Cross-modal Retrieval via Memory Network

Ge Song
sunge@nuaa.edu.cn

Xiaoyang Tan
x.tan@nuaa.edu.cn

Nanjing University of Aeronautics and Astronautics
Nanjing, P.R. China

Collaborative Innovation Center of Novel Software Technology and Industrialization
Nanjing, P.R. China

Abstract

With the explosive growth of multimedia data on the Internet, cross-modal retrieval has attracted a great deal of attention in computer vision and multimedia community. However, this task is very challenging due to the heterogeneity gap between different modalities. Current approaches typically involve a common representation learning process that maps different data into a common space by linear or nonlinear functions. Yet most of them 1) only handle the dual-modal situation and generalize poorly to complex cases; 2) require example-level alignment of training data, which is often prohibitively expensive in practical applications; and 3) do not fully exploit prior knowledge about different modalities during the mapping process. In this paper, we address above issues by casting common representation learning as a Question Answer problem via a cross-modal memory neural network (CMMN). Specifically, raw features of all modalities are seemed as 'Question', and extra discriminator is exploited to select high-quality ones as 'Statements' for storage whereby common features are desired 'Answer'. Experimental results show that CMMN can achieve state-of-the-art performance on the Wiki and CO-CO dataset and outperform other baselines on the large-scale scene dataset CMPlaces.

1 Introduction

With the popularization of social media and the explosive growth of the Internet, massive media data (e.g. image, text, and audio) have flooded our daily lives. At the same time, data retrieval, which aims to search the relevant data with a query, is becoming a very hot topic in computer vision and multimedia research community. However, in many cases, one object or topic is described simultaneously by different types of data (i.e. multi-modal data), and one may be concerned about retrieving different types of data that are relevant to the query. In this paper, we focus on this very common and normal scenario which is also known as cross-modal retrieval (CMR).

The cross-modal retrieval task is very challenging, due to the existence of heterogeneity gap between different modalities of data. To break this gap, current methods typically perform a mapping from different modalities to a unified feature space with linear or nonlinear transformations. Despite the effectiveness in some conditions, most of these methods have three main limitations. Firstly, they only focus on image-text modalities and are tedious and

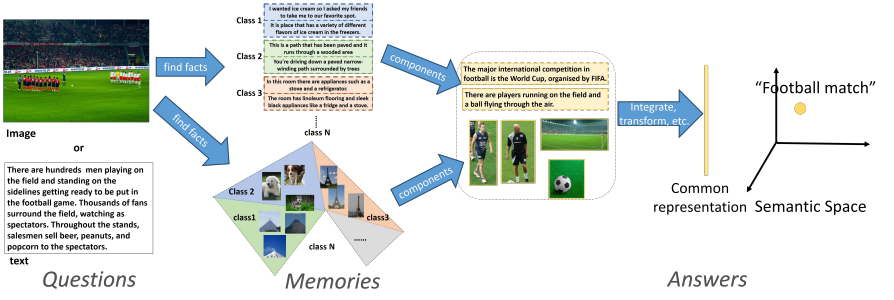


Figure 1: Illustration of our motivation. We observed that heterogeneous data could be represented by corresponding components across different modalities. We model this process as a QA problem: data with the clear semantic concept from different modalities are regard as memories while other data denote as queries, common representation act the role of answers.

inefficient to handle more complex situations (more than two modalities). Secondly, most methods require example-level aligned data (e.g. image-text pairs) for training, which is usually not available in the real-life scenario. Lastly, prior knowledge (e.g. the semantic ambiguity of image) are not fully exploited in the learning and mapping process.

In this paper, we attempt to address above issues with a memory neural network. Our main idea is to adapt the memory network (MemNN) [25, 30] which has been successfully applied in QA (Question Answer) to cross-modal retrieval. Our motivation is that: if we ask someone to imagine a scene of soccer game, firstly, he will interpret what soccer game is (e.g. 'Some people are playing football on the grass field') and search seem impressive components in his memory (e.g. person, football), and then these parts are aggregated, processed and transformed into final image of the scene. Inspired by this imagine way of human beings, we argue that integrating and mapping attention contents relevant to target object across various modalities is helpful for common representation learning in cross-modal retrieval task. Fortunately, in QA field, MemNN uses a memory to preserve input facts (statements sentence) in memory and retrieve supporting facts for input question to infer the answer. And it can learn to find relevant memories for query by an end-to-end learning way and learn to predict answer jointly. This process is quite similar with above imagine way. Therefore, we model the common representation learning process as a QA problem, as illustrated in Figure 1. The differences of MemNN between cross-modal retrieval and QA mainly lie in two aspects: the number of modalities and the role of content in memory. We propose a novel cross-modal memory network (CMMN) which can exploit the collected supporting clews (i.e. visual objects in images, textual entities in texts or others) of different modalities for common semantic concepts to alleviate the heterogeneity gap problem. Experiment results show the effectiveness of this method.

