Scholar Publishing Group
International Journal of Multimedia Computing
https://doi.org/10.38007/IJMC.2026.070101
ISSN 2789-7168 Vol. 7, Issue 1:1-8

Multimodal Learning Method for Cross-Modal Data Alignment and Retrieval

Bukun Ren

College of Engineering, University of California Berkeley, Berkeley, 94720, USA

Keywords: Cross-modal Alignment; Data Retrieval; Multi-modal Learning; Embedding Space; Semantic Alignment

Abstract: With the rapid development of information technology, cross-modal data alignment and retrieval is one of the hotspots in multimodal learning. The aim is to reduce the representation differences among different modalities and enable different representations to be applicable to data of the same modality. Cross-modal data retrieval requires based on this alignment technology to retrieve data of other modalities corresponding to a certain modality from a certain modality. This paper briefly introduces the problems and challenges of cross-modal data alignment and retrieval, and mentions the application of multimodal learning methods. It proposes many new schemes for cross-modal data alignment and retrieval, such as constructing a unified embedding space, integrating semantic perception mechanisms, information enhancement and redundancy suppression, and adopting self-supervised and transfer mechanisms and other technologies. Through case studies, the effectiveness and feasibility of these schemes are demonstrated, and the multimodal learning methods are comprehensively summarized, and the development direction is prospected.

1. Introduction

The cross-modal data alignment and retrieval technology has made remarkable progress over the past few decades, mainly in the aspects of image-text alignment and retrieval. With the development of big data and deep learning, cross-modal learning has become an important topic in the field of artificial intelligence and has been widely applied in natural language understanding, computer vision, and recommendation systems, etc. However, problems such as feature distribution differences between modalities, semantic inconsistency, and information loss still limit its accuracy and effectiveness. Therefore, more effective and precise multimodal learning methods are needed to address the core issues in cross-modal data alignment and retrieval, such as modality differences, information redundancy, and data dependency.

The main contribution of this paper is to propose a cross-modal data alignment and retrieval method based on multimodal learning, exploring how to solve the core problems in cross-modal data alignment and retrieval through techniques such as unified embedding space, semantic perception mechanism, information enhancement and redundancy suppression.

2. Overview of Cross-Modal Data Alignment and Retrieval

The deep learning techniques employed in this research mainly achieve the alignment and retrieval of cross-modal data to discover spatial representations with similarity (spatial consistency) in different modalities (text, image, video, etc.) data, thereby realizing semantic matching and retrieval. Thus, after eliminating the description differences of different modalities, information search in another mode (such as image) can be conducted based on a certain mode (such as text) through a search engine, which is widely applied in application fields such as image-text search and video recognition. The alignment and retrieval of cross-modal data not only enhance the accuracy of information retrieval but also provide theoretical support for the fusion and application of multimodal data [1]. However, there are still numerous issues to be addressed regarding the alignment and retrieval of cross-modal data. Due to the distinct feature representation methods of different modalities, such as the completely different description styles between images and text, matching them in the same semantic space becomes more challenging. Additionally, from a semantic perspective, the consistency of such matching is not absolute because the ways in which similar meanings are expressed in various modalities are not uniform, posing a challenge to the difficulty of matching.

To address the aforementioned issues, a series of new technical approaches have been proposed in recent years. For instance, through the cross-modal embedding method based on deep learning, namely the method of unified embedding dimension, cross-modal information alignment can be achieved; and by means of contrastive learning, semantic perception mechanism and other methods, the alignment accuracy and retrieval effect of cross-modal data can be improved. Besides, the introduction of self-supervised learning and transfer learning enables cross-modal retrieval models to have better generalization ability, no longer relying on massive labeled data, and enhances the universality of the models [2]. The performance comparison of different cross-modal data alignment and retrieval methods is shown in Figure 1.

