Scholar Publishing Group
International Journal of Neural Network
https://doi.org/10.38007/NN.2021.020206
ISSN 2790-2692 Vol. 2, Issue 2: 40-46

Cross-modal Neural Networks Fused with Multimedia Retrieval

Kothapalli Radhakrishnan[*]

Chandigarh University, India

[*]*corresponding author*

Keywords: Multimedia Retrieval, Neural Networks, Cross-modal, Deep Learning

Abstract: Due to the rapid development of Internet multimedia technology, the wide use of smart phones and the expansion of social networks, people can share interesting content on the Internet anytime and anywhere, makes the different mode of multimedia data on the Internet (such as text, images and video, etc.) to present the characteristics of explosive growth, huge amounts of agglomeration. Such large-scale data marks the arrival of the era of multimedia big data, and brings new oortunities and challenges to the research and alication based on multi-modal learning. This paper focuses on the cross-modal research of neural networks fused with multimedia retrieval. This paper first analyzes the basic models of machine learning and neural networks, and proposes a cross-modal multimedia semantic matching method. The simulation results show that the proposed multi-media retrieval cross-modal neural network model has certain effectiveness and feasibility.

1. Introduction

With the rapid development of Internet technology and intelligent mobile devices, multi-modal data in the network shows a blowout growth. Different types of multi-modal data are growing rapidly on various Internet platforms such as social networking sites, video websites, music platforms, news and information [1]. Multi-modal data are increasingly fused together to form a complex organizational structure and relationship. Multimedia retrieval methods mainly include unimodal data retrieval and multi-modal data retrieval. The function of "search by image" provided by search engines such as Baidu and Google is actually a unimodal retrieval task on image data [2]. The cross-modal Retrieval algorithm can simultaneously retrieve multi-modal data with the same semantics by mining the semantic correlation between multi-modal data such as text, image and video. When searching, users only need to input any one of the various modal data such as text, image, video, audio, etc., and then they can retrieve the other modal data with the same semantics, which enriches the diversity of our retrieval results to a large extent and can bring better retrieval experience. The objects retrieved by cross-modal are multi-modal data with heterogeneous underlying features. The characteristics of multi-modal data are as follows: Multi-modal data are mixed and coexist [3, 4]. The organization structure and association relationship of data are relatively complex, semi-structured and unstructured data account for most of the data, and it is difficult to store multi-modal data directly and retrieve it effectively. Multi-modal data can express the same semantic information by complementation and enhancement. It is necessary to realize the leap from one mode to another mode according to the various relationships between the modes. With the rapid growth of multi-modal data, it is difficult to cope with the existing data processing ability of human beings. Moreover, the underlying multi-modal data is doped with a large amount of noise, and the correlation relationship between different modal data is hidden, so it is not easy to mine the correlation relationship [5, 6].

In recent years, cross-recovery technology has developed rapidly and many cutting-edge technologies in this field have emerged. Like the International Conference on Computer Vision and the Conference on Computer Vision and Model Recognition. Experts in probability statistics, graph theory and pattern recognition have proposed a number of recovery methods. Cross-recovery methods can be roughly divided into shallow learning methods and deep learning methods [7]. Among the shallow learning methods, graph correlation method and dictionary learning method are favored by researchers because of their high accuracy. Among them, graph regularization can better describe the similarity within and between modes, and better mine the association between labels, so it is widely used in semi-supervised learning [8]. Because deep learning has a powerful modeling ability for nonlinear data, and does not need to manually extract features, it can automatically mine the rich internal information in the data, and deep learning technology has made a breakthrough in the cross-modal field. Deep learning can be roughly divided into traditional neural network methods and GANs [9, 10].

Across a modal data under the background of rapid development, in-depth understanding and mining data contained in the information, establish a cross modal data, the relationship between effectively from huge amounts of multimodal data retrieved modal information, you need to provide a better service for people, improve people study and work efficiency, convenient production and living of people.