The remainder of this paper describes and analyzes our CMMN in detail. We first discuss related work in section 2. The CMMN model details are presented in Section 3, experimental results are given in Section 4. And we finally draw conclusions in Section 5.

2 Related Work

Cross-Modal retrieval. Many approaches [1, 10, 13, 22] have been proposed for cross-

modal retrieval task. A representative traditional method is Canonical Correlation Analysis (CCA) [22], which aim to learn two mappings for two modalities of data so that maximally correlated, and its variants [1, 7] are proposed later. Wang *et al*. [27] bring up a joint framework for feature selection and subspace learning. Recently, encouraged by the great success of deep learning achieved, some methods based on deep architecture are proposed. Jiang *et al*. [13] exploit the existing image-text databases to optimize a pairwise ranking function which enhances both local alignment and global alignment for cross-modal retrieval. CMPlaces [3] present a method to regularize cross-modal convolutional neural networks so that they have a shared representation that is agnostic of the modality.

On the other hand, benefiting from the low storage costs and fast query speed of binary codes, cross-modal hashing methods [2, 5, 15, 24, 34] have attracted much attention from academia. Lin *et al*. [18] utilize kernel logistic regression to learn nonlinear hash functions by minimizing the Kullback-Leibler divergence between two affinity matrixes of semantic and hash code. While J. Zhou *et al*. [35] learn kernel functions for hashing via preserving intermodal similarities within an AdaBoost framework. Zhang *et al*. [32] posed SCM to integrate semantic labels into the hashing learning procedure via maximizing semantic correlations. Jiang *et al*. [12] pose to integrate feature learning and hash-code learning into a deep model and training with data pairs. Both these hashing methods devote to find nonlinear functions to mapping features from different modalities to a common Hamming space.

Memory Network. Memory networks were early proposed by Weston *et al*. [30] with the goal of prediction. Its central idea is to combine successful machine learning models with an extra memory component that can be read and written for more accurate inference. It was evaluated in the context of QA and outperforms prevalent LSTM models with the greater capability of remembering facts from the past. Later, Sukhbaatar *et al*. introduce an end-to-end neural network with a recurrent attention model over a large external memory. Miller *et al*. [21] presented a key-value memory network to be read documents more viable for answering. Meanwhile, a number of recent efforts [4, 28] have explored ways to use RNNs or LSTM-based models with memory in natural language processing field. This memory mechanism is also used in Neural Turing Machine of Graves *et al*. [8] to tackle problems of sorting and recall. Particularly, closely related work in computer vision is recently proposed Stacked Attention Networks [31] for image QA.

3 Memory for cross-modal

In this section, we give a detailed description of the proposed cross-modal memory network (CMMN) model, which is presented in Figure 2.

3.1 Cross-Modal Memory Networks

Suppose we have two modal feature $\{x_i\}_{i=1}^{N_X} \in R^{D_X}, \{y_j\}_{j=1}^{N_Y} \in R^{D_Y}$ and corresponding labels L, N_X, N_Y are numbers of data, D_X, D_Y are dimension. Memory M_1 includes N_{M2} vectors m_{1i} from X and memory M_2 contains N_{M1} vectors m_{2j} from Y.

Input embedding and memory search: We are given an input q whether from X or Y and then to find its related clues (memory vectors) for high-level inference. Firstly, both q and memory vectors will convert into a common continuous space, using embedding matrix $A \in R^{D\times V}$, $B_1 \in R^{D_X\times V}$ and $B_2 \in R^{D_Y\times V}$. Then, we compute the inner product of q and each memory m_{1i}, m_{2j} in the embedding space as their match scores. Finally, we take a soft

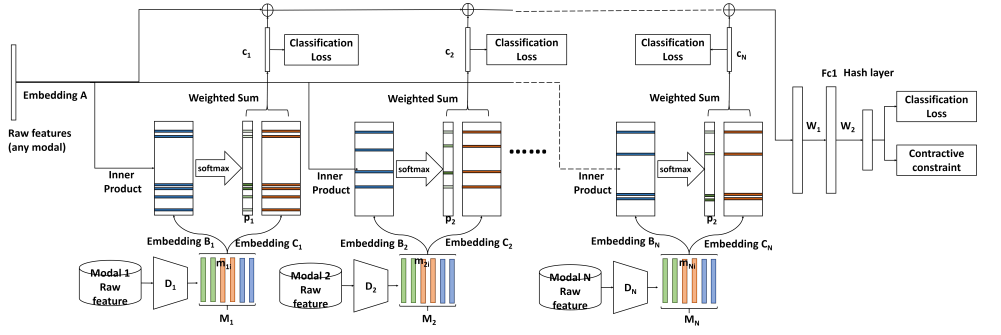


Figure 2: Cross-modal memory networks. D represents discriminator and hash layer outputs binary-like codes for fast retrieval.