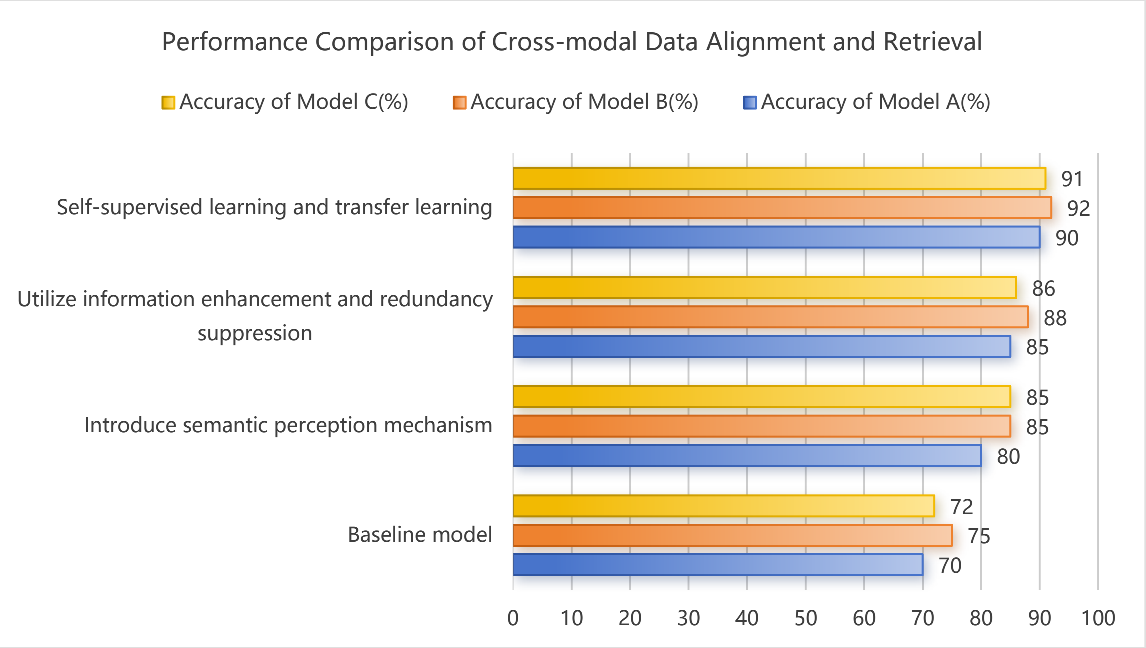

Figure 1. Bar Chart Comparing the Performance of Different Cross-modal Data Alignment and Retrieval Methods

The four groups of bar charts in Figure 1 respectively represent the accuracy rate changes of the baseline model, the model with semantic perception mechanism introduced, the model using information enhancement and redundancy suppression, and the self-supervised and transfer learning methods in the image-text retrieval task.

As can be seen from Figure 1, for Model A, the baseline model has the lowest accuracy rate, approximately 70%. After introducing the semantic perception mechanism, the accuracy rate has increased to 80%, showing an improvement in performance. After applying information enhancement and redundancy suppression, the accuracy rate further rose to 85%. By applying self-supervised and transfer learning strategies, the accuracy rate reached 90%, demonstrating the significant effect of this method in improving the retrieval accuracy. All the above-mentioned technologies are as shown in Figure 1, and they have achieved considerable improvements in the process of image-text retrieval. This indicates that with the development of the times, the accuracy of cross-modal data alignment and retrieval will gradually increase, especially with the adoption of self-supervised and transfer learning strategies, the generality of the model and the accuracy of cross-modal data alignment can be improved.

3. The main issues of cross-modal data alignment and retrieval

3.1 The differences in the distribution of modal features lead to difficulties in alignment

One of the core issues in cross-modal data alignment is the difference in feature distribution between different modalities. For instance, the information conveyed by two different modalities such as images and text has significant differences in their expression forms. The former relies on hundreds or thousands of high-level pixels, while the latter is composed of discrete letters or other symbols. Therefore, it is quite challenging to convert the corresponding features of two different types of modalities into a shared semantic space while preserving the integrity of the original information. This is not only a problem of low-level feature representation but also involves how to handle high-dimensional, complex, and sparse data.

