International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Traitement du Signal
Vol. 40, No. 5, October, 2023, pp. 2041-2049
Journal homepage: http://iieta.org/journals/ts

Multi-Modal Medical Image Matching Based on Multi-Task Learning and Semantic-Enhanced Cross-Modal Retrieval

Yilin Zhang

First Clinical Medical College, Guangxi Medical University, Nanning 530021, China

Corresponding Author Email: zhangellynn@163.com

https://doi.org/10.18280/ts.400522

Received: 22 May 2023
Revised: 18 August 2023
Accepted: 2 September 2023
Available online: 30 October 2023

Keywords:
multi-modal medical imaging, image matching, image retrieval, multi-task learning, semantic enhancement, complementary semantic information

ABSTRACT

With the continuous advancement of medical imaging technology, a vast amount of multi-modal medical image data has been extensively utilized for disease diagnosis, treatment, and research. Effective management and utilization of these data becomes a pivotal challenge, particularly when undertaking image matching and retrieval. Although numerous methods for medical image matching and retrieval exist, they primarily rely on traditional image processing techniques, often limited to manual feature extraction and singular modality handling. To address these limitations, this study introduces an algorithm for medical image matching grounded in multi-task learning, further investigating a semantic-enhanced technique for cross-modal medical image retrieval. By deeply exploring complementary semantic information between different modality medical images, these methods offer novel perspectives and tools for the domain of medical image matching and retrieval.

1. INTRODUCTION

In the medical domain, a massive volume of image data is generated daily, attributed to the rapid advancements in medical imaging technologies. This data, sourced from an array of medical imaging equipment including *MRI*, *CT*, and *X-rays* [1-3], offers invaluable insights into patient conditions and holds significant implications for disease diagnosis and treatment [4, 5]. Nevertheless, the multi-modal, high-dimensionality, and intricate structure of these medical images have rendered their management and interpretation particularly challenging, especially when engaging in image matching and retrieval tasks [6-9].

Accurate matching and retrieval of multi-modal medical images hold profound implications in the medical field [10-13]. These processes can assist clinicians in gaining a more comprehensive understanding of a patient's disease trajectory, bolstering clinical decision-making and availing precious data resources for medical research [14, 15]. With the evolution of digital medicine and artificial intelligence technologies, a focal point for researchers has become how to harness cutting-edge computational methodologies for an in-depth analysis and extraction of these medical images, aiming to elevate the quality and efficiency of healthcare services.

At present, a majority of the methodologies for medical image matching and retrieval still anchor their foundations in traditional image processing techniques [17, 18]. While these methods might demonstrate efficacy when handling single-modal medical images, their effectiveness tends to diminish with multi-modal images, as capturing intricate relationships and complementary information between varying modalities proves elusive [19, 20]. Additionally, many methodologies are tethered to manually extracted features, an approach that's not only labor-intensive but might also overlook the nuanced semantic information embedded within the images [21-23]. Crucially, these methods often falter when confronted with large-scale and high-dimensional medical image datasets.

To address these challenges, an algorithm grounded in multi-task learning for medical image matching is introduced in this study. With the incorporation of a hierarchical convolutional network designed for multiple tasks, this algorithm is capable of concurrently processing various modalities of medical images, optimally extracting interrelated information between them, thus achieving enhanced accuracy in medical image matching. Furthermore, a semantic-enhanced technique for cross-modal medical image retrieval, underpinned by an innovative multi-modal multi-granularity semantic enhancement network, is investigated. This method is designed to deeply explore the complementary semantic information between different modalities of medical images, aiming for more precise cross-modal medical image retrieval. Such investigations not only present fresh perspectives and methodologies in the medical image matching and retrieval domain but also furnish the medical field with more efficacious and intelligent tools, poised to catalyze further advancements in medical image processing technologies.

2. THE MEDICAL IMAGE MATCHING ALGORITHM BASED ON MULTI-TASK LEARNING

Within the medical domain, both images and descriptive texts serve as important data forms, offering detailed insights into patient conditions. Typically, medical images convey visual information about diseases, while descriptive texts elucidate diagnoses, treatment recommendations, and other pertinent textual descriptions. Thus, the effective matching and categorization of medical images and descriptive texts is deemed paramount.

The learning of inter-modal associations primarily delves into establishing connections between two distinct data modalities, such as images and texts. Such a learning approach assists in comprehending the mutual relationships between modalities, enabling cross-modal data matching and retrieval. Conversely, the learning of intra-modal associations is chiefly concerned with data relationships within a single modality. For instance, within the image modality, classification tasks might be centred on categorizing medical images into various disease types. Within the text modality, classification endeavours might focus on categorizing diseases into distinct classes based on descriptive texts. By concurrently engaging in both inter and intra-modal associative learning, knowledge can be garnered from two different perspectives, potentially enhancing the model's generalization capabilities and reducing the risk of overfitting. Figures 1 and 2 respectively depict the convolutional processing of medical images and their corresponding descriptive texts.

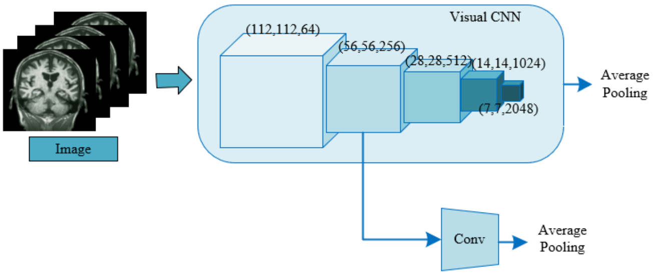


Figure 1. Convolutional processing of medical images

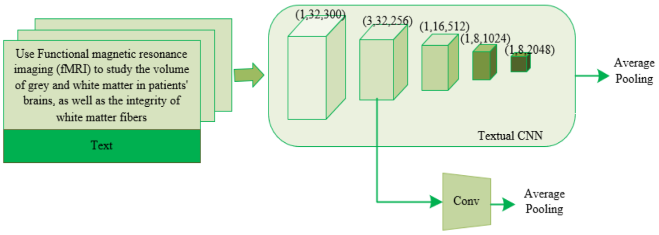

Use Functional magnetic resonance imaging (fMRI) to study the volume of grey and white matter in patients' brains, as well as the integrity of white matter fibers
Text
(1,32,300)
(3,32,256)
(1,16,512)
(1,8,1024)
(1,8,2048)
Textual CNN
Average Pooling
Conv
Average Pooling