attention mechanism to obtain the selected memories. Let variables z_1 and z_2 represents the position of M_1 and M_2 to be read. The probability of z given m_i and q is defined as follows:

$$\begin{aligned} p(z_1 = i \mid m_{1i}, q) &= \text{Softmax}(((m_{1i}B_1)^T(qA))) \\ p(z_2 = i \mid m_{2j}, q) &= \text{Softmax}(((m_{2j}B_2)^T(qA))) \end{aligned} \tag{1}$$

where $\text{Softmax}(m_i) = \frac{1}{Z}\exp(m_i)$, $Z = \sum_j \exp(m_j)$. Defined in this way p is the distribution over the memories.

Memory representation and output embedding: Each memory vectors m_i have some discriminative information to support q. We collect these salient contents and sum weighted with the distribution p, then we obtain the response context vector c for q. Because of the low discriminative of raw features, we transform memory vectors before summing by another embedding matrix $C \in R^{D\times V}$.

$$\begin{aligned} c_1 &= \sum_{i=1}^{N_{M1}} p(z_1 = i \mid m_{1i}, q)(m_{1i}C_1) \\ c_2 &= \sum_{j=1}^{N_{M2}} p(z_2 = j \mid m_{2j}, q)(m_{2j}C_2) \end{aligned} \tag{2}$$

In order to enforce the discriminative of c, it is connected to a fully-connect layer and imposed by classification loss (a fully-connect layer with the same classes' number nodes). Suppose the output of the loss layer are $f(c_1)$ and $f(c_2)$ (a softmax to produce the predicted label), N is the training sample number, the cross entropy loss of classification is formulated as below:

$$\text{Classification Loss} = -\sum_{i=1}^{N} l_i \log(f_i(c)) \tag{3}$$

With the supervised information, the obtained context vector will be more discriminative.

Cross-modal feature learning: The core inference part of MemNN [25, 30] is a black box and represented as a score function [30] or a nonlinear transformation [25]. Our target is aggregating various modalities feature with their cross-modal supporting context vectors and fusing them to a common feature space. Therefore, it contains two parts in output

component of our model: feature aggregation and fusion. We firstly integrate features by simple weighted summing them and then we adopt same way as [25] done to learn common representation with supervise information. Let r and h are the output of Fc1 and hash layer.

$$\begin{aligned} r =& \text{relu}((\alpha(qA) + \beta c_1 + \gamma c_2)W_1 + b) \\ h =& \sigma(rW_2) \end{aligned} \tag{4}$$

Where $\text{relu}(x) = max(x,0)$, $\sigma(x) = \frac{1}{1+\exp(-x)}$. Particularly, α, β and γ in formula (4) are not manually set, they are learned by training and initialized with 1.

In order to speed up retrieval, we embed a hash-layer (fully-connect layer activated with sigmoid function) after the layer $Fc1$ and impose hashing constraints as deep hashing method [16]done. Here, we didn't directly enforce the output h to be 0 or 1. Encouraged by contractive auto-encoder [23], Jacobin penalty term can naturally restrain the disturbance of data and output $0,1$ feature, we replace hashing constraint(e.g. $||h|-1|$) with it. And the final objective function is defined as following:

$$\begin{aligned} \text{Loss} =& \text{classification loss} + \lambda \|J_f(r)\|_F^2 \\ =& -\sum_{i=1}^{N}(l_i \log(f(h_i)) + \lambda \sum_{p=1}^{d_h}(h_{ip}(1-h_{ip}))^2 \sum_{q=1}^{d_r} W_{pq}^2) \end{aligned} \tag{5}$$

Where d_h and d_r denote the dimension of h and r, λ is a balance parameter. Finally, it is easy to obtain hash code with $b = sign(h > 0.5)$.

3.2 Memory Generation and Optimization

Memory generation: In QA problem, memory component storage all available facts or sentences without any prior knowledge. Generally, we, human beings always remember the general and specific characteristics of classes. And the general characteristic is helpful to distinguish intra-class while specific for inter-class. This indicates that we should collect the data that contains more general characteristic of the specific class as memories. So we propose a simple way to find those features for target class within a given modality. Suppose we have training features $x_{i}^{N}{}_{i=1}$ and corresponding labels $l_{i}^{N}{}_{i=1}$, a discriminator D which can predict the probability $p(y_i = l_i|x_i)$. The memory vectors for class C can compute by the following:

$$M = \text{Top}_k p(y_i = l_i|x_i), \quad i = \{l_i = C\} \tag{6}$$

We sort features of class C according to p and pick out top k candidates as memory content. Then, we obtain all memories for each class by this way. Besides, two unsupervised methods could be helpful, random selection or K-means algorithm. It is notable that the discriminator could be a deep neural network or Support Vector Machine (SVM) and so on.