3.2 An alignment mechanism lacking consistency at the semantic level

The cross-modal data alignment still has the problem of semantic inconsistency among modalities. Although the data of various modalities can be aligned through certain attributes, there is no effective mechanism to ensure the semantic consistency of these features [3]. The synonyms in the text and the different manifestations of the same object in the image will bring semantic inconsistency to the alignment effect.

3.3 Information asymmetry restricts the improvement of retrieval performance

For cross-modal data alignment, there exists information asymmetry, mainly manifested in the aspects of information completeness and accuracy. The information in the language domain cannot precisely represent every detail of the pictures, while the information in the pictures cannot precisely and completely represent the language information. Due to the existence of information differences, the effect of cross-modal retrieval is not perfect. In a large amount of complex data information environment, certain information between different modalities may be missing, which will not be conducive to the improvement of the retrieval effect.

3.4 The model has weak generalization ability and thus relies on a large amount of labeled data

Although existing multimodal learning models can achieve satisfactory results in specific tasks, most of these models have significant demands for large annotated datasets. The main reason is that deep learning models usually rely on a large number of labeled samples to learn the relationships between different modalities during training. However, it is obvious that labeled information is costly and the generalization ability of the models in multiple real-world scenarios is relatively low, making it difficult to handle new problems or sample transformation issues.

4.Multimodal learning methods for cross-modal data alignment and retrieval

4.1 Construct a unified embedding space to narrow the modality gap

The key to cross-modal data alignment is to map data from different modalities into a unified embedding space to reduce the gap between modalities. In this space, such patterns can be coordinated using a unified representation form, so that data with homogeneous but different time periods can be more similar. To achieve this goal, generally, multi-layer perceptrons (MLPs) or convolutional neural networks (CNNs) are used to extract features for each modality, and then contrastive or maximum similarity methods are employed to ensure that all types of modality data exist in one space. The optimization objective can be expressed by the following formula:

$$L = \sum_{i,j}(\|f(x_i) - f(x_j)\|^2 - \alpha)^2| \tag{1}$$

Among them, the feature extraction function for each modality data is defined $f(x)$, and α is a certain regularization term, aiming to ensure that the modality data are aligned in the embedding space.

In order to construct a unified embedding space, feature extraction is carried out for each modality's data first. For the image modality, a convolutional neural network (CNN) is selected to extract visual features; for the text modality, a recurrent neural network (RNN) or Transformer is chosen to extract semantic features. After the features are extracted, through contrastive learning, that is, by enhancing similarity and reducing dissimilarity, data points with the same semantics but belonging to different modalities are brought closer together in the fusion space. In addition, a regularization term α is added to avoid overfitting and improve the registration accuracy.

4.2 Enhance the alignment effect by integrating semantic perception mechanism

This semantic perception construction method utilizes the fusion of language information from different modalities during the alignment process to enhance the information matching degree across modalities. In the learning stage, in addition to learning low-level features, it is also necessary to consider the high-level semantic relationships of each modality to ensure that the established model not only performs feature matching but also can understand the similarities of different semantics within each modality. The common method is to simulate the semantic correlations between different modes by using bidirectional recurrent neural networks (Bi-RNN) or Transformer networks, etc. The alignment loss can be further enhanced the semantic consistency through the following formula:

$$L_{semantic} = \sum_{i,j}(\|g(f(x_i)) - g(f(x_j))\|^2 - \beta)^2| \tag{2}$$

Among them, the feature extraction function for each modality data is defined $f(x)$, g is the transformation function for introducing semantic perception, and β represents the weight of semantic consistency, aiming to enhance the semantic alignment capability of cross-modal data.