Figure 2. Convolutional processing of descriptive texts for medical images

Lower-level features closely resemble the original data and contain a plethora of detailed information, such as textures and colours in images, and vocabulary and basic semantics in texts. This information is deemed pivotal in preliminary classification and matching. For these lower-level features, the intrinsic structure of data can be better captured through self-supervised methods, facilitating effective representation learning even in the absence of explicit labels. In image-text matching tasks, a more meticulous capture of the correlations between the two modalities for these lower-level features becomes necessary. Soft bidirectional ranking loss ensures that both similarities and differences between the two modalities are aptly considered. Thus, for the low-level features of images and texts, self-supervised clustering loss and soft bidirectional ranking loss are respectively utilized for image-text classification and matching tasks.

Higher-level features predominantly capture abstract and higher-order information, like structures and object relations in images, and advanced semantics and context in texts. Such information proves crucial for intricate classification and matching tasks. For these higher-level features, the uniqueness and categorical characteristics of each instance become more vital during classification tasks. Instance loss ensures that, at this abstract level, each instance is correctly classified. Moreover, for these higher-level features in matching tasks, comparisons need to be made on a broader semantic scale, ensuring accurate matching of the advanced semantics of both modalities. Hence, for the high-level features of images and texts, instance loss and soft ranking loss are respectively employed for image-text classification and matching tasks. Figure 3 presents a schematic diagram of the image-text classification and matching task learning processes.

Learning visual and linguistic representations for medical images and their descriptive texts is both a complex and pivotal endeavour. Utilising a *ResNet*-50 pre-trained on *ImageNet* as the image encoder ensures a robust feature extraction capability right from the model's initiation, especially in capturing fundamental concepts and object characteristics in images. Extracting both high and low-level features ensures that information is harvested at various scales. While higher-level features tend to capture an image's holistic and abstract information, the lower-level features reflect the image's details and texture. This complementarity is believed to enhance the model's matching and classification accuracy. The selection of the output from the third residual module of text convolution as the low-level textual feature indicates an emphasis on mid-level textual data. Such features capture certain key nuances of textual data, vital for ensuring text richness and integrity.

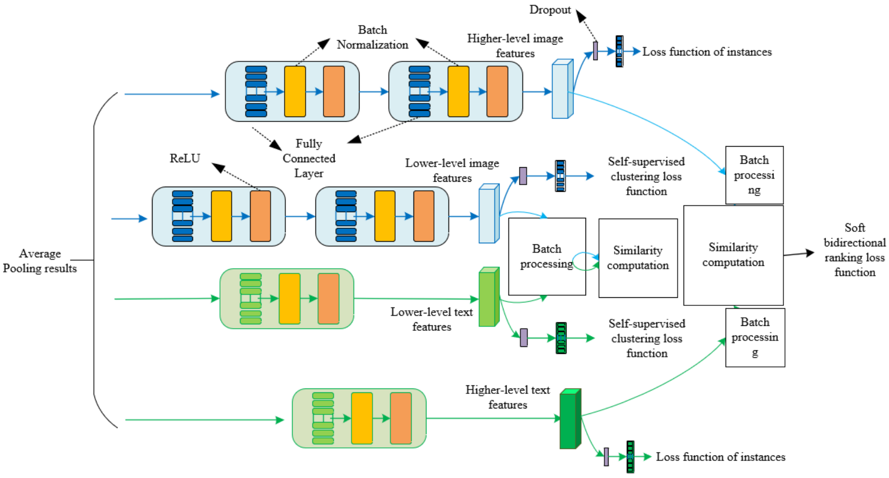

Dropout
Batch Normalization
Higher-level image features
Loss function of instances
Fully Connected Layer
ReLU
Lower-level image features
Self-supervised clustering loss function
Batch processi ng
Average Pooling results
Batch processing
Similarity computation
Similarity computation
Soft bidirectional ranking loss function
Lower-level text features
Self-supervised clustering loss function
Batch processin g
Higher-level text features
Loss function of instances

Figure 3. Learning processes for image-text classification and matching tasks

In medical imaging, minute detail variations might signify entirely distinct medical conditions. Emphasis is placed on the individuality of instances by the instance loss function, ensuring that these subtle yet pivotal differences are captured. Such a function ensures that instances within the same category are drawn closer while those from different categories are kept distinct. Effective learning in the presence of noisy data is facilitated by this loss since it prioritizes the characteristics of individual instances over entire categories. Additionally, the distinction between instances that might appear similar but belong to different categories is enhanced, an aspect that becomes particularly salient in medical imaging where numerous conditions might bear striking visual resemblances. Within the instance loss function framework, medical images and their accompanying descriptive texts are treated as cross-modal data pairs, and an individual label, denoted by uf_g, is assigned. Assuming the image concept representation of the medical image and its descriptive text is represented by $M^g{}_u$, and the textual concept representation is denoted by $M^g{}_y$, the criteria "1{}" must satisfy 1{true}=1 and 1{false}=0. The transformation matrix of the fully connected layer is represented by $Q^g \in E^{v\times 2048}$, the number of all classification possibilities and that of uf_g is represented by $O=[O^1, ..., O^v]$, and the weights of the fully connected layer are denoted by Q^g. The following definitions are given:

