International Journal of Mathematical, Engineering and Management Sciences
Vol. 7, No. 6, 964-981, 2022
https://doi.org/10.33889/IJMEMS.2022.7.6.060

Impact of Binary-Valued Representation on the Performance of Cross-Modal Retrieval System

Nikita Bhatt
U & P U. Patel Department of Computer Engineering,
CSPIT, CHARUSAT, Gujarat, India.
Corresponding author: nikitabhatt.ce@charusat.ac.in

Amit Ganatra
Devang Patel Institute of Advance Technology and Research,
CHARUSAT, Gujarat, India.
E-mail: amitganatra.ce@charusat.ac.in

Nirav Bhatt
Smt. Kundanben Dinsha Patel Department of Information Technology,
CSPIT, CHARUSAT, Gujarat, India.
E-mail: niravbhatt.it@charusat.ac.in

Purvi Prajapati
Smt. Kundanben Dinsha Patel Department of Information Technology,
CSPIT, CHARUSAT, Gujarat, India.
E-mail: purviprajapati.it@charusat.ac.in

Mrugendra Rahevar
U & P U. Patel Department of Computer Engineering,
CSPIT, CHARUSAT, Gujarat, India.
E-mail: mrugendrarahevar.ce@charusat.ac.in

Martin Parmar
U & P U. Patel Department of Computer Engineering,
CSPIT, CHARUSAT, Gujarat, India.
E-mail: martinparmar.ce@charusat.ac.in

(Received on April 09, 2022; Accepted on September 06, 2022)

Abstract
The tremendous proliferation of Multi-Modal data and the flexible need of users has drawn attention to the field of Cross-Modal Retrieval (CMR), which can perform image-sketch matching, text-image matching, audio-video matching and near infrared-visual image matching. Such retrieval is useful in many applications like criminal investigation, recommendation systems and person reidentification. The real challenge in CMR is to preserve semantic similarities between various modalities of data. To preserve semantic similarities, existing deep learning-based approaches use pairwise labels and generate binary-valued representation. The generated binary-valued representation provides fast retrieval with low storage requirement. However, the relative similarity between heterogeneous data is ignored. So, the objective of this work is to reduce the modality-gap by preserving relative semantic similarities among various modalities. So, a model named "Deep Cross-Modal Retrieval (DCMR)" is proposed, which takes triplet labels as the input and generates binary-valued representation. The triplet labels locate semantic similar data points nearer and dissimilar points far in the vector space. Extensive experiments are performed and the result is compared with deep learning-based approaches, which shows that the performance of DCMR increases by 2% to 3% for Image→Text retrieval and by 2% to 5% for Text→Image retrieval in mean average precision (mAP) on MSCOCO, XMedia, and NUS-WIDE datasets. So, the binary-valued representation generated from triplet labels preserve better relative semantic similarities than pairwise labels.

Keywords- Information retrieval, Multi-modal data, VGG-F network, Glove, Multi-layer perceptron (MLP), Mean average precision (MAP).

1. Introduction

In recent years, Artificial Intelligence (AI) is on everyone's lips, in the news, in the industries, and in politics as it implies societal, economic, and cultural challenges (Fast and Horvitz, 2017). The objective of AI is to develop systems that facilitate human life, which includes major advances like autonomous vehicles (Chabot et al., 2017), household robots (Brodeur et al., 2017), and decision support in the medical field (Litjens et al., 2017) and much more. Among the sub-fields of AI, Machine Learning (ML) has been impacted dramatically over the last thirty years. Nowadays, the growth of the internet and social media has generated a huge amount of visual data, which needs computation power for retrieval. In recent times, information is available from a collection of resources. For example, a Facebook post or any real-world article contains not only text but also contains an image, video, audio, etc. Such data is called Multi-Modal data. Figure 1 shows an example of Multi-Modal data from the XMedia dataset (Peng et al., 2016, 2018), where image and text modalities are used to explain a topic, which has a strong semantic correlation.

"A stretch of the river passes through the Hanford Site, established in 1943 as part of the Manhattan Project. The site served as a plutonium production complex, with nine nuclear reactors and related facilities located on the banks of the river".

"Until April, the Polish forces had been slowly but steadily advancing eastward. The new Latvian government requested and obtained Polish help in capturing Daugavpils".

"The barn owl is the most widely distributed species of owl. The barn owl is found almost everywhere in the world except polar and desert regions, Asia north of the Himalayas, most of Indonesia, and some Pacific islands".

Figure 1. Examples of multi-modal data from XMedia dataset.

One can easily associate vision with language and vice versa; but it is difficult by the Information Retrieval (IR system) (Kiros et al., 2014; Vendrov et al., 2016a; Yanagi et al., 2020; Zhen et al., 2019). The objective of IR is to obtain information from various resources, which is relevant to a query. The growth of deep learning has achieved a lot of success with a single modality of data. However, single modality-based searching and retrieval techniques are not applied to Multi-Modal data as different modalities have different feature representations (Jiang and Li, 2017; Xu et al., 2017). The searching and retrieval techniques for handling Multi-Modal data must store, organize and handle a variety of modalities like - depth, RGB, photo, sketch, text, visual images, etc. In recent times, Cross-Modal Retrieval (CMR) from Multi-Modal data is widely used for the task of object detection and sequence modelling. Figure 2 shows the output of CMR system, which retrieves semantic categories associated with a query image and retrieves all the documents associated with it. Figure 3 shows an example of a CMR system, where image modality is the input and text/audio modality are the output.

Query: Image Query

Semantic Categories: Flower, nature, insect, animal.

Query: Image Query

Semantic Categories: Stone, sculpture, Futuredeity, devi, kanyakumari, tamilnadu.

Figure 2. Examples of semantic categories associated with an image.

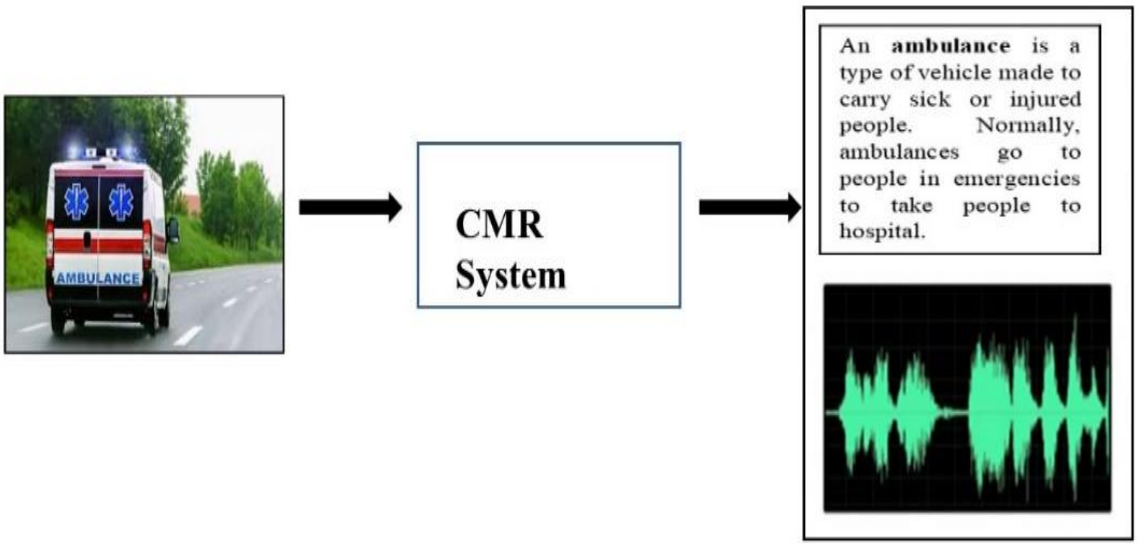

Figure 3. An example of a CMR system.