In order to establish a semantic perception mechanism, deep neural networks are employed to extract basic attributes in different modes, such as words in text and pixels in images. Additionally, bidirectional recurrent neural networks (Bi-RNN) or Transformers are utilized to simulate semantic correlations between different modes to capture long-distance relationships and semantic contents, enabling the effective aggregation of the semantics of each mode. During learning, the loss function for semantic consistency is optimized to adjust the distinctive expressions of each modality, making data of different modes with the same semantics closer and improving the matching accuracy across modes.

4.3 Enhancing information and suppressing redundancy to improve matching accuracy

The combination of information enhancement and redundancy suppression is conducive to improving the accuracy of cross-modal matching. Information enhancement involves extracting more underlying feature information for each type of modality to increase the information representation capacity of the modality, thereby enabling it to carry more underlying meanings; redundancy suppression is to reduce the redundancy in different modality data to prevent redundant information from affecting the model [4]. To achieve this, regularization methods such as L2 regularization and attention mechanisms are usually adopted to suppress redundant information. Information enhancement can be realized through the following formula:

$$\mathrm{L_{info}} = \lambda \sum_{\mathrm{i}} \left\| f(\mathrm{x_i}) \right\|^2 + \mu \sum_{\mathrm{i,j}} (\left\| f(\mathrm{x_i}) - f(\mathrm{x_j}) \right\|^2)| \tag{3}$$

Among them, the feature extraction function for each modality data is defined $f(\mathrm{x})$, λ and μ are regularization coefficients, which are used to regulate the degree of information enhancement and redundancy suppression.

During the process of information enhancement and redundancy suppression, first, a deep network is used to extract more latent features to ensure that each mode can fully describe the information content. Then, the feature dimension is increased to make each mode more detailed, thereby enhancing the machine's processing ability for complex information. The L2 normalization method is used to limit the expansion of features and introduce a focus strategy to concentrate on the information with the strongest inter-pattern correlation, reducing the interference of redundant information. Adjusting the regularization coefficients λ and μ can balance the effects of information enhancement and redundancy suppression to achieve the best matching effect.

4.4 Enhance generalization ability by adopting self-supervised and transfer mechanisms

Self-supervised learning techniques and transfer learning techniques can provide better generalization ability and reduce the reliance on a large amount of labeled information. Self-supervised learning generates labels automatically through a pre-trained model without any labeled information, thereby assisting the model in learning valuable information features. Transfer learning enhances the adaptability and generalization ability of the model by transferring knowledge learned in one domain to another. The self-supervised loss function can be expressed as:

$$\mathrm{L_{self}} = \sum_{\mathrm{i}} \left\| f(\mathrm{x_i}) - f(\hat{\mathrm{x}}_{\mathrm{i}}) \right\|^2 | \tag{4}$$

Among them, the feature extraction function for each modality data is defined $f(\mathrm{x})$, $\hat{\mathrm{x}}_{\mathrm{i}}$ is pseudo-labeled data generated through self-supervised tasks. In this way, the model can maintain good generalization ability across different datasets and tasks.

Self-supervised learning and transfer learning can effectively enhance the generalization ability of models and reduce the reliance on a large amount of labeled data. The combination of self-supervised learning and transfer learning enables models to maintain good generalization ability when facing different datasets and tasks, and enhances their effectiveness in diverse applications.

5.Case-based Empirical Research

Take a certain company as an example. It has studied the current situation and optimization schemes of cross-modal data alignment and retrieval. This company is involved in the promotion work of products on e-commerce platforms and the work of search engines. Therefore, it involves a large amount of multimodal content. Regarding this issue, using traditional methods means adopting convolutional neural networks (CNN) to extract information from images and using bag-of-words models to encode text. However, the drawback of this method is that due to imperfect semantic alignment, the query results will be inaccurate, reducing the user experience; at the same time, the applicability of this system is weak. Once encountering user-generated content with unknown tags, the accuracy of this method will drop sharply.

