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

Traitement du Signal
Vol. 40, No. 5, October, 2023, pp. 2161-2168
Journal homepage: http://iieta.org/journals/ts

Cross-Media Retrieval Based on Two-Level Similarity and Collaborative Representation

Jiahua Zhang[1,2]

[1] State Key Laboratory of Public Big Data, Guizhou University, Guiyang 550025, China
[2] College of Computer Science and Technology, Guizhou University, Guiyang 550025, China

Corresponding Author Email: gs.jhzhang20@gzu.edu.cn

https://doi.org/10.18280/ts.400533

Received: 6 May 2023
Revised: 24 August 2023
Accepted: 2 September 2023
Available online: 30 October 2023

Keywords:
collaborative representation, two-level similarity, cross-media retrieval

ABSTRACT

In the exploration of cross-media retrieval encompassing images and text, an advanced method incorporating two-level similarity and collaborative representation (TLSCR) is presented. Initially, two sub-networks were designed to handle both global and local features, facilitating enhanced semantic associations between images and textual content. Whole images, along with regional image sectors, served as representations for images, while textual content was depicted both through complete sentences and select keywords. An innovative two-level alignment approach was introduced to segregate and then amalgamate the global and local depictions of paired images and texts. Subsequently, employing collaborative representation (CR) technology, each experimental image was collaboratively reconstructed by utilising the entirety of the training images, and every experimental text by incorporating all the training texts. The collaborative coefficients derived were subsequently employed as congruent dimensional representations for both images and texts. Upon completion of these operations, cross-media retrieval between the two modalities was conducted. Experimental outcomes on datasets like Wikipedia and Pascal Sentence confirm the superior precision of the proposed method, surpassing conventional cross-media retrieval methodologies.

1. INTRODUCTION

In the wake of rapid advancements in information technology, the widespread use of multimodal data, encompassing images, text, video, and audio, on the internet has been observed [1-8]. Instances where images on web pages are accompanied by descriptive text underline this trend. Cross-media is characterised not only by the juxtaposition of diverse media objects, such as text, images, audio, and video, but also by the intricate relationships and structural formations among them. A predominant challenge in cross-media retrieval stems from the heterogeneous nature of different media types. Often, identical content information spans various multimedia formats. Comprehensive understanding of the content information embedded within these cross-modal complexes has been achieved through fusion analyses and other methodologies.

Over the past decades, growing interest in cross-media retrieval has been noted, and extensive research in this domain has been conducted. Such retrieval mechanisms have proven beneficial to a multitude of users. Nonetheless, the frequent spanning of multimodal data across diverse feature spaces introduces feature heterogeneity, posing significant challenges for cross-media retrieval endeavours. The focus of this study is centred on the cross-media retrieval between images and text, with tasks encompassing the search for text or images relevant to a pre-known query image or text.

In addressing the aforementioned feature heterogeneity, various methodologies for the universal representation of distinct modal data have been proposed by researchers. It has been noted that a significant portion of existing literature emphasises subspace learning methods. In these methods, a pair of mapping matrices is learned, allowing data from different modalities to be projected into isomorphic subspaces. This subsequently facilitates direct similarity measurements between multimodal data. Classic unsupervised subspace learning methods, such as Canonical Correlation Analysis (CCA) [9] and Partial Least Squares (PLS) [10], have been developed to learn projection matrices by optimising the correlation between multimodal data. On the other hand, supervised techniques, like the Semantic Correlation Matching (SCM) [9], revert the isomorphic subspace derived from conventional correlation analysis back to the semantic subspace. The Three View Canonical Correlation Analysis (CCA-3V) [11], by introducing a semantic viewpoint into the canonical correlation analysis, aims to enhance the differentiation of multimodal data of varied categories within the learning subspace. Furthermore, the Generalized Multiview Analysis (GMA) [12] provides a comprehensive framework for multimodal feature extraction technologies. As the field progresses, subspace learning methodologies have been categorically divided into four primary types below.

First, projection-based subspace learning. Within this approach, feature mapping is employed to discern potential subspaces across varied modal data. Delineation can be drawn between linear projection methods, such as CCA [9] and PLS [10], and nonlinear projection methodologies, represented by kernel canonical correlation analysis (KCCA) [13] and Deep Canonical Correlation Analysis (DCCA) [14].

Second, matrix decomposition-based subspace learning. This method capitalises on matrix decomposition to discern the foundational vectors of potential subspaces within different modal data. Distinctive branches emerge within this category, comprising non-negative matrix decomposition approaches, exemplified by Joint Shared Non-negative Matrix Factorization (JSNMF) [15], and methods centred on feature decomposition, such as Multi Output Regularized Feature Projection (MORFP) [16].

Third, task-based subspace learning. Through this approach, multiple tasks are concurrently learned, enhancing the overarching generalisation performance across each task. Such an approach can be further subdivided into multi-task learning methods, like alternating structure optimisation (ASO) [17] and convex multi-task feature learning (CMTFL) [18]. Additional branches include multi-label learning strategies, such as Shared Subspace Learning for Multi-Label Classification (SSLMC) [19], and multi-class learning approaches represented by Shared Structures in Multi-Class Classification (SSMCC) [20].

Finally, metric-based subspace learning. This method is explicitly devised to cultivate robust metrics between multiple modal datasets, fostering efficient similarity comparisons. It can be stratified into Euclidean distance metric methods, such as Multi Modal Distance Metric Learning (MMDML) [21], and Markov distance metric techniques, including Shared Subspace for Multiple Metric Learning (SSMML) [22].