The comprehensive and flexible retrieval from the CMR system performs text-image matching, sketch-image matching, audio-video matching, and near infrared-visual image matching, which can be useful in various applications like criminal investigations, recommendation systems, person reidentification, etc. The entire CMR system as shown in Figure 4 is separated into the training and testing phase.

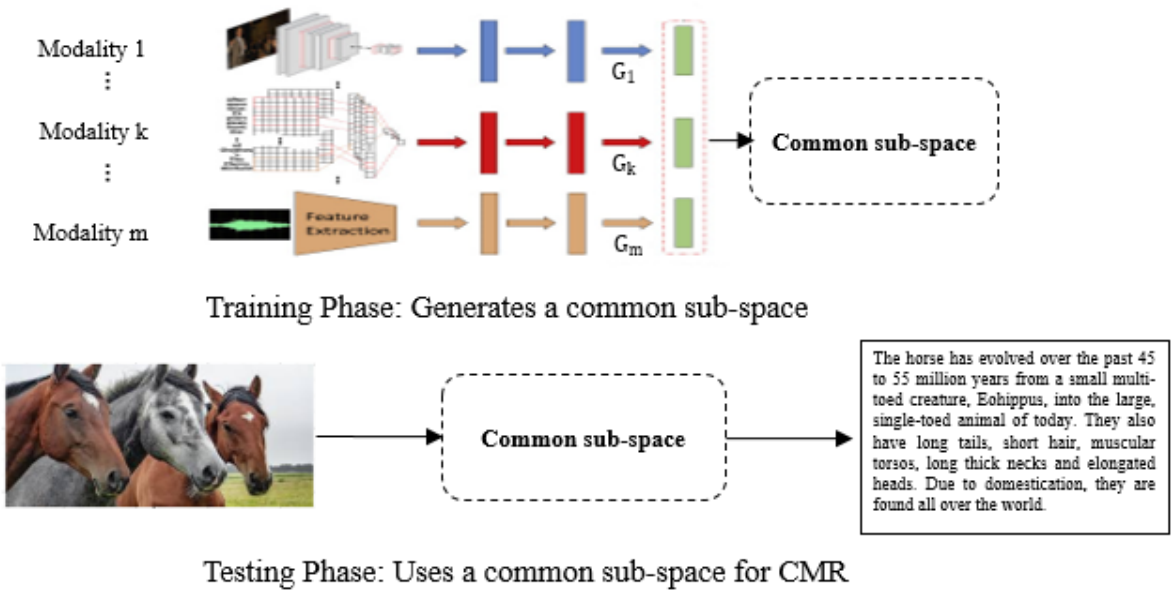

Figure 4. Phases of cross-modal retrieval system.

The training phase generates a modality-invariant common sub-space, which has an encoding form of information called binary-valued representation or actual information called real-valued representation (Cao et al., 2016; Duan et al., 2017; Wang et al., 2021). The testing phase of the CMR system uses a common sub-space for retrieval. The state-of-the-art methods generate a common sub-space by performing independent feature learning and correlation learning, which does not lead to satisfactory performance (Ding et al., 2014; Hu et al., 2019; Lin et al., 2014; Song et al., 2013; Zhang and Li, 2014). However, the capability of deep networks is widely used for the task of CMR in recent times. The existing CMR methods use different vectorization methods to map words into vectors. The local representation method generates a sparse vector, which does not preserve semantic similarities between different words. On the other hand, distributional representation generates a dense vector, which preserves semantic similarities between the words. Such representation is widely used in the field of IR. Sometimes, the user expresses the query with certain keywords, but documents have the same concept with different words. Such words are similar at the semantic level but different from the surface level, which is called a term-mismatch problem. One of the issues in the field of IR is handling the term-mismatch problem (Huang et al., 2012; Mikolov et al., 2013a, 2013b).

The objective of distributional representation is to solve the term-mismatch problem, by finding semantic similar words to a user query. The generated words are appended to a user query and an extended query is given to the search engine for retrieval. There are many methods to generate distributional representation and the selection of an appropriate method is one of the open issues in the field of NLP (Kiros et al., 2014; Mikolov et al., 2011; Wu et al., 2010). Further, like standard algorithms, CMR algorithms can be trained under (a) supervised and (b) unsupervised scenarios. The supervised algorithms use label information to generate a common sub-space, while unsupervised algorithms do not use any label information. The supervised information in deep learning-based CMR methods is in terms of pairwise labels or triplet labels. Existing deep learning-based CMR methods use pairwise labels to generate a common sub-space, which minimizes the hamming distance between corresponding data points (Cao et al., 2016; Jiang and Li, 2017; Kong and Li, 2012; Kumar and Udupa, 2011; Li et al., 2015; Liu et al., 2019; Wang et al., 2021; Yang et

al., 2017). However, the generated binary-valued representation does not preserve relative similarities among various modalities of data. So, there is a need to propose a model, which preserves relative similarities between different modalities of data. Below is the summary of our work.

- The proposed model named "Deep Cross-Modal Retrieval (DCMR)" uses hashing method for the generation of binary-valued representation, which leads to less time for retrieval. The DCMR preserves relative similarities between heterogeneous modalities using triplet labels, which put the query instance nearer to the positive instance and far from the negative instance in the vector space.
- The generation of triplet labels are computationally costly, which is resolved by generating different groups from the similarity matrix. The similarity matrix is generated based on the label information associated with each data point of image and text modality.
- The selection of the training model for each modality impacts a lot on the performance of the system. The DCMR has adopted Glove model for text modality, which preserves semantic similarities between words and VGG-F network for the image modality. The objective function of DCMR preserves intra-modal and inter-modal triplet loss to enhance the performance of the system. The intra-modal triplet loss preserves relative similarities within the modality and inter-modal triplet loss preserves relative similarities between heterogeneous data points.
- The experiments are performed and result is compared with deep learning-based pairwise approaches in terms of mean average precision (mAP). It concludes that DCMR preserves relative similarities between heterogeneous data points and increases performance by 2% to 3% for Image→Text retrieval and by 2% to 5% for Text→Image retrieval tasks.

2. Literature Survey

The CMR becomes one of the demanding topics in the field of IR. The hashing techniques are used to generate binary-valued representation. The hashing techniques are classified as (a) Uni-Modal Hashing (UMH) and (b) Multi-Modal Hashing (MMH). The MMH is more popular due to growth of Multi-Modal data. In recent times, various deep networks are widely used to generate a common sub-space. The popular model was proposed by Rajagopalan et al. (2016), where long short-term memory (LSTM) is used to generate joint representation, which can be useful for multi-view data to perform human behavior analysis. Further, the boosting method has become a very successful ensemble learning technology, which is limited to a single modality. However, the author Wang et al. (2019) has proposed a multi-modal boosting framework called MMBoost, which deals with heterogeneous modalities. The framework captures intra-modal and inter-modal semantic correlation at the same time.