$$O_{u,g} = SOFTMAX\left(Q^g c^g\right) \tag{1}$$

$$M_u^g = -\sum_{b=1}^{v} 1\left\{uf_g = b\right\} \log O_{u,g}^b \tag{2}$$

$$O_{y,g} = SOFTMAX\left(Q^g y^g\right) \tag{3}$$

$$M_y^g = -\sum_{b=1}^{v} 1\left\{uf_g = b\right\} \log O_{y,g}^b \tag{4}$$

Annotation of medical data is both costly and time-consuming. Utilization of unlabelled data for learning is facilitated by the self-supervised clustering loss, optimizing the usage of available data. Efforts are made to ensure that similar instances are clustered together, while different instances are separated. Reliance on labelled data is not required by this loss function, making it especially valuable in the realm of medical imaging where extensive labelled datasets might not be accessible. It is adept at capturing the inherent structure of data, which becomes particularly significant given the intricate and latent relationships potentially present in medical images and their descriptive texts. The *k-means* algorithm is employed by the self-supervised clustering loss function. This algorithm divides the sample data into $j(j\leq v)$ datasets, represented by $A=\{A_1, ..., A_k\}$, with the volume of training set data denoted by v. Post encoding of medical images using *ResNet*-50, the resulting feature dataset is denoted as $Z=\{\mathrm{z}_1, ..., z_v\}$, where $z_u \in E^{2048}(u \in [1,v])$. For the initialization of the clustering model, random initialization for the j cluster centers $\omega=\{\omega_1, ..., \omega_j\}$ is necessitated, where $\omega_k \in E^{2048}(k \in [1,j])$. Assuming the cluster index assigned to sample z_u is represented by x_u, the optimization objective function for clustering is given as follows:

$$K = \frac{1}{v}\sum_{u=1}^{v}\left\|z_u - \omega_{x_u}\right\|^2 \tag{5}$$

$$x_u = \underset{k}{ARGMIN}\left\|z_u - \omega_k\right\|^2 \tag{6}$$

In order to minimize the previously mentioned objective function, the cluster center i_k must be updated, and the update formula is:

$$\omega_k = \frac{\sum_{u=1}^{v} 1\left\{x_u = k\right\} z_u}{\sum_{u=1}^{v} 1\left\{x_u = k\right\}} \tag{7}$$

Specifically, an iteration termination condition of $K\leq 1e^{-5}$ was set in the experiments conducted. Constraints are also established for the visual feature class labels of medical images and their descriptive text modalities, resulting in the cluster labels $X=\{x_1, ..., x_v\}$, also referred to as appearance labels uf_l. Analogous to the structure of the instance loss function, the image self-supervised clustering loss function is denoted by $M^l{}_u$, while the text self-supervised clustering loss function is denoted by $M^l{}_y$. Assuming the appearance features are represented by c^l and y^l, and the weight matrix of the fully connected layer is represented by $Q^l \in E^{j\times 2048}$, this layer is tasked with categorizing c^l and y^l into j categories. The probabilities of c^l and y^l belonging to different categories are represented by $O_{u,l} \in E^j$ and $O_{y,l} \in E^j$ respectively, yielding the following calculation formulas:

$$O_{u,l} = SOFTMAX\left(Q^l c^l\right) \tag{8}$$

$$M_u^l = -\sum_{b=1}^{j} 1\left\{uf_l = b\right\} \log O_{u,l}^b \tag{9}$$

$$O_{y,l} = SOFTMAX\left(Q^l y^l\right) \tag{10}$$

$$M_y^l = -\sum_{b=1}^{j} 1\left\{uf_l = b\right\} \log O_{y,l}^b \tag{11}$$

The hyper-parameter is represented by η, and the cosine similarity function is denoted as $A(.,.)$. Positive sample pairs and negative sample pairs are denoted by $(c^g{}_{1o}, y^g{}_{1o})$ and $(c^g{}_{1b}, y^g{}_{1o})$ respectively. Traditional bidirectional ranking loss functions are:

$$\begin{aligned} M_{RA1}^g\left(c^g, y^g\right) = {} & MAX\left[0, \eta - A\left(c_{1o}^g, y_{1o}^g\right) + A\left(c_{1b}^g, t_{1o}^g\right)\right] \\ & + MAX\left[0, \eta - A\left(c_{1o}^g, y_{1o}^g\right) + A\left(c_{1p}^g, y_{1b}^g\right)\right] \end{aligned} \tag{12}$$

$$\begin{aligned} M_{RA1}^l\left(c^l, y^l\right) = {} & MAX\left[0, \eta - A\left(c_{1o}^l, y_{1o}^l\right) + A\left(c_{1b}^l, y_{1o}^l\right)\right] \\ & + MAX\left[0, \eta - A\left(c_{1o}^l, y_{1o}^l\right) + A\left(c_{1o}^l, y_{1b}^l\right)\right] \end{aligned} \tag{13}$$

However, within the actual dataset of matching medical images with descriptive texts, a significant majority of sample combinations might be mismatches. The difficult-negative mining strategy focuses on those negative samples currently deemed "difficult to differentiate" rather than all negative samples. Such a strategy can effectively prevent models from overly focusing on easily distinguishable samples, thus mitigating the risk of overfitting. For matching tasks between medical images and descriptive texts, models must consider the profound connections of both modalities. Such tasks involve not only recognizing and learning the correlation between images and texts but also managing a large number of negative samples and ambiguous matches. Against this backdrop, the use of the aforementioned loss function is deemed unsuitable. Assuming the mean value of all concept feature negative samples within the cluster where y^{g}_{2o} resides is represented by y^{-g}_{2b}, and the mean value of all appearance feature negative samples within the cluster where y^{l}_{2o} resides is represented by y^{-l}_{2b}, the improved bidirectional ranking loss functions are given as:

$$\begin{aligned} M^{g}_{RA1}\left(c^{g},y^{g}\right) &= MAX\left[0,\eta-A\left(c^{g}_{2o},y^{g}_{2o}\right)+A\left(c^{g}_{2b},y^{g}_{2o}\right)\right] \\ &\quad +MAX\left[0,\eta-A\left(c^{g}_{2o},y^{g}_{2o}\right)+A\left(c^{g}_{2o},\bar{y}^{g}_{2b}\right)\right] \end{aligned} \tag{14}$$

$$\begin{aligned} M^{l}_{RA2}\left(c^{l},y^{l}\right) &= MAX\left[0,\eta-A\left(c^{l}_{2o},y^{l}_{2o}\right)+A\left(c^{l}_{2b},y^{l}_{2o}\right)\right] \\ &\quad +MAX\left[0,\eta-A\left(c^{l}_{2o},y^{l}_{2o}\right)+A\left(c^{l}_{2o},\bar{y}^{l}_{2b}\right)\right] \end{aligned} \tag{15}$$