Historically, traditional subspace learning strategies for cross-media retrieval treated all image or text samples collectively, striving to learn the projection matrix, whilst often overlooking the unique characteristics inherent to each sample. The diverse distribution and representation of features across modalities necessitated a bridging of the semantic gap between such modalities. Typically, the fundamental approach involved the creation of a common subspace, into which all data were projected for further learning. Yet, certain modalities occasionally presented scenarios where multiple semantically approximate instances from differing modalities found alignments. Mere alignment of text and images through a shared subspace was recognised as insufficient. Additionally, supervised cross-media retrieval methodologies demanded considerable volumes of labelled data. Given the challenge in procuring such data, limitations on these methods' applicability became evident.

Subsequently, an unsupervised CR method for cross-media retrieval, based on two-tier networks, was introduced. This approach adeptly fused both global and local similarities, ensuring enhanced retrieval precision. By employing all training images to collaboratively reconstruct each test image, and analogously using all training texts to collaboratively reconstruct each test text, the collaboration coefficients between the two were utilised as a homogenised representation of image and text. Thus, a framework was established wherein each test sample was co-represented alongside all training samples, preserving the distinctiveness of each test sample. Comparative analyses between this novel method and conventional cross-media retrieval strategies have been conducted, and the preliminary findings underscore the efficacy of the introduced methodology.

2. CROSS-MEDIA METHODOLOGY BASED ON TWO-LEVEL MODEL

A two-level cross-media model has been proposed for the execution of retrieval tasks. Within this methodology, two distinct loss functions, namely global and local, were formulated to capture both the global and local features inherent to images and texts. Utilising the feature representations derived from these images and texts, two levels of similarity were strategically developed and aggregated, ensuring a degree of information complementarity. Such an approach manifested in augmented outcomes in cross-media retrieval tasks.

The integration of both global and local similarities by the two-level model method offers a comprehensive semantic description of images and texts. Such integration potentially facilitates a deeper exploration into the latent semantic alignment between images and texts. On one end, data from diverse modalities, when transitioned from their independent representation spaces to a shared subspace, permits the calculation of similarities across varying modal instances. This translates to a reduction in the semantic feature distance across distinct modal data. On the opposite end, drawing inspiration from ranking-based methodologies [23-25], a tripartite-based loss function was employed. This function was designed to augment the distance between varying modalities encompassing different semantic features while minimising the distance among modalities with analogous semantic features. The ultimate goal remains an enhanced alignment between image and textual data.

Intricately designed, this model not only capitalises on self-attention networks to derive global, overarching representations of images but also harnesses attention mechanisms for localised text representations. An innovative two-level similarity fusion method was introduced to facilitate mutual enhancement, thus propelling cross-media association learning towards achieving information complementarity.

2.1 Global representation processing

2.1.1 Global representation of images

Each image i_m was first resized to dimensions of 256×256, post which they were subjected to a convolutional neural network (CNN) to harness their high-dimensional information. The network, analogous in configuration to ResNet-152 [26], was pre-trained on the expansive dataset, ImageNet. The final image feature was subjected to mean-pooling, resulting in the extraction of the global image feature x. This global feature was subsequently processed through a self-attention network [27], depicted in Figure 1.

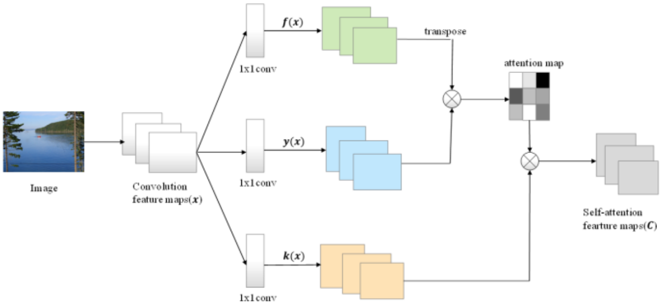


Figure 1. Structure of self-attention network

Initially, based on the pre-obtained global feature x of the image, calculations of $f_x=W_fx$ and $y(x)=W_yx$ were performed. This yielded two distinctive feature spaces, each generated by multiplying image features with different weight matrices W_f and W_y. The softmax function was employed to calculate the correlation $a_{j,i}$, which represents the degree of correlation between the image content in region j of the model and region i. The mathematical representation is provided as:

$$a_{j,i} = \frac{\exp\left(b_{ij}\right)}{\sum_{i=1}^{n}\exp\left(b_{ij}\right)} \tag{1}$$

where, $b_{ij}=f(x_i)^T y(x_j)$. Subsequently, the output c_j of the self-attention network was computed:

$$\mathbf{c}_j = \sum_{i=1}^{N} a_{j,i} k\left(\mathbf{x}_i\right) \tag{2}$$

where, $k(x_i)=W_k x_i$, and W_k represents a weight parameter matrix. $C=(c_1, c_2, \ldots, c_j, \ldots, c_n)$ plays a pivotal role, ensuring the harmonious amalgamation of information.

In the final phase, the primal features of the image were fused with the attributes of the attention layer. This convergence yielded the output $g_i=\lambda c_i+x_i$, serving as a comprehensive representation of image λ. Notably, the parameter value has been set at 0.1.

