Research on Modelling Capability of English Multimodal File Search based on Transformer

Hongjuan Li
College of Continuing Education, Pingdingshan Polytechnic College, China
hongjuanlih@outlook.com

Abstract: *With the exponential growth of file data in the multimedia era, file retrieval ability to achieve effective data management has become a hot research field. Based on people's English file search needs, this paper proposes an English multimodal file search model based on transformer. Through ablation experiments on two public data sets and comparison experiments with the benchmark model, the effectiveness and superiority of the proposed transformers algorithm model in multi-modal data processing are verified. The multi-modal fusion retrieval system can usually achieve better performance than the single-modal retrieval system. This experiment focuses on three modes: Audio, Image and Text. The experimental results show that the proposed method can not only improve the efficiency of file search, but also extract modal features and perform feature fusion better. In the future, we can further explore different types of other attention mechanisms or integrate a variety of different architectures to further enhance the feasibility and superiority of multimodal file search.*

Keywords: *Transformer, attention mechanism, multimodal, English document retrieval.*

Received July 16, 2024; accepted November 14, 2024
https://doi.org/10.34028/iajit/22/1/9

1. Introduction

As the capacities of computer storage have grown alongside the escalating demands for storage, individuals have been compelled to invest considerable effort in the retrieval of files, predominantly within the context of English-based file systems. Consequently, the capability for effective data management through efficient file retrieval has emerged as a burgeoning field of study across various sectors [3, 16]. Traditional retrieval methods fall short of accommodating the needs of users in the multimedia domain, who aspire to transcend the boundaries of disparate data types and achieve semantic content retrieval across different media, such as retrieving text through images, or video through text [10, 13]. Multimodal file retrieval, which involves the retrieval of information by processing and integrating various types of data or information, including Text, Images, Audio, and video, aims to provide a more comprehensive understanding and handling of information.

Currently, research into multimodal data retrieval primarily focuses on network multimedia documents [8], with multimodal fusion serving as a pivotal technology. This technology analyzes the relevance and complementarity between different information modalities, evaluates the significance of each modality in a query, and explores the optimal strategy for modal fusion [2, 17]. Through this strategy, a deeper understanding of the document's advanced semantics is achieved, resulting in more effective retrieval outcomes. Hyperlink Induced Topic Search (HITS) [7] and page rank [5] use the Hyperlink relationship inside the web page to establish and determine the context relationship between the web pages. Shah *et al*. [14] introduced context-enhanced search, utilizing contextual information to reorder and expand content search results. They emphasized the use of strict causality to guide the search, enabling more accurate identification of data flows between files and reducing false positives due to context switching and background noise. Wei *et al.* [18] mined semantic features from image data using convolutional networks. Zhang *et al*. [21] utilized a coupled deep fully connected network to map the feature representations of different modalities into a common subspace, enabling mixed cross-modal similarity learning. Yu *et al*. [19] extracted image and text features through bottom-up attention and Recurrent Neural Networks (RNNs), embedding text features into the image feature space, effectively reducing the heterogeneity between image and text. Research in this field continues to expand. Zhang *et al*. [22] selected multiple feature information from images to construct a graph structure and mined the relationships between feature data through graph convolution. Bianchi *et al*. [1] proposed a graph convolutional layer with an autoregressive moving average filter, exhibiting greater robustness to noise. However, the majority of the aforementioned methods are confined to the internal feature mining of individual sample data, with insufficient measurement of the consistency between the features of different sample data.

The Transformer model, which utilizes an encoder-decoder architecture grounded in attention mechanisms, has found widespread application in the realms of

Natural Language Processing (NLP) and computer vision [4, 20, 23]. In contrast to the traditional RNN and Long Short-Term Memory networks (LSTMs) models, the Transformer entirely dispenses with the autoregressive computation approach during training, thereby enabling efficient parallel training [15]. By introducing the attention mechanism to address long-term dependencies, the Transformer employs a set of weights to describe the dependencies between positions, thereby enabling it to capture global features more effectively. In translation tasks, the Transformer has demonstrated outstanding performance, showcasing faster training speeds and superior experimental results, making it a highly favored model in both contemporary research and applications. Panboonyuen *et al*. [11] proposed a high-performance anchor-free YOLO object detection method based on Feature Pyramid Network (FPN) and Transformer. This model can attend globally to the dependencies between image feature blocks, retaining sufficient spatial information for object detection through multi-head self-attention. Quan *et al*. [12] proposed a simple and effective two-stage Pairwise Convolutional neural network-Transformer (PCT) method. This model leverages the benefits of both the object detector and rich contextual information. Ilharco *et al*. [6] have improved upon the Vision Transformer (ViT) model by adopting a hierarchical construction method reminiscent of a Convolutional Neural Network (CNN), building hierarchical feature maps by merging image blocks at deeper levels.