The *BR* loss function aims to minimize the distance between matched image and text pairs while maximizing the distance between unmatched pairs. Bidirectional ranking loss considers both Image-To-Text and Text-To-Image sorting relationships, ensuring bidirectional consistency in matches. By combining the two, not only can match accuracy be improved, but also the stability and robustness of the model's matching results can be enhanced. A soft ranking loss function was constructed by integrating the *BR* loss function and bidirectional ranking loss. Unlike traditional classification loss functions, the soft ranking loss function priorities the relative relationships between samples rather than the absolute classification of each sample. This function provides a larger margin of error for models, allowing a certain level of errors while still promoting correct ordering. It also enables models to recognize and learn varying degrees of matches, ranging from complete matches to complete mismatches, rather than a simple binary match/non-match decision. For concept and appearance features, the corresponding loss function expressions are provided below:

$$M^{g}_{RA}=M^{g}_{RA1}+M^{g}_{RA2} \tag{16}$$

$$M^{l}_{RA}=M^{l}_{RA1}+M^{l}_{RA2} \tag{17}$$

3. CROSS-MODAL RETRIEVAL OF MEDICAL IMAGES BASED ON SEMANTIC ENHANCEMENT

Medical images and their accompanying textual descriptions contain vast inter-modal complementary information. This is evident as images provide a visual representation of a patient's pathological features, whereas textual descriptions offer insights into intricate details or clinical backgrounds that might be challenging to discern from the image alone. To harness this wealth of data effectively, extracting inter-modal complementary semantic information becomes paramount. In pursuit of this goal, a multi-modal, multi-granularity semantic enhancement network is proposed for the mining of inter-modal complementary semantic information. By extracting this complementary semantic information, a more precise alignment between images and textual descriptions is achieved, thereby improving retrieval accuracy. The general framework of this multi-modal, multi-granularity semantic enhancement network is illustrated in Figure 4.

Consider a medical image dataset, which is represented by $F=\{a^{1}_{u},a^{2}_{u},\ldots,a^{W}_{u},\}^{B}_{u=1}$ and describes B semantic concepts. Within this, $F_{u}=\{a^{1}_{u},a^{2}_{u},\ldots,a^{W}_{u},\}$ encapsulates W different modality samples of the u-th semantic concept. When a query item a^{s}_{u} from the s-th modality is being considered, the objective of cross-modal retrieval in medical images is to find the most relevant results in the n-th modality of medical images, that can ensure $1\leq s$, $n\leq W$ and $s\neq n$, and this is the target of the cross-modal retrieval tasks of medical images. In this study, the image s and description text n executing the medical image cross-modal retrieval tasks are referred to as the primary modalities, denoted as $ZYMT=\{a^{s}_{u},a^{n}_{u}\}^{B}_{u=1}$. The remaining W-2 types of modalities are termed auxiliary modalities, represented as $FZMT=\{a^{1}_{u},\ldots,a^{s-1}_{u},a^{s+1}_{u},\ldots,a^{n-1}_{u},a^{n+1}_{i},\ldots,a^{W}_{u},\}^{B}_{u=1}$. Subsequently, the following three matrices are defined:

Definition 1 (Primary similarity matrix *O*): it represents the similarity matrix between medical images and their associated textual descriptions. A high value in this matrix indicates a strong match between a medical image and its textual description. It is composed of the similarity between image *s* and description text *n* modalities within *ZYMT*. It's assumed that the cross-modal similarity between a^{s}_{u} and a^{n}_{u} is represented by O_{uk}.

Definition 2 (Auxiliary similarity matrix *S*): it is the similarity matrix among all auxiliary modalities. Given the presence of multiple auxiliary modalities, an auxiliary similarity matrix can be established for each. Similarities between samples from different modalities in *ZYMT* and *FZMT* give rise to this matrix, where the cross-modal similarity between the *u*-th sample from a modality in *ZYMT* and the *jk*-th sample from an auxiliary modality in *FZMT* is denoted by S_{uk}.

Definition 3 (Cross-modal affinity matrix *V*): it denotes the affinity between the primary and auxiliary modalities, capturing the complementary information between them. When the *u*-th sample from a modality and the *k*-th sample from another different modality describe the same semantic concept, then V_{uk}=+1; otherwise, V_{uk}=-1. This matrix serves the purpose of providing supervisory information for cross-modal retrieval.

In the context of multi-modal learning, leveraging information from auxiliary modalities to enhance the retrieval performance of primary modalities becomes critical. The primary modalities (medical images and their descriptive texts) might not capture all semantic nuances. Auxiliary modalities, such as patient medical records or physiological signals, might hold essential information complementary to medical images or their textual descriptions. Joint optimization of the primary and auxiliary similarity matrices aids in capturing this additional semantic information.

When dealing with intricate data, especially medical images, features of different granularities or scales are pivotal. Coarse-grained features encapsulate global information like overall shape or structure, while fine-grained features focus on local details, such as texture or edges. The proposed method, *M*²*HSE*, is designed to address both types of features by constructing global-level and local-level sub-networks. Within each sub-network, besides computing the primary similarity matrix, 2(*W*-2) auxiliary similarity matrices are also calculated. These auxiliary matrices consider the similarity between other modalities and the primary modality, thus offering a rich set of cross-modal association information.

To ensure synergy and balance among different similarity matrices, the adopted M^2HSE introduces a multi-spring balance loss function. This loss function is likely designed to ensure a balance between the primary similarity matrix and auxiliary similarity matrices during the optimization process, allowing for an equitable integration of information across different modalities and granularities. By jointly optimizing the primary and auxiliary similarity matrices, the M^2HSE aims to thoroughly exploit the complementarities between various modalities. Structural information about medical images might be provided by one modality, while another might offer functional or physiological details. The amalgamation of this information has the potential to produce a more precise and comprehensive interpretation of medical images.