2.1.2 Global representation of text

Textual input t_k was interpreted as a character sequence and processed via a character convolutional network, termed Char-CNN [28]. The final activation layer generated a representation sequence, which was then relayed to an RNN for character-level text classification. This process facilitated the extraction of high-tier abstract semantic features, eliminating the requirement for pre-trained word vectors or intricate grammatical structures. The resulting sequence from Char-CNN for each input text t_k was denoted as P, serving as the input for Long Short-Term Memory (LSTM) [29] processing. Let $H_i = \{h_1^i, \ldots, h_m^i\}$ be the output of the hidden unit, then the global representation for the text was subsequently derived, as illustrated in Eq. (3):

$$\mathbf{g}_t = \frac{1}{m}\sum_{k=1}^{m}\mathbf{h}_k^i \tag{3}$$

Char-CNN interprets each statement as a character sequence, maintaining a standardised length of 300 characters. Statements exceeding this length underwent truncation, whereas shorter statements were padded with zeroes. The Char-CNN architecture included three convolutional layers, with parameters set at (256, 4), (512, 4), and (2048, 4), with the enclosed values representing the kernel number and width respectively.

2.2 Local representation processing

2.2.1 Local representation of images

Each image i_m, once processed through Faster R-CNN [30], yielded multiple bounding boxes, facilitating the identification of all candidate image regions. Subsequent to this, the initial five regions, ranked based on their scores, were selected for further calculations. These shortlisted regions were then subjected to the ResNet-152 network, culminating in the derivation of regional features via mean pooling of the final image feature [31]. These features, indicative of n distinct regions within a given image, served as the local representation of the image $\{l_i^1, \ldots, l_i^n\}$, where i denotes the sequence number of the image.

2.2.2 Local representation of text

For effective assimilation of the text's local representation, the sole deployment of LSTM might prove ineffectual in encoding information in a bidirectional manner. Fine-grained classification demands acute attention to interactions between emotion words, degree words, and negation words, necessitating information encoding in both forward and reverse orientations. As such, the bidirectional LSTM (Bi-LSTM) [32] was employed to capture the semantic dependencies inherent to text in two directions.

For the i-th term in a particular statement $Y=\{y_1, y_2, \ldots, y_i, \ldots, y_m\}$, a search operation within the vocabulary yields its representation through the word embedding matrix W_E, as articulated in Eq. (4):

$$\mathbf{W}_E \cdot y_i = \mathbf{W}_E \omega_i, i \in [1, m] \tag{4}$$

In this instance, words were embedded into a 300-dimensional vector space. The Bi-LSTM method, benefiting from two primary components (i.e. forward LSTM and backward LSTM) was used to encapsulate word data from two directions. Specifically, the forward LSTM, traversing from ω_1 to ω_n, perused the statement Y in the n direction, whereas the backward LSTM, spanning ω_n to ω_1, interpreted the statement Y in the inverse direction, as described in Eq. (5):

$$\overrightarrow{\mathbf{h}_l} = \overrightarrow{\mathrm{LSTM}}\left(y_i\right), i \in [1, m] \tag{5}$$

The feature e_m of the terminal word is discerned by averaging the forward and backward hidden states $\overrightarrow{h_l}$ and $\overleftarrow{h_l}$, encapsulating centralised statement information around ω_i, as shown in Eq. (6):

$$\mathbf{e}_m = \frac{\left(\overrightarrow{\mathbf{h}_l} + \overleftarrow{\mathbf{h}_l}\right)}{2}, i \in [1, m] \tag{6}$$

For this methodology, the output dimension from word embedding extraction stands at 2048. Utilising the word embedding output as an input to the Bi-LSTM, the hidden unit outputs were captured, denoted as $E=\{e_1, \ldots, e_i, \ldots, e_m\}$. These outputs, representing m distinct text sections within a given statement, provide the foundational features to delineate the context of said statement.

With an assumption of n texts in play, the output from the Bi-LSTM's hidden unit, symbolised as $E' = \{e_1^n, \ldots e_i^n, \ldots, e_m^n\}$, encompasses m varied text fragments spanning n statements. Following Bi-LSTM and attention mechanism processing, the local representation $l_t = \frac{1}{m}\sum_{k=1}^{m}\sum_{j=1}^{n} a_k^t\, e_k^j$ of the n statements is procured, marking the ultimate local text representation.

To culminate the representation processes, two fully connected networks were appended to both global and local representation processing frameworks, transmuting the dimensionalities of the image and text feature vectors to 1024. These networks, in their essence, functioned as cross-media semantic alignment components, mapping heterogeneous features into a shared subspace.

2.3 Cross-media two-level alignment

The global and local representations were derived employing the triplet loss function [33]. At the core of this function lies the anchor example, supplemented by a positive and a negative example within a shared model. Through this shared model, it was observed that the anchor examples and positive examples were optimally clustered, concurrently distancing themselves from negative examples. The triplet loss function is denoted by $Loss_{triplet}$=max($d(a, p)$-$d(a, n)$+margin, 0), where a represents the anchor example, p stands for the positive example, and n signifies the negative example.

