Multimedia Tools and Applications (2019) 78:16615–16631
https://doi.org/10.1007/s11042-018-7068-0

Two-stage deep learning for supervised cross-modal retrieval

Jie Shao[1] · Zhicheng Zhao[1,2] · Fei Su[1,2]

Received: 25 April 2018 / Revised: 5 November 2018 / Accepted: 11 December 2018 /
Published online: 19 December 2018

Abstract
This paper deals with the problem of modeling internet images and associated texts for cross-modal retrieval such as text-to-image retrieval and image-to-text retrieval. Recently, supervised cross-modal retrieval has attracted increasing attention. Inspired by a typical two-stage method, i.e., semantic correlation matching(SCM), we propose a novel two-stage deep learning method for supervised cross-modal retrieval. Limited by the fact that traditional canonical correlation analysis (CCA) is a 2-view method, the supervised semantic information is only considered in the second stage of SCM. To maximize the value of semantics, we expand CCA from 2-view to 3-view and conduct supervised learning in both stages. In the first learning stage, we embed 3-view CCA into a deep architecture to learn non-linear correlation between image, text and semantics. To overcome over-fitting, we add the reconstruct loss of each view into the loss function, which includes the correlation loss of every two views and regularization of parameters. In the second stage, we build a novel fully-convolutional network (FCN), which is trained by joint supervision of contrastive loss and center loss to learn better features. The proposed method is evaluated on two publicly available data sets, and the experimental results show that our method is competitive with state-of-the-art methods.

Keywords Two-stage · 3-view · Reconstruct loss · Center loss · Contrastive loss

1 Introduction

Over the last decade there has been a massive explosion of multimedia content on the web. More and more people upload pictures tagged with texts to the Internet. Articles describing

The first two authors contributed equally to this work.

✉ Zhicheng Zhao
zhaozc@bupt.edu.cn

Jie Shao
shaojielyg@163.com

Fei Su
sufei@bupt.edu.cn

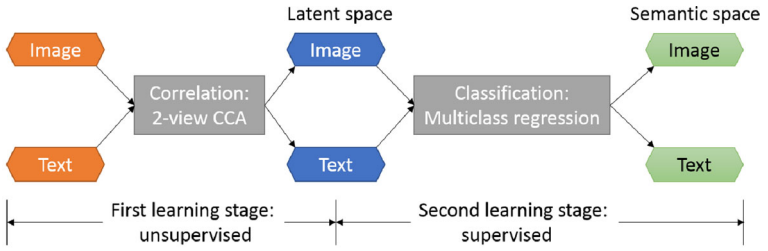
[1] School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing, China

[2] Beijing Key Laboratory of Network System and Network Culture, Beijing University of Posts and Telecommunications, Beijing, China

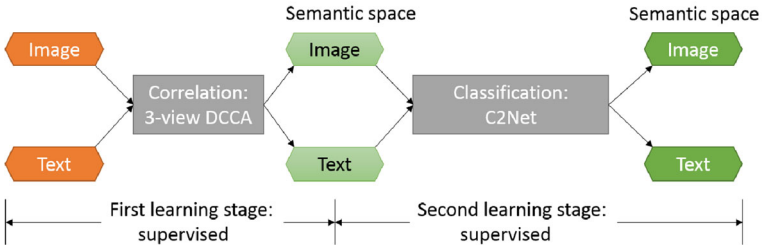
news and technologies include a lot of pictures and texts. The presence of massive multi-modal data on the Internet brings a growing demand for cross-modal retrieval, such as using a text query to search for images, and in turn using an image query to search for texts.

Many unsupervised methods [7, 8, 14, 26, 34] have been proposed for cross-modal retrieval. These methods focus on modeling intra-pair correlation between different views. However the inter-pair correlation which corresponds to the semantic consistency of retrieved results is ignored. To address this problem, several methods [16, 22, 23, 25, 41, 46] are proposed to regularize the latent space with the help of the supervised semantic information. Semantic correlation matching(SCM) [23] is a very typical method to conduct supervised cross-modal retrieval. The learning procedure of SCM can be divided into two learning stages (Fig. 1a): in the first learning stage, SCM uses 2-view canonical correlation analysis(CCA) to project image view and text view into a common latent space. In the second learning stage, SCM extracts high-level semantic representations of image view and text view based on the features in the latent space. Similarity measured in the semantic space is used for cross-modal retrieval. However, the supervised semantic is only considered in the second stage of SCM. This drawback of SCM is limited by the fact that traditional CCA is a 2-view method for building correlation. To maximize the value of supervised semantics, we propose to conduct supervised learning in both stages, as shown in Fig. 1b.