The introduction and widespread application of the Transformer model have led to significant breakthroughs in NLP tasks such as language modelling, machine translation, and question-answering systems. The Transformer model's ability to model sequences more effectively through its self-attention mechanism has supplanted the traditional role of RNNs and LSTMs in NLP tasks. Based on this premise, this paper proposes and implements an English document search model based on the Transformer model.

2. Method Introduction

In order to realize the modal fusion and retrieval of multimodal English files, how to flexibly represent the relationship between different modal data and store the data and its relationship is the basic problem of multimodal file retrieval. This paper proposes and implements a multimodal English file search model based on Transformer model.

2.1. Multimodal File Retrieval System

The idea of retrieving multimodal file data in this paper is shown in Figure 1.

Initially, multimodal data and the semantic relationships between data are modelled in graphical form and parallelly stored within a graph database. The features of each modal data are extracted and serve as indices, establishing an indexing structure. Simultaneously, semantic relationships between different modal indices are established. By analyzing the correlation between different modal indices, the correlation is added to the indexing graph in the form of edges, thereby forming a richer data semantic association. Upon a user's query submission, the query is modelled and then searched for within the indexing graph for relevant indices that contain the query. The query is further expanded to other modal indices through correlation, ultimately locating the data that contains these indices. Finally, there is modal fusion and similarity computation. The similarity between different modal data is calculated based on the modal fusion technology, which combines the association between different modal data within the indexing graph to yield the final search results.

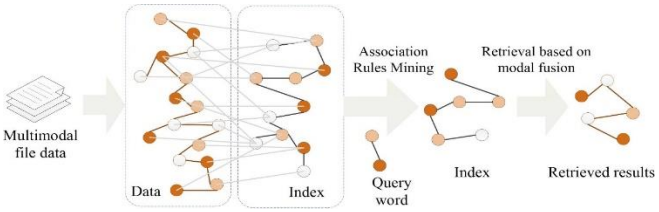


Figure 1. Multimodal file data retrieval process.

Through the aforementioned process, multimodal and cross-modal retrieval can effectively integrate the relationships between different types of data, more accurately meeting the user's retrieval needs and providing users with a richer variety of search results. Figure 2 illustrates the basic workflow of the proposed multimodal file search method based on Transformers presented in this article.

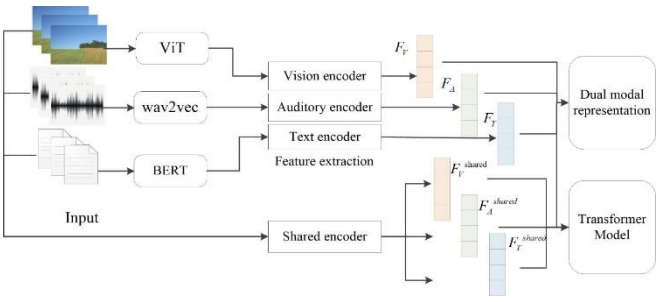


Figure 2. Overall network framework of multimode file search method.

This study primarily focuses on three modalities: Audio (a), Text (t), and Image (v). After employing the language pre-training of deep Bidir-Ectional Transformers for language understanding model (BERT), the visual pre-training model ViT, and the audio pre-training model wav2vec to extract feature sequences $\{I_a, I_t, I_v\}$ from the raw data, these sequences are fed into the encoder. To imbue the feature sequences of the visual and audio modalities with temporal information, a single-layer LSTM network is utilized to inject contextually relevant information and long-term dependencies into these two modalities' feature

sequences. Furthermore, a fully connected layer is employed to map the feature representations of the three modalities to a common dimensional space, facilitating their subsequent processing within the network model, as illustrated in Equations (1) to (4).

$$F_a = FC(sLSTM(I_a\theta_a^{lstm})) \quad (1)$$

$$F_v = FC(sLSTM(I_v\theta_v^{lstm})) \quad (2)$$

$$F_t = FC(I_t) \quad (3)$$

$$F_m \in R^{T_m \times d} \quad (4)$$

where, F_m represents the projected representations of the initial feature representations of each modality after being encoded by the long-short term memory network, mapped into a unified feature dimensionality; d represents the unified feature dimension; Tm represents the length of each respective characteristic sequence; Θ^{lstm} denote the parameters of the networks corresponding to each mode.

Figure 3 illustrates the overall workflow of the multimodal parallel loading algorithm, which includes the execution of the parallel modeling and loading algorithm within the dashed thread pool, divided into two stages: data modeling and data loading. At this juncture, the system amasses and parses multimodal data, constructing a data dictionary that delineates and elucidates the structural, formatting, and relational aspects of the data. During the data ingestion phase, the system employs the previously established Data Dictionary as a foundational framework, leveraging parallel processing capabilities to expeditiously transfer the post-modeling data into the designated system or database.