2. Cross-Modal Multimedia Retrieval based on Neural Networks

2.1. Deep Learning and Neural Networks

(1) Neural network

Deep neural network models are usually connected by a large number of neuron nodes, and each neuron processes the input from the neurons connected with it in the previous layer through a certain output function [11, 12]. As shown in Figure 1:

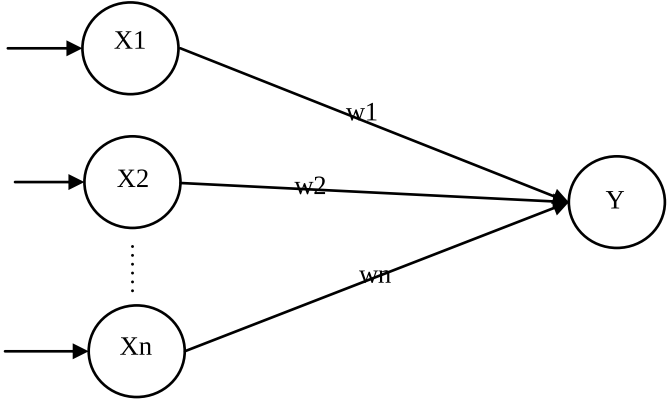

Figure 1. Schematic diagram of neuron

(2) Convolutional neural network

With the development of deep learning, more and more researchers realize that convolutional neural network has unique advantages in extracting image features, which also drives the combination of traditional visual tasks and deep learning [13]. In recent years, more and more convolutional networks have been alied to fields such as semantic separation, object detection, image classification and so on [14].

Convolutional neural network has been widely used in various Computer Vision tasks in recent years due to its powerful ability of image feature extraction, such as object detection and image classification. In general, the data processed by traditional convolutional neural networks are regular pixel lattice such as images or video frames, which have regular spatial structure [15]. However, many data in the real world do not have such regular Euclidean structure, such as social networks, knowledge graphs and other relational graph structures. Graph Convolution neural Network (GCN) is proposed to extract the features of data with topological Graph structure by using the feature extraction ability of convolutional Network [16].

(3) Cross-modal hash model based on deep learning

Cross-modal hashing algorithms based on deep learning have been greatly developed in recent years and achieved better and better retrieval results. Traditional cross-modal hashing methods map image and text data to binary space by constructing hash maing Matrix, and then use such as canonical correlation analysis (CCA) or Similarity Matrix as constraint guidance and optimization process. So of although the method can achieve good retrieval effect, but most traditional cross modal hash method USES man-made structure features as the input into the hash function (using the SIFT features such as images, text data using TF - IDF features, etc.) cannot be dynamically updated as the training feature, so we can not get the optimal hash function. And cross modal hash algorithm based on the deep learning generally use pre trained convolution neural network (CNN) recurrent neural network (RNN) to study the characteristic of image text data, and also introduces some network such as generated against network (GAN) figure convolution structure such as neural network (GCN) to further the potential information in the data mining, Therefore, significant effect improvement has been achieved [17, 18].

Although many cross-modal hashing methods based on deep learning have achieved good retrieval results, the current mainstream cross-modal retrieval algorithms at home and abroad also have some shortcomings. First of all, most algorithms mainly focus on modeling inter-modal Correlation between multi-modal data, so that the learned data features (such as image features, text features, video features, etc.) can retain the Correlation between different modal data as much as possible. However, the importance of correlation in intra-modal data is ignored. In addition, most cross-modal retrieval algorithms are based on supervised learning. By introducing category labels as the prior of training data, the learned features have more semantic information, so they can bring better retrieval results. But the vast majority of data that exists on the Internet is unlabeled.

2.2. Cross-Modal Multimedia Semantic Matching

(1) Retrieval process

The first step of cross-modal multimedia retrieval is to extract multi-modal data features. Then, based on the feature representation of the extracted multi-modal data, a cross-modal multimedia retrieval model was established to learn the common representation of different modal data. Finally, based on the retrieval results obtained by the retrieval model, other modal data that are most relevant to the queried modal data are obtained.

