Exploration of Optimization Paths for Retrieval Algorithms Based on Deep Learning Models

Zhiqiang Zhou, Fan Chen, Jie Qiu*
School of Computer Science and Engineering, Yulin Normal University, Yulin, Guangxi, China
**Corresponding Author.*

Abstract: In the era of big data, the rapid and accurate retrieval of massive information is of great significance. Traditional retrieval algorithms struggle to meet the complex and diverse retrieval requirements. This paper focuses on exploring the optimization paths for retrieval algorithms based on deep learning models, aiming to enhance the performance and efficiency of retrieval algorithms. By combing the basic theories of deep learning and retrieval algorithms, this paper deeply analyzes common retrieval algorithms based on convolutional neural networks (CNNs), recurrent neural networks (RNNs), and Transformers, clarifying their advantages and limitations. In response to the challenges at the data, algorithm, and application scenario levels, optimization paths such as data preprocessing and augmentation, model structure improvement and hyperparameter tuning, and multimodal fusion retrieval are proposed, and verified through practical cases in the fields of image retrieval and text retrieval. The research shows that the optimized retrieval algorithms have significantly improved in key indicators such as accuracy and recall rate, providing a useful reference for the further application of deep learning in the retrieval field.

Keywords: Deep Learning Models; Retrieval Algorithms; Optimization Paths; Data Preprocessing; Multimodal Fusion Retrieval

1. Introduction

1.1 Research Background and Significance

With the rapid development of information technology, the era of big data has arrived, and the amount of data has been growing explosively. According to the prediction of the International Data Corporation (IDC), the global data volume will increase from 33ZB in 2018 to 175ZB in 2025. Such a huge amount of data contains rich information value. In this context, information retrieval, as a key means of obtaining knowledge, has become increasingly important. Whether it is the review of massive literature in academic research, the mining of market data in the business field, or the search for various types of information on the Internet in daily life, efficient and accurate retrieval has become an urgent need for people.

Traditional retrieval algorithms, such as the algorithm based on Term Frequency - Inverse Document Frequency (TF - IDF), gradually reveal many limitations in the era of big data. These algorithms mainly rely on the occurrence frequency of keywords in the text and their rarity in the entire document collection for retrieval and ranking. However, with the rapid expansion of data scale, the types of data have become increasingly complex, including not only text but also multiple modalities such as images, audio, and video. Traditional algorithms have difficulty effectively processing such diverse data and are even less capable of understanding complex semantics. For example, when a user enters a relatively ambiguous or metaphorical query, traditional algorithms based on keyword matching often cannot accurately understand the user's true intentions and thus return a large number of irrelevant results. When facing the retrieval of large - scale image or audio data, traditional algorithms also lack effective feature extraction and matching methods, and their retrieval efficiency and accuracy are far from satisfactory.

The rise of deep - learning technology has brought new opportunities for the optimization of retrieval algorithms. Deep learning can automatically learn the internal features and patterns of data from a large amount of data by constructing multi - layer neural networks [1]. In the field of image retrieval, convolutional neural networks (CNNs) can automatically extract high - level semantic features of images, such as the shape, color, and texture of objects, enabling more accurate image similarity matching. In text retrieval, recurrent neural networks (RNNs) and their variants, such as Long Short - Term Memory networks (LSTMs) and Gated Recurrent Units (GRUs), can better process the context information of text and understand the semantic coherence, thereby improving the accuracy of text retrieval. In addition, the introduction of the attention mechanism further enhances the ability of deep - learning models to focus on key information, enabling the models to more accurately capture important features related to the query when processing complex data.

Optimizing retrieval algorithms based on deep-learning models has important theoretical and practical significance [1,2]. From a theoretical perspective, deep learning has opened up a new path for the research of retrieval algorithms, promoting the in - depth integration of information retrieval theory with fields such as machine learning and artificial intelligence. By deeply studying the application of deep - learning models in retrieval, the internal mechanism of information retrieval can be further revealed, enriching and improving the theoretical system of information retrieval. In practical applications, the optimized retrieval algorithms can significantly improve retrieval efficiency and accuracy, providing users with better retrieval services. In the academic field, researchers can obtain relevant research literature more quickly, accelerating the progress of scientific research; in the business field, enterprises can analyze market data more accurately, grasp market trends, and formulate more effective marketing strategies; in Internet search engines, users can find the information they need more quickly, enhancing the user experience.

1.2 Research Status at Home and Abroad

Abroad, the research on deep - learning - based retrieval algorithms started early and has achieved fruitful results. Google, a giant in the search - engine field, began to apply deep - learning technology to the optimization of search algorithms many years ago. Google uses deep - learning models to semantically understand web - page content. By constructing large - scale neural networks, it learns the semantic features and context relationships in the text, enabling it to more accurately judge users' search intentions. For example, when processing complex natural - language queries, Google's deep - learning model can identify the semantic associations between keywords and the implicit semantic information in the query statement, and then return search results that better meet users' needs. In the field of image retrieval, Google's image search engine uses convolutional neural networks (CNNs) to extract and classify image features, realizing content - based image retrieval. When a user uploads a picture, the search engine can quickly find other similar pictures. This technology has been widely applied in fields such as image copyright management and image material search.

