Cross-modal semantic correlation learning by Bi-CNN network

Chaoyi Wang | Liang Li | Chenggang Yan | Zhan Wang | Yaoqi Sun | Jiyong Zhang

1 Hangzhou Dianzi University, Hangzhou, China
2 Institute of Computing Technology, CAS, Beijing, China
3 RTInvent Technology Co., Ltd, Beijing, China

Correspondence
Jiyong Zhang, Hangzhou Dianzi University, Hangzhou, China.
Email: jzhang@hdu.edu.cn

Abstract
Cross-modal retrieval can retrieve images through a text query and vice versa. In recent years, cross-modal retrieval has attracted extensive attention. The purpose of most now available cross-modal retrieval methods is to find a common subspace and maximize the different modal correlation. To generate specific representations consistent with cross-modal tasks, this paper proposes a novel cross-modal retrieval framework, which integrates feature learning and latent space embedding. In detail, we proposed a deep CNN and a shallow CNN to extract multi-level semantic representation of text. Meanwhile, we enhance the semantic manifold by constructing cross-modal ranking and within-modal discriminant loss to improve the division of semantic representation. Moreover, the most representative samples are selected by using online sampling strategy, so that the approach can be implemented on a large-scale data. This approach not only increases the discriminative ability among different categories, but also maximizes the relativity between different modalities. Experiments on three real word datasets show that the proposed method is superior to the popular methods.

1 INTRODUCTION

Under the booming of internet society, online multimedia information is growing at an amazing speed, and the content is usually presented in multiple ways. Soon afterwards, different forms of representation learning are essential for many practical applications, such as image captions and video captions, etc. In this paper, we are absorbed in cross-modal retrieval, which uses image queries to search text, and vice versa. The core of cross media retrieval is to explore semantic consistency between different modes. The efficacious retrieval of cross-media plays an important role in our daily applications. Figure 1 gives two demonstration of cross-modal. Differ from single-modal, cross-modal retrieval can provide multi-modal results. The key issue of cross-modal retrieval is that the distribution of different modalities is not consistent, so we cannot directly measure the cross-modal similarity.

In order to retrieve cross-media information, a great deal of researchers concentrate on building the common space models bridge different information modalities. Canonical Correlation Analysis (CCA) [1] and its variants [2–6] are the most popular methods. The basic concept of these methods is to obtain the common subspace by learning the maximum relativities among different modalities. For instance, [2] applies CCA to maximize the correlation among modalities. In addition, some works add additional supervised information to improve the CCA performance. In details, [3] jointly learns multiple additional information to improve the CCA performance. In details, [3] jointly learns multiple additional information to improve the CCA performance. In details, [3] jointly learns multiple additional information to improve the CCA performance. In details, [3] jointly learns multiple additional information to improve the CCA performance. In details, [3] jointly learns multiple additional information to improve the CCA performance. In details, [3] jointly learns multiple additional information to improve the CCA performance. In details, [3] jointly learns multiple additional information to improve the CCA performance. In details, [3] jointly learns multiple additional information to improve the CCA performance. In details, [3] jointly learns multiple additional information to improve the CCA performance. In details, [3] jointly learns multiple additional information to improve the CCA performance. In details, [3] jointly learns multiple additional information to improve the CCA performance. In details, [3] jointly learns multiple additional information to improve the CCA performance. In details, [3] jointly learns multiple additional information to improve the CCA performance. In details, [3] jointly learns multiple additional information to improve the CCA performance. In details, [3] jointly learns multiple additional information to improve the CCA performance. In details, [3] jointly learns multiple additional information to improve the CCA performance. In details, [3] jointly learns multiple additional information to improve the CCA performance. In details, [3] jointly learns multiple additional information to improve the CCA performance. In details, [3] jointly learns multiple additional information to improve the CCA performance. In details, [3] jointly learns multiple additional information to improve the CCA performance. In details, [3] jointly learns multiple additional information to improve the CCA performance. In details, [3] jointly learns multiple additional information to improve the CCA performance. In details, [3] jointly learns multiple additional information to improve the CCA performance. In details, [3] jointly learns multiple additional information to improve the CCA performance. In details, [3] jointly learns multiple additional information to improve the CCA performance. In details, [3] jointly learns multiple additional information to improve the CCA performance. In details, [3] jointly learns multiple additional information to improve the CCA performance. In details, [3] jointly learns multiple additional information to improve the CCA performance. In details, [3] jointly learns multiple additional information to improve the CCA performance. In details, [3] jointly learns multiple additional information to improve the CCA performance. In details, [3] jointly learns multiple additional information to improve the CCA performance. In details, [3] jointly learns multiple additional information to improve the CCA performance. In details, [3] jointly learns multiple additional information to improve the CCA performance. In details, [3] jointly learns multiple additional information to improve the CCA performance. In details, [3] jointly learns multiple additional information to improve the CCA performance. In details, [3] jointly learns multiple additional information to improve the CCA performance. In details, [3] jointly learns multiple additional information to improve the CCA performance. In details, [3] jointly learns multiple additional information to improve the CCA performance. In details, [3] jointly learns multiple additional information to improve the CCA performance. In details, [3] jointly learns multiple additional information to improve the CCA performance. In details, [3] jointly learns multiple additional information to improve the CCA performance. In details, [3] jointly learns multiple additional information to improve the CCA performance. In details, [3] jointly learns multiple additional information to improve the CCA performance. In details, [3] jointly learns multiple additional information to improve the CCA performance. In details, [3] jointly learns multiple additional information to improve the CCA performance. In details, [3] jointly learns multiple additional information to improve the CCA performance. In details, [3] jointly learns multiple additional information to improve the CCA performance. In details, [3] jointly learns multiple additional information to improve the CCA performance. In details, [3] jointly learns multiple additional information to improve the CCA performance. In details, [3] jointly learns multiple additional information to improve the CCA performance. In details, [3] jointly learns multiple additional information to improve the CCA performance. In details, [3] jointly learns multiple additional information to improve the CCA performance. In details, [3] jointly learns multiple additional information to improve the CCA performance. In details, [3] jointly learns multiple additional information to improve the CCA performance. In details, [3] jointly learns multiple additional information to improve the CCA performance. In details, [3] jointly learns multiple additional information to improve the CCA performance. In details, [3] jointly learns multiple additional information to improve the CCA performance. In details, [3] jointly learns multiple additional information to improve the CCA performance. In details, [3] jointly learns multiple additional information to improve the CCA performance. In details, [3] jointly learns multiple additional information to improve the CCA performance. In details, [3] jointly learns multiple additional information to improve the CCA performance. In details, [3] jointly learns multiple additional information to improve the CCA performance. In details, [3] jointly learns multiple additional information to improve the CCA performance. In details, [3] jointly learns multiple additional information to improve the CCA performance. In details, [3] jointly learns multiple additional information to improve the CCA performance. In details, [3] jointly learns multiple additional information to improve the CCA performance. In details, [3] jointly learns multiple additional information to improve the CCA performance. In details, [3] jointly learns multiple additional information to improve the CCA performance. In details, [3] jointly learns multiple additional information to improve the CCA performance. In details, [3] jointly learns multiple additional information to improve the CCA performance. In details, [3] jointly learns multiple additional information to improve the CCA performance. In details, [3] jointly learns multiple additional information to improve the CCA performance.
Deep neural network have been shown to be effective in many tasks [9–11], by provides scalable nonlinear transformation for feature representations in unimodal scenarios. An increasing number of researchers have deployed deep neural network in the cross-media retrieval [12–14], which then specifically exploit nonlinear ear correlations in learning generic subspace representations. The existing deep neural network based cross-media retrieval models can be mainly divided into two main phases. The first learning phase is to generate a separate representation for each modality, and the second learning phase is to obtain a generic representation across modalities. Although the best results can be ensured in the corresponding phase, it may produce suboptimal results for the cross-modal tasks. First, cross-modal generic representations may not be optimally compatible with visual and textual representations. Secondly, representation of different modalities are extracted independently of each other, thus ignoring the semantic correlation between the modalities. Third, the two phases of learning may lead to cumulative errors in which the representations of each modality are indistinguishable.