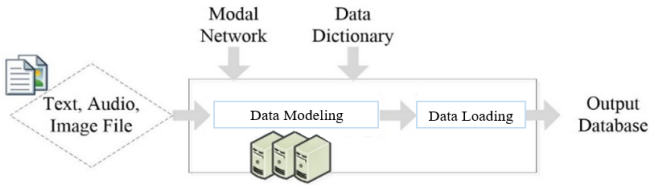


Figure 3. Flow chart of multimodal parallel loading algorithm.

This approach employs a cross-modal Transformer module to simultaneously capture interactions among one modality and other two modalities, in order to achieve the representation of three modalities-Text, Image, and Audio. Within the bimodal representation generation module, the final modality representation is generated by combining the private and shared representations of modalities, and the bimodal representation is regenerated after fusing each pair of modal data. In essence, this method utilizes modality neural networks to analyze the composition of multimodal data and the semantic relationships among them, thereby establishing a modality network instance. The modality network, represented in a graph form, illustrates the semantic relationships among different modalities, enabling us to better comprehend the dependency relations among modal data.

Following the collection of multimodal file data, we utilize modality neural networks to analyze the diverse compositions of the data modalities and their semantic relationships. By modeling the data, a modality network instance is obtained, as shown in Figure 4. This instance can be divided into two parts, namely the data storage graph and the corresponding modality network. Different modalities and their semantic relationships are depicted in the graph in the form of nodes and edges. Each node represents a modality, while edges signify the semantic relationships among modalities, fundamentally revealing the interdependence among different modal data.

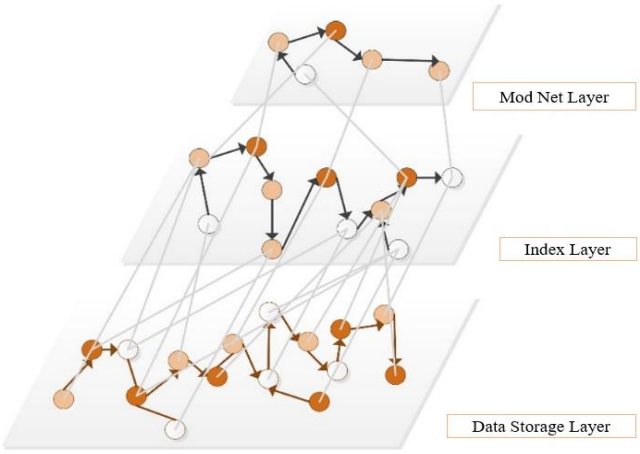


Figure 4. Overall structure of modal network, index and data storage.

2.2. Collaborative Attention Cross-Modal Transformer Module

The encoder of the Transformer model consists of multiple identical layers, each comprising two parts: Multi-Head Attention and Feed-Forward Network (FFN) [9]. Within the Multi-Head Attention mechanism, the input sequence is first transformed into query, key, and value vectors through three distinct linear mappings. Subsequently, the dot product of the query and key is calculated, scaled, and applied with the Softmax function to obtain attention weights, thereby assigning varying levels of significance to each value. The Multi-Head Attention allows the model to simultaneously focus on information from different positions in the sequence, enhancing its ability to represent various features and relationships through the computation of multiple attention heads. Following each attention mechanism, the attention output is processed through a feed-forward neural network. This feed-forward neural network typically includes an activation function between two linear transformation layers, such as ReLU. Such a structure aids the model in learning complex nonlinear relationships and capturing higher-order features. By stacking multiple such encoder layers, the Transformer encoder effectively captures various characteristics and relationships within input sequences, thereby enhancing the model's ability to model sequential data and significantly improving its

performance in tasks such as machine translation and language modeling.

The self-attention mechanism of the Transformer is a mechanism that models interdependencies among elements in sequences, as illustrated in Figure 5. Through this mechanism, the model dynamically adjusts attention weights based on the correlation between each element and other elements to capture long-range dependencies within the sequence. In the Transformer, each input sequence element is represented as a vector, and its correlation with other elements in the sequence is calculated via the self-attention mechanism.

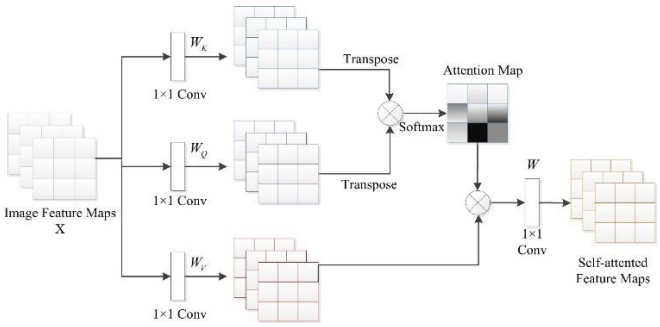


Figure 5. Schematic diagram of self-attention structure.