Let the global-level joint optimization objective function be denoted by K^H, and the local-level joint optimization objective function also be represented by K^M. The loss function is denoted by $G(\cdot)$, with the global-level primary similarity matrix and global-level auxiliary similarity matrices being represented as O^H and $\{S^H{}_u|u=1,...,2(W-2)\}$, respectively. The local-level primary similarity matrix and local-level auxiliary similarity matrices are denoted as O^M and $\{S^M{}_u|u=1,...,2(W-2)\}$. Parameters for the two subnetworks are denoted by Φ^H and Φ^M, while the multi-modal complementarity adjustment coefficients for the two subnetworks are denoted by hyperparameters $\{\beta_u|u=1,...,2(W-2)\}$ and $\{\alpha_u|u=1,...,2(W-2)\}$, leading to the Eq. (18):

$$\begin{aligned} K^H &= G\left(O^H,\Phi^H\right)+\sum_{u=1}^{2(W-2)}\beta_u G\left(S_u^H,\Phi^H\right) \\ K^M &= G\left(O^M,\Phi^M\right)+\sum_{u=1}^{2(W-2)}\alpha_u G\left(S_u^M,\Phi^M\right) \end{aligned} \tag{18}$$

A multi-modal approach is adopted in this study, leveraging various feature extraction techniques to enhance the cross-modal retrieval performance between medical images and descriptive text. Convolutional Neural Networks (*CNN*) have demonstrated superiority in image processing, capturing local patterns and hierarchical structures within images. Meanwhile, Bidirectional Gated Recurrent Units (*Bi-GRU*) are adept at handling sequential data, capturing contextual information in text. The combination of these two networks indicates the capability to extract high-quality feature representations from both images and text. Scale-Invariant Feature Transform (*SIFT*) serves as a classical visual feature extraction method, enabling the extraction of stable keypoints under varying scales. The Bag-of-Words (*BoW*) method facilitates the encoding of these keypoints, producing a fixed-length feature vector. The choice of *SIFT-BoW* as an auxiliary modality might aim to provide visual features distinct from those extracted by *CNN*, thereby bolstering the model's robustness and tapping into complementary information.

Substituting the three modalities into the equation yields:

$$\begin{aligned} K^H &= G\left(O^H,\Phi^H\right)+\beta_1 G\left(S_1^H,\Phi^H\right)+\beta_2 G\left(S_2^{\mathrm{H}},\Phi^H\right) \\ K^M &= G\left(O^M,\Phi^M\right)+\beta_1 G\left(S_1^M,\Phi^M\right)+\alpha_2 G\left(S_2^M,\Phi^M\right) \end{aligned} \tag{19}$$

where, O^H, $S^H{}_1$, and $S^H{}_2$ can be calculated based on the three coarse-grained features through the global-level cross-modal similarity computation module set in the model. Similarly, O^M, $S^M{}_1$, and $S^M{}_2$ can be computed based on the three fine-grained features through the local-level cross-modal similarity computation module set in the model. The optimal network parameters for the two subnetworks can be determined through the following formula by minimizing K^H and K^M:

$$\begin{aligned} \tilde{\Phi}^H &= ARG\underset{\Phi^H}{MIN}\,K^H \\ \tilde{\Phi}^M &= ARG\underset{\Phi^M}{MIN}\,K^M \end{aligned} \tag{20}$$

Let the optimal primary similarity matrices be represented by $O^{\sim H}$ and $O^{\sim M}$, which can further be obtained through computation. The optimization process for the two subnetworks is achieved through gradient descent. Assuming that the multi-granularity complementarity adjustment coefficients are represented by ϕ_1 and ϕ_2, the subsequent formula provides the gradient computation for K^H and K^M:

$$\begin{aligned} \frac{\partial K^H}{\partial \Phi^H} &= \frac{\partial G\left(O^H,\Phi^H\right)}{\partial \Phi^H}+\beta_1\frac{\partial G\left(S_1^H,\Phi^H\right)}{\partial \Phi^H}+\beta_1\frac{\partial G\left(S_2^H,\Phi^H\right)}{\partial \Phi^H} \\ \frac{\partial K^M}{\partial \Phi^M} &= \frac{\partial G\left(O^M,\Phi^M\right)}{\partial \Phi^M}+\alpha_1\frac{\partial G\left(S_1^M,\Phi^M\right)}{\partial \Phi^M}+\alpha_1\frac{\partial G\left(S_2^M,\Phi^M\right)}{\partial \Phi^M} \end{aligned} \tag{21}$$

$$\tilde{O}=\varphi_1\tilde{O}^H+\varphi_2\tilde{O}^M \tag{22}$$

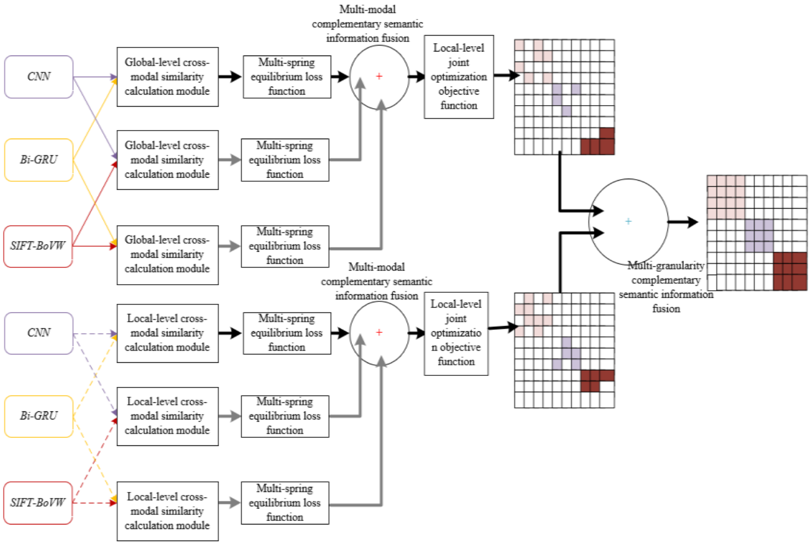

Multi-modal complementary semantic information fusion
CNN
Bi-GRU
SIFT-BoVW
Global-level cross-modal similarity calculation module
Global-level cross-modal similarity calculation module
Global-level cross-modal similarity calculation module
Multi-spring equilibrium loss function
Multi-spring equilibrium loss function
Multi-spring equilibrium loss function
+
Local-level joint optimization objective function
Multi-modal complementary semantic information fusion
CNN
Bi-GRU
SIFT-BoVW
Local-level cross-modal similarity calculation module
Local-level cross-modal similarity calculation module
Local-level cross-modal similarity calculation module
Multi-spring equilibrium loss function
Multi-spring equilibrium loss function
Multi-spring equilibrium loss function
+
Local-level joint optimizatio n objective function
+
Multi-granularity complementary semantic information fusion