In this paper, we fully explicate category information to generate semantic representations of images and text in cross-media retrieval. We design a two-way CNN model named Cross-modal Semantic correlation learning by Bi-CNN Network (CSBN) which generates multi-level textual feature and generates discriminative feature of images. For the problem of cumulative errors, the bidirectional CNN employs probabilistic functions directly on textual and visual features, which can produce semantic representation of different modalities. Based on the two-way CNN model, we propose a novel ranking loss function for modelling semantic relationship across modalities. In contrast to existing cross-media retrieval algorithms, which only focus on the relevance of paired items, the proposed ranking loss function does not require image-text pair as input. In addition, two label predictors are learned so that they jointly perform label prediction and preserve the underlying cross-modal semantic structure in data. Therefore, the proposed framework ensures that the learned representations are both distinguishable within modalities and invariant between modalities.

As shown in Figure 2, the proposed framework generate semantic representation of images and texts. On the top half of the figure, we generate semantic representations of images via ResNet50 [9]. Obviously, words describe objects image, phrases describe attributes or activities of objects, and the whole sentence expresses a complete meaning covering semantic information of the entire image. Therefore, CSBN use kernels of multiple sizes to extract word and phrase features. In addition, max-pooling is used for word and phrase features to generate a semantic representation of the sentences.

The remainder of this paper is organised as follows. In Section 2 related work is discussed. In Section 3, we provide details of the CSBN and its target optimisation. In Section 4, we present experimental results and discussions for three real data sets. Finally, conclusions and future work are given in Section 5.

2 | RELATED WORK

Measuring semantic similarity between visual and textual features is the key problem in cross-media retrieval. The common idea is to learn a joint embedding space where heterogeneous data representation can be directly compared. Traditional methods and deep neural network based methods are two main streams of cross-media retrieval.

For the traditional approach, several works [1–4, 12, 15–20] focus on learning correlations between different modalities. These methods aim to learn the common subspace of images and text by maximising the correlation between the different transmitted feature vectors of different modalities. Alternative technique [7, 8, 21] adapts the cross-media retrieval to a manifold alignment problem. These methods make the assumption that they embedded high dimensional data in a low dimensional intrinsic space. The joint embedding representation can be achieved by learning an underlying manifold. To transfer
images and texts into a common semantic space, a number of ranking based approaches [22, 23] have been proposed in recent years. Ref. [24] projects heterogeneous data into a semantic space by applying multiclass logistic regression to preserve the semantic similarity. These methods learn semantic models by using bi-directional training samples, which combine the advantages of both directions of retrieval, thus improving generation performance. Lastly, to exploit label information, some algorithm [5, 6, 25–28] directly minimize the cross-entropy loss between the predictions and the class labels. [29, 30] are combined with fine-grained [31] calculate the semantic similarity between the instances by both the label information and feature information of instances. The correlation between modalities is formulated by class labels. Therefore, class information helps to learn discriminative potential spaces.

