Hindawi
Wireless Communications and Mobile Computing
Volume 2023, Article ID 9891758, 1 page
https://doi.org/10.1155/2023/9891758

*Retraction*

# Retracted: Graph Convolutional Networks for Cross-Modal Information Retrieval

**Wireless Communications and Mobile Computing**

Received 25 July 2023; Accepted 25 July 2023; Published 26 July 2023

This article has been retracted by Hindawi following an investigation undertaken by the publisher [1]. This investigation has uncovered evidence of one or more of the following indicators of systematic manipulation of the publication process:

(1) Discrepancies in scope
(2) Discrepancies in the description of the research reported
(3) Discrepancies between the availability of data and the research described
(4) Inappropriate citations
(5) Incoherent, meaningless and/or irrelevant content included in the article
(6) Peer-review manipulation

The presence of these indicators undermines our confidence in the integrity of the article's content and we cannot, therefore, vouch for its reliability. Please note that this notice is intended solely to alert readers that the content of this article is unreliable. We have not investigated whether authors were aware of or involved in the systematic manipulation of the publication process.

Wiley and Hindawi regrets that the usual quality checks did not identify these issues before publication and have since put additional measures in place to safeguard research integrity.

We wish to credit our own Research Integrity and Research Publishing teams and anonymous and named external researchers and research integrity experts for contributing to this investigation.

The corresponding author, as the representative of all authors, has been given the opportunity to register their agreement or disagreement to this retraction. We have kept a record of any response received.

## References

[1] X. Yang and W. Zhang, "Graph Convolutional Networks for Cross-Modal Information Retrieval," *Wireless Communications and Mobile Computing*, vol. 2022, Article ID 6133142, 8 pages, 2022.

Hindawi
Wireless Communications and Mobile Computing
Volume 2022, Article ID 6133142, 8 pages
https://doi.org/10.1155/2022/6133142

*Research Article*

# Graph Convolutional Networks for Cross-Modal Information Retrieval

**Xianben Yang and Wei Zhang**

*College of Computer Science and Technology, Beihua University, Jilin, 132000 Jilin, China*

Correspondence should be addressed to Wei Zhang; allenlov@21cn.com

Received 3 November 2021; Revised 2 December 2021; Accepted 8 December 2021; Published 6 January 2022

Academic Editor: Shalli Rani

In recent years, due to the wide application of deep learning and more modal research, the corresponding image retrieval system has gradually extended from traditional text retrieval to visual retrieval combined with images and has become the field of computer vision and natural language understanding and one of the important cross-research hotspots. This paper focuses on the research of graph convolutional networks for cross-modal information retrieval and has a general understanding of cross-modal information retrieval and the related theories of convolutional networks on the basis of literature data. Modal information retrieval is designed to combine high-level semantics with low-level visual capabilities in cross-modal information retrieval to improve the accuracy of information retrieval and then use experiments to verify the designed network model, and the result is that the model designed in this paper is more accurate than the traditional retrieval model, which is up to 90%.

## 1. Introductions

With the development of image acquisition and sharing technology, the amount of image data that people experience has increased significantly, and the biggest problem facing today is how to effectively and accurately obtain the images that users are interested in [1, 2]. Cross-modal information retrieval technology has made great progress in recent years, but it still cannot meet human needs. The main reason is that due to the semantic gap between image visual features and high-level semantic concepts, the accuracy of cross-modal information retrieval cannot meet the demand, and the image feature vectors used in CBIR usually result in slow retrieval speed [3, 4]. TBIR uses only textual information to index and search images. Compared with visual information, text information basically uses low-dimensional and simple concepts to describe the content of images, which is easy for humans to understand, but TBIR often requires manual semantic annotation, which is only suitable for small special libraries [5, 6]. In recent years, the development of social networks has made image data grow exponentially, but this kind of semantic information-based retrieval is very random, contains a lot of noise, and is incomplete [7, 8].