In China, with the rapid development of artificial - intelligence technology, the research on deep - learning - based retrieval algorithms has also shown a booming trend. Baidu, a leading enterprise in China's search - engine field, actively explores the application of deep learning in search algorithms. Baidu has realized semantic understanding and intelligent question - answering functions using deep - learning technology. Baidu's knowledge graph, combined with deep - learning algorithms, can understand users' questions and find accurate answers from massive knowledge. For example, when a user asks a question like "What is the capital of China?", Baidu's intelligent question - answering system can directly give the accurate answer "Beijing", instead of just returning relevant web - page links. In the field of image search, Baidu has also made significant progress. By using deep - learning models to extract and match image features, it has achieved high - precision and high - efficiency image search.

Although many achievements have been made in the field of deep - learning - based retrieval algorithms at home and abroad, there are still some research gaps and deficiencies. On the one hand, in the field of multimodal data fusion retrieval, although some studies have tried to fuse data of multiple modalities such as text, images, and audio, the current fusion methods are not mature enough. They cannot fully explore the internal relationships between multimodal data, resulting in a need for further improvement in retrieval performance. For example, in text - image fusion retrieval, how to better fuse the visual features of images and the semantic features of text to achieve more accurate retrieval remains an urgent problem to be solved. On the other hand, the interpretability problem of deep - learning models has always been an important factor restricting their wide application in the retrieval field. Most current deep - learning models are complex black - box models, making it difficult to understand the decision - making process and basis of the models. In some scenarios where high interpretability of results is required, such as medical information retrieval and financial information retrieval, this restricts the application of deep - learning - based retrieval algorithms. In addition, when dealing with large - scale and high - dimensional data, the efficiency and scalability of deep - learning - based retrieval algorithms also face challenges. How to improve the operation efficiency of algorithms and reduce the consumption of computing resources while ensuring retrieval accuracy is also a key point that future research needs to focus on.

2. Basics of Deep Learning and Retrieval Algorithms

2.1 Introduction to Common Deep Learning Models

In the field of deep learning, various models have their unique structural features and applicable scenarios, providing diverse tools for solving different types of problems.

The Convolutional Neural Network (CNN) is designed specifically for processing data with a grid - like structure, such as images and audio. Its core components include convolutional layers, pooling layers, and fully - connected layers [3]. The convolutional layer performs convolution operations by sliding a convolution kernel over the data, automatically extracting local features. Meanwhile, the characteristic of weight sharing significantly reduces the number of model parameters and lowers the computational cost.

The Recurrent Neural Network (RNN) is mainly used for processing sequential data, such as text, speech, and time series [4]. It has feedback connections and can remember the input information from previous time steps, thus capturing the temporal dependencies in the data. In text processing, the RNN can understand the semantics of the current word based on the content of the previous text, and then process the semantics of the entire sentence or passage.

The Transformer model has had a significant impact in the field of deep learning in recent years. Based on the self - attention mechanism, it completely abandons the recurrent and convolutional structures and can process the entire sequence in parallel, greatly improving the computational efficiency [5]. The self - attention mechanism allows the model to directly focus on any position in the input sequence, better capturing the global dependencies in the sequential data.

2.2 Basics of Retrieval Algorithms

As the core component of an information retrieval system, the retrieval algorithm aims to quickly and accurately find information relevant to the user's query from massive data resources. With the continuous development of information technology, retrieval algorithms have evolved from traditional to modern ones, forming various types, each with its unique principles and application scenarios.

The Vector Space Model (VSM) is a classic information retrieval model [6]. It was first proposed by G. Salton et al. in the late 1960s and applied in the SMART information retrieval system [7]. The basic idea of this model is to represent both documents and queries as vectors in a vector space, and measure the relevance between a document and a query by calculating the similarity between the vectors. In the vector space model, each dimension represents a feature term (usually a word), and the value of the vector indicates the weight of the feature term in the document or query. The advantage of the vector space model is that it is simple and intuitive, easy to understand and implement, and can effectively handle text information retrieval tasks. However, it also has some limitations. For example, it has insufficient understanding of semantic relationships in text and cannot accurately capture the semantic associations between words, resulting in poor retrieval performance when dealing with some semantically complex queries.

Term Frequency - Inverse Document Frequency (TF - IDF) is a widely used weighting technique in information retrieval and text mining, often used to calculate the importance of words in a text [8]. TF represents term frequency, which is the number of times a certain word appears in a document, reflecting the local importance of the word in the document. IDF represents inverse document frequency, which measures the rarity of a word in the entire document collection. If a word appears in many documents, its IDF value is low, indicating that the word has a low discriminatory power. Conversely, if a word appears in only a few documents, its IDF value is high, indicating that the word has a strong discriminatory power. The TF - IDF value is the product of TF and IDF. It comprehensively considers the frequency of a word in a document and its rarity in the entire document collection, and can more accurately measure the importance of a word in a document. In a search engine, when a user enters a query word, the system calculates the TF - IDF value of the query word in each document, and then sorts the documents based on these values. The documents with higher TF - IDF values are ranked higher and returned to the user as results with a higher relevance to the query. The TF - IDF algorithm is simple and effective and has been widely applied in tasks such as text classification, keyword extraction, and text similarity calculation. However, it also has some drawbacks, such as insufficient consideration of word order and semantic relationships in text and the inability to handle complex semantic problems like polysemy.