Deep learning models have achieved great success in representation learning recently. Deep models use a cascade of multiple layers of nonlinear processing units for feature extraction and possess significantly greater representation power than traditional shallow models. However, deep learning methods are prone to over-fitting because of the added layers of abstraction, which allow them to model rare dependencies in the training data. To address this problem, several methods [6, 37, 43] are proposed to learn better features with



(a) Framework of SCM.



(b) Framework of our method.

Fig. 1 Frameworks of SCM and our method

multiple supervisory signals. Multiple supervisory signals are beneficial for enhancing the discriminative power of the deeply learned features, i.e., center loss is added to reduce the intra-class features variations in face recognition. What's more, these supervisory signals can be viewed as the regularization term for each other, thus overcome overfitting. Inspired by this fact, we propose to combine correlation learning with representation learning in the first stage. In the second stage, the deep network named C2Net is trained with the joint supervision of contrastive loss [13] and center loss [43].

1.1 Related work

Many approaches have been proposed to develop solutions to cross-modal retrieval tasks. We classify these approaches into two categories: conventional shallow models and deep models.

Shallow models Grangier et al. [12] propose a passive-aggressive model, which is the first attempt to address the problem of ranking images retrieved by text queries. Rasiwasia et al. [23] propose correlation matching to map the features of images and texts into a common latent space using CCA. Complete introduction and recent extensions about CCA can be found in [35, 36]. Based on CCA, various variants [2, 5, 27, 48] are proposed to model the multi-modal correlations. Gong et al. [11] expand two-view CCA to three-view CCA by incorporating a third view that captures high-level image semantics, represented either by a single category or multiple non-mutually-exclusive concepts. Wu et al. [45] formalize the retrieval task as a ranking problem(Bi-CMSRM) similar to [12] and try to learn a common latent space for images and texts as CCA. The latent space embedding of Bi-CMSRM is discriminatively learned by the structural large margin learning for optimization with certain ranking criteria (mean average precision) directly. Other algorithms are also proposed to deal with cross-modal retrieval problems, such as partial least square (PLS) [24], Bilinear Model(BLM) [18, 21, 38] and etc.

Deep models Srivastava et al. [34] propose to learn a generative model of the joint space of image and text inputs using Deep Belief Network(DBN), which consists of multiple stacked restricted boltzmann machine(RBM [28]). Gaussian RBM [42] and replicated softmax RBM [15] are used to model the real-valued feature vectors for image and the discrete sparse word count vectors for text, respectively. Based on DBN, Feng et al. [6] propose to learn the latent space of image and text inputs by correspondence autoencoder(Corr-AE). Corr-AE defines a novel optimal objective, which minimizes a linear combination of representation learning errors for each modality and correlation learning error between hidden representations of two modalities. Moreover, we notice several deep learning methods [7, 8, 29, 44] for learning a joint embedding space of image and text very recently. Andrew et al. [1] build a deep architecture (DCCA)to learn complex nonlinear transformations of two views of data such that the resulting representations are highly linearly correlated. DCCA can be viewed as a nonlinear extension of traditional linear CCA. Though the representation power improves, DCCA is easy to over-fitting, especially when the datasets are not big enough. Peng et al. [22] propose the cross-media multiple deep network (CMDN) to exploit the complex cross-media correlation by hierarchical learning. Huang et al. [16] propose cross-modal deep metric learning with multi-task regularization (CDMLMR). Recently, we also notice several impressive methods [9, 10, 30–33, 39, 40, 47] for cross-modal retrieval and image analysis.