(2) Semantic matching

Semantic matching is to project different modal data into a common semantic space. Taking the mutual inspection of image and text as an example, I represents a graph sample, and T represents a text sample. Furthermore, VIp×n represents the underlying features of the image, VTq×n represents the underlying features of the text, where P and q represent the dimensions of the underlying features of the image and the underlying features of the text, and N represents the number of samples (image and text samples correspond one to one). Semantic matching is to map the underlying features of image and text to their common semantic space, as shown below:

$$M_I : V_I^{p \times n} \to S^{k \times n} \tag{1}$$

$$M_T : V_T^{q \times n} \to S^{k \times n} \tag{2}$$

In semantic space, there is a semantic dictionary L=(L1, L2,... Lk) such as architecture class and biology class, where k is the number of semantic categories.

(3) Cross-modal multimedia semantic matching based on deep neural network

Because polymorphic data usually have different statistical characteristics and inconsistent distribution, it is difficult to compare them directly for cross-retrieval. The fragmentation of cross functions of neural networks tries to learn the functions of specific functions and maps the corresponding functions in a common space where the codes fragmentation of different ways can be directly compared with each other. Then calculate the similarity between the different templates according to the received global hash code to retrieve deleted templates. In addition, in the Hamming common area, the similarity of the same sample category shall be greater than that of the different sample categories.

After projecting the underlying features of the image and text into the semantic space of the two, a query text is obtained. The retrieval model can obtain its semantic representation×Tkˆ Sk, and the image most relevant to the query text can be obtained by finding the minimum distance between the image and the image to be retrieved:

$$D(T, I) = dis\tan ce(\pi_T^k, \pi_I^k) \tag{3}$$

Similarly, the same is true for image query related text. The pseudo-code of cross-modal multimedia semantic matching algorithm based on deep neural network is shown below

Input: Training the underlying features and labels of images and text, testing the underlying features of images and text

Output: cross-modal multimedia retrieval results

Create image network NI and text network NT;

The underlying features and tags of the training images are input into NI and the network is trained, and the underlying features and tags of the training text are input into NT and the network is trained.

The bottom feature of the test image is input into the trained network NI, and the bottom feature of the test text is input into the trained network NT.

The top-level output of network NI and NT is obtained, which is the common semantic space of image and text.

3. Model Simulation Experiment

3.1. Introduction to the Dataset

This paper uses two widely used datasets: Wikipedia and NUS-Wide-10K, both of which are composed of labeled image text pairs.

3.2. Comparative Experiment

In order to evaluate the performance of the proposed model, the following methods are selected for comparison, including the classical cross-modal hashing methods CVH and SMFH. And the recent cross-modal hashing methods SSAH and DSSAH based on deep learning. In order to be fair, these methods uniformly adopt the same experimental Settings as this experiment. In this paper, the code provided by the original author or the running method of the original paper is used to obtain the retrieval results matching the published results, and the results of the model in this paper are compared.

4. Analysis of Experimental Results

4.1. Wikipedia Data Set

Table 1. Wikipedia dataset image retrieval text task comparison

	16	32	64
CVH	0.217	0.246	0.254
SMFH	0.337	0.390	0.364
SSAH	0.473	0.502	0.498
DSSAH	0.557	0.561	0.578
Our model	0.595	0.603	0.619