With the development of deep - learning technology, new retrieval algorithms based on deep learning have emerged. The image retrieval algorithm based on the Convolutional Neural Network (CNN) utilizes the powerful image feature extraction ability of the CNN to automatically learn the local and global features of images. In image retrieval, the image is first input into the CNN model. Through a series of operations such as convolutional layers and pooling layers, the feature vector of the image is extracted. Then, extract features and find similar images. In an image search engine, when a user uploads a landscape photo, the CNN - based retrieval algorithm can quickly find similar landscape photos from a massive image database. The text retrieval algorithm based on the Recurrent Neural Network (RNN) and its variants (such as LSTM and GRU) is mainly used to process the sequential information of text. In text retrieval, the RNN can understand the semantic relationships between words according to the context information of the text, thus better matching the user's query. When processing a news text, the RNN can accurately determine the theme and content of the text by learning the order and semantic associations of the words in the text. When the user enters a news - related query, the RNN can find the relevant news text based on the learned semantic information. The Transformer model performs well in retrieval tasks in the field of natural language processing. Based on the self - attention mechanism, it can better capture the global dependencies in the text. In text retrieval, the Transformer model can simultaneously focus on different positions in the input text, understand the overall semantics of the text, and thus improve the accuracy of retrieval. In an intelligent question - answering system, the Transformer model can find the most relevant answers from a large amount of text data according to the user's questions. These new retrieval algorithms based on deep learning can automatically learn the features and patterns of data, have obvious advantages in processing complex data and semantic understanding, and have brought new breakthroughs and developments to information retrieval.

3. Analysis of Retrieval Algorithms Based on Deep Learning Models

3.1 Analysis of Common Retrieval Algorithms Based on Deep Learning

3.1.1 Image retrieval algorithm based on Convolutional Neural Network (CNN)

In the field of image retrieval, the algorithm based on the Convolutional Neural Network (CNN) has become one of the mainstream technologies. Its powerful feature extraction ability provides strong support for achieving efficient and accurate image retrieval. Taking an image search engine as an example, the workflow of such algorithms covers multiple key links from image feature extraction to similarity matching.

Image feature extraction is one of the core steps of the CNN - based image retrieval algorithm. In this process, CNN models are widely used to automatically learn the rich features of images. Take the classic AlexNet model as an example. It contains multiple convolutional layers and pooling layers. When an image is input, it first enters the convolutional layer. The convolution kernels in the convolutional layer perform convolution operations by sliding over the image to extract local features of the image, such as edges and textures. Convolution kernels with different sizes and parameters can capture feature information at different scales and types. For example, smaller convolution kernels may be better at extracting detailed features in the image, while larger convolution kernels help capture the overall structural features of the image. In AlexNet, after being processed by multiple convolutional layers, the image is transformed into a series of feature maps with different features. Subsequently, the pooling layer downsamples these feature maps. Through operations such as max - pooling or average - pooling, it reduces the resolution of the feature maps while retaining the main features, reducing the computational load. After the alternating processing of convolutional layers and pooling layers, the high - level semantic feature vectors of the image are finally obtained. These vectors contain rich information such as the shape, color, and texture of the objects in the image and can effectively represent the content of the image.

In addition to AlexNet, the VGGNet is also a CNN model widely used in image feature extraction [3]. VGGNet has a deeper network structure. By stacking multiple convolutional layers to increase the depth of the network, it can learn more advanced and abstract image features. Its convolutional layers usually use smaller convolution kernels (such as 3×3) and gradually extract image features through multi - layer convolution operations. For example, when processing a natural landscape image, the convolutional layers of VGGNet can start from low - level features such as the edges and textures of the image, gradually learn intermediate - level features such as mountains, rivers, and the sky, and finally extract the high - level semantic features representing the entire landscape scene. This deep network structure makes VGGNet perform outstandingly in image feature extraction and can provide more representative feature vectors for image retrieval.

After obtaining the feature vectors of the images, the CNN - based image retrieval algorithm needs to perform similarity matching to find the images most similar to the query image. Commonly used similarity measurement methods include Euclidean distance, cosine similarity, etc. The Euclidean distance measures the similarity of two feature vectors by calculating their geometric distance in space. The smaller the distance, the more similar the two vectors are. The cosine similarity measures the similarity by calculating the cosine value of the angle between two feature vectors. The closer the cosine value is to 1, the more similar the directions of the two vectors are, that is, the more similar the image contents are. For example, in a database containing a large number of commodity images, when a user uploads a query image, the system first uses the CNN model to extract the feature vector of the query image, and then calculates the cosine similarity between this feature vector and the feature vectors of all images in the database. According to the calculation results of the cosine similarity, the images with higher similarity are returned to the user as retrieval results. These images are the most similar to the query image in content and may be different styles or different - angle shooting images of the same type of commodity.

The CNN - based image retrieval algorithm has achieved remarkable results in practical applications. In the field of image copyright management, by performing CNN - based feature matching between the image to be detected and the images in the copyright image database, it is possible to quickly and accurately determine whether there are copyright issues with the image to be detected. In image material search, designers can upload a reference image and use the CNN - based image retrieval algorithm to find images with similar styles and contents in a massive image material library, providing rich material resources for design work. However, this algorithm also has some limitations. For example, for some images with complex semantics, similar contents but large differences in details, inaccurate retrieval may occur. In addition, the training of CNN models requires a lot of labeled data and computing resources, resulting in high training costs. Moreover, the interpretability of the models is poor.