Building upon the foundations of the triplet loss, an objective function was subsequently formulated:

$$\begin{gathered}
L_{\text{global}} = \frac{1}{N}\sum_{n=1}^{N} L_1\left(i_+^n, t_+^n, t_-^n\right) + L_2\left(t_+^n, i_+^n, i_-^n\right) \\
L_1\left(i_+^n, t_+^n, t_-^n\right) = \max\left(0, a - d\left(\mathrm{g}_{i+}^n, \mathrm{g}_{t+}^n\right) + d\left(\mathrm{g}_{i+}^n, \mathrm{g}_{t-}^n\right)\right) \\
L_2\left(t_+^n, i_+^n, i_-^n\right) = \max\left(0, a - d\left(\mathrm{g}_{i_+}^n, \mathrm{g}_{t+}^n\right) + d\left(\mathrm{g}_{i-}^n, \mathrm{g}_{t+}^n\right)\right)
\end{gathered} \tag{7}$$

where, L_1 and L_2 symbolise the similarity between the globally matched image-text pairs encountered during the model training phase. The objective was to maximise the disparity between the similarities of these matched pairs and their mismatched counterparts. The function $d(\ \cdot)$ embodies the dot product of the image-text pairs, translating to their similarity. $(\mathrm{g}_{i+}^n, \mathrm{g}_{t+}^n)$ epitomises the matched image-text pairs, while $(\mathrm{g}_{i+}^n, \mathrm{g}_{t-}^n)$ and $(\mathrm{g}_{i-}^n, \mathrm{g}_{t+}^n)$ denote the counts of the mismatched image-text pairs. The variable n corresponds to the count of image-text pairs, and α is the marginal parameter with N being the total number of triplets extracted from the training dataset. The principal aim of local alignment was discerning the optimal alignment between the text's local representation l_t and the myriad local representations $\{l_i^1, \dots, l_i^n\}$ exhibited by an image pair. This translates to selecting the K nearest neighbours from the manifold of image local representations corresponding to each text local representation. It was determined that a K value of 3 optimally aligns the local representations of both images and text. The governing objective function was:

$$L_{\text{local}} = \max\left(0, a - \frac{1}{K}\sum_{k=1}^{K} d\left(\mathbf{l}_{t+}, \mathbf{l}_{i+}^k\right) + \frac{1}{K}\sum_{k=1}^{K} d\left(\mathbf{l}_{t+}, \mathbf{l}_{i-}^k\right)\right) \tag{8}$$

To encapsulate the entirety of the cross-media alignment, a comprehensive similarity metric between image i_m and text t_k was conceived. This metric amalgamates both global and local similarities, computing them within a unified 1024-dimensional subspace:

$$\begin{gathered}
\text{sim}1 = d\left(\mathrm{g}_i, \mathrm{g}_t\right); \text{sim}2 = \frac{1}{K}\sum_{k=1}^{K} d\left(\mathbf{l}_i^k, \mathbf{l}_t\right); \\
\text{sim}\left(i_m, t_k\right) = \theta \cdot \text{sim}1 + (1-\theta) \cdot \text{sim}2
\end{gathered} \tag{9}$$

For this computation, both the global and local features resulting from the image-text transformation were harnessed, yielding the overarching similarity θ, a parameter introduced in this study, was restricted to values ranging between 0.3 and 0.7.

3. CR TECHNOLOGY

CR technology employs all training images for the collective reconstruction of individual test images. Similarly, every training text is harnessed for the CR of each test text. The collaboration coefficient between the two is utilised as a congruent dimensional representation for both image and text, enabling the execution of cross-media retrieval between them. This study proceeds to elucidate both the CR technology and the Collaborative Representation-based Classification (CRC). For a dataset comprising n training samples, the entire training set is denoted as a dictionary X:

$$X = \left[X_1, X_2, \dots, X_k\right] = \left[x_{1,1}, x_{1,2}, \dots, x_{k,n_k}\right] \tag{10}$$

where, k is identified as the aggregate count of semantic categories, while the representation of the i-th class example is rendered as:

$$X_i = \left[x_{i,1}, x_{i,2}, \dots, x_{i,n_i}\right] \in R^{m \times n_i} \tag{11}$$

where, m stands for the dimensionality of sample features, and n_i is the count of i-class samples. For a given test sample y, it was expressed as a linear combination of all training samples:

$$y = Xp \in R^m \tag{12}$$

where, $p = \left[0, \dots 0, p_{i,1}, p_{i,2}, \dots, p_{i,n_i}, 0, \dots, 0\right]^T$ is discerned as an n-dimensional coefficient vector, and all constituent elements are zero save for those associated with class i. The solution for CR Technology was found using the subsequent regularized least squares methodology [23-27]:

$$\hat{p} = \underset{p}{\operatorname{argmin}} \| y - X \cdot p \|_2^2 + \lambda \| p \|_2^2 \tag{13}$$

where, $\| \ \|_2$ is the norm of l_2, and λ is the regularization parameter. Deriving the partial derivative of the above equation and equating it to zero provided:

$$\hat{p} = \left(X^T X + \lambda \cdot I\right)^{-1} X^T y \tag{14}$$

where, I is recognised as the identity matrix. With P=$(X^TX+\lambda \cdot I)^{-1}X^T$, it became evident that P and y are mutually exclusive. Consequently, P was pre-calculated and characterised as a projection matrix. For every class i, the coefficient vector exclusive to class i was denoted as $\hat{P}_i$. Employing solely the elements associated with class i, the estimated value for the test sample y was articulated as $X_i \hat{p}_i$. The residual between y and the estimated values for all categories was computed:

$$r_i = \left\| y - X_i \hat{p}_i \right\|_2 / \left\| \hat{p}_i \right\|_2 \tag{15}$$

Subsequently, y was categorised into the class yielding the minimal residual:

$$\text{Identity}(y) = \arg\min_i \left\{ r_i \right\} \tag{16}$$

The CRC algorithm was presented as follows, as referenced in studies [34-38]:

Algorithm 1: CRC algorithm

Input: Training sample matrix $X = [X_1, X_2, \dots, X_k] \in$

$R^{m\times n}$, one test sample *y*.