Within the self-attention module, the input sequence data is denoted as $x \in R_n \times d$, where n represents the length of the sequence and d signifies the dimension of the input vectors. The value vectors V, key vectors K, and query vectors Q are derived through linear transformations applied to the input sequence.

$$V = xW^v \tag{5}$$

$$K = xW^K \tag{6}$$

$$Q = xW^Q \tag{7}$$

Compute the dot product of the query vector Q and the key vector K to obtain the attention score. Normalize the scaled attention score using the Softmax function to obtain the attention weights. Multiply the attention weights by the value vector to obtain the weighted value vector representation.

$$Attent(Q,K,V) = Softmax\left(\frac{QK^T}{\sqrt{d}}\right)V \tag{8}$$

$$MultiHead(Q,K,V) = Concat(head_1, \ldots, head_h)W^0 \tag{9}$$

$$head_1 = Attention(QW_i^Q, KW_i^K, VW_i^V) \tag{10}$$

where, W^K and W^Q are weight matrices; d represents the dimensionality of the model.

The FFN typically consists of two linear transformations (fully connected layers) and the activation function between them (commonly ReLU). Refer to Equation (11) for the computation, where x denotes the input.

$$FFN(x) = max(0, XW_1 + b_1)W_2 + b_2 \tag{11}$$

The function of residual connections in Transformers is similar to that in CNNs, both aiming to address the issues of vanishing or exploding gradients in deep neural networks. Each sublayer (encoder or decoder) in a Transformer contains residual connections, which help simplify the learning process, facilitate smooth gradient propagation, and mitigate the difficulty of training deep networks. Additionally, there exist normalization layers in each sublayer to maintain the distribution of data flowing through the network, aiding in alleviating gradient vanishing problems and accelerating model convergence.

During training, Transformers typically obtain all prediction results simultaneously. However, during inference, Transformers generate output words one by one. To ensure consistent performance in both training and inference processes, we can introduce a Mask module to handle this. The computational formula is as follows:

$$a_{ij} = Softmax(score(v_1 + v_1)Softmax(-\infty) = 0 \tag{12}$$

Here, v_i serves as the encoder for the i-th image block, while v_j acts as the decoder for the j-th image block.

As depicted in Figure 6, the multi-head attention mechanism in the Transformer is employed to address the issue of multiple representation subspaces. By utilizing the multi-head attention mechanism, a model can concurrently process various information focusing on different parts and dimensions. Within the framework of the multi-head attention structure, the input word vectors are projected into distinct representation subspaces through multiple sets of Q (query), K (key), and V (value) matrices, allowing the model to attend to the input from multiple perspectives. Typically, the original 512-dimensional input data is projected through 8 different linear projections, with each projection matrix having a dimension of 64. Consequently, for each head of attention mechanism, a 64-dimensional output is obtained.

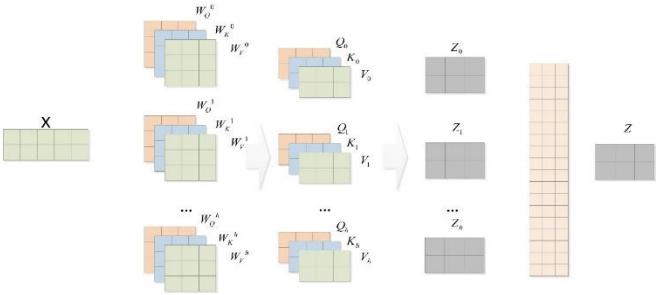


Figure 6. Multi-head attention projected into quantum space.

In the Transformer model, in order to enable the model to automatically learn positional information from the input sequences, cosine and sine functions are typically utilized to encode positional information. The formula is as follows:

$$\begin{cases} PE(pos, 2d_{pos}) = sin\left(\dfrac{pos}{10000^{2d_{pos}+\frac{1}{d_{model}}}}\right) \\ PE(pos, 2d_{pos}+1) = cos\left(\dfrac{pos}{10000^{2d_{pos}+\frac{1}{d_{model}}}}\right) \end{cases} \tag{13}$$

In this context, pos denotes the position of the current object in the current dimensional sequence; d_{pos} represents the dimension of position pos; and $10000^{\frac{2d_{pos}}{d_{model}}}$ signifies the frequency. Incorporating positional encoding into the shared representation of each pattern enables the model to capture sequential information about the sequence, as illustrated in Equation (14).

$$X_m = F_m^{shard} P_m \tag{14}$$

$$F_m^{shard} = Enc^{shard}(F_m \theta^{shard}) \tag{15}$$

$$P_m \in R^{T_m \times d_{map}} \tag{16}$$

$$X_m \in R^{T_m \times d_{map}} \tag{17}$$

where, P_m encodes the position of each pattern; X_m serves as the positional encoding of each pattern, intended for input into subsequent network models for multi-pattern fusion.

