series of possible target regions are extracted by Region recommendation network (RPN) to obtain the location information of each target in the image (in this paper, a four-dimensional coordinate is used to display the location of each target region). According to the four-dimensional coordinates, the main target area is determined, and then the features of the main target area are extracted by using the convolutional neural network (GoogleNet). In order to generate the hash coding, the hash layer is added after the full connection layer, and the main target area features represent the whole image to be retrieved in the database.

![figure 1. Basic implementation framework](image)

2. Model

2.1. Regional recommendation module

The purpose of region recommendation is to find some potential targets on the image and form candidate regions. There are many perfect methods in the field of regional recommendation, such as edgeboxes, geodesic and selective search, which can be roughly divided into two categories: grouping method and window scoring method. The former first breaks up the images and then aggregates them, such as selective search, which does not need to learn, and then aggregates them according to the artificially defined distance. The latter is to generate a large number of candidate boxes and score them, then filter out the low score candidate boxes, and finally sort them, such as Bing, edgeboxes. For the
purpose of this project, through the analysis and comparison of the existing regional recommendation algorithms, RPN network is selected to apply to the multi label image retrieval of this project.

RPN network is the representative of classical framework in the field of target detection, and the core part of fast r-cnn \[^2\]. RPN network inputs a picture, and its output is divided into two channels, one of which outputs the probability of target and non target, \( P \in [0,1] \). If the probability of a region is greater than 0.5, it is considered that the region belongs to one of the specific categories. These target regions selected by the network are called ROI (region of interests), that is, regions of interest. The other output is the approximate position of these regions of interest, and each target region is represented by a four-dimensional coordinate \( S_i = (x, y, w, h) \).

As shown in Figure 2, after the image passes through the RPN network, a series of possible target regions are generated. The distance \( D_i \) between the center point \( (x_i, y_i) \) of the ith target region and the center point \( (x_0, y_0) \) of the image is calculated, and the target region with the smallest distance value is taken as the main target region of the query image.

\[
di = 1 - \frac{\sqrt{(x_i - x_0)^2 + (y_i - y_0)^2}}{\sqrt{x_0^2 + y_0^2}}
\]

(1)

For the multi label image of this subject, the application of RPN network can effectively extract the main object of the image, reduce the interference, and get the results more in line with the user's query needs.

Figure 2. target area selection module

2.2. Convolutional neural network model

Googlenet has achieved very good results in the task of classification, with the error rate of 6.67%, and achieved good results in the field of target detection, with the average retrieval accuracy of 43.9%. Compared with the previous Alexnet, VGG and other structures, it can achieve better training effect by increasing the number of network layers. It can improve the training effect from another perspective. A large number of \( 1 \times 1 \) convolution kernels are added in the perception module to reduce the dimension and prevent over fitting, and more features can be extracted under the same amount of calculation. The parameters of the whole googlenet model are significantly reduced.

The retrieval algorithm in this paper uses the main target region features of the image for image retrieval, rather than the whole image as the query object. Therefore, this paper uses googlenet as the middle layer to extract the main target region image features, and according to the actual needs, refine tune the network.

The information of the input image through the full connection layer can be used for retrieval and classification, but the full connection layer has thousands of dimensional features, and the dimension of feature vector is very high, so the retrieval will be very time-consuming. By adding a hash layer in the middle of the last two fully connected layers, the image features are mapped from high dimension to low dimension, and binary hash code is generated, which reduces the retrieval cost.

Because the hash layer (fc8) is added to the new network. There are \( h \) nodes in this layer. Because the image propagates forward through the hash layer, there will be \( h \) outputs, and the output value is between (0,1). Set a threshold(0.5) to change the output value into binary.
2.3. Feature comparison module

According to the characteristics of the main target area of the image, the image similar to the query is stored in the database. The process of image retrieval is hierarchical retrieval, which is divided into rough retrieval and detailed retrieval. Rough retrieval is to compare the similarity with the hash code features of the image. After the similarity is measured by hamming distance, the similar candidate set is obtained, and then the detailed retrieval is carried out in the candidate set. The detailed retrieval uses the features of the full connection layer, and the similarity is measured by Euclidean distance. The smaller the distance, the higher the similarity between the two images. On the contrary, the lower the similarity between the two images, the higher the ranking of the images, and select the top 10 images with high score as the retrieval results.

The feature similarity between images is measured, and the Hamming distance of hash code between two image categories is calculated

$$C(a_z,a_q) = \sum_{i=1}^{h} a_z \oplus a_q$$

Where "\(\oplus\)" is the XOR operator, \(a_z\) is the hash code corresponding to the target region in the query image, and \(a_q\) is the hash code corresponding to the database image.