Figure 4. Overall framework of the multi-modal multi-granularity semantic enhancement network

4. EXPERIMENTAL RESULTS AND ANALYSIS

A medical image matching model based on multi-task learning has been proposed. By incorporating multi-task hierarchical convolutional networks, the algorithm can handle multiple modalities of medical images simultaneously, ensuring that their associative information is thoroughly explored, leading to more precise medical image matching. In Table 1, the performance of different image-text matching models on the task of matching medical images with descriptive text is displayed. The metrics R@1, R@5, R@10 represent the retrieval performance of the models, with higher values indicating superior performance. It can be seen that *DeViSE* exhibits moderate performance on the task of matching descriptive text and medical images, with an *mR* value of 63.5, indicating its overall average performance. *VSRN*, being a visual-semantic reasoning model, demonstrates improved performance on the text matching task, but it only achieves 9.6 for R@10 in the medical image, suggesting a potential outlier. On the whole, its average performance is recorded at 74.2. *DPC* is a dual-path convolutional image-text embedding model. It is observed that it exhibits commendable performance on both the text and medical image matching tasks, with an *mR* of 77.9, illustrating its balanced cross-modal performance. The model proposed in this research achieves the best results across all metrics. Notably, the R@10 for descriptive text reaches 96.6, suggesting that there's a 96.6% chance of finding the actual matching medical image among the top 10 retrieval results. Moreover, the R@1 for medical images is also notable at 51.8, implying that over half the top-matching retrieval results are correct. Overall, its average performance stands at 81.4, outperforming the other models.

From Figure 5, a clear contrast in performance between the single-task learning model and the multi-task learning model proposed in this research is evident on different retrieval metrics (R@1, R@5, R@10). It is observed that the proposed model, when tasked with matching medical images to descriptive text, outperforms the single-task learning model in both image-text retrieval and text-image retrieval. This strongly indicates the effectiveness and superiority of the proposed model. Outstanding performance is witnessed on the R@1 metric, suggesting high accuracy within the most relevant search results. As previously discussed, the model incorporates multiple loss functions, difficult-negative mining strategies, and a multi-modal multi-granularity semantic enhancement network. These design elements may be crucial factors in its outperformance over the single-task model.

In the Table 2, various cross-modal retrieval models and their performance in image-text matching tasks are presented. Two retrieval tasks, "Image-To-Text" and "Text-To-Image", are listed, with three evaluation metrics each: R@1, R@5, R@10. These metrics represent the hit rates in the top 1, 5, 10 search results, respectively. The highest score of 78.1 on the R@1 metric is achieved by the proposed model. Additionally, leading scores of 93.6 and 96.1 are recorded on R@5 and R@10 respectively. Similarly, for the "Text-To-Image" task, the highest scores on R@1, R@5, and R@10 are 57.9, 83.8, and 91.2 respectively. Thus, it can be concluded that the proposed model clearly excels when compared with other cross-modal retrieval models. This further attests to the advantages of the multi-modal multi-granularity semantic enhancement network in handling cross-modal retrieval tasks for medical images.

Table 3 displays the performance comparison of ablation models in cross-modal retrieval. The importance of each component of the model is evaluated by removing them one by one. It is observed that the model integrates various techniques for cross-modal retrieval, each playing a pivotal role in enhancing the overall performance. Performance deterioration is noted upon the removal of any individual component. Essential roles of *CNN*, *Bi-GRU*, and the similarity computation module in capturing the deep semantic relationships between images and text are identified. Although some metrics occasionally show a slight improvement in certain ablation experiments, it does not undermine the importance of the corresponding components. Instead, it suggests their significance in handling specific types of data or tasks. Collectively, the proposed model offers an efficient and robust cross-modal retrieval solution. The ablation experiments further confirm the significance of each component, establishing the rationale and efficacy behind the model's design.

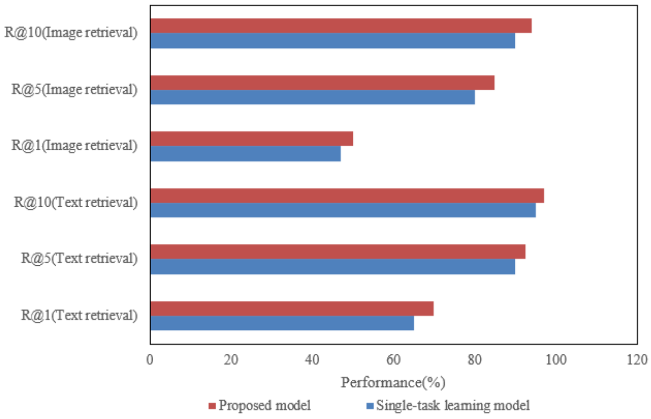
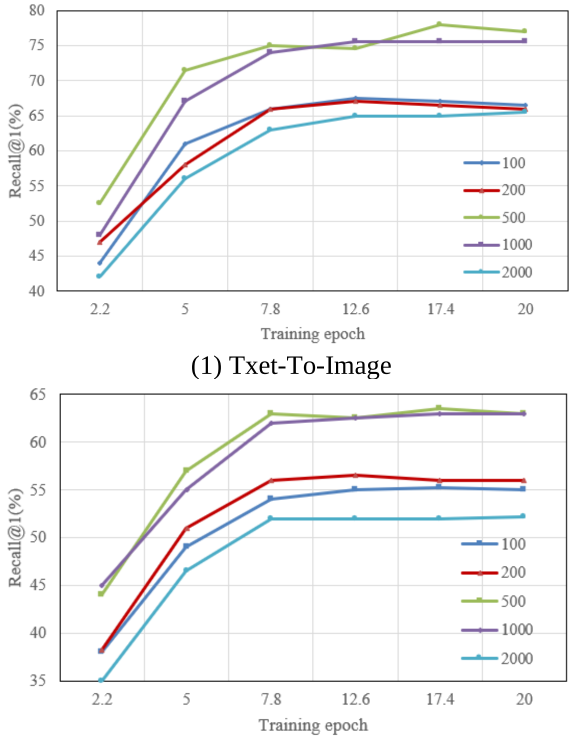