In the collaborative attention cross-modal Transformer based on the text modality, the multi-head attention mechanism plays a pivotal role. Through the multi-head attention, the model is able to simulate the interactions between Text, Audio, and Visual patterns. Specifically, this involves projecting the representation of the text modality as the query sequence, concatenating and projecting the representations of the Audio and Visual modalities as the key and value sequences. This enables the model to compute the relational similarities between each word in the text modality sentence and the Audio and Visual features of each frame. Subsequently, the output of the cross-modal attention undergoes processing by a feedforward neural network, yielding the output of the collaborative attention cross-modal Transformer layer. This design effectively integrates information from diverse modalities, enhancing the model's performance in multimodal data processing tasks.

3. Experiment and Analysis

3.1. Data Set and Evaluation Index

This section aims to validate the effectiveness of the proposed Transformer-based multimodal document retrieval model. Two publicly available datasets, namely Carnegie Mellon University-Multimodal Opinion-level Sentiment and Intensity (CMU-MOSI) and Carnegie Mellon University-Multimodal Opinion Sentiment and Emotion Intensity (CMU-MOSEI), are employed in this study. These datasets consist of Text, Visual, and Audio modalities, as illustrated in Table 1. A series of experiments are conducted on these two benchmark datasets in the present study.

Within the multimodal Transformer module, each component is comprised of four Transformer layers, where the number of attention heads within the self-attention mechanism is four. The specific parameter settings during the training process can be referenced in Table 2.

Table 1. Composition and partitioning of data sets.

Data set	CMU-MOSI	CMU-MOSEI
Training set	1028	16048
Validation set	381	1706
Test set	790	5102

Table 2. Related parameter settings.

Items	Value
Optimizer	Adam
learning rate	0.001
batch size	{16,32,64}
epoches	50

This text employs Accuracy (Acc) and F1-score as the evaluation metrics for model performance. The specific calculation formulas are shown in Equations (17) and (18).

$$Acc = \frac{TP + TN}{TP + FP + FN + TN} \tag{18}$$

$$F_1 = \frac{2TP}{2TP + FP + FN} \tag{19}$$

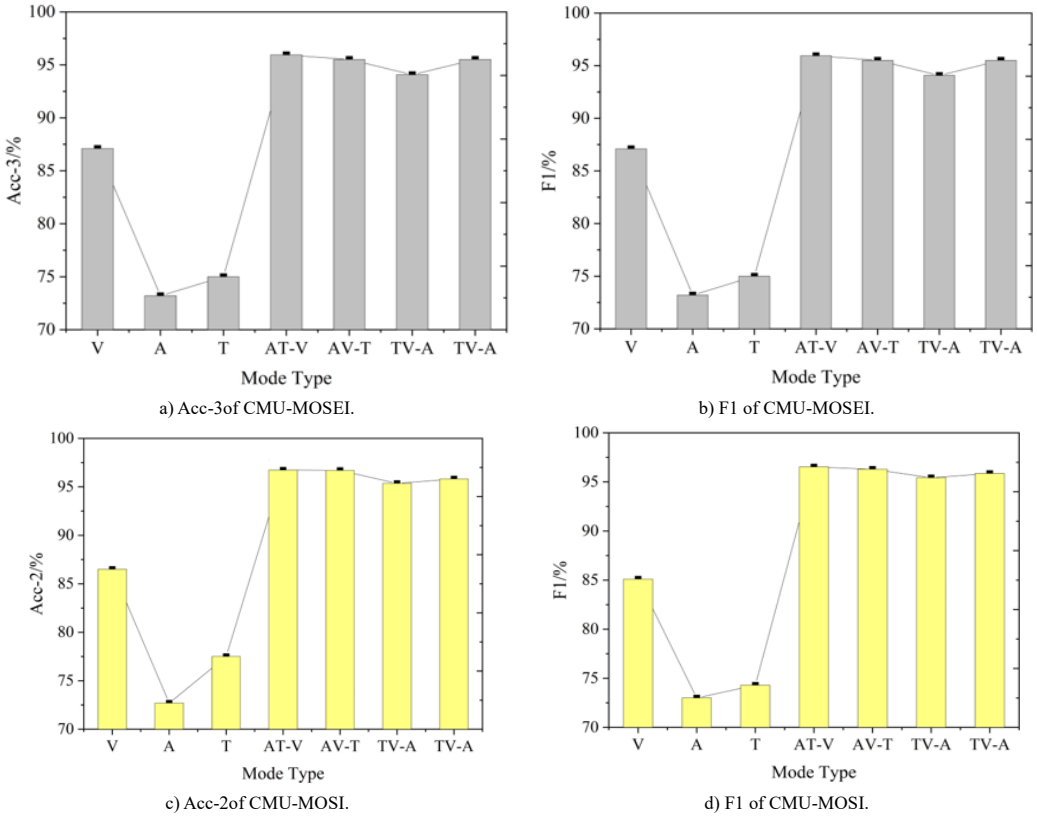
Among these, *FP* represents the number of samples actually negative but predicted as positive; *TN* represents the number of samples with both actual and predicted values negative; *FN* represents the number of samples actually positive but predicted as negative.