Regarding the research of image retrieval technology, some researchers believe that as people's demand for restoration continues to grow, content-based image retrieval systems may not be able to meet people's needs. At present, the derivation based on semantic image retrieval technology can better represent the local attributes of image semantic information. However, there are still some problems; that is, the complex image descriptors obtained by deep learning algorithms (such as mapping and grouping) may not be able to fully express the semantic information expressed by the complex graphics itself. Therefore, in real applications, the more complex the image structure, the worse the robustness of the semantic search method [9]. Some researchers derived the image mean and standard deviation as general color features in the YUV color space and obtained a binary image bitmap, thereby deriving its local color features. Then, the shape of the image is derived through more compact and instant Clauchuk feature variables, and then, the texture characteristics of the image are obtained through the improved four-pixel concurrent matrix algorithm. Finally, multiple feature points are integrated to calculate the similarity between the inspected image and the image in the image library, thereby returning a highly similar image [10]. When studying deep learning technology, some researchers found that in traditional content-based image retrieval algorithms, the choice of predesigned image feature extraction algorithms is often a priori image content. Therefore, SIFT algorithm is usually more conducive to deriving characteristic scene images, while HOG algorithm is more able to measure pedestrian characteristics and behavior. The main disadvantages of this selection method are low efficiency and long debugging time. Deep learning technology overcomes these shortcomings. Deep learning is a feature extraction process that does not require human intervention at all. This is a deep framework using unsupervised learning algorithms. In-depth training automatically learns the low-level characteristics of the image in the original input image, gradually subtracts, maps, and combines this feature, and finally obtains the high-level semantic function and then analyzes or detects the real-time high-level semantic feature. Higher-order semantic features are more expressive than those obtained in traditional feature extraction algorithms and also help to improve search accuracy [11]. In summary, there are many research results on image retrieval technology, but there are few researches on the fusion of convolutional neural network technology and cross-modal image retrieval technology, and there are still some problems in the retrieval process.

This paper studies the graph convolutional network for cross-modal information retrieval. Based on the literature, it puts forward the basic ideas of cross-modal information retrieval in this paper and summarizes the application of convolutional neural networks. The cross-modal information retrieval of the neural network is designed, and then, the designed network model is verified, and conclusions are drawn through experiments.

## 2. Cross-Modal Information Retrieval and Convolutional Network Research

*2.1. The Basic Idea of Cross-Modal Information Retrieval.* With the development of social networks in recent years, it is easier to obtain text information related to the image from the user, such as comments and tags, from users during the process of sharing pictures on the Internet [12]. Artificial annotation semantics technology allows the extraction of semantic concepts related to images from textual information. In fact, the traditional way of manually labeling semantics has been socialized through Internet technology. For example, photo sharing sites provide users with annotation and tagging functions. Users viewing photos can add descriptive text paragraphs to photos at any time and select one or more tags (birds, trees, sea, etc.) for the photos. However, due to the high randomness of comments and points, the generated text contains a lot of noise. Some tags may be assigned to images that are unrelated to them. In addition, some images may receive incomplete labels. This means that some important semantic information described in the image has not been recorded.

Both content-based image search and text-based search have disadvantages, but they compensate for each other. For example, it is difficult to infer high-level semantics directly from image search features, but the semantics of text-based searches are more accurate, so they can use their complementarity to partially filter text semantic noise. In addition, some semantic thinking can be inferred more accurately from the content. In this case, the use of content-based editing technology to extract semantics can be used as a supplement to image semantics. Therefore, this article focuses on cross-graphic information retrieval. The main reason why images and texts can learn from each other is that they are related to each other at the semantic level, while the mutual search of images and text requires a high degree of generalization and content representation. The association of images and text requires the creation of a cross link between the two features of graphics and text. In-depth understanding is the key to crossing the two characteristics of graphics and text. For example, an image search system can search for images in a clear sky by assigning "l blue dots," while a text search system can search for the sky in the text using the "sky" keyword. Therefore, the cross information retrieval system also needs to understand the correlation between the text "sky" and the visual feature "blue." Therefore, when editing images or text, you need to go through many hidden layers to remove semantic noise and extract features between different modes.