Optimization: Because the overall model from input to output is smooth, it is easy to compute gradients and back-propagate through it. In our CMMN, the parameters mainly lie in three parts, the embedding matrix A for input feature, each memory has two embedding matrices B_i, C_i, and other mapping matrices. In particular, papers [25, 30] initialize memories before training and without any update during training, which is unreasonable. In fact, human beings' memories will be updated according to the external stimulation. Therefore, both parameters and memories are jointly learning by minimizing the object loss (5). Training is performed using stochastic gradient descent (SGD).

3.3 Extensions

As we can see, the proposed CMMN can be easily extended to cases with more than two modalities. If a new modality data are available, we firstly generate corresponding memory M_{new} with formula (6). Then we assemble it to CMMN and update the model. Finally, in training processes, a new context vector can be calculated by the formula (2) and the learned representation (4) will contain new discriminative information of additional modality.

4 Experiments

4.1 Experimental Settings

To validate the proposed CMMN, we conduct experiments on three multi-modal datasets.

Wiki [22] is a image-text dataset and consists of 2,866 image-text pairs. For each pair, the image is represented by the 128-dimensional SIFT descriptor vector and the text as a 10-dimensional topic vector. Besides, each pair is annotated with one of 10 semantic labels. Total 693 pairs as the query set and the rest 2,173 pair as the retrieval set. We also evaluate the deep representation performance by extracting `fc7` product of AlexNet per-trained on ImageNet with Caffe [11] as [27] done.

CMPlaces [3] is a large-scale cross-modal places dataset, which includes five different modalities: Natural Images (NAT) from Place 205 database [33] (2.4 million training and 20,500 validation), Line Drawings (LDR) contains 14,830 training and 2,050 validation sketches, Descriptions (DSC) is composed of 9,752 training and 2050 validation detailed texts description of the scene, Clip Art (CLP) consists of 11,372 training and 1,954 validation cartoon images, Spatial Text (SPT) contains 456,300 training and 2,050 validation synthetic spatial text images. Each example is annotated with a unique label of 205 scene categories. In experiments, we take training set as retrieval database, test examples as the query. Particularly, if NAT is the retrieval database, 1,000 testing data are random as query set. The DSC are represented by average-pooling the 4800-D Skip-thought vectors [14] of each sentence.

Microsoft COCO [17] is a large-scale common object dataset, contains 82,783 training images and 40,504 testing images. Each image is associated with five sentences, belonging to 90 categories. After pruning images with no category, we generate 82,081 image-sentence pairs as training set and random 4,956 pairs as the testing set. In experiments, image are represented by 2,048 deep features extracted form ResNet [9] pre-trained on ImageNet and description is represented by 4800-D Skip-thought vectors [14].

In our model, the input dimension of different modalities should be same, and the prior information $p(y = l|x)$ of the feature must be estimated. Therefore, we use deep neural networks to complete above two tasks simultaneously. For pixel based modalities, we use Alexnet to produce `fc7`. While for description, we use an MLP on vectors to produce same dimensional representation as `fc7`. Details as follow, suppose input data is $x \in R^D$: $input(D) \rightarrow fc1(D) \rightarrow fc2(42364) \rightarrow pool5(size:3\times3, stride:2) \rightarrow fc6(4096) \rightarrow fc7(4096) \rightarrow fc8(class\ number)$, all layers are activated by ReLU nonlinearities. Finally, all raw features will share the same dimension, i.e. $D_X = D_Y$. And we empirically set the model parameter λ in the objective function of CMMN (i.e. formula(5)) as 0.001 for all datasets. The memory size k in formula(6) is set according to the modality data size (k set to 10 in general, while for NAT of CMPlaces is set as 90).

We initialized the weights of layers of CMMN using a Gaussian distribution with $std = 0.1$. And we trained model with $learning\_rate = 0.01$, $batch\_size = 32$ and $epoches = 200$.

Method	Image query *v.s.* Text	Text query *v.s.* Image	Average
LSCMR(2013) [20]	0.2021	0.2229	0.2125
DCN-S(2016) [19]	0.2139	0.2253	0.2196
DCN-C(2016) [19]	0.2268	0.2461	0.2365
CCA-3V(2014) [2]	0.2752	0.2242	0.2497
SliM^2(2013) [36]	0.2548	0.2021	0.2285
M^{3}R(2014) [29]	0.2298	0.2677	0.2488
LCFS(2013) [26]	0.2798	0.2141	0.2470
JFSSL(2016) [27]	**0.3063**	0.2275	0.2669
CMMN$_{\text{real}-\text{value}}$	0.2655	**0.6199**	**0.4427**
CCA-3V+CNN [27]	0.4049	0.3651	0.3850
LCFS+CNN [27]	0.4123	0.3845	0.3984
JFSSL+CNN [27]	**0.4279**	0.3957	0.4118
CMMN$_{\text{real}-\text{value}}$ + CNN	0.3919	**0.6948**	**0.5433**

Table 1: MAP of different real-valued representation learning methods on Wiki dataset.