3.2. Experiments on Test Sets

This section conducts experiments on the CMU-MOSI and CMU-MOSEI benchmark datasets to compare the performance of Transformer-based multimodal methods with baseline methods. The detailed analysis is combined with the experimental results illustrated in Figure 7. By contrasting the experimental outcomes, a clearer understanding of the superior performance of our proposed method in multimodal document processing tasks can be achieved.

The study constructs a multimodal recognition model using three modalities: Audio (A), text (T), and image (V). To validate the effectiveness of different modality combinations in the proposed multimodal framework, four modality fusion approaches are explored: Audio-Text-Image (AT-V), Audio-Image-Text (AV-T), Text-Video-Audio (TV-A), and Audio-Text-Image (A-T-V). Discussions on these modality combinations aid in assessing the roles and relationships of different modalities in multimodal tasks.

As shown in Figure 7, the fusion of Audio (A) and Text (T) features in the Audio-Text-Image (AT-V) combination yields superior recognition results on both datasets compared to the other three schemes. This outcome indicates that among the three modalities studied, the fusion of audio and text is the optimal combination.



a) Acc-3of CMU-MOSEI. b) F1 of CMU-MOSEI.

c) Acc-2of CMU-MOSI. d) F1 of CMU-MOSI.

Figure 7. Comparison of results of three-mode combination schemes.

3.3. Modal Ablation Experiment

To validate the applicability of the network model proposed in this article, several modal combinations disintegration experiments were conducted on the CMU-MOSEI dataset. The specific experimental results are shown in Figure 8.

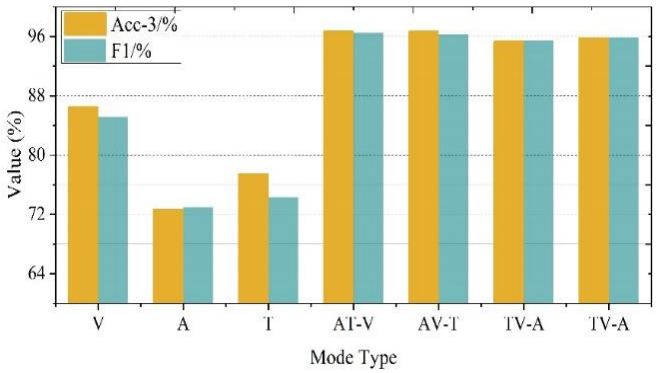


Figure 8. Experimental results of ablation under different modes.

By observing Figure 8, it is evident that the file search capability based on image recognition excels in the unimodal file search experiment, achieving an accuracy of 88.01% and an F1-score of 87.76%. In the trimodal scenario, both the accuracy and F1-score surpass those of unimodal and bimodal situations, showcasing optimal performance. These ablation experiments not only validate the significance of leveraging Audio, Text, and Image modalities in multimodal file search and identification.

Experimental evidence suggests that incorporating semantic information across audio and textual modalities can mutually reinforce each other. This outcome underscores the significance of effectively leveraging information from different modalities within a single target domain to enhance the performance and efficacy of cross-modal tasks. In the case of three target modalities, the accuracy of file search is optimized, yet the complexity of interactions among multiple modalities may somewhat diminish the model's performance. Overall, harnessing multimodal information can augment the efficiency and accuracy of search and retrieval, offering users a more enriched set of retrieval outcomes.

3.4. Comparative Experiment

In this section, the author compares the multi-modal analysis methods based on Transformers with benchmark approaches on the CMU-MOSI and CMU-MOSEI benchmark datasets. The experimental results are presented in Figures 9 and 10.