*2.2. Application of Convolutional Network.* CNN has long been used in the field of image classification. Compared with other theories, convolutional neural networks achieve better classification accuracy on large data sets. Similarly, important discoveries have been made in reducing the size of the filter and increasing the depth of the filter. For many image sorting tasks in image classification, creating a hierarchical classifier structure is a common strategy of convolutional neural networks. By using a hierarchical structure to exchange information between related classes, classification performance can also be improved without many training examples. The strategy to improve the classification accuracy is to build a tree structure to more accurately record the detailed features in subcategory recognition. There is a theory that the development of convolutional networks is not only reflected in accuracy but also in layering. In this theory, different categories are grouped according to similarities and organized voluntarily to form different levels.

Image classification and image positioning methods are common techniques in cross-modal image analysis. But in fact, image target discovery also solves the semantic process of discovering the target first and then discovering the target with the idea of sorting. At the same time, in the entire convolution model, semantic extraction only needs to obtain the final semantic label, not the semantic label coordinates of the image or the covered area. For example, an image includes people, cars, and trees. The first thing to do is to extract features representing people, cars, and trees from the image, but you do not need to know the relationship between the

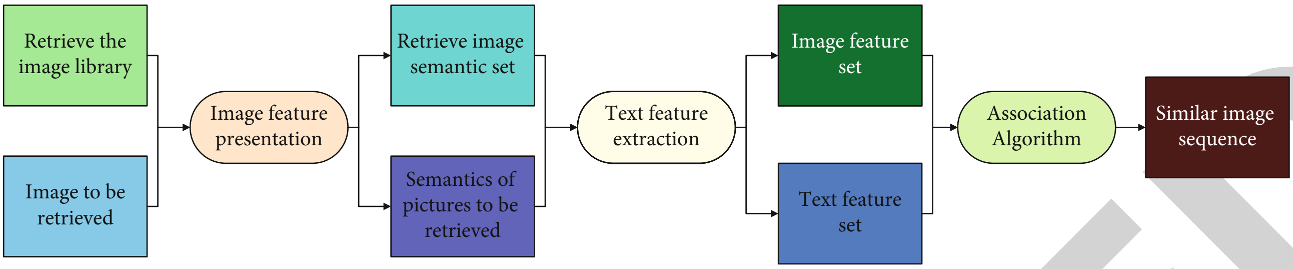

Figure 1: Cross-modal semantic information acquisition model.

coordinates and positions of people, cars, and trees. In addition, for some special tags, it is difficult to identify the areas covered by the tags, such as family and carnival tags.

*2.3. Convolutional Neural Network Algorithm.* The neural network model is mainly composed of the following three parts: the connection between the networks, the weighted summation of the input information, and the activation mapping function. Assuming that the input sample is composed of four elements $(x_1, x_2, x_3, +1)$, the output result of the neuron is expressed as:

$$h_{w,b}(x) = f\left(w^T x\right) = f\left(\sum_{i=1}^{3} w_i x_i + b\right). \quad (1)$$

Among them, $f$ is the activation mapping function of this neuron, and $b$ is a bias.

CNN is a supervised learning network model with multiple hidden layers. The hidden layers include convolutional layers and downsampling layers. The two operations are performed alternately to complete the main part of extracting features. The network structure can be represented by the following function:

$$h_w(x) = f_1(\cdots f_2(f_1(x, w_1), w_2) \cdots, w_l). \quad (2)$$

Generally, a convolutional neural network includes three stages: convolution, nonlinear transformation, and downsampling. Each hidden layer of the convolutional neural network has an input $x_i$ and a weight parameter $w_i$. The output $x_{i+1}$ of each layer is the input of the next layer, and the parameter $w = (w_1, w_2, \cdots, w_i)$ is the weight of the convolution kernel of each layer.

## 3. Cross-Modal Information Retrieval Based on Convolutional Neural Network

*3.1. Cross-Modal Information Retrieval Analysis.* Combining high-level semantics with low-level visual capabilities to perform image search more accurately and efficiently is a hot spot in the current image search research field. At present, in view of the overall high-level semantics and low-level features of the image, there is no mature semantic information collection model, which is difficult to meet the needs of today's massive image information retrieval. This article constructs a basic semantic information acquisition model to extract the semantics of the objects in the image from the basic features of the image information and conduct a series of exploratory research. Figure 1 is a schematic diagram of a cross-modal semantic information acquisition model.