Figure 5. Comparison of performance between the single-task learning model and the proposed model



(1) Txet-To-Image

(2) Image-To-Txet

Figure 6. Recall@1 for different retrieval sample sizes

Table 1. Performance comparison of different image-text matching models

Algorithm	Descriptive Text			Medical Image			*mR*
	R@1	R@5	R@10	R@1	R@5	R@10	
DeViSE	41.2	72.6	83.4	31.2	67.9	81.5	63.5
VSRN	55.6	83.4	9.6	44.5	82.3	91.3	74.2
DPC	64.3	88.9	94.3	46.3	78.6	91.4	77.9
Proposed Model	72.5	92.4	96.6	51.8	85.2	92.8	81.4

Table 2. Performance comparison of different cross-modal retrieval models

Model	Image-To-Txet			Txet-To-Image			*sumR*
	R@1	R@5	R@10	R@1	R@5	R@10	
Bimodal AE	53.1	81.2	88.7	39.6	71.2	78.9	412.2
CCA	66.9	91.3	94.3	47.3	78.2	84.3	445.3
Deep CCA	73.5	92.8	95.5	52.1	78.4	87.3	478.2
GMA	72.4	91.5	95.6	53.8	82.3	87.5	485.2
IMTL	72.8	91.2	96.1	54.2	78.4	85.4	479.3
CMD-VAE	72.3	91.7	95.7	54.8	82.1	88.3	488.2
ACRM	72.4	92.8	95.3	54.3	81.6	86.9	485.3
Proposed Model	78.1	93.6	96.1	57.9	83.8	91.2	512.4

Table 3. Performance comparison of ablation models in cross-modal retrieval

Model	Image-To-Txet			Txet-To-Image		
	R@1	R@5	R@10	R@1	R@5	R@10
Before introducing the multi-spring balanced loss function	75.2	92.4	96.8	57.9	83.9	88.9
Before introducing *CNN*	73.5	92.3	96.4	56.3	82.6	88.5
Before introducing *Bi-GRU*	73.1	92.9	95.3	54.9	82.4	88.6
Before introducing *SIFT-BoVW*	74.9	93.1	95.4	58.1	82.6	88.7
Before introducing the similarity computation module	76.9	93.5	96.3	57.6	85.2	91.2
Proposed Model	76.2	93.6	95.8	57.2	84.6	89.3

Table 4. Recall@1 for different retrieval sample sizes

		100	200	500	1000	2000
Image-To-Text	*Epoch*-1	56.9	55.9	67	57	51.2
	Epoch-5	68.9	76.4	72.3	73.1	71.6
	Epoch-10	76.5	76.3	73.4	78.4	76.5
	Epoch-15	76.2	75.8	75.8	78.6	76.3
	Epoch-20	75	75.9	75.9	77.9	76.8
		100	200	500	1000	2000
Txet-To-Image	*Epoch*-1	45.6	47.8	52.9	51.2	42.7
	Epoch-5	58.2	62.5	57.4	57.6	55.3
	Epoch-10	61.9	62.3	62.8	62.8	62.4
	Epoch-15	62	63.1	62.7	63.7	61.7
	Epoch-20	62.8	63.4	63.4	64.2	63

Table 4 showcases the Recall@1 performance of the Image-To-Text and Text-To-Image tasks at different training epochs and retrieval sample sizes. As the training epochs increase, an improvement in the model's Recall@1 performance across different retrieval sample sizes is noted, confirming the model's adaptability and effectiveness over continuous training. High performance at specific retrieval sample sizes underscores the model's ability to maintain excellence across varying dataset sizes. Even though there's a slight drop in performance after certain training epochs, the model consistently showcases superior performance in most scenarios. This data further emphasizes the model's effectiveness and robustness across different retrieval tasks, particularly with varying sample sizes and training epochs.

Figure 6 depicts the change in Recall@1 performance for Text-To-Image and Image-To-Text tasks across training epochs. Both tasks display remarkable Recall@1 performance, particularly in the initial training epochs. While some decline in performance for certain sample sizes is noticed as training progresses, the model consistently delivers stable results across most scenarios. This reiterates the model's robustness and efficiency across datasets of different sizes.

5. CONCLUSIONS

A study detailing a multi-task learning-based medical image matching algorithm has been presented. This algorithm, capitalizing on multi-modal information, achieves effective matching of medical images. Moreover, an in-depth exploration into a semantic-enhanced method for cross-modal medical image retrieval has been carried out. This method employs an innovative multi-modal multi-granularity semantic enhancement network, aiming to unearth the complementary semantic information shared among different modalities of medical images. Experimental results indicate that, when pitted against other cross-modal retrieval models, the proposed model showcased superior performance in both

Image-To-Text and Text-To-Image tasks. Particularly, its performance exceeded other comparative models in metrics such as Recall@1, Recall@5, and Recall@10. Through a comparison with the ablation models, the significance of various components like the multi-spring balance loss function, *CNN*, *Bi-GRU*, *SIFT-BoVW*, and similarity calculation modules has been observed. Each component's contribution to the overall performance has been elucidated. In the Recall@1 experiment involving varied retrieval sample quantities, the robustness and efficacy of the proposed model under varying data volumes have been demonstrated. Notably, during the later training phases, the model managed to maintain commendable performance.

This research has successfully introduced a medical image matching algorithm based on multi-task learning, displaying excellence across multiple evaluation metrics. The further exploration into the semantic-enhanced cross-modal medical image retrieval has also shown its potency and utility. Experimental outcomes have conclusively affirmed the pioneering position of the proposed method within the realm of cross-modal medical image retrieval, offering valuable insights and guidelines for future investigations in related domains.