For Wiki, all training examples of both image and text are used for training. While for CMPlaces, we random 38,950 NAT, 18,450 SPT from their training data and overall training examples of LDR, CLP, DSC to construct training set. And for COCO, we random 5,000 pairs for training all methods.

we adopt the commonly-used Mean Average Precision (mAP) as the performance metric. $P@n = \frac{\#\{relevant\ images\ in\ top N\}}{N}, AP = \frac{\sum_n P@n \times I\{image\ n\ is\ relevant\}}{\#\{retrieved\ relevant\ image\}}, mAP = \frac{1}{Q}\sum_i AP_i$. where # is count function, I is indicator function, Q represents the total number of queries.

4.2 Experimental Results

Results on Wiki. We compare CMMN with variours state-of-the-art cross-modal real-valued representation learning methods. It includes LSCMR [20], CCA-3V [2], SliM^2 [36], M^{3}R [29], LCFS [26], JFSSL [27] and two deep models contains modified DCMH [12] refered as DCN-C and extended DSH [19] for two modalities refered as DCN-S. On the other hand, the state-of-the-art cross-modal hashing methods CMSSH [2], CVH [15], IMH [24], LSSH [34], CMFH [6], KSH-CV [35],SCM-Seq [32] and SePH$_{\text{km}}$ [18] are also taken as baselines. The mAP results are reported in Table 1 and Table 2. Mentation that, CMMN$_{real-value}$ represents we do retrieval with continuous value feature h and similarity is measured by Euclidean distance, while CMMN$_{bin}$ using hash code b with Hamming distance. Methods with '+CNN' denotes the image feature is 4096-dim deep representation.

From the experimental results in Table 1, we can find that CMMN can outperform all other non-hashing methods. In more detail, the state performance is greatly improved by CMMN at least 30% in the case the text to image retrieval, while the case the image to text retrieval, CMMN obtains comparable results 26.5% which yields JFSSL 4.1%. We analyze this result and find that most other methods exploit the inter-modal similarity information can boost the discriminative of the poorer modality feature. On the other hand, because CNN features contain more high-level semantic information of images, the performance of traditional methods can be improved. The poor performance of DCN-C and DCN-S may be caused by inadequate training samples.

From Table 2, we can find that CMMN$_{bin}$ is inferior to SePH$_{\text{km}}$ in all cases. Because SePH$_{\text{km}}$ is a kernel-based method, we guess the better performance of SePH$_{\text{km}}$ mainly comes

	Imgae query *v.s.* Text				Text query *v.s.* Image			
Method	16bits	32bits	64bits	128 bits	16bits	32bits	64bits	128 bits
CMSSH [2]	0.1877	0.1771	0.1646	0.1552	0.1630	0.1617	0.1539	0.1517
CVH [15])	0.1257	0.1212	0.1215	0.1171	0.1185	0.1034	0.1024	0.0990
IMH [24]	0.1573	0.1575	0.1568	0.1651	0.1463	0.1311	0.1290	0.1301
LSSH [34]	0.2141	0.2216	0.2218	0.2211	0.5031	0.5224	0.5293	0.5346
CMFH [5]	0.2132	0.2259	0.2362	0.2419	0.4884	0.5132	0.5269	0.5375
KSH-CV [35]	0.1965	0.1839	0.1701	0.1662	0.1710	0.1665	0.1696	0.1576
SCM-Seq [32]	0.2210	0.2337	0.2442	0.2596	0.2134	0.2366	0.4479	0.2573
SePH_{km} [18]	0.2787	0.2956	0.3064	0.3134	0.6318	0.6581	0.6646	0.6709
CMMN_{bin}	0.2372	0.2448	0.2655	0.2604	0.5931	0.6078	0.6199	0.6235
$\text{CMMN}_{\text{bin}}+\text{CNN}$	**0.3636**	**0.3879**	**0.3905**	**0.3875**	**0.7054**	**0.7019**	**0.6957**	**0.6801**

Table 2: MAP comparison of different cross-modal hashing methods on the Wiki dataset with 16,32,64 and 128 bits code length.