*3.2. Data Collection*

(1) Through Baidu, Google, and Tencent three map software, in real-time street view, according to the operation of the time machine, obtain the street shop signs in this city. In fact, the main source of the screenshots here is Baidu Time Machine, because Google's time machine in mainland China is limited, and Tencent Maps captures a relatively small number of images while retaining the GPS location

(2) Filter and clean up the collected data set. Before cleaning the image, this article first determines the main area of the image used in this article. The main areas of the logo mainly include main text (shop name), shop logo, and related main auxiliary explanatory text

(3) After receiving the data set, the size of the data has a greater impact on the recognition of image semantics, in order to facilitate the continuation of the work of this paper. This article uses data improvement strategies to increase the amount of data. Common data expansion strategies include zooming, scaling the image according to a specific ratio, rotating, rotating the image at a specific angle or direct mirroring, and adding noise, such as adding specific noise or Gaussian motion blur

*3.3. Image Text Extraction*

*3.3.1. Image Extraction.* For image processing, this article uses a convolutional neural network. The output network consists of 10 cohesive layers, 4 concentrated layers, and 3 fully connected layers. The size of the torsion core of the first rotating layer is 5∗5. For a long time, the number of feature maps retrieved has been 64. The rest of the twist core size is 3∗3, the feature map data gradually increases from 64, 128, 256, and 512, and the size of all concentrated layers is 2∗2. The input image size is 256 ∗ 256 pixels, the spacing of all distortion layers is 1, the spacing of all concentrated layers is 2, the size of the feature map obtained after the first

Table 1: Comparison of model retrieval accuracy.

| | Top 1 | Top 2 | Top 4 | Top 6 | MAP |
|---|---|---|---|---|---|
| LMNN (1) | 47.41 | 51.12 | 55.56 | 58.34 | 53.401 |
| LMNN (2) | 50.52 | 56.01 | 60.12 | 63.80 | 57.721 |
| NetVlad | 65.56 | 67.21 | 71.00 | 76.11 | 70.136 |
| Deephash | 46.43 | 50.17 | 53.24 | 55.52 | 51.657 |
| CNN (image) | 77.35 | 79.56 | 80.10 | 82.24 | 79.931 |
| CNN (full) | 85.37 | 87.21 | 88.76 | 90.11 | 87.612 |

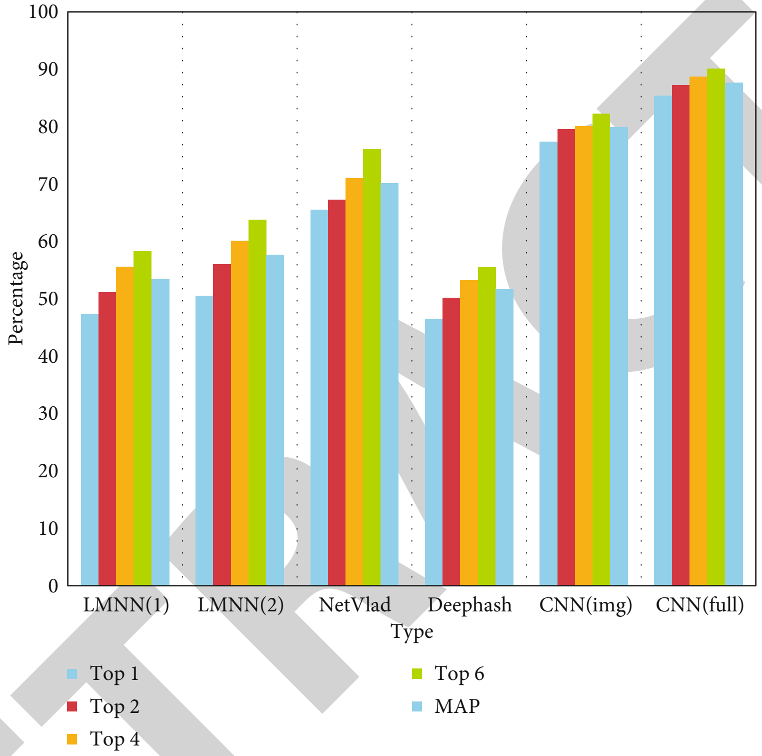

Figure 2: Comparison of model retrieval accuracy.