In response to this, corresponding improvements were made based on the aforementioned multimodal learning methods. Firstly, a shared embedding space was constructed to map both image and text data into the same semantic space, thereby reducing the differences between the two modalities. Then, semantic perception was introduced, combined with the Transformer network to enhance the semantic consistency among different modalities, ensuring more precise semantic associations between images and texts. After solving the aforementioned problems, information enhancement and redundancy elimination were achieved. By using regularization and attention mechanisms to reduce redundant features, the matching accuracy was improved. To enhance the generalization ability of the system, self-supervised learning and transfer learning were introduced, reducing the reliance on labeled data and improving the model's adaptability in different scenarios.

Table 1. Comparison Before and After Optimization

Indicators	Before optimization	After optimization
Image-text matching accuracy	72%	87%
User satisfaction	68%	82%
System generalization ability	70%	85%
Search response time	1.5 seconds	1.2seconds

By applying multimodal learning methods, the performance in image-text matching accuracy and user satisfaction rate has been improved. The improved model can better understand the semantic relationship between images and texts at a deeper level and find more relevant information. Moreover, it can effectively expand the generalization ability of the entire system and handle the content generated by unmarked users effectively, thereby enhancing the intelligent recommendation capability of the platform [5]. Through these methods, the overall performance of Company X in cross-modal data alignment and retrieval tasks has been significantly improved, bringing higher efficiency and user experience.

Conclusion: This paper explores the multimodal learning methods for cross-modal data alignment and retrieval, proposes a series of technical solutions, and addresses the main issues in cross-modal alignment and retrieval. By constructing a unified embedding space, introducing semantic perception mechanisms, enhancing information and suppressing redundancy, as well as adopting self-supervised and transfer learning strategies, the accuracy of data alignment and the performance of retrieval are improved. After being tested in various practical application scenarios, these new technical methods have been proven to be effective and highly applicable for tasks such as image-to-text search, emotion judgment, speech-to-text search, and other tasks.

References:

[1] Li Z, Xie Y. BCRA: bidirectional cross-modal implicit relation reasoning and aligning for text-to-image person retrieval. Multimedia Systems, 2024, 30(4).

[2] Ma H, Fan B, Ng B K, et al. VL-Few: Vision Language Alignment for Multimodal Few-Shot Meta Learning. Applied Sciences-Basel, 2024, 14(3):19.

[3] Fang J, Yan X. MDSEA: Knowledge Graph Entity Alignment Based on Multimodal Data Supervision. Applied Sciences (2076-3417), 2024, 14(9).

[4] Ma H, Fan B, Ng B K, et al. VL-Few: Vision Language Alignment for Multimodal Few-Shot Meta Learning. Applied Sciences (2076-3417), 2024, 14(3).

[5] Cui J, He Z, Huang Q, et al. Structure-aware contrastive hashing for unsupervised cross-modal retrieval. Neural Networks, 2024, 174(000):10.

[6] Xiu, L. (2025, June). Research on Personalized Recommendation Algorithms in Modern Distance Education Systems. In 2025 IEEE 3rd International Conference on Image Processing and Computer Applications (ICIPCA) (pp. 2019-2024). IEEE.

[7] Xu, D. (2025). Integration and Optimization Strategy of Spatial Video Technology in Virtual Reality Platform. International Journal of Engineering Advances, 2(3), 131-137.

[8] Huang, J. (2025). Adaptive Reuse of Urban Public Space and Optimization of Urban Living Environment. International Journal of Engineering Advances, 2(4), 9-17.

[9] Zhou, Y. (2025). Using Big Data Analysis to Optimize the Financing Structure and Capital Allocation of Energy Enterprises. Economics and Management Innovation, 2(7), 8-15.

[10] Zhang, Q. (2025). Use Computer Vision and Natural Language Processing to Optimize Advertising and User Behavior Analysis. Artificial Intelligence and Digital Technology, 2(1), 148-155.