Deep learning has achieved state-of-the-art performance in some tasks of single-modal scenario, such as image classification [9], object detection [11] and 3D layout [32]. These excellent works have inspired researchers to model the intrinsic semantic correlations between different modalities through deep neural networks. Ref. [23] seeks an effective common subspace by adopting adversarial learning, in which adversarial learning is implemented as an interaction between two processes. To address large-scale cross-media retrieval, [12] propose a deep multi-network framework that decomposes large matrices into many sub-matrices and devises novel sampling strategies to select the most representative samples to construct cross-pattern ranking losses. Ref. [33] constructs different semantic fragments from words by using a multimodal convolutional neural network, and then the fragments are interacted with image at different levels. Although m-CNN considers multi-level matching relations, it uses different matching levels as separate steps, thus ignores the intrinsic relation from words to sentences. Refs. [34, 35, 58] incorporates domain adaptation to transform synthetic data into realistic images. Ref. [36] proposes a framework that does not require any adjustment of parameters or thresholds. Ref. [57] transformed the language information into different expressions to improve the accuracy of prediction. Ref. [37] uses a bi-modal CNN to train a holistic scene classifier in two modalities, and then semantic relevance of the sub-concepts included in the images are used to improve overall scene recognition. Ref. [13] seeks an effective public subspace based on adversarial learning. Adversarial learning is implemented as an interplay between the two processes. The first process attempts to generate a modality-invariant representation in the common subspace, while the other process attempts to distinguish between different modalities based on generated representation. Ref. [38] proposes an approach, the modality-specific and modality-shared features are jointly explored and leveraged, such that the complementarity and correlation information is effectively used for the retrieval task.

Generally, cross-modal retrieval tasks require low storage capacity and high retrieval efficiency, and recently, hashing has received more and more attention. Ref. [39] uses the semantic tag information provided by data set, combining the classifier learning and matrix decomposition method. Ref. [40] uses a hybrid deep structure, which constitutes a visual semantic fusion network, is used to learn hash functions to generate binary codes for generational comparisons. Furthermore, the architecture combines joint multi-modal embedding and cross-modal hashing with advantage, and it is based on a novel combination of Convolutional Neural Networks on images. Recurrent Neural Networks over text and a structured maximum margin targets to learn high quality hash codes. A triplet-based deep hash network for cross-modal retrieval is proposed in [41], which utilises a triplet table and establishes a multi-view loss function for hash codes. In addition, graph regularisation is introduced to preserve the original semantic similarity between hash codes in Hamming space. Ref. [42] proposed a new hash learning method that combines multi-view and deep learning methods.
Ref. [43] proposed an efficient discrete optimal algorithm to directly learn discrete hash code matrix with closed-form solution instead of learning them bit-by-bit. Ref. [44] employed the adversarial learning, it utilised two adversarial networks to maximise the consistency of semantic associations and representations between different modalities. Moreover, high-level semantic information in the form of multi-label annotations is also available via a self-supervised semantic network.

3 | THE PROPOSED METHOD

In this section, we introduce CSBN for cross-media retrieval, as shown in Figure 2. The CSBN is a bi-directional convolution neural network, one deep convolution neural network on images and one shallow convolution neural network on sentences. Furthermore, a novel ranking loss and a label projector is deployed to learn high quality inconsistent representation.

3.1 | Image model

In this paper, we extract image feature \( f_I \) from an original image \( I \) by applying ResNet. Specifically, the first step is to rescale the each image to be 224×224 pixels. We alter the last fully connected layer in order to fit our dataset. Then we obtain the features \( f_I \) from the last fully connected layer. Therefore, the dimension of \( f_I \) is \( c \), where \( c \) is the category number of dataset. Finally, in order to generate the category probability of image, we apply a softmax function to the feature vector.

\[
V = \text{softmax}(f_I)
\]  

(1)

The image semantic representation is denoted by \( V = \{v_1, v_2, \ldots, v_n\} \), where \( v_i \) represents the probability of image belong to the \( i \)th category.

3.2 | Textual model

The pre-trained word vectors are used as embedding layer, and each row of the embedding matrix is a word vector. We perform multikernel convolution on the chopped sentences. Different size of convolution kernels gets different feature size, so we use pooling function for each feature map so that they have the same dimensionality.

Word-level feature. We define \( W^v_{\text{embedding}} \in \mathbb{R}^{n \times k} \) as a word embedding matrix, where \( n \) is the size of vocabulary and \( k \) is the dimension of word embedding. Let \( w_i \in \mathbb{R}^k \) be the \( k \)-dimensional word vector corresponding to the \( i \)th word in the text. Text with length \( n \) is represented as

\[
W^v = w_1 \oplus w_2 \oplus \ldots \oplus w_n
\]  

(2)

where \( \oplus \) is the concatenation operator. Here, we initialise the word embedding matrix using pre-trained word2vec and fine-tune it by back-propagation.

Phrase-level feature. Generally, let \( w_{i:i+b} \) refer to the concatenation of words \( w_i, w_{i+1}, \ldots, w_{i+b} \). A 1-D convolution operation involves a filter \( W_{\text{Conv}} \in \mathbb{R}^{n \times k} \), which is applied to a window of \( b \) words to produce a \( b \)-gram feature. For instance, a feature \( p_i \) is generated from a window of words \( w_{i:i+b} \) by

\[
p_i = f(W_{\text{Conv}} \cdot w_i \!: i + b - 1 + b)
\]  

(3)

Here \( b \in R \) is a bias term and \( f \) is a non-linear function such as ReLU. This filter is applied to words in each possible window of in the text \( \{w_1, b, w_2, b+1, \ldots, w_n-b+1:n\} \) to produce a feature map

\[
p = [p_1, p_2, \ldots, p_n - b + 1]
\]  

(4)

With \( p \in \mathbb{R}^{n-b+1} \). Then we apply a max-pooling operation over the feature map and take the maximum value \( p^* = \max\{p\} \) as the feature corresponding to this particular filter.