Table 2: Comparison of model retrieval accuracy.

| | Top 1 | Top 2 | Top 4 | Top 6 | MAP |
|---|---|---|---|---|---|
| LMNN (1) | 72.21 | 77.42 | 85.58 | 92.36 | 82.141 |
| LMNN (2) | 76.53 | 82.21 | 88.12 | 95.01 | 85.221 |
| NetVlad | 82.51 | 87.53 | 93.02 | 96.19 | 89.45 |
| Deephash | 71.41 | 77.01 | 83.45 | 89.74 | 80.342 |
| CNN (image) | 89.11 | 90.41 | 93.12 | 98.51 | 92.573 |
| CNN (full) | 96.01 | 97.0 | 98.76 | 99.38 | 97.423 |

rotation is 261 ∗ 261, the number is 64, and the number of sides is 64, so the size of the feature map will not be adjusted after the second rotation. The number of feature maps after sampling reduction is 128, the size is 108∗108, and the feature maps are rotated. They are all angular, the feature map does not adjust the image size, and the next two rotations do not fill the edges until the size becomes 27 ∗ 27 after two samples. The fill edge becomes 24∗24 after sampling. The feature map size before entering the fully connected layer is 12∗12∗512. The fully connected layer has three layers, the first layer has 4096 outputs, and the output of the last two layers is adjustable. The reason for the design of fully connected 3 layers here is that different sorted image data sets have different image types, that is, different numbers of image tags. When the number of layers is different and the corresponding pictures are the same, but the number of labels corresponds to the same, this requires a training set with different output numbers in the last fully connected

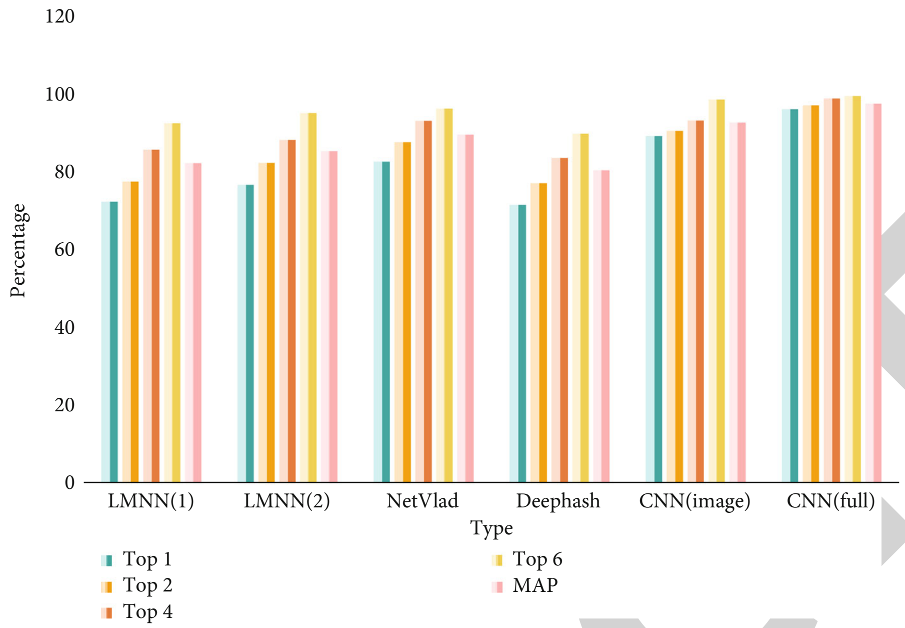

Figure 3: Comparison of model retrieval accuracy.

network layer, and there is a big gap in the number of labels between different data sets. The fully coupled layer ensures that the latter can smoothly adapt to different label numbers. Define the second fully connected layer and at the same time make the number of outputs of the second fully connected layer variable.