3.1.2 Text retrieval algorithm based on Recurrent Neural Network (RNN)

In the field of text retrieval, the algorithm based on the Recurrent Neural Network (RNN) provides important support for achieving accurate text retrieval with its effective processing ability for text sequence information. Taking an intelligent customer service system as an example, this algorithm fully demonstrates its unique advantages and working mechanisms in the process of processing user questions and matching relevant answers.

After understanding the semantics of the user's question, the RNN - based text retrieval algorithm needs to perform matching in the knowledge base to find the answers related to the question. The knowledge base stores a large amount of text data, including frequently asked questions and their answers, product descriptions, technical documents, etc. The algorithm determines which texts are related to the user's question by calculating the similarity between the semantic representation of the user's question and the semantic representation of each text in the knowledge base. For example, the cosine similarity is used to calculate the cosine value of the angle between the user's question vector and the text vectors in the knowledge base. The higher the cosine value, the more similar the semantics of the two are. Suppose there are detailed introductions about the camera functions of Apple and Huawei mobile phones in the knowledge base. When the user asks the above - mentioned question, the algorithm can find the text paragraphs most relevant to the question by calculating the similarity and return them to the user as answers.

The RNN - based text retrieval algorithm has important application value in intelligent customer service systems. It can quickly and accurately understand the user's questions, find relevant answers from the massive knowledge base, and improve the response efficiency and service quality of customer service. In e - commerce customer service, it can quickly answer users' questions about product information and purchase processes. In technical support customer service, it can provide users with professional technical solutions. However, this algorithm also has some disadvantages. When processing a large - scale knowledge base, the retrieval efficiency may be affected because it is necessary to calculate the similarity for each text in the knowledge base, resulting in a large amount of computation. In addition, for some questions with ambiguous or polysemous semantics, the algorithm may have difficulty accurately understanding the user's intentions, leading to inaccurate retrieval results.

3.1.3 Multimodal retrieval algorithm based on transformer

With the increasing richness of data types, multimodal retrieval has become a research hotspot in the field of information retrieval. The multimodal retrieval algorithm based on Transformer can effectively integrate information from multiple modalities, providing an innovative solution for achieving more accurate and comprehensive retrieval. Taking cross - media retrieval applications as an example, this algorithm demonstrates powerful information integration and retrieval capabilities when processing multiple - modality data such as text, images, and audio.

The multimodal retrieval algorithm based on Transformer first needs to extract features from data of different modalities [9]. For text data, it usually used to convert text into vector representations. BERT learns rich language knowledge and semantic information through pre - training on a large - scale corpus and can map each word or term in the text into a high - dimensional vector. These vectors contain the semantic, syntactic, and context information of the text. As mentioned above, CNN can automatically learn the local and global features of images. Through the processing of convolutional layers and pooling layers, the image is transformed into a feature vector. For example, when using the ResNet model to extract features from an image of a cat, ResNet can extract features such as the shape, color, and texture of the cat in the image and generate a feature vector representing the visual content of the image. In the processing of audio data, the audio signal is usually first converted into a representation form such as a spectrogram, and then a specialized audio processing model, such as a model based on a convolutional neural network or a recurrent neural network, is used to extract audio features. For example, after converting a cat meowing audio into a spectrogram, a convolutional neural network is used to extract features from the spectrogram to obtain a vector representing the audio features. After extracting the data features of different modalities, the multimodal retrieval algorithm based on Transformer needs to perform information fusion. The Transformer model, based on the self - attention mechanism, can effectively capture the correlations between features of different modalities. A common fusion method is to use the feature vectors of different modalities as inputs and perform interaction and fusion through the self - attention mechanism of the Transformer. When processing text - image retrieval tasks, the text feature vector and the image feature vector are simultaneously input into the Transformer. The self - attention mechanism in the Transformer can calculate the attention weights between each word in the text and each region in the image, thus achieving the in - depth fusion of text and image information. For example, when a user enters a text query "A cat is on the grass" and provides an image containing grass at the same time, the Transformer model can, through the self - attention mechanism, find the regions in the image related to the text description, such as the grass and the possible position of the cat, and fuse the features of the image and the text to obtain a more comprehensive semantic representation. Another fusion strategy is to use multiple independent Transformer encoders to process the input data of different modalities respectively, and then fuse the outputs of the encoders. For text and image data, one Transformer encoder can be used to process the text sequence, and another Transformer encoder can be used to process the image features. Then, the outputs of the two encoders are fused through methods such as concatenation and weighted summation. This method can give full play to the processing capabilities of the Transformer for different - modality data and maintain the independence of different - modality data, facilitating subsequent analysis and processing.

After completing the information fusion, the multimodal retrieval algorithm based on Transformer performs retrieval by calculating the similarity between the query and the multimodal data in the database. Similar to retrieval algorithms based on a single modality, commonly used similarity measurement methods include Euclidean distance, cosine similarity, etc. By calculating the similarity between the multimodal feature vector of the query and the feature vectors of each multimodal data in the database, the data with higher similarity are returned to the user as retrieval results. For example, in a cross - media retrieval system, when a user enters a query containing a text description and an image example, the system can, by calculating the similarity between the query and the multimodal data in the database, find multimedia resources that contain relevant text information and have similar image content, such as news reports and photo collections.

The multimodal retrieval algorithm based on Transformer has achieved remarkable results in applications such as cross - media retrieval. It can make full use of the complementary information between multiple - modality data to improve the accuracy and comprehensiveness of retrieval. However, this algorithm also faces some challenges. The acquisition and annotation of multimodal data are difficult and require a large amount of manpower and time costs. In addition, how to further optimize the structure and parameters of the Transformer model to better integrate multimodal information and improve retrieval efficiency and performance remains the focus and difficulty of current research.