Output: Category *Identity*(*y*) of test samples

(1) The dataset X was normalised;

(2) $\hat{p} = Py$ was computed, where $P = (X^T X + \lambda \cdot I)^{-1} X^T$;

(3) Regularised residuals $r_i = ||y - X_i \hat{p}_i||_2 / ||\hat{p}_i||_2$ were calculated;

(4) The category *y* was output, with $Identity(y) = arg\min_i\{r_i\}$.

4. CROSS-MEDIA RETRIEVAL VIA TLSCR

Building on the previously elaborated theory, a novel method for cross-media retrieval, premised on TLSCR, was proposed in this study. To commence, global and local features from both images and texts were extracted. Leveraging these feature representations, two strata of similarity were meticulously crafted and integrated. Subsequent stages encompassed the utilization of all training images for the CR of every test image and the analogous process for text samples. Through the collaboration coefficients derived from the image and text via the CR classifier, a unified dimensional representation was achieved, enabling cross-media retrieval between the image and text. It was observed that such a method facilitated the co-representation of each test sample with all training samples, preserving the distinctive nature of each test sample to a degree.

An algorithmic representation of the cross-media retrieval predicated on TLSCR is detailed below:

Algorithm 2: Cross-media retrieval via TLSCR

Input: Training image set $I_{tr} \in R^{p\times n}$, testing image set $I_{te} \in R^{p\times e}$, training text set $T_{tr} \in R^{q\times n}$, testing text set $T_{te} \in R^{q\times e}$, regularization parameter *λ*, where *m* and *e* are the number denote the counts of training and testing samples respectively, and *p* and *q* are the dimensions of image and text features, respectively.

Output: Image synergy coefficient *α*, text synergy coefficient *β*.

(1) The dataset was normalized as I_{tr}, I_{te}, T_{tr} and T_{te};

(2) For each $i_i \in I_{te}$, $\alpha_i = \left({I_{tr}}^T I_{tr} + \lambda \cdot I\right)^{-1} {I_{tr}}^T i_i$;
ENDFOR

(3) For each $t_j \in T_{te}$, $\beta_j = \left({T_{tr}}^T T_{tr} + \lambda \cdot I\right)^{-1} {T_{tr}}^T t_j$;
ENDFOR

(4) $\alpha=[\alpha_1, \alpha_2, ..., \alpha_i, ..., \alpha_e]$, $\beta=[\beta_1, \beta_2, ..., \beta_i, ..., \beta_e]$.

In Algorithm 2, it was noted that the dimension of the collaboration coefficient was solely contingent upon the number of training samples, devoid of any association with the dimensions of image and text features. As a result, the synergy coefficient between the image and text was derived, leading to a unified dimensional representation for both. Subsequently, an isomorphism of image and text features into a shared subspace was realized, as mathematically represented:

$$I_{te}^{p\times e} \to \alpha^{n\times e} \qquad (17)$$

$$T_{te}^{q\times e} \to \beta^{n\times e} \qquad (18)$$

To culminate the cross-media retrieval task, a similarity measure was instituted. The crux of this measure was to ascertain the distance between each sample in *α* and every sample in *β*. A smaller distance implied heightened similarity between samples.

5. EXPERIMENTAL ANALYSIS: CROSS-MEDIA RETRIEVAL BASED ON TLSCR

5.1 Dataset selection and evaluation indices

To rigorously assess the algorithm articulated in this study, comparisons were drawn with six benchmark cross-media retrieval techniques. These encompass methods grounded in statistical correlation analysis: CCA [39] and JFSSL [40]. Additionally, methods underpinned by deep learning paradigms such as CMDN [41], ACMR [42], DSCMR [43], and SSACR [44] were also incorporated into the comparative framework. Validation was undertaken on two publicly accessible datasets: the Wikipedia dataset and the Pascal Sentence dataset [45].

The Wikipedia dataset, recognised as a preeminent dataset for cross-media retrieval, comprises 2,866 image-text pairings distributed across 10 diverse categories including, but not limited to, history and biology. From this dataset, 2,173 pairs were selected at random for training purposes, whilst the remaining 693 pairs were designated for testing.

The Pascal Sentence dataset, sourced from the 2008 Pascal development kit, encompasses 1,000 images, categorised, on average, into 20 distinct categories. Every image in this dataset was annotated via the Amazon Mechanical Turk platform. Subsequent to this annotation, five unique sentences, each generated by a distinct annotator, were amalgamated to constitute a single document. From this consolidated dataset, a random selection process led to the demarcation of 800 image-document pairings for training, with the resitwo 200 pairings allocated for testing.

In evaluating the efficiency and accuracy of the aforementioned algorithm, two predominant evaluation indices in the realm of cross-media retrieval were utilised: the mean average precision (MAP) and precision recall (PR) [9, 28]. The empirical results unequivocally underscored the superior efficacy of the method detailed in this study.