*3.3.2. Export Text.* The word segmentation tool used in this article will continue to use the stuttering word segmentation tool to segment the document. Then, use the LDA model to model the split words of each document in the data set. Finally, a word distribution can be realized in the topic, and then, a text topic distribution can be realized according to the topic distribution of the word.

*3.3.3. The Relevance of Images and Text.* Image and text information carriers are different in function, but both can express similar semantic information in content. Therefore, there is a correlation between information carriers in multiple formats and the characteristic data of each medium. This document uses the CCA algorithm to analyze the correlation of various data attributes and realizes the correlation analysis of vector graphics and text attribute tables.

## 4. Cross-Modal Information Retrieval Verification Based on Convolutional Neural Network

*4.1. Experimental Data.* The experiment in this paper is carried out on the Chinese signboard dataset. This paper selects the first 55 street signs of the Chinese signboard dataset for training and the last 15 street signs for testing. Because the data set is filtered, all the first 55 streets are finally obtained. The ratio of the total number of signs to the total number of signs in the last 15 streets is 5: 4, which is relatively reasonable.

Table 3: Test set error of different core network structures.

| | Test set error rate\% | Time\s |
|---|---|---|
| 16-16-16 | 2.56 | 65 |
| 16-16-32 | 2.12 | 76 |
| 32-32-32 | 1.21 | 123 |
| 32-32-64 | 1.31 | 132 |
| 64-64-64 | 1.1 | 211 |

*4.2. Experimental Design*

(1) After selecting some basic comparison models, this article will conduct experiments in the following situations (Top-$n$ represents the accuracy percentage of the results returned by the first $n$ search results. MAP is the percentage of Top1-Top6 accuracy) to compare the recovery results of the CNN model in this article with the results of LMNN, NetVlad, and Deephash. First of all, since this article uses a complete data set (including Chinese and English license plates), the GPS information of each license plate is not considered. The experimental results are shown in Table 1

Top-$n$ in Figure 2 represents the accuracy percentage of the results returned by the first $n$ search results. MAP is the percentage of Top1-Top6 accuracy. LMNN (1) means that LMNN is restored to 48 dimensions, and LMNN (2) means that LMNN is restored to 96 dimensions. CNN (image) only corresponds to the search results of the model image subnet features in this article, and CNN (full) corresponds to the

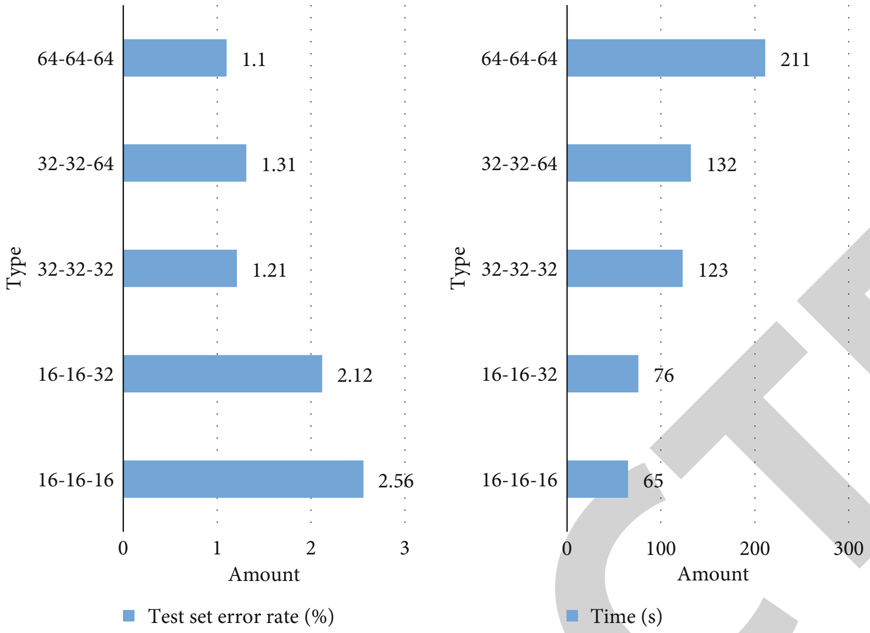

Figure 4: Test set error of different core network structures.