3.2 Limitations of Deep Learning Models in Retrieval Algorithms

Although deep learning models show many advantages in retrieval algorithms, their limitations cannot be ignored. These limitations restrict their further application and development in the retrieval field to a certain extent.

Deep learning models have extremely high requirements for computing resources, which is one of the main challenges they face [10]. In the model training stage, powerful computing hardware support is needed, such as high - performance Graphics Processing Unit (GPU) clusters. Taking the training of a large - scale Transformer model as an example, it may require hundreds or even thousands of GPUs to work together, and the training process may last for weeks or even months. This not only requires a huge investment in hardware procurement and maintenance costs but also places extremely high demands on the power supply and cooling systems of data centers. In practical applications, in some resource - constrained scenarios, such as mobile devices or small - enterprise servers, it is difficult to afford such high computing resource costs, resulting in limitations in the deployment and application of deep learning models. In image retrieval, when using a convolutional neural network for feature extraction and similarity matching, the processing of large - scale image datasets involves a huge amount of computation, which may lead to slow retrieval speeds and inability to meet real - time requirements. In text retrieval, deep - learning - based models also consume a large amount of computing resources when processing a large amount of text data, affecting the retrieval efficiency.

Deep learning models have poor interpretability and are usually regarded as black - box models [10,11]. This becomes an obstacle in some retrieval scenarios with high requirements for the interpretability of results. Deep learning models are highly dependent on data, and the quality and quantity of data directly affect the performance of the models. In terms of data quality, if there is noise, annotation errors, or biases in the training data, the model may learn incorrect patterns and features, thus affecting the accuracy of retrieval. In image retrieval, if there are errors in the image annotations of the training data, the model may establish a connection between incorrect features and image categories, resulting in incorrect matching results during retrieval. In text retrieval, if there are a large number of misspelled words, grammatical errors, or semantically ambiguous texts in the training data, the model may not be able to accurately learn the semantic features of the text, thus reducing the retrieval performance. In terms of data quantity, deep learning models usually require a large amount of training data to learn sufficient features and patterns. For some specific fields or niche scenarios, it may be difficult to obtain a sufficient amount of high - quality data, resulting in an impact on the generalization ability and accuracy of the models. In the medical information retrieval of some rare diseases, due to the limited number of cases, it is difficult to collect enough training data, making it difficult for deep - learning - based retrieval models to accurately retrieve relevant medical information.

4. Optimization Paths for Retrieval Algorithms Based on Deep Learning Models

4.1 Data Preprocessing and Augmentation

4.1.1 Data leaning and normalization

Data cleaning and normalization are crucial steps in the data preprocessing stage, which are of great significance for improving the performance of retrieval algorithms based on deep learning models. Data cleaning aims to remove noise, errors, and outliers from the data to enhance the quality and reliability of the data. In image data, noise can manifest as salt - and - pepper noise, Gaussian noise, etc. These noises can interfere with the feature extraction and recognition of images. Methods such as median filtering and Gaussian filtering can effectively remove the noise in images. Median filtering replaces the value of a pixel with the median value of the pixels in its neighborhood, which can remove salt - and - pepper noise while retaining the edge information of the image. Gaussian filtering, on the other hand, uses the Gaussian function to perform a weighted average on the image, which can smooth the image and remove Gaussian noise. In text data, noise may include misspelled words, grammatical errors, garbled characters, etc. Spelling - checking tools and grammar analyzers can be used to correct and clean the text. Spelling - checking tools based on statistical language models can automatically correct misspelled words in the text according to the occurrence probability of words and context information. Data normalization is to transform the data into a unified scale and distribution, eliminating the dimensional differences between data features, improving the training effect and stability of the model, and having a certain degree of robustness to outliers. For datasets with many numerical features, standardization can be used to make different features comparable on the same scale, improving the training effect of the model.

4.1.2 Data augmentation techniques

Data augmentation techniques are important means to expand the dataset and enhance the generalization ability of the model. In the field of images, operations such as rotation, scaling, and cropping are commonly used data augmentation methods. By rotating an image, different shooting angles can be simulated, increasing the diversity of the image. Randomly rotating an image by a certain angle (such as ±15°) enables the model to learn the features of objects at different angles, improving the model's adaptability to angle changes. The scaling operation can change the size of the image, allowing the model to learn the features of objects at different scales. Scaling the image by a certain proportion (such as 0.8 - 1.2 times) can enhance the model's robustness to changes in the size of objects. The cropping operation can extract different local regions from the image, enriching the content information of the image. Using random cropping to crop regions of different sizes and positions from the image can enable the model to learn the local features of objects, improving the model's recognition ability.

4.2 Model Optimization Strategies

4.2.1 Model structure improvement

Improving the model structure is one of the key paths to enhance the performance of retrieval algorithms based on deep learning models. In the design of neural network architectures, researchers are constantly exploring and innovating to build more efficient models. The MobileNet series is a lightweight convolutional neural network architecture designed for resource - constrained scenarios. MobileNet uses depth - wise separable convolution. Depth - wise convolution is responsible for performing independent convolution operations on each channel to extract local features, while point - wise convolution is used to integrate the features of different channels. This structural design greatly reduces the number of model parameters and the amount of computation. In image retrieval tasks, compared with traditional convolutional neural networks, MobileNet can significantly improve the retrieval speed and reduce the consumption of computing resources while maintaining a certain retrieval accuracy, making it more suitable for running in resource - limited environments such as mobile devices or embedded systems.