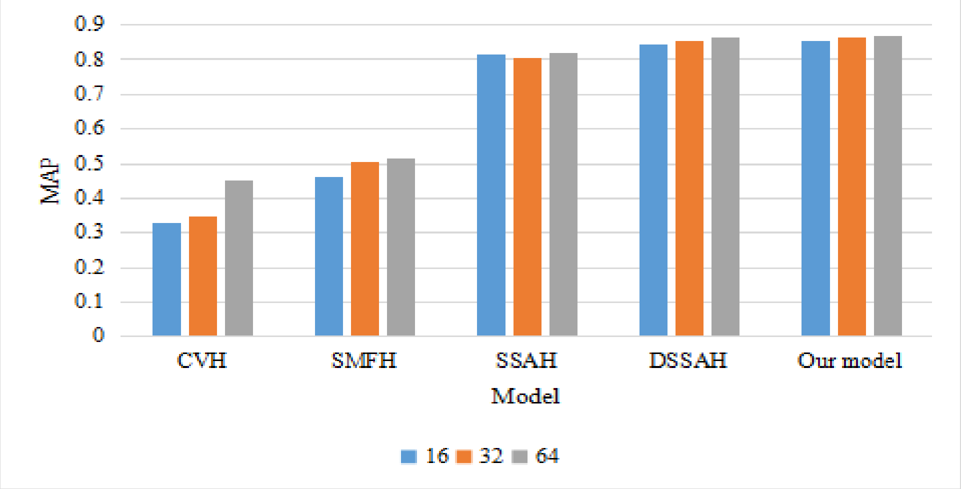

Figure2. Wikipedia dataset text retrieval image task comparison

As shown in Table 1 and Figure 2, in the text work of image recovery and the image work of text recovery, the SAH and dssah cross-fragmentation methods based on deep learning and the model proposed in this work are better than the classical CVH cross-fragmentation methods; and smfh. In Wikipedia's dataset, taking as an example the length of 64-bit hash code, the model improves text recovery performance at least 0.040 compared to other methods.

4.2. NUS - WIDE - 10 K Data Set

As shown in Table 2 and Figure 3, in the task of image-to-text retrieval and text-to-image retrieval on NUS-Wide-10K dataset, taking the length of 64-bit hash codes as an example, the proposed model is improved by at least 0.042 and 0.022, respectively. The results show that the proposed model is a cross-modal retrieval method with both accuracy and efficiency.

Table 2. Comparison of image retrieval text tasks in NUS-Wide-10K dataset

	16	32	64
CVH	0.354	0.341	0.337
SMFH	0.432	0.445	0.447
SSAH	0.623	0.659	0.661
DSSAH	0.714	0.725	0.730
Our model	0.746	0.758	0.772

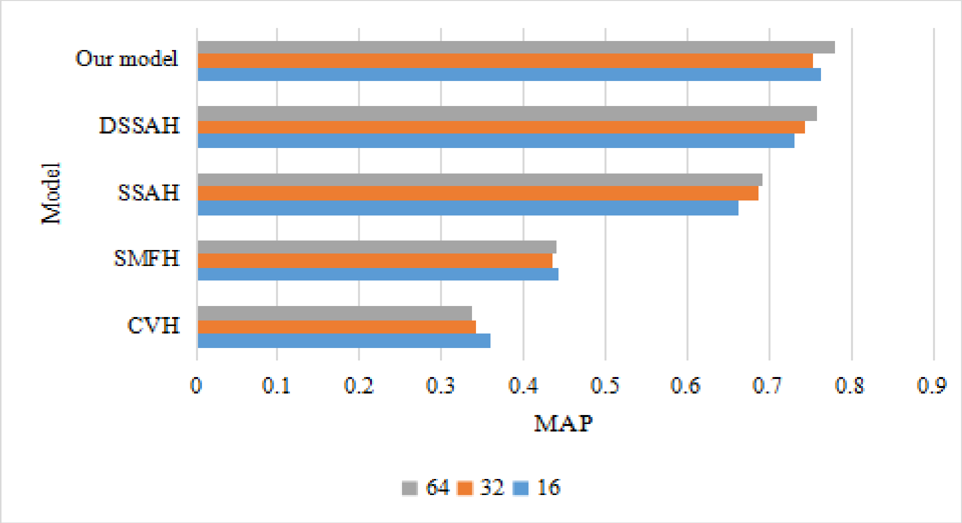

Figure 3. Comparison of text retrieval image tasks in NUS-Wide-10K dataset

5. Conclusion

Large-scale multimedia retrieval algorithm can realize the mutual retrieval of different multimedia data (such as image, text, video, etc.). Compared with traditional unimodal retrieval algorithm, it can provide users with more comprehensive retrieval experience. In this paper, a cross-modal hashing algorithm based on neural networks is proposed, which can achieve similar or even better retrieval performance than supervised methods while reducing the consumption of human resources. It is of great value and significance to study the learning methods of cross-modal retrieval and classification. In the cross-modal learning task studied in this paper, there are many directions that can be further studied and explored. In practical alication scenario, tend to lack the tag data, sample contains a lot of noise and quality of the collected data is not reliable, because of the different modal contain information not identical, therefore, how to different modal based on collaborative learning migration and promote each other between information and knowledge for model performance is a research direction in the future.

Funding

This article is not supported by any foundation.