Query	NAT				CLP				SPT				LDR				DSC				mean
Target	CLP	SPT	LDR	DSC	NAT	SPT	LDR	DSC	NAT	CLP	LDR	DSC	NAT	CLP	SPT	DSC	NAT	CLP	SPT	LDR	mAP
Bl [3]	17.8	15.5	10.1	0.8	11.4	13.1	9.0	0.8	9.0	10.1	5.6	0.8	4.9	7.6	6.8	0.8	0.6	0.9	0.9	0.9	6.4
A [3]	14.0	29.8	6.2	18.4	9.2	17.6	3.7	12.9	21.8	15.9	6.2	27.7	3.7	3.1	6.6	5.4	5.2	3.5	10.5	2.1	11.2
B [3]	17.8	23.7	9.5	5.6	13.4	18.1	8.9	4.6	16.7	16.2	8.8	5.3	6.2	8.1	**9.4**	3.3	3.0	4.1	4.6	2.8	9.5
C [3]	14.3	**32.1**	5.4	22.1	10.0	**19.1**	3.8	14.4	24.4	17.5	5.8	32.7	3.3	3.4	6.0	4.9	**15.1**	**12.5**	**32.6**	4.6	14.2
Our	**34.9**	20.8	**38.4**	**36.1**	**14.3**	18.8	**39.4**	**36.9**	**26.3**	**38.7**	**40.9**	**40.5**	**7.5**	**19.8**	6.1	**18.6**	6.5	10.3	5.2	**9.5**	**23.5**

Table 3: MAP comparison of different methods on CMPlaces dataset. Each column shows a different query-target pair. On the far right, we average over all pairs.

from the kernel embedding. Moreover, Wiki is a small-scale dataset, there are no sufficient training samples to feed deep neural networks, while kernel method is more suitable for this situation. However, $\text{CMMN}_{\text{bin}}+\text{CNN}$ can outperform SePH_{km} with CNN features, which indicates that deep representation can be a potential manner to promote the performance of CMMN.

Results on CMPlaces. Because most past approaches are designed for two modalities or require sample-level alignments, which is missing in CMPlaces dataset. Therefore, we compare against three deep baselines of paper [3]. The mAP results are shown in Table 3. Quality retrieval examples are presents in Figure 3.

From Table 3, we can obtain the following observations. 1) Overall, our method outperforms the best configuration of [3] by 9%. and for most cross-modalities pairs (e.g. SPT query LDR) CMMN improves the performance at most 35%. 2) However, in several cases, especially for DSC as the query , the result of ours is inferior to [3]. We analyse the learned feature of each modality and find that the discriminative of raw feature belong to DSC is poor (classification accuracy is 4.5%), which means that its memory data contains more noise than prior knowledge. 3) Comparing NAT-CLP with CLP-NAT, we can find that the performances are asymmetric, we guess that it caused by the imbalance of training data size across modalities or the difference of discriminative.

Results on COCO. We compare CMMN with CMFH [5], SCM-Seq [32], unsupervised deep model corAE [6] and extended DSH [19] referred as DCN-S. Specifically, for alignment training samples in CMMN, we fusion features from different modalities to obtain unified hash codes as $b = sign(h_{img} + h_{txt} > 1)$. The mAP@500 results are reported in Table 4. We can observe that CMMN outperforms other baselines which well demonstrates its effectiveness. But DCN-S is inferior to both deep and shallow methods with ResNet features (corAE

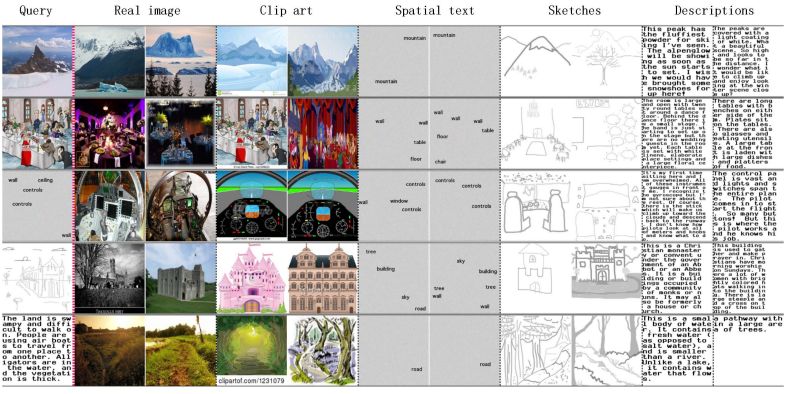


Figure 3: Quality retrieval examples on CMPlaces. The first column represents the query, and top 2 results for each modality are shown.

Method	Image query *v.s.* Text	Text query *v.s.* Image	Average
SCM-Seq [32]	0.4056	0.4439	0.4248
CMFH [8]	0.5309	0.6042	0.5675
DCN-S [19]	0.3750	0.4154	0.3952
corAE [6]	0.5179	0.6171	0.5675
CMMN	**0.5318**	**0.6831**	**0.6075**

Table 4: MAP@500 of different cross-modal hashing methods with 32 bits on COCO.