Text-level feature. In order to make full use of the semantic information of the word sequence \( W \). We concatenate the output feature of each filter as \( \hat{p} = [p^*_1, \ldots, p^*_c] \). Then, the phrase-level vector \( p^* \) is feed into a single layer perceptron, which can combine the phrase-level feature to generate the discriminative feature of text.

\[
T = \text{softmax}(W_T \hat{p} + b_T)
\]  

(5)

Where \( W_T \) is an transformation matrix and \( b_T \) is a bias term. In a word, our text representation is a step-by-step progressive from low-level word features to high-level text features. After gaining the multi-modal representation, we apply the label predictor to learn discriminatory representation with a modality and propose a rank loss to model the correlation between modalities.

3.3 | Label prediction

To ensure discrimination of the intra-modal data, two classifiers are deployed to predict the category labels of the samples in the common space. For this target, there is a cross-entropy loss function on the top of each representation output. The semantic representations of the instance of images or texts are taken as training data by the classifier and the classifier generates a probability distributions \( p^*(v, T) \). The formulation of intra-modal classification loss as follow:

\[
L_n = -\frac{1}{n} \sum_{i=1}^{n} y_i \cdot \log p_{*, i}, * \in \{V, T\}
\]  

(6)

\( L_n \) denote the cross-entropy loss of image or text category classification of all the samples. Here, \( n \) express the number of samples within each mini-batch, \( y_i \) denotes the groundtruth of each instance, while \( p^* \) is the generated probability distribution per image or text.
3.4 Cross-modal ranking loss

If two samples are close in the primitive space, they should approach each other in the common subspace. In the cross-media retrieval tasks, we usually need to rank the retrieved results so that samples have same semantics close to each other. Hence, ranking loss is naturally suitable for this mission, which can learn to rank the retrieved results by distance or similarity. Inspired to [12, 45], our model proposes a new and original triplet loss to evaluate the relative ranking distance between given query and a certain retrieved samples so that more differences can be surveyed in contrasts between retrieved samples.

Here we want to make sure that an image \( v^a \) (anchor) is closer to all other texts (positive) from the same category, rather than any text \( v^p \) (negative) with different semantic labels. This is visualized in Figure 3. We formulate the relationship as follow

\[
\| v^a_i - v^p_j \|_2^2 + \alpha \leq \| v^a_i - v^n_j \|_2^2 \forall (v^a_i, v^p_j, v^n_j) \in T
\]  

where, \( \alpha \) is a margin that is enforced between positive and negative pairs. \( T \) is the set of all possible triplets in the mini-batch. As mentioned above, there is a relationship between the text anchor and the corresponding image sample.

\[
\| v^a_i - v^p_j \|_2^2 + \alpha \leq \| v^a_i - v^n_j \|_2^2 \forall (v^a_i, v^p_j, v^n_j) \in T
\]  

Therefore the triplet loss constraints can be written as

\[
L_{\text{triplet}} = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{k=1}^{N} \left( I[k_{\text{via}} = v^a_i, k_{\text{tpk}} = v^p_j, k_{\text{ink}} = v^n_k] + \alpha \right) + \left( I[k_{\text{via}} = v^a_i, k_{\text{tpk}} = v^p_j, k_{\text{ink}} = v^n_k] + \alpha \right)
\]

Because of the large number of samples, it is impossible to sample triplets in the whole instance space is impractical. We perform online triplet sampling from labelled instances in each mini-batch. First, the distance matrix between the cross modal samples is calculated. Second, the distance matrix is used to select all samples as anchor points, positive samples and negative samples. Finally, we calculate the triplet loss by anchors, negative samples and positive samples.

3.5 Formulation of CSBN

On the basis of the above, the loss function of the CSBN is expressed as the combination of the inter-modal ranking loss and the intramodal discrimination loss. The variables of the two-way CNN are learnt by end to end. Our approach not only increases the discriminative ability among different categories, but also maximizes the correlation between different modalities. The jointly loss can be formulated as:

\[
\min L_V + L_T + \gamma L_{\text{triplet}} \theta
\]

where, \( \theta \) is the model variables which will be optimized. The hyperparameters \( \gamma \) control the contributions of the three terms.

4 EXPERIMENTS

In this section, we introduce the experimental results of our proposed CSBN and other comparison methods (image query text and text query image) for cross media retrieval. The comparisons are implemented on three public datasets Wikipedia [2], Pascal Sentence [46] and NUS-WIDE [47]. Although we used these three datasets for our experiments, the properties of these three datasets are not the same. Their text modality are not the same and their categories do not coincide. Wikipedia is the article, PASCAL Sentence is the captions and NUS-WIDE is the tags. So the features of different modalities are totally different.

4.1 Datasets

Wikipedia dataset is first propose by [2] for cross-media task. The dataset consists of 2866 text-image pairs, annotated with a label from 10 semantic classes vocabulary. We randomly split following [48]: 2173 pairs for training, 231 pairs for validation and 462 pairs for testing.

Pascal Sentence dataset [46] derives from 2008 PASCAL development kit. This dataset consists of 1000 images which are categorized into 20 categories, and each image has 5 corresponding sentences. For each category, we select 40 image-text pairs for training, five pairs for testing and five pairs for validation following [48].