5.2 Analysis of retrieval results

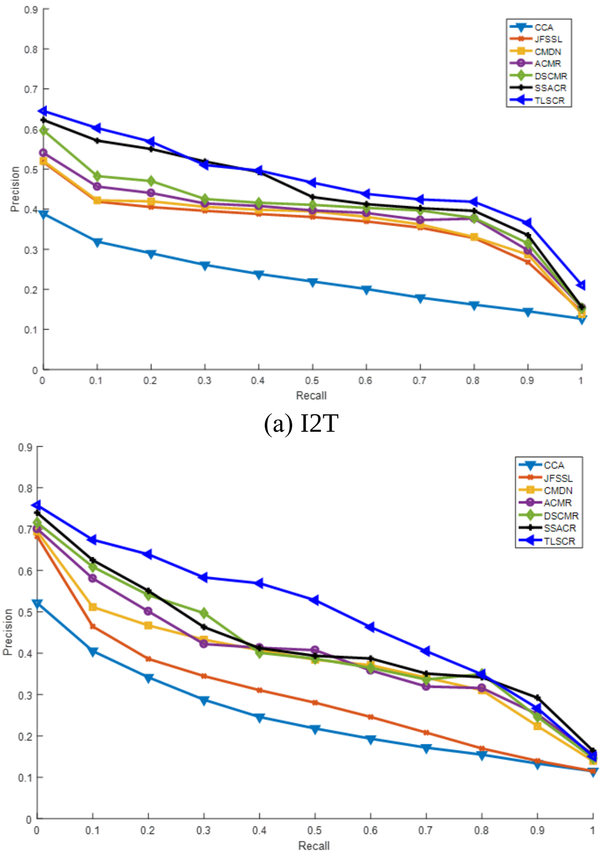
Table 1 presents the MAP values for six benchmark cross-media retrieval methods alongside the performance of the TLSCR methodology when applied to the Wikipedia dataset. Further visual representation can be gleaned from Figure 2, which delineates the PR curve for all seven techniques.

From the acquired results, it is discernible that the TLSCR method surpasses its counterparts in varying retrieval tasks, encompassing image retrieval text, text retrieval image, and the average of both, as well as in the recall.

In Table 2, the MAP values for the seven methodologies are exhibited when subjected to the Pascal Sentence dataset. Similarly, Figure 3 illustrates the PR graphs for these methods under the same dataset.

Table 1. MAP performance on the Wikipedia dataset

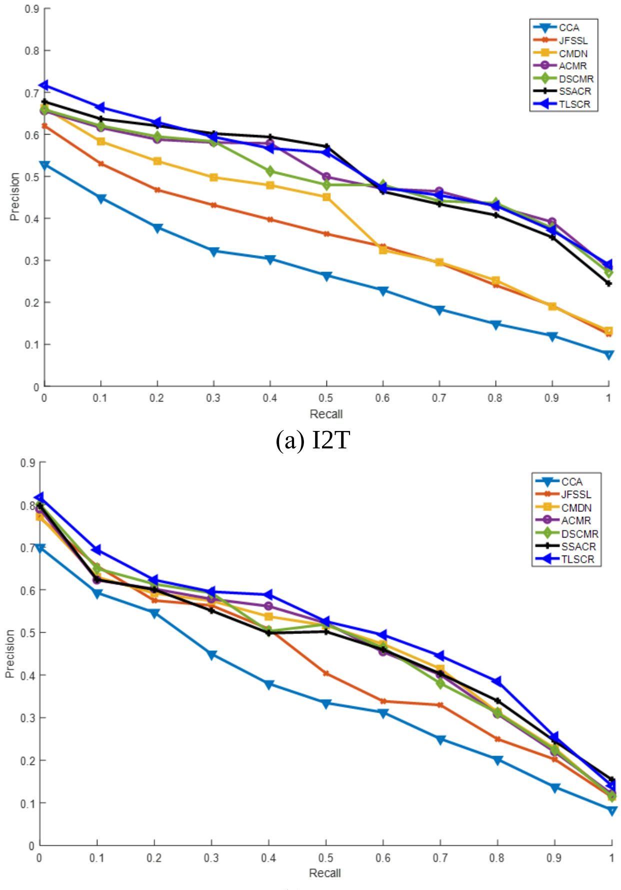
Methods	I2T	T2I	Avg
CCA	0.310	0.316	0.313
JFSSL	0.392	0.381	0.387
CMDN	0.429	0.352	0.391
ACMR	0.513	0.439	0.476
DSCMR	0.506	0.458	0.482
SSACR	0.509	0.461	0.485
TLSCR	**0.532**	**0.481**	**0.507**



(a) I2T

(b) T2I

Figure 2. Recall curve on the Wikipedia dataset



(a) I2T

(b) T2I

Figure 3. Recall curve on the Pascal Sentence dataset

Table 2. MAP performance on the Pascal Sentence dataset

Methods	I2T	T2I	Avg
CCA	0.337	0.439	0.388
JFSSL	0.406	0.401	0.404
CMDN	0.512	0.418	0.465
ACMR	0.638	0.491	0.565
DSCMR	0.644	0.496	0.571
SSACR	0.665	0.493	0.579
TLSCR	**0.701**	**0.506**	**0.604**

Through a careful examination of the results, it becomes evident that the TLSCR method exhibits a marked superiority over the other six traditional techniques across different retrieval operations and in recall indices.

Upon scrutiny of the experimental results, significant advantages of the TLSCR cross-media retrieval method over the other six classical techniques can be ascertained.

6. CONCLUSIONS

In the study presented, a cross-media retrieval approach grounded in TLSCR has been introduced. Initially, a two-level cross-media model was posited, leveraging both global and local representations. This duality was instrumental in encapsulating diverse facets of cross-media relevance learning. Subsequently, two distinct strata of cross-media alignment were proffered, with a meticulous integration scheme devised for the pairing of similarity levels. CR was employed to derive collaborative coefficients, thereby facilitating a dimensionally consistent representation for both image and text. This method has been shown, through rigorous empirical studies on two datasets, to bolster the efficacy of cross-media retrieval considerably.