ShuffleNet is also an innovative lightweight neural network architecture. It effectively solves the problem of poor information flow between channels caused by grouped convolution by introducing the channel shuffle operation. In ShuffleNet, the input feature map is first subjected to grouped convolution, and then through the channel shuffle operation, the channels of different groups are rearranged and combined, enabling full information exchange between channels of different groups. This structural design not only reduces the amount of computation but also improves the feature extraction ability of the model. In practical applications, ShuffleNet shows good performance in tasks such as image retrieval and object detection, and can achieve efficient image information retrieval and analysis under low - computing - resource conditions.

The introduction of the attention mechanism has brought more powerful information - processing capabilities to deep learning models. In image retrieval, the attention mechanism allows the model to focus more on the key regions in the image, thereby extracting more representative features. The SENet (Squeeze - and - Excitation Network) models the channel dimension of the feature map by introducing the squeeze - and - excitation module, automatically learning the importance weights of each channel. When processing an image, SENet can weight the features of different channels according to the image content, highlighting key information and suppressing irrelevant information, thus improving the accuracy and effectiveness of image feature extraction. When retrieving an image containing multiple objects, SENet can, through the attention mechanism, allocate more attention to the channels where the objects related to the query are located, extract more accurate features, and thus improve the retrieval accuracy.

In text retrieval, the self - attention mechanism in the Transformer model enables the model to simultaneously focus on different positions in the input text, better capturing the semantic relationships and context information in the text. BERT (Bidirectional Encoder Representations from Transformers), during the pre - training process, uses the self - attention mechanism to learn from a large - scale corpus, enabling it to understand the complex semantic relationships between words in the text, such as grammatical structures and semantic dependencies. When applied to text retrieval, BERT can accurately understand the query intention according to the user's query and find relevant information from a large amount of text data. When processing a news report on technological development, BERT can, through the self - attention mechanism, capture key information such as technological achievements, application fields, and development trends in the text. When the user queries relevant technological information, it can quickly and accurately retrieve the news report.

4.2.2 Hyperparameter tuning

Hyperparameter tuning is an important means to improve the performance of deep learning models. By reasonably adjusting hyperparameters, the model can achieve optimal performance in retrieval tasks. Grid search is a simple and intuitive hyperparameter tuning method. It exhaustively lists the specified parameter combinations, calculates the performance of each set of parameters on the validation set, and finally selects the parameter combination with the best performance. When training an image retrieval model based on a convolutional neural network, hyperparameters such as the learning rate, batch size, and number of convolutional kernels need to be tuned. Suppose the value range of the learning rate is [0.001, 0.01, 0.1], the value range of the batch size is [16, 32, 64], and the value range of the number of convolutional kernels is [32, 64, 128]. Then grid search will train and evaluate all possible combinations of these three hyperparameters, that is, a total of 27 experiments will be carried out. By comparing the accuracy, recall rate, and other indicators of these 27 experiments on the validation set, the hyperparameter combination with the best performance is selected, and this set of hyperparameters is determined as the optimal hyperparameter combination.

Random search, on the other hand, randomly selects parameter combinations in the parameter space for training and evaluation. Different from grid search, it does not test all values between the upper and lower bounds but finds the optimal solution through random sampling. When dealing with high - dimensional parameter spaces, random search can find parameter combinations close to the optimal solution in a relatively short time. When training a text retrieval model based on a recurrent neural network, hyperparameters such as the number of hidden - layer neurons and the Dropout probability are randomly searched. Set the value range of the number of hidden - layer neurons to [100, 500] and the value range of the Dropout probability to [0.2, 0.5]. Random sampling is carried out within this range to generate multiple sets of hyperparameter combinations for training. After a certain number of random searches, the hyperparameter combination with the best performance is selected according to the performance of the model on the validation set. The advantage of random search is that it has high computational efficiency and can avoid falling into local optimal solutions to a certain extent. However, its results have a certain degree of randomness, and it cannot guarantee to find the global optimal solution.

In addition to grid search and random search, there are some more advanced hyperparameter tuning methods, such as Bayesian optimization. Bayesian optimization learns the shape of the objective function and uses the results of previous experiments to guide the next parameter selection, thus more efficiently searching for the optimal solution. It is suitable for optimization problems with high - dimensional, high - cost, and limited samples and has good application prospects in the hyperparameter tuning of deep learning models. Population - based optimization methods, such as genetic algorithms, optimize hyperparameters by simulating the genetic, mutation, and selection mechanisms in the biological evolution process. Genetic algorithms regard hyperparameter combinations as biological individuals, generate new individuals through crossover and mutation operations, and then select according to the performance of individuals on the validation set, retaining individuals with better performance and eliminating individuals with poor performance. After multiple generations of evolution, the optimal hyperparameter combination is gradually found.