Motivated by the great power and success of deep learning, a variety of approaches have been proposed to generate a common sub-space. These approaches perform simultaneous feature learning and correlation learning in the same framework. In comparison with shallow architectures, deep networks can learn better latent feature representation and capture better semantic information of data. The correlation between Multi-Modal data is learned using a Deep auto-encoder (DAE) and a common sub-space is generated using a Restricted Boltzmann Machine (RBM) in an unsupervised way(Ngiam et al., 2011). The graphical-based model called Deep Boltzmann machine (DBM) is used, which does not need supervised data for training, and each layer of the Boltzmann machine adds more level of abstract information (Srivastava and Salakhutdinov, 2012). Further, the boosting method has become a very successful ensemble learning technology, which is limited to a single modality. However, the author Wang et al. (2019) has proposed a multi-modal boosting framework called MMBoost, which deals with heterogeneous modalities. The framework captures intra-modal and inter-modal semantic correlation at the same time. Still, an ensemble learning approach needs to be explored in the field of CMR.

In recent years, coordinated representation-based methods extract modality-specific features and are coordinated by correlation measures like cosine similarity, and distance function, which are widely used in Multi-Modal retrieval, translation, and zero-shot learning (Ranjan et al., 2015). Like standard algorithms, coordinated representation-based CMR methods can typically be trained under supervised and unsupervised scenarios. Because it can achieve greater performance than unsupervised hashing, supervised hashing has gotten a lot of interest (Cao et al., 2016; Hua et al., 2016; Wang et al., 2017; Xu et al., 2017). The supervised information can be given in two different forms: pairwise labels and triplet-based labels. The CMR approaches based on pairwise labels preserve similarity between corresponding data points. Deep pairwise-supervised hashing (DPSH) (Li et al., 2015), Pairwise relationship-guided deep hashing (PRDH) (Yang et al., 2017), and Deep Visual Semantic Hashing (DVSH) (Cao et al., 2016) perform simultaneous feature learning and hash-code generation. Deep Cross-Modal hashing (DCMH) (Jiang and Li, 2017) is another pairwise deep learning-based CMR method, which uses Bag-of-words (BoW) model for text modality and a deep network for image modality. Deep Supervised Cross-Modal Retrieval (DSCMR) (Zhen et al., 2019) was proposed, which has adopted the word2vec model for text modality and generates real-valued representation. Ranking-based deep cross-modal hashing (RDCMH) (Liu et al., 2019a) is proposed, where the modality gap is bridged using label and feature information from different modalities, and a common sub-space is generated by introducing an adversarial modality discriminator. Further, Semantic-Preserving Hashing based on Multi-scale Fusion (SPHMF) was proposed (Zhang and Pan, 2021), which preserves the semantic similarities between different words using the Text Pooling Model for Multi-scale Fusion (TPMSF). a model was proposed for image and text modality, which uses pairwise labels to generate real-valued representation and preserves similarities between corresponding data points (Bhatt and Ganatra, 2021). All above approaches preserve similarity between corresponding data points but fail to preserve relative similarity between heterogeneous modalities. On the other hand, triplet labels contain query instances, similar instances, and dissimilar instances, which puts query instances nearer to similar instances and far from negative instances. The Deep Supervised Hashing with Triplet Labels (Lai et al., 2015; Wang et al., 2016) and Deep Triplet Quantization (DTQ) (Liu et al., 2019) are proposed, which use triplet labels to preserve relative similarity between data points. However, these approaches preserve the relative similarities between instances of image modality only.

3. Proposed Model

The proposed model named "Deep Cross-Modal Retrieval (DCMR)" generates a common sub-space using triplet labels, which preserves relative similarity between heterogeneous data points. The triplet label has a query instance from text modality, a positive (similar) instance, and a negative (dissimilar) instance from image modality, as shown in Figure 5. The DCMR has five main components (a) The triplet labels are given as input to the deep network, which is a selection from similarity matrix S. (b) Each modality has a separate training model. The VGG-F Network is used for learning deep representations from image modality (c) The Glove model is used as an embedding technique to map words into vectors. (d) The objective function of DCMR reduces the intra-modal and inter-modal triplet loss to preserve relative similarities between heterogeneous data points. The triplet loss is used for pulling together similar pairs (bold line) pushing away dissimilar pairs (dotted line); and (e) Generation of binary-valued coordinated representation B.

3.1 Problem Formulation and Proposed Architecture

There is a collection of n training instances denoted by triplet labels. For each instance, represents the text feature vector and represents the image feature vector, where d and f are the dimensionalities of text and image modalities, respectively. Each pair of image and text modality has label information associated with it, which generates similarity matrix S. The common sub-space is generated using triplet labels, which are

selected from S. The triplet label is selected from the group, where is a query instance, is a similar instance and is a dissimilar instance. The set of similar instances is denoted by, and the set of dissimilar instances is denoted by. Given triplet label as the input, deep network for image modality and text modality learn representation. Moreover, and are a text feature matrix, image feature matrix, and similarity matrix of all the training instances. The binary hash code is generated using and for text and image modalities, respectively, where b is the length of the binary code. Further, binary code is generated using and for text and image modality, respectively. The hash functions generate binary code such that where represents the Hamming distance between binary code, contains binary code of query instance, contains binary code of similar instance and contains binary code of dissimilar instances.

The DCMR works for image and text modalities and a separate network are adopted for each modality. Due to large Multi-Modal datasets, a huge amount of triplet labels is generated and it is computationally costly to work with all the triplets. The problem is resolved by adopting the idea, that several groups are created randomly from similarity matrix S for the selection of triplet labels (Ding et al., 2014; Kiros et al., 2014). Both similar and dissimilar instances of image modality are given to the image network. The image network adopts VGG-F (Zhang and Li, 2014) architecture due to promising performance achieved in the domain of computer vision, which is pre-trained on the ImageNet dataset (Lin et al., 2015). The last layer of VGG-F uses the identity function and the remaining layers use Rectified Linear Unit (ReLU) as an activation function. The ReLU activation function has the property of differentiable, which provides optimization in the VGG-F network. The last fully-connected layer (fc8) is replaced with a fully connected hash (fch), which maps learned image features into binary hash code using the sign (.) function. The sign function outputs 1 for similar instances and -1 for dissimilar instances. The DCMR adopts Glove as a vectorization method for text modality, which puts semantic similar words nearer to each other in the vector space. The generated vectors from Glove are given as input to Multi-Layer Perceptron (MLP) to extract textual features. The MLP has three fully connected layers (*fc*), where all possible connections are from one layer to other layers, but the within-layer connection is not possible. The MLP has used mini-batch SGD, which uses various data points and performs the derivation to update the weights. The activation function used by MLP is ReLU, which is defined as. The advantage of ReLU over other activation functions is that it does not suffer from a vanishing gradient. Another reason for using ReLU is that derivation of ReLU is simpler, which is defined as

$$\frac{\partial f_{ReLU}}{\partial z} = \begin{cases} 0 & if\ z < 0 \\ 1 & if\ z > 0 \end{cases} \quad (1)$$

The last layer is replaced with a fully connected hash (fch), which mapped text features into binary hash code in the same way as image modality. The generated binary code for each instance reduces intra-modal and inter-modal triplet loss and generates binary-valued representation. The DCMR contains two main components: (a) Feature Learning and (b) Generation of binary-valued representation, which is discussed in the next section.