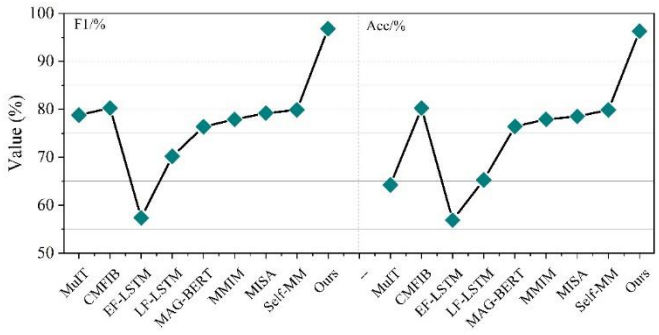


Figure 9. Performance comparison results of different model based on CMU-MOSI.

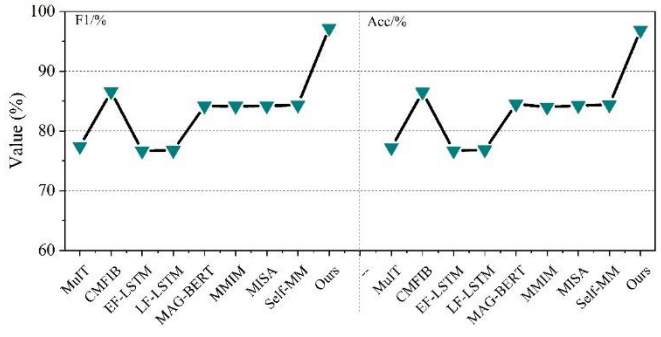


Figure 10. Performance comparison results of different models based on CMU-MOSEI.

On the CMU-MOSI dataset, this approach achieved an accuracy of 84.91% and an F1 score of 85.57% in the classification task, outperforming other baseline models significantly. On the CMU-MOSEI dataset, this method obtained the best results in binary classification tasks in terms of both accuracy and F1-score, further confirming the importance of eliminating redundant information to enhance the accuracy of multimodal emotion analysis. In conclusion, the experimental results demonstrate that the performance of this approach on both datasets clearly surpasses that of other baseline methods.

4. Conclusions

This paper presents a Transformer-based English multimodal document retrieval model. By comparing the precision and recall of the system, it is demonstrated that the approach proposed in this paper can offer query services for different modal files while maintaining good precision and recall. Through ablation experiments on two public datasets and comparative experiments with baseline models, the effectiveness of the proposed Transformer algorithm in extracting modal features and performing feature fusion to enhance the accuracy of document retrieval has been validated. Additionally, this study further demonstrates that multimodal fusion retrieval systems generally outperform single-modal retrieval systems. In the future, we can further explore the modeling capability of multimodal document retrieval in specific domains to realize more application value.

References

[1] Bianchi F., Grattarola D., Livi L., and Alippi C., "Graph Neural Networks with Convolutional Arma Filters," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 7, pp. 3496-3507, 2022. DOI:10.1109/TPAMI.2021.3054830

[2] Bruch S., Gai S., and Ingber A., "An Analysis of Fusion Functions for Hybrid Retrieval," *ACM Transactions on Information Systems*, vol. 42, no. 1, pp. 1-35, 2023. https://dl.acm.org/doi/10.1145/3596512

[3] Guo J., Fan Y., Pang L., Yang L., Ai Q., Zamani H., Wu C., Croft W., and Cheng X., "A Deep Look into Neural Ranking Models for Information Retrieval," *Information Processing and Management*, vol. 57, no. 6, pp. 102067, 2020. https://doi.org/10.1016/j.ipm.2019.102067

[4] Han K., Wang Y., Chen H., Chen X., Guo J., and Liu Z., "A Survey on Vision Transformer," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 1, pp. 87-110, 2022. DOI:10.1109/TPAMI.2022.3152247

[5] Haveliwala T., "Topic-Sensitive PageRank: A Context-Sensitive Ranking Algorithm for Web Search," *IEEE Transactions on Knowledge and Data Engineering*, vol. 15, no. 4, pp. 784-796, 2003. DOI:10.1109/TKDE.2003.1208999

[6] Ilharco G., Wortsman M., Gadre S., Song S., Hajishirzi H., Kornblith S., Farhadi A., and Schmidt L., "Patching Open-Vocabulary Models by Interpolating Weights," *in Proceedings of the Advances in Neural Information Processing Systems*, New Orleans, pp. 1-50, 2022. https://arxiv.org/abs/2208.05592

[7] Kleinberg J., "Authoritative Sources in a Hyperlinked Environment," *Journal of the ACM*, vol. 46, no. 5, pp. 604-632, 1999. https://doi.org/10.1145/324133.324140

[8] Larson R., "A Fusion Approach to XML Structured Document Retrieval," *Information Retrieval*, vol. 8, pp. 601-629, 2005. https://link.springer.com/article/10.1007/s10791-005-0749-0

[9] Liu H., Chen W., "Re-Transformer: A Self-Attention-based Model for Machine Translation," *Procedia Computer Science*, vol. 189, pp. 3-10, 2021. https://doi.org/10.1016/j.procs.2021.05.065