4.3 Retrieval Strategy Optimization

4.3.1 Multimodal fusion retrieval

Multimodal fusion retrieval aims to integrate data from multiple modalities to provide more comprehensive and accurate retrieval results. In image - text retrieval, early fusion and late fusion are two commonly used multimodal fusion methods. Early fusion fuses data from different modalities at the feature extraction stage to generate a comprehensive feature representation. In image - text joint retrieval, the visual features of the image and the semantic features of the text can be directly concatenated into a new feature vector after feature extraction. Specifically, for an image of an apple, a convolutional neural network is used to extract its visual features, such as color, shape, texture, and other feature vectors. For the text "red apple" describing the apple, natural language processing technology is used to extract its semantic feature vector. Then these two feature vectors are concatenated to obtain a comprehensive feature vector containing image and text information. During retrieval, the similarity between the comprehensive feature vector of the query and all the comprehensive feature vectors in the database is calculated to find the relevant images and texts. The advantage of this method is that it can make full use of the complementary information between multimodal data and improve the accuracy of retrieval. However, due to the early fusion, some modality - specific detailed information may be lost, affecting the retrieval effect.

Late fusion, on the other hand, processes and retrieves data from each modality separately and then combines the results at the decision - making stage. For image and text data, the image retrieval algorithm based on a convolutional neural network and the text retrieval algorithm based on a recurrent neural network are used for retrieval respectively to obtain their respective retrieval results. Then, the two retrieval results are fused through methods such as voting mechanisms and weighted combinations. In the voting mechanism, for each candidate retrieval result, votes are given according to the results of image retrieval and text retrieval respectively, and the result with the most votes is taken as the final retrieval result. In the weighted combination, different weights are assigned to the retrieval results of image retrieval and text retrieval according to their reliability, and then the weighted results are combined. The advantage of late fusion is that it can give full play to the advantages of each modality and can be flexibly adjusted according to the reliability of different modalities during the fusion process. However, due to the late fusion, the information fusion may be insufficient, affecting the overall performance of retrieval.

4.3.2 Semantic understanding and deep matching

Deep learning models have unique advantages in achieving semantic understanding and deep matching. Taking the pre - trained model based on Transformer as an example, it can capture the semantic relationships and context information in the text through the self - attention mechanism, thus achieving more accurate semantic understanding. In text retrieval, when a user enters a query statement, the Transformer - based model can deeply analyze the query statement, understand its grammatical structure, semantic dependencies, and implicit semantic information. For the query statement "What are the latest products released by Apple?", the model can, through the self - attention mechanism, focus on the semantic relationships between keywords such as "Apple", "latest release", and "products", and accurately grasp the user's query intention. Then, when processing text data, the model can analyze the associations between each word in the text and other words, learning the semantic associations and context dependencies between words. When processing a news report on Apple's product launch, the model can understand the relationships between the information such as the product description, performance characteristics, and release time in the report, so that it can accurately match this news report with the user's query during retrieval.

In terms of deep matching, deep learning models can achieve deep semantic matching between queries and documents by constructing complex neural network structures. In deep - learning - based image retrieval, the model can not only match the surface features of images but also deeply understand the semantic content of images, achieving more accurate image retrieval. For an image containing a cat, the model can accurately determine whether there is a cat in the image through the learned feature pattern of the cat and match it with other images containing cats. In text retrieval, deep learning models can extract and match the features of queries and documents through multi - layer neural networks, exploring the deep - level semantic associations between them. By mapping the query and the document into the same semantic space and calculating their distances or similarities in the semantic space, deep matching can be achieved. When processing the retrieval of a scientific and technological paper, the model can deeply match the user's query "The application of artificial intelligence in the medical field" with the relevant content in the paper. It can not only match the paragraphs that directly mention "artificial intelligence" and "medical field" but also find the content that is semantically related to the query although it does not directly mention the keywords, such as the research on the application of artificial intelligence technology in disease diagnosis and medical image analysis, thus improving the accuracy and comprehensiveness of retrieval.

5. Conclusions and Prospects

5.1 Summary of Research Results

This research deeply explored the optimization paths for retrieval algorithms based on deep learning models and achieved a series of results with important theoretical and practical significance. At the level of algorithm analysis, it comprehensively analyzed the image retrieval algorithm based on the convolutional neural network (CNN), the text retrieval algorithm based on the recurrent neural network (RNN), and the multimodal retrieval algorithm based on the Transformer. Through in - depth research on these algorithms, their working mechanisms and advantages in feature extraction, information fusion, and semantic understanding were clarified. In image retrieval, the CNN can automatically learn the local and global features of images and achieve efficient image similarity matching. In text retrieval, the RNN and its variants can effectively process the sequential information of text and understand the semantic relationships in the text. The multimodal retrieval algorithm based on the Transformer can integrate information from multiple modalities and achieve more accurate cross - media retrieval.

In response to the multi - level challenges faced by the optimization of retrieval algorithms, such as those related to data, algorithms, and application scenarios, a series of targeted optimization paths were proposed. In terms of data preprocessing and augmentation, through data cleaning, normalization, and data augmentation techniques, the quality and diversity of data were effectively improved, providing better - quality data support for subsequent model training. In image data cleaning, methods such as median filtering and Gaussian filtering were used to remove noise, making image feature extraction more accurate. In text data cleaning, spelling - checking tools and grammar analyzers were used to correct misspelled words and grammatical errors, improving the semantic understanding ability of the text. In model optimization strategies, by improving the model structure, such as adopting lightweight architectures like MobileNet and ShuffleNet, and introducing the attention mechanism, the performance and efficiency of the model were significantly enhanced. In image retrieval, the depth - wise separable convolution structure of MobileNet reduced the number of model parameters and the amount of computation, improving the retrieval speed. In text retrieval, the self - attention mechanism of the Transformer model enhanced the ability to capture semantic relationships in the text, improving the retrieval accuracy. Through hyperparameter tuning, methods such as grid search and random search were used to find the optimal hyperparameter combination of the model, further improving the model's performance.