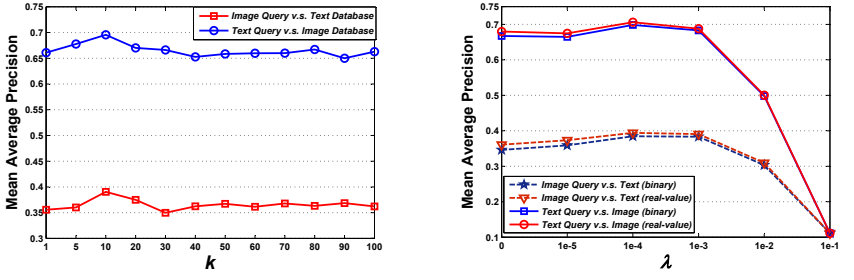
and CMFH), which implies that the performance gap between deep and shallow methods mainly lie in the representative of features.

4.3 Effect of Model Configuration

Here we carry out additional experiments to analyze the effects of the introduction of memory and Jacobin penalty. All experiments are conducted on Wiki dataset and training parameters are $epoch = 200, learning_rate = 0.001, batch\_size = 64, code_length = 64$.

Effect of Memory. In previous experiments, the k in formula (6) is empirically set as 10. We will vary k from 1 to 100 (almost half of training data as memory) with fixing $\lambda = 0.001$ to learn common features, and then measure their impact with cross-modal retrieval performance. Figure 4 (a) illustrates the effects of k. It can be observed that with k increasing, the performance of learned feature for Wiki (whether Image or Text) firstly increases and then decreases to flat. However, this result is not similar with MemNN for QA (the more memories the better performance). It indicates that the appropriate number of memories can help to learn a common representation of cross-modalities while no more useless.

Effect of Jacobi penalty term. The λ of model balance the effect of Jacobin term in formula (5), we vary it in $\{0, 10^{-4}, 10^{-3}, \ldots, 1\}$ with fixing $k = 10$ and observe its influence for performance. Figure 4 (b) shows the results. As we can see that with λ increasing, the quality of learned hash codes increases firstly after declines, besides, the performance gap between real-value and binary code are narrowed. This result is reasonable, because a suitable λ is useful to reduce the quantization loss and make the learnt feature (e.g. h in



(a) (b)

Figure 4: Effect of affecting factors on the proposed CMMN. Subfigure (a) illustrates the effects of Memory size k. (b) show the effects of hyperparameter λ for Jacobi penalty term.

formula (4)) near to the binary code, while a large λ may lead the optimization process to focus less on minimizing the classification loss and thus learn codes that cannot well preserve semantic information.

5 Conclusion

In this paper, we proposed a memory network termed CMMN for cross-modal retrieval. CMMN exploits memory mechanism to pre-store discriminative private features of potential relevant components across available modalities as memories. Then, given a query feature of a special modality, CMMN can find supporting facts from memories and learn common representation through aggregating and transforming these features. We compared CMMN with several state-of-the-arts on three datasets and achieves superior retrieval performance.

Acknowledgements: This work is partially supported by National Science Foundation of China (61373060, 61672280) and Qing Lan Project. The authors thank the anonymous reviewers for their helpful comments.

References

[1] Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. Deep canonical correlation analysis. In *ICML*, pages 1247–1255, 2013.

[2] Michael M. Bronstein, Alexander M. Bronstein, Fabrice Michel, and Nikos Paragios. Data fusion through cross-modality metric learning using similarity-sensitive hashing. In *CVPR*, pages 3594–3601, 2010.

[3] LluÍłs Castrejíőn, Yusuf Aytar, Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Learning aligned cross-modal representations from weakly aligned data. In *CVPR*. IEEE, 2016.

[4] Jianpeng Cheng, Li Dong, and Mirella Lapata. Long short-term memory-networks for machine reading. *Eprint Arxiv*, 2016.

[5] Guiguang Ding, Yuchen Guo, and Jile Zhou. Collective matrix factorization hashing for multimodal data. In *CVPR*, pages 2083–2090, 2014.

[6] Fangxiang Feng, Xiaojie Wang, and Ruifan Li. Cross-modal retrieval with correspondence autoencoder. pages 7–16, 2014.

[7] Yunchao Gong, Qifa Ke, Michael Isard, and Svetlana Lazebnik. A multi-view embedding space for modeling internet images, tags, and their semantics. *IJCV*, 106(2): 210–233, 2014.

[8] Alex Graves, Greg Wayne, and Ivo Danihelka. Neural turing machines. *Computer Science*, 2014.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[10] D. W. Jacobs, H. Daume, A. Kumar, and A. Sharma. Generalized multiview analysis: A discriminative latent space. In *CVPR*, pages 2160–2167, 2012.

[11] Jia, Yangqing, Shelhamer, Evan, Donahue, Jeff, Karayev, Sergey, Long, and Jonathan. Caffe: Convolutional architecture for fast feature embedding. *Eprint Arxiv*, 2014.