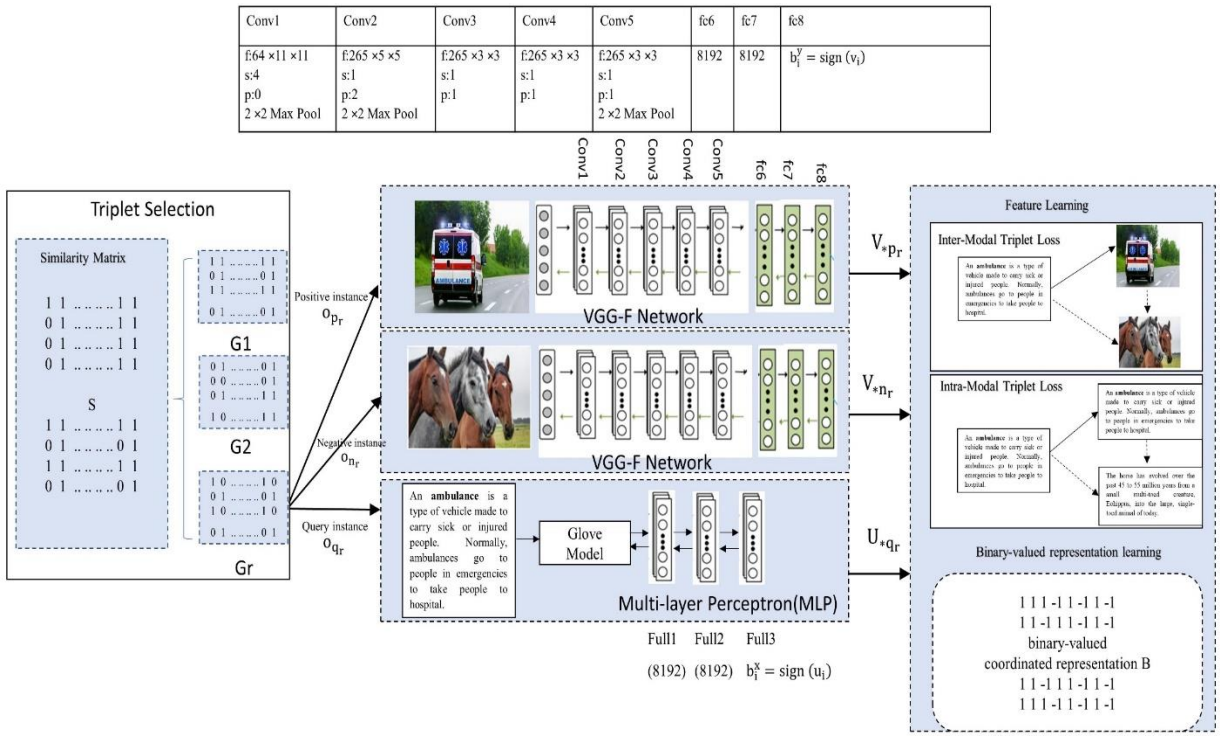

Figure 5. The proposed model (DCMR) has five main components (a) triplet selection from similarity matrix S (b) a standard VGG-F Network for learning deep representations from image modality (c) a Glove model as an embedding technique to map words into vectors (d) triplet loss for pulling together similar pairs (bold line) pushing away dissimilar pairs (dotted line); and (e) Generation of binary-valued coordinated representation B.

3.2 Feature Learning

The triplet labels preserve relative semantic similarity between heterogeneous data points. $U \in \mathbb{R}^{d\times n}$ Contains text representation from the MLP, $V \in \mathbb{R}^{f\times n}$ contains image representation from VGG-F network and τ is triplet labels. Suppose query instance is a text and similar and dissimilar instances are from image modality. So, triplet label likelihood is defined as Eq. (2),

$$p(\tau|U,V,V) = \prod_{r=1}^{R}(p\big((q_r,p_r,n_r)\big|U,V,V\big) \quad (2)$$

where,

$$p\big((q_r,p_r,n_r)\big|U,V,V\big) = \sigma(\theta_{q_r^x p_r^y} - \theta_{q_r^x n_r^y} - \alpha)$$

Here σ is a sigmoid function, which is defined using $\sigma(x) = \frac{1}{1+e^{-x}}$. $\theta_{q_r^x p_r^y} = \frac{1}{2}U_{*q_r}^T V_{*p_r}$ and $\theta_{q_r^x n_r^y} = \frac{1}{2}U_{*q_r}^T V_{*n_r}$ .The threshold α is a margin parameter, which is enforced between similar and dissimilar instances of image and text modalities. Given a query instance, the objective function should minimize inter-modal and intra-modal triplet loss for similar instances and maximize inter-modal and intra-modal triplet loss for dissimilar instances. The inter-modal triplet loss with heterogeneous data points for image → text and text → image retrieval is defined using Eq. (3) and Eq. (4),

$$J_1 = -\log p(\tau|U,V,V)$$
$$= -\sum_{r=1}^{R} \log p\left((q_r,p_r,n_r)\middle|U,V,V\right)$$
$$= -\sum_{r=1}^{R}(\theta_{q_r^x p_r^y} - \theta_{q_r^x n_r^y} - \alpha - \log(1+e^{\theta_{q_r^x p_r^y} - \theta_{q_r^x n_r^y} - \alpha}) \quad (3)$$

$$J_2 = -\log p(\tau|V,U,U)$$
$$= -\sum_{r=1}^{R} \log p\left((q_r,p_r,n_r)\middle|V,U,U\right)$$
$$= -\sum_{r=1}^{R}(\theta_{q_r^y p_r^x} - \theta_{q_r^y n_r^x} - \alpha - \log(1+e^{\theta_{q_r^y p_r^x} - \theta_{q_r^y n_r^x} - \alpha}) \quad (4)$$

The total inter-modal triplet loss is preserved by,

$$J_{inter-modal} = J_1 + J_2$$

(a) Intra-Modal Triplet Loss: The intra-modal semantic similarity for image modality is defined as Eq. (5),

$$J_3 = -\log p(\tau|U)$$
$$= -\sum_{r=1}^{R} \log p\left((q_r,p_r,n_r)\middle|U\right)$$
$$= -\sum_{r=1}^{R}(\theta_{q_r^x p_r^x} - \theta_{q_r^x n_r^x} - \alpha - \log(1+e^{\theta_{q_r^x p_r^x} - \theta_{q_r^x n_r^x} - \alpha}) \quad (5)$$

Similarly, the intra-modal semantic similarity for text modality is defined as Eq. (6),

$$J_4 = -\log p(\tau|V)$$
$$= -\sum_{r=1}^{R} \log p\left((q_r,p_r,n_r)\middle|V\right)$$
$$= -\sum_{r=1}^{R}(\theta_{q_r^y p_r^y} - \theta_{q_r^y n_r^y} - \alpha - \log(1+e^{\theta_{q_r^y p_r^y} - \theta_{q_r^y n_r^y} - \alpha}) \quad (6)$$

The intra-modal triplet loss is preserved by,

$$J_{intra-modal} = J_3 + J_4$$

The feature learning stage preserves relative similarity between heterogeneous data points.

3.3 Generation of Binary-Valued Representation

The binary-valued representation is generated by the sign(.) function, which is denoted by B^X and B^Yrespectively.

$$B^X = \text{sign}(U) \text{ and } B^Y = \text{sign}(V)$$

Both B^x and B^y must preserve the similarity with U and V respectively, denoted as Eq. (7),

$$\gamma\left(\left\|B^{(X)}-U\right\|_F^2+\left\|B^{(Y)}-V\right\|_F^2\right) \quad (7)$$

where, F is the Frobenius norm of a matrix and γ is the hyper-parameter used to balance the weight of each part. For training points, both BX and BY is the same which is represented as:

$$B = B^X = B^Y$$

Eq. (7) can be rewritten as Eq. (8),

$$J_{binary} = \gamma\left(||B-U||_F^2+||B-V||_F^2+\eta\left(\|U\|_F^2+\|V\|_F^2\right)\right. \quad (8)$$

where, η is a hyper-parameter used to balance each bit of hash code. The generated binary code preserves similarities with feature representations. The final objective function for DCMR is denoted as Eq. (9),

$$\min_{B,\theta_x,\theta_y} J = \min_{B,\theta_x,\theta_y} J_{inter-modal}+J_{intra-modal}+\gamma\left(\|B-U\|_F^2+\|B-V\|_F^2\right)+\eta\left(\|U\|_F^2+\|V\|_F^2\right) \quad (9)$$

4. Learning Algorithm

The network parameters (θ_x, θ_y) and binary matrix B is optimized by the objective function of DCMR using a mini-batch SGD algorithm. Here, only one parameter is updated by keeping two parameters fixed.