NUS-WIDE dataset [47] consists of 270,000 images with their tags categorized into 81 classes. We follow [49] which prune the original train-test split of the NUS dataset by keeping the pairs which belongs to 10 largest classes. Each image-tags pair belongs to a unique class.

4.2 Compared methods

The proposed CSBN is compared with several related methods, such as CCA [1], 3view-CCA [5], SM [2], SCM [2], GMMFA [3], GMLDA [3] and LGCFL [15]. CCA focuses on learning a common subspace and analysing the relevance between different modalities in the subspace. 3view-CCA uses CCA as a...
starting point, combined with a third view capturing high-level image semantics that can be represented by a single category or multiple non-mutually exclusive concepts. SM learns two logistic regression variables to generate probability for each sample, and uses category probability to represent samples. SCM applies the CCA to learn two maximally related subspaces, and then studies the logistic regressors in each subspace. GMA constructs a general discriminant subspace and proposes a generalized multi-view analysis. GMMFA and GMLDA all use the framework of GMA. LGCFL learns the consistent features by applying the valuable class information on the cross-media retrieval.

### 4.3 Implementation details

For CSBN, we extract features directly from the original pixels of the image using a residual network. With respect to sentences, we transformed the word sequences in each sentence into a word index as input to the text part. For all the methods compared in our experiments, we extracted feature vectors of the same size from the images by means of the same residual network. Word vectors are obtained by adopting word2vec pre-trained on extra dataset for each sentence, and then we calculate the average of all word vectors as the final feature vector.

We implement our method in Tensorflow framework. For the training of the network, we employ ResNet-50 for image, and the convolutional network for text. To initialize our network, we apply the pre-trained ResNet-50 variables and word2vec word embedding. Back-propagation is applied to fine-tune the ResNet and text-CNN and train the new layers. Due to the new layer is trained from scratch, we first adopt the fixing lower layers to train it. Then, we adjust slightly all layers and reduce the learning rate to one tenth of training new layers. The proposed CSBN is implemented using PyTorch. During training, we use SGD optimizer with 0.9 momentum. The maximum number of words in each sentence for Wikipedia training, we use SGD optimizer with 0.9 momentum. The proposed CSBN is implemented using PyTorch. During training, we use SGD optimizer with 0.9 momentum.

### 4.4 Experiment settings

We utilize the precision-recall (PR), scope-precision (SP) and mean average precision (MAP) to measure the retrieval results. MAP is the mean value of the average precision (AP) for each query in the query set. AP is defined as:

\[
P = \frac{1}{N} \sum_{k=1}^{n} \frac{R_k}{k} \times rel_k \tag{11}\]

where, \(n\) is the size of test set, \(N\) is the number of retrieval results, \(R_k\) is the \(k\)th accumulated positive samples in the retrieval list. \(rel_k = 1\) if the \(k\)th sample at the returned retrieval list is relevant, and \(rel_k = 0\) otherwise.

In particular, we measure the retrieval performance by adopting MAP@R [50] after the fixed number of back sampling. If the first 50 samples are retrieved, \(R\) is set to 50, if all samples are retrieved, \(R\) is set to "all". In addition, the precision-recall curves and scope-precision curves are also given for all methods. The scope is specified by the number of samples with the highest ranking when the retrieved samples are ranked on the basis of the similarities between them and the query.

The parameters \(\gamma\) and \(\alpha\) need to be set for our CSBN framework. We set \(\gamma\) and \(\alpha\) according to five-fold cross validation result on the Pascal Sentence dataset. Then, the parameter is used for all of the experiments. In detail, the parameter \(\gamma\) are set 1.0, the \(\alpha\) is set 5.0. As shown Figure 4, our CSBN is not very sensitive to \(\alpha\) parameter.

### 4.5 Image-text retrieval

Tables 1, 2 and 3 report the MAP@50 and MAP@All of the different methods on the test set of Wikipedia, Pascal Sentence and NUS-WIDE, respectively. From these tables, we conclude the following:

As shown in Table 1, CSBN achieves the best results in both bi-modal retrieval tasks on Wikipedia dataset. The excellent performance of CSBN is due to the addition of the labels predictors and cross-modal ranking loss, which can produce distinctive feature and preserve the original data structure.

First of all, CSBN performs better than the relevant comparison method on either task tasks. CCA has the worst performance, because CCA focuses on the common subspace learning and ignores the semantic correlation among different modalities. CSBN utilizes a new ranking loss function to model cross-modal semantic relations so it works better than CCA. Second, SM adopt two logistic regressors to generate samples category probability, which achieves comparable results on both tasks. CSBN chooses to use softmax and label prediction, which has a significant improvement over SM. Third, 3-View CCA, SCM, GMMFA and GMLDA have higher MAP scores than the CCA. In fact, these three methods are based on the CCA, and they introduce the semantic modelling in the CCA framework. This proves the significance of the semantic correlation constraint. Fourth, LGCFL excels over traditional methods by learning consistent features for cross-modal retrieval. LGCFL models both paired and unpaired image-text loss, but it still has the disadvantage of not being able to cluster samples with the same class well in the common space. So, CSBN uses triplet loss for semantic clustering and achieves better results. As a result, the MAP@All of CSBN is higher than LGCFL by 1.6% and 7.2% on text-query-images and image-query-texts retrieval tasks. Ref. [51] reduces the distribution differences between different modes to enhance the similarity. In contrast to it, We are directly optimizing the distance between feature vectors to reduce their differences. In comparison, the average Map@ALL scores of our method on the NUS-WIDE dataset is about 0.12 higher than it.