3. experiment

Pascal VOC2012\textsuperscript{[3]}, which contains 11530 images of 20 kinds of objects, such as aircraft, automobile and horse, and more than 2000 images of 10 kinds of objects, such as food and elephant, are used to train and verify the proposed algorithm. Pascal VOC2012 contains an average of 2.8 tags per image. 9000 of them are randomly selected as the training set, and the remaining more than 2000 are selected as the test set.

Using the new network of RPN and googlenet, which is based on Ubuntu 16.04 operating system, training and testing under the Caffe \textsuperscript{[4]} framework, the node number of the newly added hash layer is 48, that is, the 48 bit hash code describing the image features can be learned. Using the end-to-end training method, after 50000 iterations, the weights of each layer in the network can be updated through back propagation. The experiment uses ndcg, ACG and ACG MAPW is used as an evaluation index to test the performance of the algorithm\textsuperscript{[5]}

2000 images in voc2012 database are randomly selected for experiments, and the NDCG and ACG values are calculated when different numbers of retrieval images are returned. The results are shown in Table 1.

It can be seen from table 1 that with the increase of the number of returned images, the value of NDCG shows an increasing trend. When the number of returned images is 1000, the value of NDCG is 0.9085. On the contrary, with the increase of the number of returned images, the value of ACG decreases gradually. When the number of returned images is 1000, the value of ACG is 0.8481.

| k     | NDCG@k | ACG@k |
|-------|--------|-------|
| 200   | 0.8501 | 1.1704|
| 400   | 0.8620 | 1.1085|
| 600   | 0.8776 | 1.0081|
| 800   | 0.8876 | 0.9167|
| 1000  | 0.9085 | 0.8481|

The first 10 images are visualized, and figure 3 is the image returned from the multi tag image retrieval of human riding. It can be seen that the first 10 images are more similar to the query image and more in line with human visual characteristics. Figure 4 shows the retrieval results of sheep. The
returned image contains the front, side and long-distance images of sheep, so the retrieval results are more comprehensive.

Fig. 3 partial retrieval

Fig. 4 partial retrieval results of sheep

In the algorithm, the Hamming distance is calculated by using hash code for retrieval. Compared with the real valued feature vector retrieval by removing the hash layer and using the full connection layer, the time of using hash code is less, the online retrieval speed is faster, and it is suitable for large-scale image retrieval. In the experiment, the specific time of two different methods is compared: the retrieval time with hash code is 0.8s, and the retrieval time without hash code is 1.78s.

1000 images in voc2012 dataset are randomly selected as test images. 101 test images in reference [6] algorithm have errors when returning the first 10 retrieval images, and the total correct rate is 89.9%. In this paper, 45 test images have errors, and the total correct rate is 95.5%, which is about 5 times higher than that of reference [6] algorithm Percentage points of the total.

Through the above experiments, it can be seen that the new network combining region recommendation network (RPN) and googlenet can effectively retrieve similar images in the database, eliminate the interference of complex background, and accurately capture the target in the image.

4. Conclusion

In the information age, there are many kinds of image data and complex information, so multi label image retrieval has become a hot research direction. This paper proposes a new network which combines regional recommendation network (RPN) with googlenet. The network model is obtained by training the new network with voc2012 image data set. After a network forward propagation, the image position information, feature vector and hash code are extracted, and similar images are retrieved in the database. The two images are sorted by calculating the Hamming distance of the hash code, and the retrieval results are returned. The experimental results show that the first 10 retrieval results returned by our algorithm on two datasets, and compare the results of our algorithm with other retrieval algorithms based on ACG, ndcg and MAPW. The experimental results show that the full accuracy of our algorithm on 1000 randomly selected test images reaches 95.5%, which can effectively retrieve similar images in the database and eliminate the interference of complex background to accurately capture the target in the image.

References

[1] Lin, Kevin, et al. (2015). Deep learning of binary hash codes for fast image retrieval. CVPRW IEEE.
[2] Ren S, He K, Girshick R, et al.(2015). Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 39(6):1137-1149.
[3] Everingham M, Gool L V, Williams C K I, et al.(2010) ThePascalVisual Object Classes (VOC) Challenge[J]. International Journal of Computer Vision, 88(2):303-338.
[4] Jia Y, Shelhamer E, Donahue J, et al.(2014) Caffe: Convolutional Architecture for Fast Feature Embedding[J].
[5] Gong Y, Lazebnik S, Gordo A, et al. (2013) Iterative Quantization: A Procrustean Approach to Learning Binary Codes for Large-Scale Image Retrieval[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,35(12):2916-2929.
[6] SALVAIDOR, AMAIA, GIRO-I-NIETO, et al. (2016). Faster R-CNN features for instance search[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, NJ: IEEE.