(I)Updating B

The parameters θ_x and θ_y are fixed and parameter B is updated until the model is optimized or maximum iterations are reached. The objective function in Eq. (10) is expanded while fixing θ_x and θ_y as follows:

$$\min_B \gamma tr(B^TB-UB^T-VB^T) \quad (10)$$

where, $B \in \{-1,1\}^{c\times n}$

The derivation of Eq. (10) with respect to B is defined as,

$$B = sign(U+V)$$

(II)Updating θ_x

The parameters B and θ_y are fixed and parameter θ_x is updated using the backpropagation algorithm.

$$\frac{\partial J}{\partial U_{*i}} = \frac{\partial J_{inter-modal}}{\partial U_{*i}}+\frac{\partial J_{int-modal}}{\partial U_{*i}}+\frac{\partial J_{binary}}{\partial U_{*i}}$$
$$= -\frac{1}{2}\sum_{r:(i,p_r,n_r)}^{R}(1-\sigma(\theta_{ip_r^y}-\theta_{in_r^y}-\alpha))(V_{*p_r}-V_{*n_r})+2\gamma(U-B)+2\eta U1 \quad (11)$$

(III)Updating θ_y

The parameters B and θ_x are fixed and parameter θ_y is updated using the backpropagation algorithm.

$$\frac{\partial J}{\partial V_{*i}} = \frac{\partial J_{inter-modal}}{\partial V_{*i}}+\frac{\partial J_{int-modal}}{\partial V_{*i}}+\frac{\partial J_{binary}}{\partial V_{*i}}$$
$$= -\frac{1}{2}\sum_{r:(i,p_r,n_r)}^{R}(1-\sigma(\theta_{ip_r^x}-\theta_{in_r^x}-\alpha))(U_{*p_r}-U_{*n_r})$$

$$-\frac{1}{2}\sum_{r:(i,p_r,n_r)}^{R}(1-\sigma(\theta_{ip_r^y}-\theta_{in_r^y}-\alpha))(V_{*p_r}-V_{*n_r})\ +\ 2\gamma\ (V-B)+2\eta V1 \quad (12)$$

Once the training is over, network parameters θ_x and θ_y are optimized for text and image modality, respectively. Given any query image x_q, feature representation is performed using image network f and it is represented as,

$$U_q = f(x_q, \theta_x)$$

The encoding is performed on learned image representation by,

$$b_q = \text{sign}(U_q)$$

The hamming distance is calculated between a query point and database points. The DCMR retrieves all the documents from the text modality, given image as a query. The below section provides combative result analysis with existing deep learning-based methods.

4.1 Evaluation Measure, Hyper-Parameter Settings and Result Discussion

The experiments are conducted on well-known Multi-modal datasets like NUS-WIDE, MSCOCO and XMedia. The MSCOCO, NUS-WIDE and XMedia datasets have 349000 images, 269648 images and 8000 images, respectively with associated captions. The mean average precision (mAP) is selected as a statistical measure, similar to most of the previous work (Cao et al., 2016; Jiang and Li, 2017; Wang et al., 2021; Yang et al., 2017). The selection of hyper-parameters affects a lot in the performance of the system. In DCMR, γ and η are hyperparameters. The Grid search is the most basic hyperparameter tunning method, which is used in this research to fine-tune the hyper-parameter. Experiments are performed with (a) $\gamma = \eta = 1$ (b) $\gamma = \eta = 0.1$ and (c) $\gamma = \eta = 0.01$. Figure 6 and Figure 7 show mAP of DCMR on the MSCOCO dataset and XMedia dataset, respectively with different hyper-parameter values for the task of image→text retrieval.

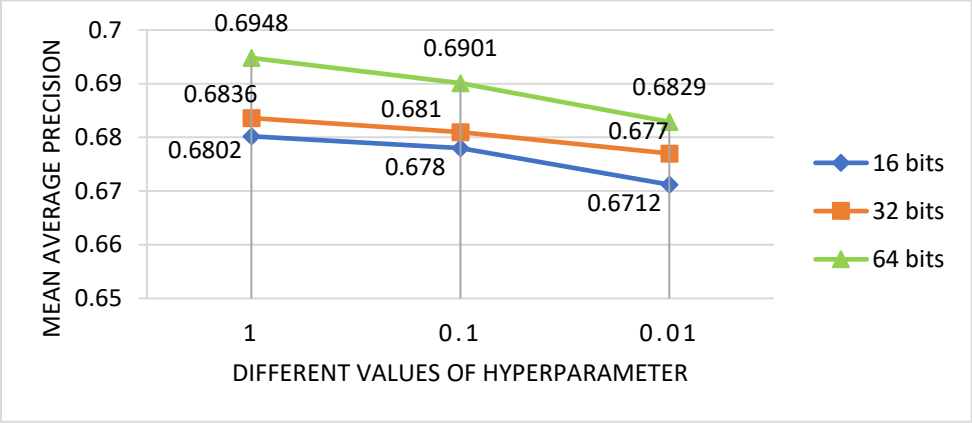

Figure 6. Performance of DCMR with different hyper-parameter values on MSCOCO dataset.

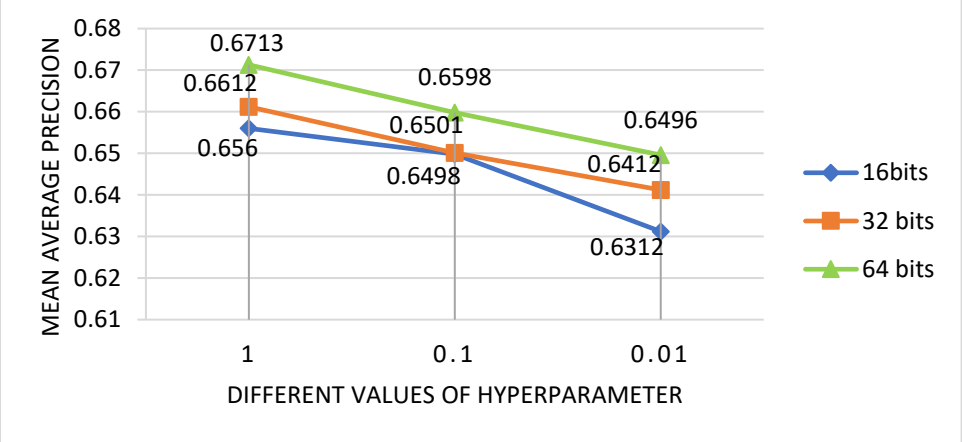

Figure 7. Performance of DCMR with different hyper-parameter values on XMedia dataset.

Both experiments conclude that $\gamma = \eta = 1$ is an optimized hyper-parameter value. Hence, it is considered in the subsequent experiments. The hyper-parameter selection problem can be resorted to experimentation to figure out which values of hyper-parameters works best. The Grid search is the most basic hyper-parameter tuning method, which is used in this research to fine tune the hyper-parameter. The Grid search technique builds a model for given possible combinations of the hyper-parameter, evaluating the model and select the architecture, which produces the best results.