[12] Qing Yuan Jiang and Wu Jun Li. Deep cross-modal hashing. *Eprint Arxiv*, 2016.

[13] Xinyang Jiang, Fei Wu, Xi Li, Zhou Zhao, Weiming Lu, Siliang Tang, and Yueting Zhuang. Deep compositional cross-modal learning to rank via local-global alignment. In *ACM International Conference on Multimedia*, pages 69–78, 2015.

[14] Ryan Kiros, Yukun Zhu, Ruslan Salakhutdinov, Richard S Zemel, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. Skip-thought vectors. *Computer Science*, 2015.

[15] Shaishav Kumar and Raghavendra Udupa. Learning hash functions for cross-view similarity search. In *International Joint Conference on Artificial Intelligence*, pages 1360–1365, 2011.

[16] Kevin Lin, Huei Fang Yang, Jen Hao Hsiao, and Chu Song Chen. Deep learning of binary hash codes for fast image retrieval. In *CVPR Workshops*, pages 27–35, 2015.

[17] Tsung Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollĺćr, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755, 2014.

[18] Zijia Lin, Guiguang Ding, Mingqing Hu, and Jianmin Wang. Semantics-preserving hashing for cross-view retrieval. In *CVPR*, 2015.

[19] Haomiao Liu, Ruiping Wang, Shiguang Shan, and Xilin Chen. Deep supervised hashing for fast image retrieval. In *CVPR*, pages 2064–2072, 2016.

[20] Xinyan Lu, Fei Wu, Siliang Tang, Zhongfei Zhang, Xiaofei He, and Yueting Zhuang. A low rank structural large margin method for cross-modal ranking. In *International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 433–442, 2013.

[21] Alexander Miller, Adam Fisch, Jesse Dodge, Amir Hossein Karimi, Antoine Bordes, and Jason Weston. Key-value memory networks for directly reading documents. *Eprint Arxiv*, 2016.

[22] E. Coviello G. Doyle G. R. Lanckriet R. Levy N. Rasiwasia, J. Costa Pereira and N. Vasconcelos. A new approach to cross-modal multimedia retrieval. In *International Conference on Multimedea 2010*, pages 251–260, 2010.

[23] Salah Rifai, Pascal Vincent, Xavier Muller, Xavier Glorot, and Yoshua Bengio. Contractive auto-encoders: Explicit invariance during feature extraction. In *ICML*, 2011.

[24] Jingkuan Song, Yang Yang, Yi Yang, Zi Huang, and Heng Tao Shen. Inter-media hashing for large-scale retrieval from heterogeneous data sources. In *ACM SIGMOD International Conference on Management of Data*, pages 785–796, 2013.

[25] Sainbayar Sukhbaatar, Arthur Szlam, Jason Weston, and Rob Fergus. End-to-end memory networks. *Computer Science*, 2015.

[26] Kaiye Wang, Ran He, Wei Wang, Liang Wang, and Tieniu Tan. Learning coupled feature spaces for cross-modal matching. In *ICCV*, pages 2088–2095, 2013.

[27] Kaiye Wang, Ran He, Liang Wang, Wei Wang, and Tieniu Tan. Joint feature selection and subspace learning for cross-modal retrieval. *PAMI*, 38(10):2010, 2016.

[28] Mingxuan Wang, Zhengdong Lu, Hang Li, and Qun Liu. Memory-enhanced decoder for neural machine translation. *Eprint Arxiv*, 2016.

[29] Yanfei Wang, Fei Wu, Jun Song, Xi Li, and Yueting Zhuang. Multi-modal mutual topic reinforce modeling for cross-media retrieval. In *International Conference on Multimedia*, pages 307–316, 2014.

[30] Jason Weston, Sumit Chopra, and Antoine Bordes. Memory networks. *Eprint Arxiv*, 2014.

[31] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. Stacked attention networks for image question answering. In *CVPR*, pages 21–29, 2016.

[32] Dongqing Zhang and Wu Jun Li. Large-scale supervised multimodal hashing with semantic correlation maximization. In *AAAI*, pages 2177–2183, 2014.

[33] Bolei Zhou, Agata Lapedriza Garcia, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. Learning deep features for scene recognition using places database. pages 487–495, 2014.

[34] Jile Zhou, Guiguang Ding, and Yuchen Guo. Latent semantic sparse hashing for cross-modal similarity search. In *Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 415–424, 2014.

[35] Jile Zhou, Guiguang Ding, Yuchen Guo, Qiang Liu, and Xin Peng Dong. Kernel-based supervised hashing for cross-view similarity search. In *IEEE International Conference on Multimedia and Expo*, pages 1–6, 2014.

[36] Y. Zhuang, Y. Wang, F. Wu, Y. Zhang, and W. Lu. Supervised coupled dictionary learning with group structures for multi-modal retrieval. In *AAAI*, 2013.