Data Availability

Data sharing is not applicable to this article as no new data were created or analysed in this study.

Conflict of Interest

The author states that this article has no conflict of interest.

References

[1] Talreja F, Valenti M C, Nasrabadi N M. Error-Corrected Margin-Based Deep Cross-Modal Hashing for Facial Image Retrieval. IEEE Transactions on Biometrics Behavior and Identity Science. (2020) (99):1-1.

[2] Dorfer M, Jr J H, Arzt A, et al. Learning Audio–Sheet Music Correspondences for Cross-Modal Retrieval and Piece Identification. Transactions of the International Society for Music Information Retrieval. (2018) 1(1):22.

[3] Oleinik A L, Kukharev G A. Algorithms for Face Image Mutual Reconstruction by Means of Two-Dimensional Projection Methods. SPIIRAS Proceedings. (2018) 2(57):45.

[4] Mandal D, Rao P, Biswas S. Semi-Supervised Cross-Modal Retrieval with Label Prediction. IEEE Transactions on Multimedia. (2019) (99):1-1.

[5] Talreja F, Valenti M C, Nasrabadi N M. Error-Corrected Margin-Based Deep Cross-Modal Hashing for Facial Image Retrieval. IEEE Transactions on Biometrics Behavior and Identity Science. (2020) (99):1-1.

[6] Moreno-Schneider J, Martinez P, Martinez-Fernandez J L. Combining Heterogeneous Sources in an Interactive Multimedia Content Retrieval Model. Expert Systems with Alications. (2017) 69(mar.):201-213.

[7] Marzban E N, Eldeib A M, Yassine I A, et al. Alzheimer's disease Diagnosis From Diffusion Tensor Images Using Convolutional Neural Networks. PLoS ONE. (2020) 15(3):e0230409.

[8] Fonseca R, Guarnizo O, Suntaxi D, et al. Convolutional Neural Network Feature Extraction Using Covariance Tensor Decomposition. IEEE Access. (2021) (99):1-1.

[9] Reiser P, Eberhard A, Friederich P. Graph Neural Networks in Tensorflow-Keras with Raggedtensor Representation (kgcnn). Software Impacts. (2021) 9(4):100095.

[10] Haridas P, Chennupati G, Santhi N, et al. Code Characterization with Graph Convolutions and Capsule Networks. IEEE Access. (2020) (99):1-1.

[11] Vuuturi A, Gupta A, Ghosh N. MCA-DN: Multi-Path Convolution Leveraged Attention Deep Network for Salvageable Tissue Detection in Ischemic Stroke from Multi-Parametric MRI. Computers in Biology and Medicine. (2021) 136(5):104724.

[12] Samani Z R, Parker D, Alaatt J A, et al. Nimg-21. Differentiating Tumor Types Based on the Peritumoral Microenvironment Using Convolutional Neural Networks. Neuro-Oncology. (2020) 22(Sulement_2):ii151-ii151.

[13] Taguchi H, Liu X, Murata T. Graph Convolutional Networks for Graphs Containing Missing Features. Future Generation Computer Systems. (2021) 117(5):155-168.

[14] Kh. U R, Solovyev V V, Rogozin I B. Recognition of Prokaryotic and Eukaryotic Promoters Using Convolutional Deep Learning Neural Networks. PLoS ONE. (2017) 12(2):e0171410.

[15] Roy P, Song S L, Krishnamoorthy S, et al. NUMA-Caffe: NUMA-Aware Deep Learning Neural Networks. ACM Transactions on Architecture and Code Optimization. (2018) 15(2):1-26.

[16] Na, Ta, Hanshuang, et al. Mining Key Regulators of Cell Reprogramming and Prediction Research Based on Deep Learning Neural Networks. IEEE Access. (2020) (99):1-1.

[17] Srinivasu P N, Sivasai J G, Ijaz M F, et al. Classification of Skin Disease Using Deep Learning Neural Networks with MobileNet V2 and LSTM. Sensors. (2021) 21(8):2852.

[18] Srisailam E. Household Load forecasting using Deep Learning neural networks. Turkish Journal of Computer and Mathematics Education (TURCOMAT). (2021) 12(2):788-794.