4.2 Comparative Analysis of DCMR with Non-Deep Learning-Based Methods

The performance of DCMR is equated with the following Non-Deep Learning-Based Methods (1) Traditional Supervised Cross-Modal hashing methods: SePH (Lin et al., 2014), SCM (Zhang and Li, 2014). (2) Traditional Unsupervised Cross-Modal hashing methods: CMFH(Ding et al., 2014), LSSH(Zhou et al., 2014), IMH (Song et al., 2013), CVH (Kumar and Udupa, 2011). Table 1 and Table 2 show the performance of DCMR with hand-crafted-based CMR methods on MSCOCO and XMedia datasets, respectively for the task of image→ text and text→image retrieval. Here different length of binary code is selected for the experiment. SePH performs better than other hand-crafted CMR methods as it uses non-linear modelling methods and logistic regression methods, which can handle the complex structure of Multi-Modal data.

Table 1. Performance comparison of DCMR with hand-crafted based CMR methods on MSCOCO dataset in terms of mAP

	image→text			**text→image**		
	16bits	32 bits	64 bits	16bits	32 bits	64 bits
DCMR	0.6802	0.6836	0.6948	0.7223	0.7328	0.731
SePH	0.6034	0.6281	0.635	0.6341	0.6428	0.644
SCM	0.5932	0.612	0.622	0.628	0.631	0.6388
CMFH	0.4234	0.4428	0.4873	0.465	0.4676	0.5023
LSSH	0.3965	0.4231	0.4455	0.4122	0.4439	0.465
IMH	0.408	0.4398	0.4566	0.445	0.4652	0.4876
CVH	0.2237	0.2359	0.2456	0.3922	0.4233	0.443

Table 2. Performance comparison of DCMR with hand-crafted based CMR methods on XMedia dataset in terms of mAP.

	image→text			**text→image**		
	16bits	32 bits	64 bits	16bits	32 bits	64 bits
DCMR	0.656	0.6612	0.6713	0.6612	0.687	0.7092
SePH	0.5622	0.5718	0.5827	0.6123	0.625	0.6315
SCM	0.5514	0.5622	0.5735	0.5834	0.5878	0.5954
CMFH	0.4387	0.4456	0.4492	0.5124	0.5198	0.5342
LSSH	0.4064	0.4176	0.4279	0.4712	0.479	0.483
IMH	0.4154	0.4238	0.4396	0.498	0.5135	0.5236
CVH	0.2345	0.2436	0.2564	0.453	0.467	0.476

The performance comparison of DCMR with hand-crafted based CMR methods shows that the performance of DCMR increases at least by 9% for image→text and by 12% in mAP for text→image retrieval tasks on MSCOCO and XMedia datasets. In DCMR, a deep network is used for each modality, which does feature learning and generation of binary-valued representation in the same model. The state-of-the-art methods ignore the correlation when modality-wise feature learning is performed. Further, the performance of DCMR is compared with following deep learning-based CMR methods: SPHMF (Zhang and Pan, 2021), PRDH (Yang et al., 2017), DVSH (Cao et al., 2016), DCMH(Jiang and Li, 2017) and DSCMR(Wang et al., 2021), which uses pairwise labels for the generation of a common sub-space. Source codes of all these methods are provided by respective authors. Table 3, 4 and 5 show the performance of DCMR with deep learning-based methods on MSCOCO, XMedia and NUS-WIDE datasets, respectively.

Table 3. Performance comparison of DCMR on MSCOCO dataset.

	image→text			**text→image**		
	16bits	32 bits	64 bits	16bits	32 bits	64 bits
DCMR	0.6802	0.6836	0.6948	0.7223	0.7328	0.731
SPHMF	0.669	0.673	0.6812	0.7036	0.7165	0.7201
DSCMR	0.6523	0.6621	0.6742	0.6982	0.7012	0.7123
PRDH	0.6412	0.6492	0.6626	0.6522	0.6823	0.6828
DCMH	0.622	0.644	0.6535	0.648	0.6679	0.6714
DVSH	0.6124	0.6286	0.6452	0.6424	0.6532	0.6524

Table 4. Performance comparison of DCMR on XMedia dataset.

	image→text			**text→image**		
	16bits	32 bits	64 bits	16bits	32 bits	64 bits
DCMR	0.656	0.6612	0.6713	0.6612	0.687	0.7092
SPHMF	0.638	0.6414	0.6518	0.6449	0.652	0.6816
DSCMR	0.623	0.6372	0.6448	0.6412	0.6654	0.6872
PRDH	0.598	0.5962	0.6218	0.6321	0.652	0.6616
DCMH	0.581	0.5897	0.601	0.6245	0.6414	0.6514
DVSH	0.5762	0.578	0.5924	0.6278	0.6322	0.643

The DCMR reduces the modality gap by preserving the semantic similarities among heterogeneous modalities. The triplet labels have the capability to put the semantic similar words nearer and dissimilar words far in the vector space. Hence, the hamming distance between similar data points are minimized, which improves the result of the proposed model. The generated results from the proposed model shows that performance of DCMR increases by 2% to 3% for Image→Text retrieval and by 2% to 5% for Text→Image retrieval in mAP on MSCOCO, XMedia and NUS-WIDE datasets. The deep learning based methods use pairwise labels to generate a common sub-space, which preserves similarity between corresponding data points but fails to preserve relative similarities. Further, mAP of text→image retrieval

is more as text modality contains more background information that cannot be presented by its corresponding image. Further, experiments are conducted with varying lengths of binary code on MSCOCO and XMedia datasets. Figure 8 and Figure 9 show the performance of image→text and text→image retrieval on MSCOCO and XMedia datasets, respectively. The relative performance enhancement of proposed model (DMCR) with state-of-the-art deep learning based model called SPHMF.

Table 5. Performance comparison of DCMR on NUS-WIDE dataset.

	image→text			**text→image**		
	16 bits	32 bits	64 bits	16 bits	32 bits	64 bits
DCMR	0.7124	0.728	0.7392	0.7393	0.7432	0.762
SPHMF	0.695	0.699	0.7126	0.7065	0.7235	0.7341
DSCMR	0.6621	0.6782	0.6897	0.701	0.7114	0.7228
PRDH	0.6532	0.6652	0.6779	0.6679	0.6923	0.6914
DCMH	0.6335	0.658	0.6624	0.6533	0.6786	0.689
DVSH	0.6231	0.6352	0.6543	0.6431	0.6522	0.6681

The objective of the proposed model is to preserve the relative semantic similarities, which is one of the open challenges in CMR. The state-of-the-art methods does faster retrieval than proposed model but does not preserve the relative semantic similarities between heterogeneous modalities. So, the constraint in both the approaches differs in terms of retrieval and time.