Figure 5(a) and (b) show the precision-recall and precision scope curves of the eight methods. The scope (i.e., the top
Cross-modal retrieval results by cross-validation on the training set of the Pascal Sentence dataset for our CSBN

### TABLE 1  MAP scores on Wikipedia dataset

| Tasks | $R = 50$ | $R = \text{All}$ |
|-------|----------|-----------------|
|       | Text query | Image query | Average | Text query | Image query | Average |
| CCA [1] | 0.398 | 0.382 | 0.390 | 0.294 | 0.326 | 0.310 |
| SM [2] | 0.352 | 0.418 | 0.375 | 0.266 | 0.371 | 0.319 |
| SCM [2] | 0.468 | 0.416 | 0.442 | 0.350 | 0.381 | 0.366 |
| GMMFA [3] | 0.259 | 0.372 | 0.316 | 0.185 | 0.350 | 0.265 |
| GMLDA [3] | 0.237 | 0.413 | 0.325 | 0.182 | 0.355 | 0.269 |
| 3view-CCA [5] | 0.468 | 0.420 | 0.444 | 0.336 | 0.371 | 0.354 |
| LGCFL [15] | 0.498 | 0.440 | 0.469 | 0.364 | 0.405 | 0.385 |
| CSBN | 0.526 | 0.466 | 0.496 | 0.380 | 0.477 | 0.429 |

### TABLE 2  MAP scores on Pascal sentence dataset

| Tasks | $R = 50$ | $R = \text{All}$ |
|-------|----------|-----------------|
|       | Text query | Image query | Average | Text query | Image query | Average |
| CCA [1] | 0.442 | 0.395 | 0.419 | 0.405 | 0.360 | 0.383 |
| SM [2] | 0.413 | 0.452 | 0.433 | 0.320 | 0.370 | 0.345 |
| SCM [2] | 0.434 | 0.425 | 0.430 | 0.419 | 0.409 | 0.414 |
| GMMFA [3] | 0.396 | 0.459 | 0.428 | 0.385 | 0.456 | 0.421 |
| GMLDA [3] | 0.387 | 0.455 | 0.421 | 0.381 | 0.450 | 0.416 |
| 3view-CCA [5] | 0.422 | 0.498 | 0.460 | 0.406 | 0.475 | 0.441 |
| LGCFL [15] | 0.517 | 0.508 | 0.513 | 0.506 | 0.502 | 0.504 |
| CSBN | 0.523 | 0.565 | 0.544 | 0.532 | 0.566 | 0.549 |

Precision-recall curves and precision-scope curves for the image-query-texts and text-query-images experiment on Wikipedia dataset
TABLE 3  MAP scores on NUS-WIDE dataset

| Methods       | $R = 50$   | $R = \text{All}$ |
|---------------|------------|------------------|
|               | Text query | Image query | Average | Text query | Image query | Average |
| CCA [1]       | 0.410      | 0.416       | 0.413   | 0.251      | 0.276       | 0.264   |
| SM [2]        | 0.762      | 0.459       | 0.611   | 0.439      | 0.492       | 0.466   |
| SCM [2]       | 0.448      | 0.379       | 0.414   | 0.315      | 0.309       | 0.312   |
| GMMFA [3]     | –          | –           | –       | –          | –           | –       |
| GMLDA [3]     | –          | –           | –       | –          | –           | –       |
| 3view-CCA [5] | 0.540      | 0.481       | 0.510   | 0.274      | 0.307       | 0.291   |
| LGCFL [15]    | 0.806      | 0.538       | 0.672   | 0.511      | 0.589       | 0.550   |
| CSBN          | 0.787      | 0.502       | 0.645   | 0.523      | 0.603       | 0.563   |

FIGURE 6  Precision-recall curves and precision-scope curves for the image-query-texts and text-query-images Experiment on Pascal Sentence

K retrieved samples) of precision-scope curve diff from 100 to 500. We can observe that our method achieves higher precision at almost all levels, whether it is text-query-images retrieval or image-query-texts retrieval. For precision-scope curve, the CSBN performance is superior to the compared methods at all levels. Hence, the superiority of CSBN for cross-modal retrieval is further validated by curves.

The Pascal Sentence is widely used to assess the effectiveness of cross-modal retrieval algorithms. The results on Pascal Sentence are shown in Table 2. The trends of the results on Pascal Sentence are consistent with Wikipedia datasets. Figure 6(a,b) are the precision-recall and precision-scope curves of image-query-texts and text-query-images respectively. From Figure 6, we can find the performance of CSBN is unsatisfactory at low recall level. The purpose of cross-modal rank loss is to learn discriminative semantic representation of different modalities, which leads to sub-optimal performance at low recall level.

Moreover, we also try to run the codes of GMA on NUS-WIDE. But our device cannot meet their memory requirements, thus we do not report their results. As shown on Table 3, CSBN does not achieve the highest scores on MAP@50, which may result from the anchor of cross-modal ranking loss not sensitive to the neighbor samples. However, our CBSN approach can stably benefit from two-way CNN framework and keeps the almost best among all compared methods. Figure 7 shows the precision-recall curves and precision-scope curves of bi-modal retrieval tasks on NUS-WIDE datasets, which can further verify the effectiveness of our CSBN approach.

In Figure 8, to visually determine the search results, we give two retrieved instances of image-query-texts and text-query-images. For each instance, we display the query on the left, and the five most relevant retrieved results are displayed on the right. We can observe that our method achieves higher precision at almost all levels, whether it is text-query-images retrieval or image-query-texts retrieval. For precision-scope curve, the CSBN performance is superior to the compared methods at all levels. Hence, the superiority of CSBN for cross-modal retrieval is further validated by curves.