Table 4: Error comparison of different convolution kernel sizes.

| | Training set error rate\% | Test set error rate\% |
|---|---|---|
| 3 × 3 | 0.31 | 1.12 |
| 5 × 5 | 0.65 | 1.76 |
| 7 × 7 | 0.73 | 2.02 |

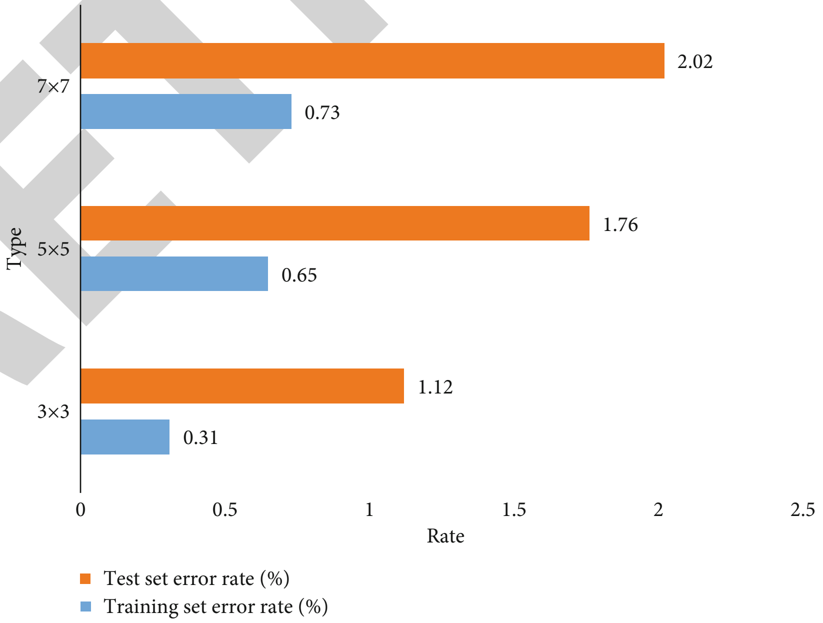

Figure 5: Error comparison of different convolution kernel sizes.

search results of the complete model in this article. In summary of the retrieval experiment results, the model in this article can achieve a good retrieval effect.

(2) This article also conducts experiments on the GPS information (1 km) combined with the signboard. The results are shown in Table 2

It can be seen from Figure 3 that when the GPS information is used to narrow the search range, the model data dimension is restored to 99, so the CNN (full) model in this article basically reaches the commercial level. In actual commercial use, only the general GPS range of the current query image is obtained, which greatly improves the speed and accuracy of the query.

*4.3. The Influence of the Choice of Convolution Parameters.* Based on the CNN convergence network model designed in this article, this article conducted various comparative experiments on the Caffe platform to determine the impact of different configuration settings on the performance of the convolutional neural network. The computer's CPU frequency is 5.6 GHz, and the memory is 8 G. The designed CNN model consists of three processes: sample assembly, activation, and execution; after inserting the complete combination layer into the softmax classifier, 10 categories are derived.

(1) Choice of audit

To see how the number of convolution cores in each twist layer affects network performance, just change the number of twist cores at each level according to the CNN network structure. The core number structure is 16-16-16, 16-16-32, 32-32-32, 32-32-64, and 64-64-64. In this paper, the converted core structure model is used to compare the test set errors of different core network structures. In the test, the training time batch size is 200, and the number of repetitions is 20. The results are shown in Table 3.

It can be seen from Figure 4 that the number of cores increases within a certain range, the performance is improved, and the time increases with the increase of the number of shrinking cores. Comparing the structure 32-32-32 with the 16-16-16 structure, the network time is more than twice the training time, and the test time is also increased. Therefore, different structures can be selected according to the actual needs of precision grading or time efficiency.