REFERENCES

[1] Dawood, T.A., Hashim, A.T., Nasser, A.R. (2023). Automatic skull stripping of MRI head images based on adaptive gamma transform. Mathematical Modelling of Engineering Problems, 10(1): 304-310. https://doi.org/10.18280/mmep.100136

[2] Battula, B.P., Balaganesh, D. (2020). Medical image data classification using deep learning based hybrid model with CNN and encoder. Revue d'Intelligence Artificielle, 34(5): 645-652. https://doi.org/10.18280/ria.340516

[3] Mahdi, H.A., Shujaa, M.I., Zghair, E.M. (2023). Diagnosis of medical images using Fuzzy Convolutional Neural Networks. Mathematical Modelling of Engineering Problems, 10(4): 1345-1351. https://doi.org/10.18280/mmep.100428

[4] Mouhni, N., Elkalay, A., Chakraoui, M., Abdali, A., Ammoumou, A., Amalou, I. (2022). Federated learning for medical imaging: An updated state of the art. Ingénierie des Systèmes d'Information, 27(1): 143-150. https://doi.org/10.18280/isi.270117

[5] Parvathy, V.S., Pothiraj, S., Sampson, J. (2020). Optimal deep neural network model based multimodality fused medical image classification. Physical Communication, 41: 101119. https://doi.org/10.1016/j.phycom.2020.101119

[6] Meher, B., Agrawal, S. (2020). A multimodality medical image fusion method using region based approach. In Proceedings of the 1st International Conference on Data Science, Machine Learning and Applications (ICDSMLA 2019), pp. 1156-1162. https://doi.org/10.1007/978-981-15-1420-3_126

[7] Parvathy, V.S., Pothiraj, S., Sampson, J. (2021). Multimodality medical image fusion based on non-sub-sampled contourlet transform. International Journal of Computer Applications in Technology, 65(4): 358-367. https://doi.org/10.1504/IJCAT.2021.117279

[8] Parvathy, V.S., Pothiraj, S., Sampson, J. (2020). A novel approach in multimodality medical image fusion using optimal shearlet and deep learning. International Journal of Imaging Systems and Technology, 30(4): 847-859. https://doi.org/10.1002/ima.22436

[9] Liu, X., Mei, W., Du, H. (2018). Detail-enhanced multimodality medical image fusion based on gradient minimization smoothing filter and shearing filter. Medical and Biological Engineering and Computing, 56(9): 1565-1578. https://doi.org/10.1007/s11517-018-1796-1

[10] Jaber, M.M., Yussof, S., Elameer, A.S., Weng, L.Y., Abd, S.K., Nayyar, A. (2022). Medical image analysis using deep learning and distribution pattern matching algorithm. Computers, Materials and Continua, 72(2): 2175-2190.

[11] Belfedhal, A.E. (2023). Multi-modal deep learning for effective malicious webpage detection. Revue d'Intelligence Artificielle, 37(4): 1005-1013. https://doi.org/10.18280/ria.370422

[12] Shermadurai, P., Thiyagarajan, K. (2023). Deep learning framework for classification of mental stress from multimodal datasets. Revue d'Intelligence Artificielle, 37(1): 155-163. https://doi.org/10.18280/ria.370119

[13] Pinapatruni, R., Chigarapalle, S.B. (2022). Adversarial image reconstruction learning framework for medical image retrieval. Signal, Image and Video Processing, 16(5): 1197-1204. https://doi.org/10.1007/s11760-021-02070-6

[14] Duan, Y., Li, Y., Lu, L., Ding, Y. (2022). A faster outsourced medical image retrieval scheme with privacy preservation. Journal of Systems Architecture, 122: 102356. https://doi.org/10.1016/j.sysarc.2021.102356

[15] Liu, C., Ding, W., Cheng, C., Tang, C., Huang, J., Wang, H. (2022). DenseHashNet: A novel deep hashing for medical image retrieval. IEEE Journal of Radio Frequency Identification, 6: 697-702. https://doi.org/10.1109/JRFID.2022.3209986

[16] Guan, A., Liu, L., Fu, X., Liu, L. (2022). Precision medical image hash retrieval by interpretability and feature fusion. Computer Methods and Programs in Biomedicine, 222: 106945. https://doi.org/10.1016/j.cmpb.2022.106945

[17] Hu, B., Vasu, B., Hoogs, A. (2022). X-MIR: EXplainable medical image retrieval. In Proceedings - 2022 IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2022, pp. 1544-1554.

[18] Han, J., Cao, Y., Xu, L., Liang, W., Bo, Q., Wang, J., Wang, C., Kou, Q.Q., Liu, Z., Cheng, D. (2022). 3D reconstruction method based on medical image feature point matching. Computational and Mathematical Methods in Medicine, 2022: 9052751. https://doi.org/10.1155/2022/9052751

[19] Yudha, E.P., Suciati, N., Fatichah, C. (2021). Preprocessing analysis on medical image retrieval using one-to-one matching of SURF keypoints. In Proceedings -5th International Conference on Informatics and Computational Sciences, ICICos 2021, pp. 160-164. https://doi.org/10.1109/ICICoS53627.2021.9651782

[20] Yang, Y., Cao, S., Huang, S., Wan, W. (2021). Multimodal medical image fusion based on weighted local energy matching measurement and improved spatial frequency. IEEE Transactions on Instrumentation and Measurement, 70. https://doi.org/10.1109/TIM.2020.3046911

[21] Mojica, M., Pop, M., Ebrahimi, M. (2021). Medical

image alignment based on landmark- and approximate contour-matching. Journal of Medical Imaging, 8(6).

[22] Saygili, G. (2020). Predicting medical image registration error with block-matching using three orthogonal planes approach. Signal, Image and Video Processing, 14(6): 1099–1106. https://doi.org/10.1007/s11760-020-01650-2

[23] Tang, Y., Chen, Y., Xiong, S. (2022). Deep semantic ranking hashing based on self-attention for medical image retrieval. In 2022 26th International Conference on Pattern Recognition, ICPR 2022, pp. 4960-4966. https://doi.org/10.1109/ICPR56361.2022.9956369