[10] Navarro G., "Spaces, Trees, and Colors: The Algorithmic Landscape of Document Retrieval on Sequences," *ACM Computing Surveys (CSUR)*, vol. 46, no. 4, pp. 1-47, 2014. https://doi.org/10.1145/2535933

[11] Panboonyuen T., Thongbai S., Wongweeranimit W., Santitamnont P., Suphan K., and Charoenphon C., "Object Detection of Road Assets Using Transformer-based YOLOX with Feature Pyramid Decoder on Thai Highway Panorama," *Information*, vol. 13, no. 1, pp. 1-12, 2021. https://doi.org/10.3390/info13010005

[12] Quan H., Lai H., Gao G., Ma J., Li J., and Chen D., "Pairwise CNN-Transformer Features for Human-Object Interaction Detection," *Entropy*, vol. 26, no. 3, pp. 205-217, 2024. https://doi.org/10.3390/e26030205

[13] Sandhu M., Ahmed M., Hussain M., Head S., and Khan I., "Protecting Sensitive Images with Improved 6-D Logistic Chaotic Image Steganography," *The International Arab Journal of Information Technology*, vol. 21, no. 6, pp. 1064-1073, 2024. Doi: 10.34028/iajit/21/6/10

[14] Shah S., Soules C., Ganger G., and Noble B., "Using Provenance to Aid in Personal File Search," *in Proceedings of the USENIX Annual Technical Conference*, California, pp. 171-184, 2007. http://usenix.org/events/usenix07/tech/full_papers/shah/shah.pdf

[15] Shang Y., Ma C., Yang K., and Tan D., "Regenerative Braking Control Strategy Based on Multi-Source Information Fusion under Environment Perception," *International Journal of Automotive Technology*, vol. 23, no. 3, pp. 805-815, 2022. https://link.springer.com/article/10.1007/s12239-022-0072-4

[16] Sherstinsky A., "Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) Network," *Physica D: Nonlinear Phenomena*, vol. 404, pp. 132306, 2020. https://doi.org/10.1016/j.physd.2019.132306

[17] Singh D., Reddy S., Hamilton W., Dyer C., and Yogatama D., "End-to-End Training of Multi-Document Reader and Retriever for Open-Domain Question Answering," *in Proceedings of the 35th International Conference on Neural Information Processing System*, Online, pp. 25968-25981, 2021. https://dl.acm.org/doi/10.5555/3540261.3542249

[18] Wei Y., Zhao Y., Lu C., Wei S., Liu L., and Zhu Z., "Cross-Modal Retrieval with CNN Visual Features: A New Baseline," *IEEE Transactions on Cybernetics*, vol. 47, no. 2, pp. 449-460, 2017. DOI:10.1109/TCYB.2016.2519449

[19] Yu H., Ma R., Su M., An P., and Li K., "A Novel Deep Translated Attention Hashing for Cross-Modal Retrieval," *Multimedia Tools and Applications*, vol. 81, no. 18, pp. 26443-26461, 2022. https://link.springer.com/article/10.1007/s11042-022-12860-w

[20] Zand M., Nasab M., Sanjeevikumar P., Maroti P., and Holm-Nielsen J., "Energy Management Strategy for Solid-State Transformer-based Solar Charging Station for Electric Vehicles in Smart Grids," *IET Renewable Power Generation*, vol. 14, no. 18, pp. 3843-3852, 2020. https://doi.org/10.1049/iet-rpg.2020.0399

[21] Zhang C., Song J., Zhu X., Zhu L., and Zhang S., "HCMSL: Hybrid Cross-Modal Similarity Learning for Cross-Modal Retrieval," *ACM Transactions on Multimedia Computing Communications and Applications (TOMM)*, vol. 17, no. 1s, pp. 1-22, 2021. https://doi.org/10.1145/3412847

[22] Zhang Q., Chang J., Meng G., Xu S., Xiang S., and Pan C., "Learning Graph Structure Via Graph Convolutional Networks," *Pattern Recognition*, vol. 95, pp. 308-318, 2019. https://doi.org/10.1016/j.patcog.2019.06.012

[23] Zhu C., Ping W., Xiao C., Shoeybi M., Goldstein T., Anandkumar A., and Catanzaro B., "Long-Short Transformer: Efficient Transformers for Language and Vision," *in Proceedings of the 35th International Conference on Neural Information Processing Systems*, Virtual, pp. 17723-17736, 2021. https://dl.acm.org/doi/10.5555/3540261.3541617

Hongjuan Li graduated from Henan University of Technology in 2013 and currently work at Pingdingshan Polytechnic College. Her research interests include Cross-Cultural Communication, English and American Literature, Applied Linguistics, and more.