4.6 Semantic representations

In this section, we attempt to illustrate the difference in recognition power of low-dimensional feature representations learned by different methods. Based on Pascal Sentence, we extracted 50 sample pairs each from the ‘bicycle’ category and the ‘aeroplane’ category to construct a toy dataset. In Figure 9, we adopt the t-SNE [52] algorithm to project the semantic representations from the two-way network and the embedding features of the different methods are transformed into a mixture of images and textual representations in a 2D distribution. The red circles denote the distribution of the ‘aeroplane’ class, and green circles represent the ‘bicycle’ class.

From Figure 9, we first conclude the proposed CSBN maximizes the correlation between different modalities, and simultaneously dataset increases the discriminative ability among different categories. However, the second best result (i.e., LGCLF) only unifies the semantic representations of same-class samples. Moreover, the CSBN also differs from the other
FIGURE 7  Precision-recall curves and precision-scope curves for the image-query-texts and text-query-images Experiment on NUS-WIDE dataset

| Query                                                                 | Results                                                                 |
|-----------------------------------------------------------------------|-------------------------------------------------------------------------|
| ![Query Image](image1.png)                                            | Someone is putting something in the white **dogs mouth**                |
| ![Query Image](image2.png)                                            | Close up of black and white **bulldog face**                           |
| ![Query Image](image3.png)                                            | A small **puppy** being fed a chocolate treat                          |
| ![Query Image](image4.png)                                            | A **dog** looking out of a window while standing on its hind legs       |
| ![Query Image](image5.png)                                            | A man holds a ball in a **puppies mouth**                               |
| ![Query Image](image6.png)                                            | ![Image Examples](image7.png)                                          |
| ![Query Image](image8.png)                                            | ![Image Examples](image9.png)                                          |

FIGURE 8  Two examples of image-query-texts (in upper half) and text-query-images (in bottom half) on Pascal Sentence dataset. For image-query-texts, the query image depicts the scene about “dog”, “mouth” and “face”. For sentence-query-images, the query sentences describes the semantics about “car”, “road”. The incorrect retrieved results are shown in the red frame.

FIGURE 9  Semantic representations of images and texts from ‘aeroplane’ and ‘bicycle’ classes on Pascal Sentence dataset.
methods in the range of sample distribution, which is spread over the same range of coordinates. These results validate that CSBN ensures the consistent structures between image and text spaces such that the low-dimension representation is enhanced with stronger discrimination.

### 4.7 Quality of word embedding

To better comprehend the quality of learned word embeddings, we provide empirical analysis by showing the nearest neighbours for each given word in Table 4. CSBN adopts an end-to-end training mode, the word embeddings matrix is regarded as one part of model variables, which can be learned during training. Then we measure the similarity between words by using cosine distance, like word2vec model [53, 54].

As shown in Table 4, we find that the closet entities of each given word are clearly relevant to each other. For instance, when we input the word “women”, some related words like “man”, “people” and “men” appear at the top positions of the ranking list. We think that the advantages of CSBN benefit from the two-way CNN framework, which coherently combines the deep ResNet and shallow text-CNN layers for images and words feature respectively. As is known to all, ResNet is the most effective way of deep image feature extraction [55] and CNN layers show powerful ability in text classification [56]. In order to enable the semantic information in images and text to be mutually guided, this is achieved by unifying the architectures of two different CNNs. As a result, the word embeddedds will be updated in a more informative direction so that the relevant words are likely to be close to each other in embedding space.

### 5 CONCLUSION

In this paper, we propose a cross-media retrieval framework named CSBN based on two-way convolutional neural networks. In CSBN, images and text representation are generated by the two-way CNN, which is proposed in this paper. Specifically, the deep CNN (i.e., ResNet) is applied for images representation and the shallow multisize kernels CNN generates multi-level texts feature. Further, the cross-modal ranking loss reduces differences in the distribution of categories across modalities. Two label predictors are adopted to generate discriminative sample representation. Experimental results on three public cross-modal datasets show that CSBN works better than the related method on tasks. In future, we prioritise improving the effectiveness of our method and plan to make the CSBN more practical.