[11] Wang, Y. (2025). Research on Early Identification and Intervention Techniques for Neuromuscular Function Degeneration. Artificial Intelligence and Digital Technology, 2(1), 163-170.

[12] Wu, H. (2025). The Challenges and Opportunities of Leading an AI ML Team in a Startup. European Journal of AI, Computing & Informatics, 1(4), 66-73.

[13] Shen, D. (2025). Innovative Application of AI in Medical Decision Support System and Implementation of Precision Medicine. European Journal of AI, Computing & Informatics, 1(4), 59-65.

[14] Liu, X. (2025). Use Generative Al and Natural Language Processing to Improve User Interaction Design. European Journal of AI, Computing & Informatics, 1(4), 74-80.

[15] Liu, F. (2025). Localization Market Expansion Strategies and Practices for Global E-commerce Platforms. Strategic Management Insights, 2(1), 146-154.

[16] Hu, Q. (2025). Research on the Combination of Intelligent Management of Tax Data and Anti-Fraud Technology. Strategic Management Insights, 2(1), 139-145.

[17] Hua, X. (2025). Key Indicators and Data-Driven Analysis Methods for Game Performance Optimization. European Journal of Engineering and Technologies, 1(2), 57-64.

[18] Hui, X. (2025). Research on the Application of Integrating Medical Data Intelligence and Machine Learning Algorithms in Cancer Diagnosis. International Journal of Engineering Advances, 2(3), 101-108.

[19] Hui, X. (2025). Utilize the Database Architecture to Enhance the Performance and Efficiency of Large-Scale Medical Data Processing. Artificial Intelligence and Digital Technology, 2(1), 156-162.

[20] Jingzhi Yin. Research on Financial Time Series Prediction Model Based on Multi Attention Mechanism and Emotional Feature Fusion. Socio-Economic Statistics Research (2025), Vol. 6, Issue 2: 161-169

[21] Dingyuan Liu. Measuring the Sensitivity of Local Skill Structures to AI Substitution Risks Based on Occupational Task Decomposition. Socio-Economic Statistics Research (2025), Vol. 6, Issue 2: 177-184

[22] Yiting Hong. An Efficient Federated Graph Neural Network Framework for Cross-Enterprise Business Analysis. Socio-Economic Statistics Research (2025), Vol. 6, Issue 2: 170-176.

[23] Jiahe Sun. Research on Financial Systemic Risk Measurement Based on Investor Sentiment and Network Text Mining. Socio-Economic Statistics Research (2025), Vol. 6, Issue 2: 185-193.

[24] Thanh-Huyen Truong. Research on the Mechanism of E-commerce Model Innovation Driven by Digital Technology. International Journal of Big Data Intelligent Technology (2025), Vol. 6, Issue 2: 171-178

[25] Qi, Y. (2025). Data Consistency and Performance Scalability Design in High-Concurrency Payment Systems. European Journal of AI, Computing & Informatics, 1(3), 39-46.

[26] Fu, Y. (2025). The Push of Financial Technology Innovation on Derivatives Trading Strategy Optimization. European Journal of Business, Economics & Management, 1(4), 114-121.

[27] Li, J. (2025). High-Performance Cloud-Based System Design and Performance Optimization Based on Microservice Architecture. European Journal of AI, Computing & Informatics, 1(3), 77-84.

[28] Chuying Lu. Object Detection and Image Segmentation Algorithm Optimization in High-Resolution Remote Sensing Images. International Journal of Multimedia Computing (2025), Vol. 6, Issue 1: 144-151.

[29] Xia Hua. User Stickiness and Monetization Strategies in the Release of Global Game Projects. International Journal of Business Management and Economics and Trade (2025), Vol. 6, Issue 1: 188-195.

[30] Junchun Ding. Cross-Functional Team Collaboration and Project Management in the Automotive Industry. International Journal of Social Sciences and Economic Management (2025), Vol. 6, Issue 2: 162-170.