By maintaining the specific character of individual test samples, this methodology has managed to offer a nuanced yet robust solution to cross-media retrieval challenges. The congruence and effectiveness of this approach underscore its potential for broader applications and further explorations in the realm of cross-media data processing.

REFERENCES

[1] Zhang, L., Ma, B., Li, G., Huang, Q., Tian, Q. (2016). Cross-modal retrieval using multiordered discriminative structured subspace learning. IEEE Transactions on Multimedia, 19(6): 1220-1233. https://doi.org/10.1109/TMM.2016.2646219

[2] Peng, Y.X., Zhu, W.W., Zhao, Y., Xu, C.S., Huang, Q.M., Lu, H.Q., Zheng, Q.H., Huang, T.J., Gao, W. (2017). Cross-media analysis and reasoning: Advances and directions. Frontiers of Information Technology & Electronic Engineering, 18(1): 44-57. https://doi.org/10.1631/FITEE.1601787

[3] Aslam, N., Khan, I.U., Albahussain, T.I., Almousa, N.F., Alolayan, M.O., Almousa, S.A., Alwhebi, M.E. (2022). MEDeep: A deep learning based model for memotion analysis. Mathematical Modelling of Engineering Problems, 9(2): 533-538. https://doi.org/10.18280/mmep.090232

[4] Murugesan, L.J., Chettiar, Shanmugasundaram, S.R.S. (2021). Design and implementation of intelligent classroom framework through light-weight neural

networks based on multimodal sensor data fusion approach. Revue d'Intelligence Artificielle, 35(4): 291-300. https://doi.org/10.18280/ria.350403

[5] Shivanna, P., Venkatesiah, S.S. (2021). Secure multimodal authentication scheme for wireless sensor networks. International Journal of Safety and Security Engineering, 11(6): 653-661. https://doi.org/10.18280/ijsse.110605

[6] 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

[7] Rayavarapu, S.M., Prasanthi, T.S., Kumar, G.S., Rao, G.S., Singham, A. (2023). Employing generative networks for synthetic phonocardiogram and electrocardiogram signal creation: A privacy-ensured approach to data augmentation in heart diagnostics. Ingénierie des Systèmes d'Information, 28(4): 869-875. https://doi.org/10.18280/isi.280408

[8] Alimi, S., Kuyoro, A.O., Eze, M.O., Akande, O. (2023). Utilizing deep learning and SVM models for schizophrenia detection and symptom severity estimation through structural MRI, Ingénierie des Systèmes d'Information, 28(4): 993-1002. https://doi.org/10.18280/isi.280419

[9] Rasiwasia, N., Costa Pereira, J., Coviello, E., Doyle, G., Lanckriet, G.R., Levy, R., Vasconcelos, N. (2010). A new approach to cross-modal multimedia retrieval. In Proceedings of the 18th ACM International Conference on Multimedia, pp. 251-260. https://doi.org/10.1145/1873951.1873987

[10] Rosipal, R., Krämer, N. (2005). Overview and recent advances in partial least squares. In International Statistical and Optimization Perspectives Workshop "Subspace, Latent Structure and Feature Selection", pp. 34-51. https://doi.org/10.1007/11752790_2

[11] Gong, Y., Ke, Q., Isard, M., Lazebnik, S. (2014). A multi-view embedding space for modeling internet images, tags, and their semantics. International Journal of Computer Vision, 106: 210-233. https://doi.org/10.1007/s11263-013-0658-4

[12] Sharma, A. (2012). Generalized multiview analysis: A discriminative latent space. IEEE Conference on Computer Vision and Pattern Recognition, IEEE Computer Society, pp. 2160-2167.

[13] Hardoon, D.R., Szedmak, S., Shawe-Taylor, J. (2004). Canonical correlation analysis: An overview with application to learning methods. Neural Computation, 16(12): 2639-2664. https://doi.org/10.1162/0899766042321814

[14] Andrew, G., Arora, R., Bilmes, J., Livescu, K. (2013). Deep canonical correlation analysis. In International Conference on Machine Learning, pp. 1247-1255.

[15] Gupta, S.K., Phung, D., Adams, B., Tran, T., Venkatesh, S. (2010). Nonnegative shared subspace learning and its application to social media retrieval. In Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1169-1178. https://doi.org/10.1145/1835804.1835951

[16] Yu, S., Yu, K., Tresp, V., Kriegel, H.P. (2006). Multi-output regularized feature projection. IEEE Transactions on Knowledge and Data Engineering, 18(12): 1600-1613. https://doi.org/10.1109/TKDE.2006.194

[17] Ando, R.K., Zhang, T. (2005). A framework for learning predictive structures from multiple tasks and unlabeled data. Journal of Machine Learning Research, 6(3): 1817-1853.

[18] Argyriou, A., Evgeniou, T., Pontil, M. (2008). Convex multi-task feature learning. Machine Learning, 73: 243-272. https://doi.org/10.1007/s10994-007-5040-8

[19] Kong, X., Ng, M.K., Zhou, Z.H. (2011). Transductive multilabel learning via label set propagation. IEEE Transactions on Knowledge and Data Engineering, 25(3): 704-719. https://doi.org/10.1109/TKDE.2011.141

[20] Amit, Y., Fink, M., Srebro, N., Ullman, S. (2007). Uncovering shared structures in multiclass classification. In Proceedings of the 24th International Conference on Machine Learning, pp. 17-24. https://doi.org/10.1145/1273496.1273499

[21] Xie, P., Xing, E.P. (2013). Multi-modal distance metric learning. International Joint Conference on Artificial Intelligence, pp. 1806-1812.