**ORCID**

Chaoyi Wang https://orcid.org/0000-0002-1422-2451

**REFERENCES**

1. Hardoon, D.R., Szedmak, S., Shawetaylor, J.: Canonical correlation analysis: An overview with application to learning methods. Neural Comput. 16(12), 2639–2664 (2004).
2. Rasiwasia, N., et al.: A new approach to cross-modal multimedia retrieval. In: Proceedings of the 18th ACM international conference on Multimedia, Firenze, Italy (2010).
3. Sharma, A., et al.: Generalized multiview analysis: A discriminative latent space. In: IEEE Conference on Computer Vision and Pattern Recognition, Providence (2012).
4. Pereira, J.C., et al.: On the role of correlation and abstraction in cross-modal multimedia retrieval. TPAMI 36(3), 521–535 (2014).
5. Gong, Y., et al.: A multi-view embedding space for modeling internet images, tags, and their semantics. ICCV 106(2), 210–223 (2014).
6. Ranjan, V., Rasiwasia, N., Jawahar, C.: Multi-label cross-modal retrieval. In: IEEE International Conference on Computer Vision (ICCV), Santiago, Chile (2015), pp. 4094–4102.
7. Mao, X., et al.: Parallel field alignment for cross media retrieval. In: Proceedings of the 21st ACM international conference on Multimedia, Amsterdam (2013), pp. 897–906.
8. Mahadevan, V., et al.: Maximum covariance unfolding: Manifold learning for bimodal data. In: Conference on Neural Information Processing Systems (NIPS), Granada, Spain (2011), pp. 918–926.
9. He, K., et al.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas (2016), pp. 770–778.
10. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, Massachusetts (2015), pp. 3431–3440.
31. Yan, C., et al.: Multimodal convolutional neural networks for matching image and sentence. In: Proceedings of the IEEE International Conference on Multimedia and Expo (ICME), Hong Kong (2017), pp. 1374–1379.
32. Zhang, L., et al.: Cross-modal retrieval using multi-tiered discriminative structured subspace learning. IEEE Trans. Multimedia 19(6), 1220–1233 (2017).
33. Zhang, L., et al.: Adaptively unified semisupervised learning for cross-modal retrieval. In: International Conference on Artificial Intelligence, Melbourne, Australia (2017), pp. 3406–3412.
34. Zou, L., et al.: Research on subspace cross modal retrieval method based on high-order semantic correlation. Data Anal. Knowl. Discovery 4(5), 84–91 (2020).
35. Yan, C., et al.: Depth image denoising using nuclear norm and learning graph model. ACM Tran. Multimedia Comput. Commun. Appl. 16(4), 1–17 (2020).
36. Zhang, L., et al.: PI-ranking: a novel ranking method for cross-modal retrieval. In: Proceedings of the 2016 ACM on Multimedia Conference, Amsterdam (2016), pp. 1335–1346.
37. Wang, L., Li, Y., Lazebnik, S.: Learning deep structure-preserving image-text embeddings. In: Proceedings of the IEEE conference on computer vision and pattern recognition, Las Vegas (2016), pp. 5005–5013.
38. Wang, K., et al.: Semantic Boosting Cross-Modal Hashing for efficient multimedia retrieval. Inf. Sci. 330, 199–210 (2016).
39. Wang, K., et al.: Joint feature selection and subspace learning for cross-modal retrieval. IEEE Trans. Pattern Anal. Mach. Intell. 38(10), 2010–2023 (2016).
40. Wang, K., et al.: Learning coupled feature spaces for cross-modal matching. In: IEEE International Conference on Computer Vision, Sydney (2013), pp. 2088–2095.
41. Zhai, X., Peng, Y., Xiao, J.: Learning cross-modal joint representation with sparse and unsupervised regularization. IEEE Trans. Circuits Syst. Video Technol. 24(6), 965–978 (2014).
42. Dong, H., et al.: Knowledge base enrichment by relation learning from social tagging data. Inf. Sci. 526, 203–220 (2020).
43. Min, S., et al.: Multi-objective matrix normalization for fine-grained visual recognition. IEEE Trans. Image Process. 29, 4996–5009 (2020).
44. Liu, C.: Bidirectional attention recognition model for fine-grained object classification. IEEE Trans. Multimedia 22(7), 1785–1795 (2020).
45. Yao, T., et al.: Efficient discrete supervised hashing for large-scale cross-modal retrieval. Neurocomputing 385, 358–367 (2020).
46. Yan, C., et al.: 3D room layout estimation from a single RGB image. IEEE Trans. Multimedia 22(11), 3014–3024 (2020).
47. Ma, L., et al.: Multimodal convolutional neural networks for matching image and sentence. In: Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile (2015), pp. 2623–2631.
48. Wang, Q., et al.: Pixel-wise crowd understanding via synthetic data. Int. J. Comput. Vision 129, 225–245 (2021).
49. Wang, Q., et al.: Learning from synthetic data for crowd counting in the wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition CVPR, Long Beach (2019).
50. Wang, Q., et al.: Detecting coherent groups in crowd scenes by multiview clustering. IEEE Trans. Pattern Anal. Mach. Intell. 42(1), 46–58 (2018).
51. Zhan, M., et al.: Cross-media retrieval with semantics clustering and feature and classifier learning. Pattern Recognit Artif. Intell. 33(2), 160–165 (2020).
52. Rashtchian, C., et al.: Supervised coupled dictionary learning with group structures for multi-modal retrieval. In: Proceedings of the 27th AAAI Conference on Artificial Intelligence, Bellevue, Washington USA pp. 1070–1076 (2013).
53. Wu, F., et al.: Cross-media semantic representation via bi-directional learning to rank. In: Proceedings of the 21st ACM international conference on Multimedia. Amsterdam (2013), pp. 877–886.
54. Zhan, M., et al.: Cross-media retrieval with semantics clustering and enhancement. In: 2017 IEEE International Conference on Multimedia and Expo (ICME), Hong Kong (2017).
55. Maaten, L.v ..d.., Hinton, G.: Visualizing data using T-SNE. J. Mach. Learn. Res. 9(Nov), 2579–2605 (2008).
56. Mikolov, T., et al.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013).
57. Mikolov, T., et al.: Distributed representations of words and phrases and their compositionality. arXiv preprint arXiv:1310.4546 (2013).
58. Kim, J.-H., et al.: Multimodal residual learning for visual qa[j]. arXiv preprint arXiv:1606.01455 (2016).
59. Kim, Y.: Convolutional neural networks for sentence classification. arXiv preprint arXiv:1408.5882 (2014).
60. Long, Y., et al.: RepeatPadding: Balancing words and sentence length for language comprehension in visual question answering. Inf. Sci. 529, 166–178 (2020).
61. Min, S., et al.: Domain-aware visual bias eliminating for generalized zero-shot learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle (2020).