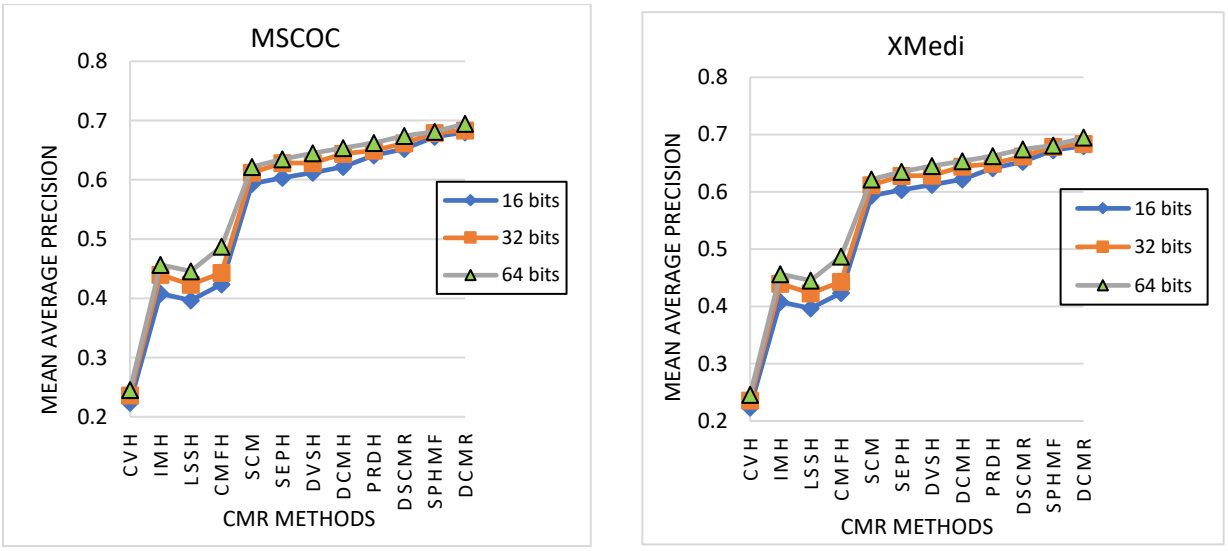

Figure 8. mAP of CMR methods with different length of binary code on MSCOCO and XMedia dataset.

The DCMR having 64-bit length binary code achieves better performance than 16-bit and 32-bit length binary code. The larger length of the binary code has the capability to preserve more semantic similarities between heterogeneous modalities.

Further, the training time and testing time of DCMR with other methods are compared. Table 6 shows a performance comparison in terms of training and testing time for MSCOCO, XMedia and NUS-WIDE datasets.

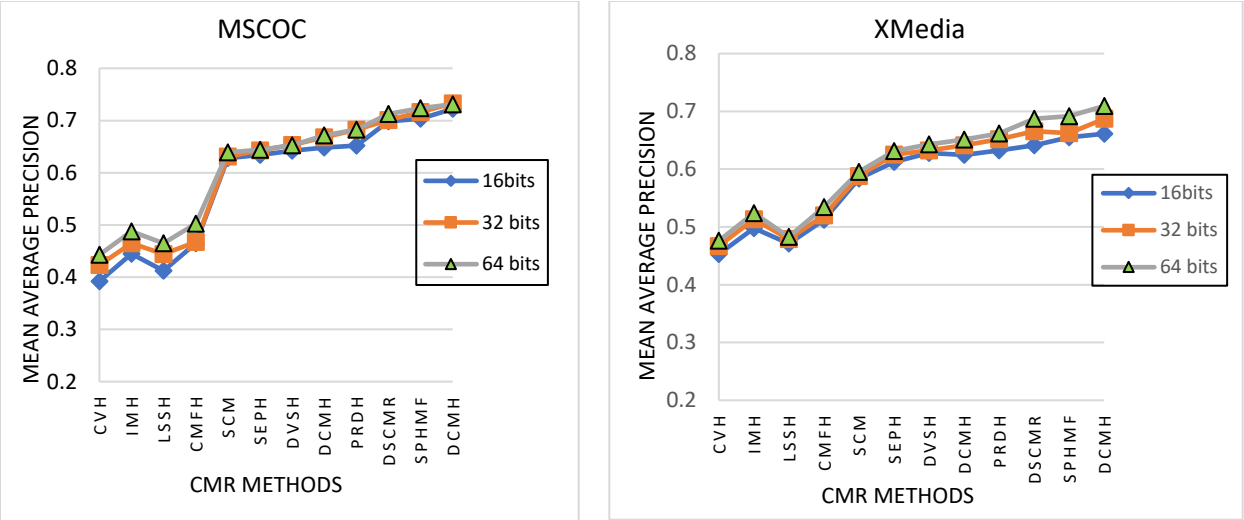

Figure 9. mAP of CMR methods with different length of binary code on MSCOCO and XMedia dataset.

Table 6. Performance comparison of DCMR with deep learning-based CMR methods in terms of training time, testing time, and mAP.

	CMR Methods	Nature of supervised information provided to CMR	Training Time (seconds)	Testing Time (seconds) text→image	mAP
MSCOCO	DCMR	Triplet labels	4296	12.860	0.731
	SPHMF	Pairwise labels	3444	11.88	0.7201
	DCMH	Pairwise labels	3258	9.655	0.6714
XMedia	DCMR	Triplet labels	1432	8.712	0.7092
	SPHMF	Pairwise labels	1258	9.382	0.6816
	DCMH	Pairwise labels	1254	7.510	0.6514
NUS-WIDE	DCMR	Triplet labels	5321	14.877	0.762
	SPHMF	Pairwise labels	4237	13.118	0.7341
	DCMH	Pairwise labels	3980	11.765	0.689

The DCMR takes more training time in comparison with other deep learning-based CMR methods as triplets labels are provided as the input to generate the common sub-space. The testing time of DCMR is more for text→image retrieval tasks. The DCMR uses the Glove model as an embedding model, which captures the relevance among various words in the text modality. The extracted multiple features are given as input to MLP for feature learning. The DCMH uses the BoW model to map words into vectors, which is given as input to MLP for feature extraction. However, DCMR outperforms in terms of mAP.

5. Conclusion

One of the open challenges in CMR is to reduce the modality gap by preserving the semantic similarities between heterogeneous modalities. Existing approaches generate binary-valued representation, which provides faster retrieval but fails to preserve relative semantic similarities. The proposed DCMR uses triplet labels to generate binary-valued representation. The triplet labels locate similar data points nearer and dissimilar words far in the vector space, which improves the performance of the retrieval. The performance of DCMR is compared with (a) hand-crafted-based CMR systems, and (b) deep learning-based systems, which concludes that the performance of DCMR increases at least by 2% to 3% for Image→Text retrieval and by 2% to 5% for Text→Image retrieval in mAP on MSCOCO, NUS-WIDE, and XMedia datasets.

However, the training time of DCMR is more in comparison with pairwise deep learning-based CMR methods. So, triplet labels are preferable when the result of retrieval is a constraint whereas pairwise labels are preferable when time is a constraint.

Conflict of Interest

The authors confirm that there is no conflict of interest to declare for this publication.

Acknowledgments

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. The authors would like to thank the editor and anonymous reviewers for their comments that help improve the quality of this work.

References

Bhatt, N., & Ganatra, A. (2021). Improvement of deep cross-modal retrieval by generating real-valued representation. *PeerJ Computer Science*, *7*, e491. https://doi.org/10.7717/peerj-cs.491.

Brodeur, S., Perez, E., Anand, A., Golemo, F., Celotti, L., Strub, F., Rouat, J., Larochelle, H., & Courville, A. (2017). Home: A household multimodal environment. *arXiv preprint arXiv:1711.11017*. https://doi.org/10.48550/arXiv.1711.11017.

Cao, Y., Long, M., Wang, J., Yang, Q., & Yu, P.S. (2016). Deep visual-semantic hashing for cross-modal retrieval. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1445-1454). https://doi.org/10.1145/2939672.2939812.

Chabot, F., Chaouch, M., Rabarisoa, J., Teuliere, C., & Chateau, T. (2017). Deep manta: A coarse-to-fine many-task network for joint 2d and 3d vehicle analysis from monocular image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2040-2049).

Ding, G., Guo, Y., & Zhou, J. (2014). Collective matrix factorization hashing for multimodal data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2075-2082).