[22] Quan, X.Z., Chen, J. (2021). Multi-source data fusion and target tracking of heterogeneous network based on data mining. Traitement du Signal, 38(3): 663-671. https://doi.org/10.18280/ts.380313

[23] Faghri, F., Fleet, D.J., Kiros, J.R., Fidler, S. (2017). Vse++: Improving visual-semantic embeddings with hard negatives. arxiv preprint arxiv:1707.05612. https://arxiv.org/abs/1707.05612

[24] Huang, Y., Wu, Q., Song, C., Wang, L. (2018). Learning semantic concepts and order for image and sentence matching. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6163-6171.

[25] Lee, K.H., Chen, X., Hua, G., Hu, H., He, X. (2018). Stacked cross attention for image-text matching. In Proceedings of the European Conference on Computer Vision (ECCV), pp. 201-216.

[26] He, K., Zhang, X., Ren, S., Sun, J. (2016). Deep resitwo learning for image recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770-778.

[27] Zhang, H., Goodfellow, I., Metaxas, D., Odena, A. (2019). Self-attention generative adversarial networks. In International Conference on Machine Learning, pp. 7354-7363.

[28] Zhang, X., Zhao, J., LeCun, Y. (2015). Character-level convolutional networks for text classification. Advances in Neural Information Processing Systems, 28: 649-657.

[29] Hochreiter, S., Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8): 1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735

[30] Ren, S., He, K., Girshick, R., Sun, J. (2015). Faster R-CNN: Towards real-time object detection with region proposal networks. Advances in Neural Information Processing Systems, 28: 91-99.

[31] Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L. (2018). Bottom-up and top-down attention for image captioning and visual question answering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6077-6086.

[32] Huang, Z., Xu, W., Yu, K. (2015). Bidirectional LSTM-CRF models for sequence tagging. http://arxiv.org/abs/1508.01991

[33] Hoffer, E., Ailon, N. (2015). Deep metric learning using triplet network. In Similarity-Based Pattern Recognition: Third International Workshop, SIMBAD 2015, Copenhagen, Denmark, October 12-14, 2015. Proceedings 3, pp. 84-92. https://doi.org/10.1007/978-3-319-24261-3_7

[34] Wright, J., Yang, A.Y., Ganesh, A., Sastry, S.S., Ma, Y. (2008). Robust face recognition via sparse representation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 31(2): 210-227. https://doi.org/10.1109/TPAMI.2008.79

[35] Zhang, L., Yang, M., Feng, X. (2011). Sparse representation or collaborative representation: Which helps face recognition? In 2011 International Conference on Computer Vision, pp. 471-478. https://doi.org/10.1109/ICCV.2011.6126277

[36] Sun, Z., Hu, Z., Wang, M., Zhao, S. (2017). Adaptive joint block-weighted collaborative representation for facial expression recognition. Turkish Journal of Electrical Engineering and Computer Sciences, 25(5): 3699-3712. https://doi.org/10.3906/elk-1606-115

[37] Akhtar, N., Shafait, F., Mian, A. (2017). Efficient classification with sparsity augmented collaborative representation. Pattern Recognition, 65: 136-145.

[38] Sui, P., Guo, Y., Zhang, K.F., Li, H. (2017). Frequency-hopping transmitter fingerprint feature classification based on kernel collaborative representation classifier. Wireless Communications and Mobile Computing, 2017: 9403590. https://doi.org/10.1155/2017/9403590

[39] Hardoon, D.R., Szedmak, S., Shawe-Taylor, J. (2004). Canonical correlation analysis: An overview with application to learning methods. Neural Computation, 16(12): 2639-2664. https://doi.org/10.1162/0899766042321814

[40] Wang, K., He, R., Wang, L., Wang, W., Tan, T. (2015). Joint feature selection and subspace learning for cross-modal retrieval. IEEE Transactions on Pattern Analysis and Machine Intelligence, 38(10): 2010-2023. https://doi.org/10.1109/TPAMI.2015.2505311

[41] Peng, Y., Huang, X., Qi, J. (2016). Cross-media shared representation by hierarchical learning with multiple deep networks. In Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI '16). Palo Alto, CA: AAAI, pp. 3846-3853.

[42] Wang, B., Yang, Y., Xu, X., Hanjalic, A., Shen, H.T. (2017). Adversarial cross-modal retrieval. In Proceedings of the 25th ACM International Conference on Multimedia, pp. 154-162. https://doi.org/10.1145/3123266.3123326

[43] Zhen, L., Hu, P., Wang, X., Peng, D. (2019). Deep supervised cross-modal retrieval. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10394-10403.

[44] Liu, C., Du, J., Zhou, N. (2021). A cross media search method for social networks based on adversarial learning and semantic similarity. Science China Information Sciences, 51(5): 779-794.

[45] Wei, Y., Zhao, Y., Zhu, Z., Wei, S., Xiao, Y., Feng, J., Yan, S. (2016). Modality-dependent cross-media retrieval. ACM Transactions on Intelligent Systems and Technology (TIST), 7(4): 1-13. https://doi.org/10.1145/2775109