(2) Convolution kernel size

The size of the convolution kernel is used to extract image features into the rotation layer. Its size will definitely affect the quality of its functions. In order to study the impact of convolution kernel size on classification performance, in the structure of the CNN network designed in this article, the size of each convolution kernel is set to $3 \times 3$, $5 \times 5$, and $7 \times 7$. The present study uses database 1 and 2 to run experiments. The experimental results are shown in Table 4.

It can be seen from Figure 5 that the smaller the core size, the lower the network error rate of the training set and test set, and the better the network performance. The larger the core size, the higher the network error rate of the training set and test set, and the lower the network performance. However, if the size is too small (for example, $1 \times 1$), the performance will not be as good as $3 \times 3$ due to increased noise.

## 5. Conclusions

This paper focuses on the research of graph convolutional network for cross-modal information retrieval. After understanding the relevant theories, the cross-modal information retrieval based on convolutional neural network is designed, and then, the designed network model is verified. Experimental results are concluded that the accuracy of the convolutional neural network model designed in this article is high, with the highest accuracy reaching 90% and the traditional model reaching 80%. Then, there are still shortcomings in the research process of this article, mainly due to the incomplete detection of the model.

## Data Availability

Data available on request from the authors.

## Conflicts of Interest

There is no potential conflict of interest in our paper.

## Authors' Contributions

All authors have seen the manuscript and approved to submit to your journal.

## References

[1] G. Tolias, Y. Avrithis, and H. Jégou, "Erratum to: image search with selective match kernels: aggregation across single and multiple images," *International Journal of Computer Vision*, vol. 116, no. 3, pp. 262–262, 2016.

[2] L. Dan, X. Liu, and X. Qian, "Tag-based image search by social re-ranking," *IEEE Transactions on Multimedia*, vol. 18, no. 8, pp. 1628–1639, 2016.

[3] R. Liu, G. Lin, and C. Shen, "CRF learning with CNN features for image segmentation," *Pattern Recognition*, vol. 48, no. 10, pp. 2983–2992, 2015.

[4] Z. Cheng and J. Shen, "On very large scale test collection for landmark image search benchmarking," *Signal Processing*, vol. 124, no. Jul., pp. 13–26, 2016.

[5] Z. Ji, Y. Pang, Y. Yuan, and J. Pan, "Relevance and irrelevance graph based marginal Fisher analysis for image search reranking," *Signal Processing*, vol. 121, no. Apr., pp. 139–152, 2016.

[6] M. Behrisch, B. Bach, M. Hund et al., "Magnostics: image-based search of interesting matrix views for guided network exploration," *IEEE Transactions on Visualization & Computer Graphics*, vol. 23, no. 1, pp. 31–40, 2017.

[7] G. Tolias, Y. Avrithis, and H. Jégou, "Image search with selective match kernels: aggregation across single and multiple images," *International Journal of Computer Vision*, vol. 116, no. 3, pp. 247–261, 2016.

[8] J. Lu, V. E. Liong, and Z. Jie, "Deep hashing for scalable image search," *IEEE Transactions on Image Processing*, vol. 26, no. 5, pp. 2352–2367, 2017.

[9] F. Shen, Y. Yang, L. Liu, W. Liu, D. Tao, and H. T. Shen, "Asymmetric binary coding for image search," *IEEE Transactions on Multimedia*, vol. 19, no. 9, pp. 2022–2032, 2017.

[10] J. Music, T. Marasovic, V. Papic, I. Orovic, and S. Stankovic, "Performance of compressive sensing image reconstruction

for search and rescue," *IEEE Geoscience & Remote Sensing Letters*, vol. 13, no. 11, pp. 1739–1743, 2016.

[11] P. Budikova, M. Batko, and P. Zezula, "ConceptRank for search-based image annotation," *Multimedia Tools & Applications*, vol. 77, no. 7, pp. 8847–8882, 2018.

[12] F. Çalışır, M. Baştan, O. Ulusoy, and U. Güdükbay, "Mobile multi-view object image search," *Multimedia Tools & Applications*, vol. 76, no. 10, pp. 12433–12456, 2017.