Duan, L., Zhao, C., Miao, J., Qiao, Y., & Su, X. (2017). Deep hashing based fusing index method for large-scale image retrieval. *Applied Computational Intelligence and Soft Computing*, *2017*, 9635348. https://doi.org/10.1155/2017/9635348.

Fast, E., & Horvitz, E. (2017). Long-term trends in the public perception of artificial intelligence. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 31, No. 1). https://doi.org/10.1609/aaai.v31i1.10635.

Hu, M., Yang, Y., Shen, F., Xie, N., Hong, R., & Shen, H.T. (2019). Collective reconstructive embeddings for cross-modal hashing. *IEEE Transactions on Image Processing*, *28*(6), 2770-2784.

Hua, Y., Wang, S., Liu, S., Cai, A., & Huang, Q. (2016). Cross-modal correlation learning by adaptive hierarchical semantic aggregation. *IEEE Transactions on Multimedia*, *18*(6), 1201-1216. https://doi.org/10.1109/TMM.2016.2535864.

Huang, E.H., Socher, R., Manning, C.D., & Ng, A.Y. (2012, July). Improving word representations via global context and multiple word prototypes. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 873-882). Association for Computational Linguistics. Jeju, Republic of Korea.

Jiang, Q.Y., & Li, W.J. (2017). Deep cross-modal hashing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 3232-3240).

Kiros, R., Salakhutdinov, R., & Zemel, R.S. (2014). Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539*. https://doi.org/10.48550/arXiv.1411.2539.

Kong, W., & Li, W.J. (2012). Isotropic hashing. *Advances in Neural Information Processing Systems*, *25*, 1-9.

Kumar, S., & Udupa, R. (2011). Learning hash functions for cross-view similarity search. In *Twenty-Second International Joint Conference on Artificial Intelligence* (vol. 22, pp. 1360).

Lai, H., Pan, Y., Liu, Y., & Yan, S. (2015). Simultaneous feature learning and hash coding with deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 3270-3278). https://doi.org/10.1109/CVPR.2015.7298947.

Li, W.J., Wang, S., & Kang, W.C. (2015). Feature learning based deep supervised hashing with pairwise labels. *arXiv preprint arXiv:1511.03855*.

Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C.L. (2014). Microsoft COCO: Common Objects in Context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds) *Computer Vision – ECCV 2014.* Lecture Notes in Computer Science (vol. 8693). Springer, Cham. https://doi.org/10.1007/978-3-319-10602-1_48.

Lin, Z., Ding, G., Hu, M., & Wang, J. (2015). Semantics-preserving hashing for cross-view retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 3864-3872). Boston.

Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., van der Laak, J.A.W.M., Van Ginneken, B., Sánchez, C.I. (2017). A survey on deep learning in medical image analysis. *Medical Image Analysis*, *42*, 60-88. https://doi.org/10.1016/j.media.2017.07.005.

Liu, B., Cao, Y., Long, M., Wang, J., & Wang, J. (2018, October). Deep triplet quantization. In *Proceedings of the 26th ACM International Conference on Multimedia* (pp. 755-763). https://doi.org/10.1145/3240508.3240516.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013a). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mikolov, T., Deoras, A., Kombrink, S., Burget, L., Cernocký, J. (2011). Empirical evaluation and combination of advanced language modeling techniques. *Interspeech,* 605-608. https://doi.org/10.21437/interspeech.2011-242.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, *26*.

Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., & Ng, A.Y. (2011). Multimodal deep learning. *In Proceedings of the 28th International Conference on Machine Learning*. Bellevue, WA, USA.

Peng, Y., Huang, X., & Zhao, Y. (2018). An overview of cross-media retrieval: Concepts, methodologies, benchmarks, and challenges. *IEEE Transactions on Circuits and Systems for Video Technology*, *28*(9), 2372-2385.

Peng, Y., Zhai, X., Zhao, Y., & Huang, X. (2016). Semi-supervised cross-media feature learning with unified patch graph regularization. *IEEE Transactions on Circuits and Systems for Video Technology*, *26*(3), 583-596.

Rajagopalan, S.S., Morency, L.P., Baltrusaitis, T., & Goecke, R. (2016). Extending long short-term memory for multi-view structured learning. In *European Conference on Computer Vision* (pp. 338-353). Springer, Cham. https://doi.org/10.1007/978-3-319-46478-7_21.

Ranjan, V., Rasiwasia, N., & Jawahar, C.V. (2015). Multi-label cross-modal retrieval. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 4094-4102).

Song, J., Yang, Y., Yang, Y., Huang, Z., & Shen, H.T. (2013). Inter-media hashing for large-scale retrieval from heterogeneous data sources. In *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data* (pp. 785-796). https://doi.org/10.1145/2463676.2465274.

Srivastava, N., & Salakhutdinov, R.R. (2012). Multimodal learning with deep boltzmann machines. *Advances in Neural Information Processing Systems*, *25*, 1-9.

Vendrov, I., Kiros, R., Fidler, S., & Urtasun, R. (2015). Order-embeddings of images and language. *arXiv preprint arXiv:1511.06361*.

Wang, C., Yang, H., & Meinel, C. (2016). A deep semantic framework for multimodal representation learning. *Multimedia Tools and Applications*, *75*(15), 9255-9276. https://doi.org/10.1007/s11042-016-3380-8.

Wang, L., Sun, W., Zhao, Z., & Su, F. (2017). Modeling intra-and inter-pair correlation via heterogeneous high-order preserving for cross-modal retrieval. *Signal Processing*, *131*, 249-260.

Wang, S., Dou, Z., Chen, D., Yu, H., Li, Y., & Pan, P. (2019). Multimodal multiclass boosting and its application to cross-modal retrieval. *Neurocomputing*, *357*, 11-23.

Wang, X., Hu, P., Zhen, L., & Peng, D. (2021). Drsl: Deep relational similarity learning for cross-modal retrieval. *Information Sciences*, *546*, 298-311.

Wu, L., Hoi, S.C., & Yu, N. (2010). Semantics-preserving bag-of-words models and applications. *IEEE Transactions on Image Processing*, *19*(7), 1908-1920.

Xu, X., Shen, F., Yang, Y., Shen, H.T., & Li, X. (2017). Learning discriminative binary codes for large-scale cross-modal retrieval. *IEEE Transactions on Image Processing*, *26*(5), 2494-2507.

Yanagi, R., Togo, R., Ogawa, T., & Haseyama, M. (2020). Enhancing cross-modal retrieval based on modality-specific and embedding spaces. *IEEE Access*, *8*, 96777-96786.

Yang, E., Deng, C., Liu, W., Liu, X., Tao, D., & Gao, X. (2017). Pairwise relationship guided deep hashing for cross-modal retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 31, No. 1, pp. 1618-1625). https://doi.org/10.1609/aaai.v31i1.10719.

Zhang, D., & Li, W.J. (2014, June). Large-scale supervised multimodal hashing with semantic correlation maximization. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 28, No. 1). https://doi.org/10.1609/aaai.v28i1.8995.

Zhang, H., & Pan, M. (2021). Semantics-preserving hashing based on multi-scale fusion for cross-modal retrieval. *Multimedia Tools and Applications*, *80*(11), 17299-17314.

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). California.

Zhou, J., Ding, G., & Guo, Y. (2014, July). Latent semantic sparse hashing for cross-modal similarity search. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval* (pp. 415-424). Association for Computing Machinery, New York.

Publisher's Note- Ram Arti Publishers remains neutral regarding jurisdictional claims in published maps and institutional affiliations.