In terms of retrieval strategy optimization, the proposed multimodal fusion retrieval and semantic understanding and deep - matching methods effectively improved the comprehensiveness and accuracy of retrieval. In multimodal fusion retrieval, the early - fusion and late - fusion methods made full use of the complementary information between different - modality data, improving the quality of retrieval results. In image - text joint retrieval, early fusion could make full use of the complementary information of multimodal data by directly concatenating the features of images and texts at the extraction stage, improving the retrieval accuracy. Late fusion, after separately retrieving images and texts, fused the results through methods such as voting mechanisms or weighted combinations, giving full play to the advantages of each modality. In semantic understanding and deep - matching, the pre - trained model based on the Transformer, through the self - attention mechanism, could deeply understand the semantic relationships and context information of the text, achieving more accurate semantic matching and retrieval.

5.2 Prospects for Future Research Directions

In the future, the research on retrieval algorithms based on deep learning models has broad exploration space in several key directions. In terms of model interpretability, researchers will be committed to developing visualization tools to intuitively display the decision - making basis of deep learning models during the retrieval process. By visualizing the feature - extraction process, it is possible to clearly show how the model extracts key features from image or text data, helping users understand how the model understands the data. In image retrieval, the visualization tool can display the image features extracted by the convolutional neural network at different convolutional layers, such as the visual expressions of edge and texture features, allowing users to intuitively see how the model recognizes the image content. In text retrieval, the visualization tool can display the working process of the self - attention mechanism in the Transformer model, graphically presenting the degree of attention the model pays to the semantic associations between different words in the text, helping users understand how the model understands the text semantics. In addition to visualization tools, interpretive algorithms can also be developed to explain the decision - making process of the model from the perspectives of mathematical principles and logical reasoning. By analyzing the parameters and calculation process of the model, the key factors and reasoning paths for the model's decision - making can be found, providing users with a more in - depth and accurate explanation.

The expansion of cross - domain applications is also one of the important future research directions. As the applications of deep - learning - based retrieval algorithms in fields such as images and texts gradually mature, expanding them to more fields, such as medicine, finance, and education, will bring new development opportunities to these fields. The in - depth study of multimodal fusion technology will further improve the performance of retrieval algorithms. Future research will pay more attention to the deep fusion of different - modality data and explore more effective fusion strategies and model architectures. In cross - media retrieval, combining video, audio, and text information can achieve more comprehensive and accurate information retrieval. By deeply fusing the image features, audio features, and text descriptions in the video, deep learning models can better understand the video content and achieve accurate retrieval based on the video content. In the field of smart homes, fusing voice, image, and sensor data can achieve more intelligent home control and information retrieval. In the future, the research on retrieval algorithms based on deep learning models will continue to innovate and break through in multiple directions, bringing more development opportunities and application prospects to the field of information retrieval.

Acknowledgments

The Research and Practice on the Reform of Network Information Security Talent Training through the Integration of Industry and Education and Promoting Learning through Competitions in 2024 Higher Education Undergraduate Teaching Reform Project of Yulin Normal University (No. 2024xjjg03); The Design and Application Research of Student Reward Information Management System Based on Retrieval Algorithms in the 2022 College Students' Innovation and Entrepreneurship Project of Yulin Normal University; The Research on Small Target Detection Methods Based on Deep Learning in the 2022 College Students' Innovation and Entrepreneurship Project of Yulin Normal University.

References

[1] LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. Nature 521, 436–444 (2015). https://doi.org/10.1038/nature14539

[2] Suresh K K, Sundaresan S, Nishanth R, et al. Optimization and Deep Learning–Based Content Retrieval, Indexing, and Metric Learning Approach for Medical Images. Computational Analysis and Deep Learning for Medical

Care: Principles, Methods, and Applications, 2021: 79-106.

[3] Chua L O. CNN: A vision of complexity. International Journal of Bifurcation and Chaos, 1997, 7(10): 2219-2425.

[4] Koutnik J, Greff K, Gomez F, et al. A clockwork rnn//International conference on machine learning. PMLR, 2014: 1863-1871.

[5] Popel M, Bojar O. Training tips for the transformer model. arXiv preprint arXiv:1804.00247, 2018.

[6] Erk K. Vector space models of word meaning and phrase meaning: A survey. Language and Linguistics Compass, 2012, 6(10): 635-653.

[7] Tai, Xiaoying, Fuji Ren and Kenji Kita. "An information retrieval model based on vector space method by supervised learning." Inf. Process. Manag. 38 (2002): 749-764.

[8] Ramos J. Using tf-idf to determine word relevance in document queries// Proceedings of the first instructional conference on machine learning. 2003, 242(1): 29-48.

[9] Han K, Wang Y, Chen H, et al. A survey on vision transformer. IEEE transactions on pattern analysis and machine intelligence, 2022, 45(1): 87-110.

[10]Janiesch C, Zschech P, Heinrich K. Machine learning and deep learning. Electronic Markets, 2021, 31(3): 685-695.

[11]Menghani G. Efficient deep learning: A survey on making deep learning models smaller, faster, and better. ACM Computing Surveys, 2023, 55(12): 1-37.