1.2 Contributions

The main contributions of this work are summarized as follows:

1) Inspired by SCM, we propose a two-stage deep learning method for supervised cross-modal retrieval. We expand CCA from 2-view to 3-view to conduct supervised learning in both stages, which could maximize the value of supervised semantics.
2) Multiple supervisory signals are successfully combined in both stages to learn better feature representations, and meanwhile, overcome the overfitting.
3) The promising results on two public datasets demonstrate the effectiveness of the proposed algorithm.

2 Methods

In this section we describe details of 3-view DCCA in the first learning stage and details of C2Net in the second learning stage.

2.1 3-view DCCA

The first learning stage of SCM is unsupervised due to the fact that traditional CCA is a 2-view method for building correlation between image view and text view. To reinforce the value of supervised semantics, we expand CCA from 2-view to 3-view and maximize the correlation among 3 views simultaneously. To learn nonlinear correlation between these 3 views, we embed 3-view CCA into a deep architecture. Meanwhile, in order to overcome over-fitting of deep network, we add the reconstruct loss of each view into the loss function.

In this section, we first give the difference between 2-view CCA and 3-view CCA. Next, we introduce the architecture of 3-view DCCA and the formulation details.

2.1.1 3-view CCA

Practical models for cross-modal retrieval tasks should meet two requirements. First, the top one result should be accurate. It is a big challenge because image features are noisy and text features are ambiguous. Second, the top n results should be relevant to the query. In other words, image and text features which are coherent in semantics should be close to each other in the latent space for cross-modal retrieval. 2-view CCA (Fig. 2a) which only

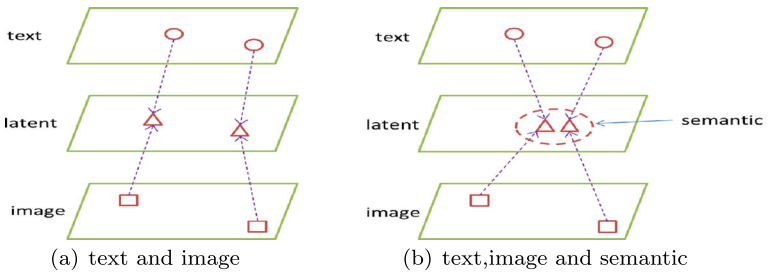


Fig. 2 **a** Paired points in text and image spaces are mapped into a common latent space independently in 2-view CCA. **b** Mapped points from relevant pairs are clustered in the latent space according to their semantic information in 3-view CCA

maximize the correlation between image and corresponding text does not meet the second requirement. To address this problem, We use the 3rd semantic view to unite all the relevant paired samples (Fig. 2b).

2.1.2 Model architecture

As shown in Fig. 3, 3-view DCCA includes three subnets(feedforward neural network), each of them corresponds to different view(image view, text view and semantic view) of data. The semantic view, is added to regularize the common latent space of text view and image view. Each subnet includes four kinds of layers: input layer, hidden layer, output layer and reconstruct layer. The representations learned from the output layer are furnished to the second learning stage.

2.1.3 Formulation

Assuming that we have n training images and each image is associated with a k1-dimensional visual feature vector, a k2-dimensional text feature vector, and a k3-dimensional semantic feature vector. The respective vectors are stacked as rows in matrices $X1 \in \Re^{n*k1}$, $X2 \in \Re^{n*k2}$, and $X3 \in \Re^{n*k3}$. Traditional two-view CCA tries to find matrices $W1 \in \Re^{k1*k}$ and $W2 \in \Re^{k2*k}$ where text and image are projected into a k-dimensional latent space so that the correlation between text and image could be maximized. This procedure is equal to minimizing the distance between text and image in the latent space. The loss function for two-view CCA is given by:

$$L_{cca(1,2)} = -corr(X_1 * W_1, X_2 * W_2) \tag{1}$$

Therefore, the loss function for our three-view CCA model includes the correlation loss of every two-view:

$$L_{cca(1,2,3)} = L_{cca(1,2)} + L_{cca(2,3)} + L_{cca(3,1)} \tag{2}$$

$$L_{cca(i,j)} = -corr(X_i * W_i, X_j * W_j) \tag{3}$$

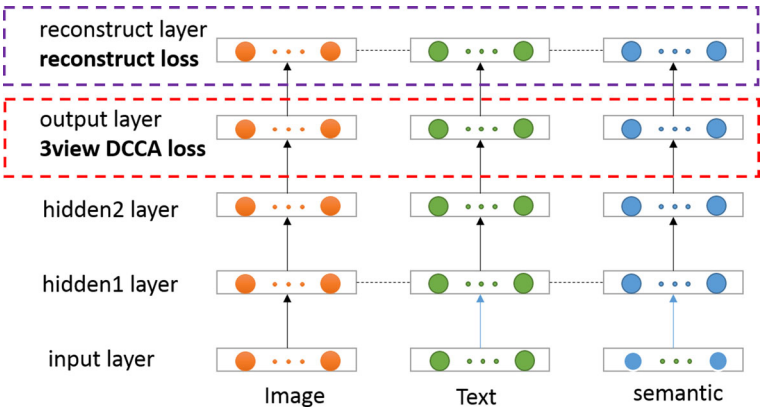


Fig. 3 The architecture of 3view DCCA

To handle complex transformations, we define three nonlinear functions $f_1(X_1, \Theta_1)$, $f_2(X_2, \Theta_2)$, $f_3(X_3, \Theta_3)$ and update the equations above:

$$L_{cca(i,j)} = -corr(f_i(X_i, \Theta_i), f_j(X_j, \Theta_j)) \tag{4}$$

In our deep model, $C_i = f_i(X_i, \Theta_i)$ denotes the representation of the output layer in the ith subnet. Meanwhile, in order to overcome over-fitting, we add the reconstruction loss and regularization penalty:

$$\begin{aligned} Loss &= L_{cca(1,2,3)} + \alpha \cdot \left(L_{1,1} + L_{2,2} + L_{3,3}\right) \\ &\quad + \lambda \left(\|\Theta_1\|_F^2 + \|\Theta_2\|_F^2 + \|\Theta_3\|_F^2\right) \\ L_{i,i} &= \left\|X_i - \hat{X}_i\right\|_F^2 \quad i = 1, ..., 3 \end{aligned} \tag{5}$$

where, $\hat{X}_i$ is the reconstruction data from X_i. The parameters $\Theta_1, \Theta_2, \Theta_3$ are trained to optimize this quantity using gradient-based optimization. In the output layer, the gradient of $corr(C_i, C_j)$with respect to C_i is calculated following the solution in [1]. Let $\overline{C}_i = C_i - \frac{1}{n}C_i$ be the centered data matrix and define $\hat{\sum}_{12} = \frac{1}{n-1}\overline{C}_1\overline{C}_2'$, $\hat{\sum}_{11} = \frac{1}{n-1}(\overline{C}_1\overline{C}_1' + r_1 I)$ for regularization constant r_1, and $\hat{\sum}_{22} = \frac{1}{n-1}(\overline{C}_2\overline{C}_2' + r_2 I)$ for regularization constant r_2. The total correlation $corr(C_1, C_2)$ is the sum of the singular values of the matrix $T = \hat{\sum}_{11}^{-\frac{1}{2}}\hat{\sum}_{12}\hat{\sum}_{22}^{-\frac{1}{2}}$. If the singular value decomposition of T is $T = UDV'$, then

$$\frac{\partial corr(C_1, C_2)}{\partial C_1} = \frac{1}{n-1}\left(-\hat{\sum}_{11}^{-\frac{1}{2}} UDU' \hat{\sum}_{11}^{-\frac{1}{2}} \overline{C}_1 + \hat{\sum}_{11}^{-\frac{1}{2}} UV' \hat{\sum}_{22}^{-\frac{1}{2}} \overline{C}_2\right) \tag{6}$$

The gradient of $L_{i,i}$ with respect to $\hat{X}_i$ in the reconstruct layer is:

$$\begin{aligned} \frac{\partial L_{i,i}}{\partial \hat{X}_i} &= \frac{\partial \left\|X_i - \hat{X}_i\right\|_F^2}{\partial \hat{X}_i} \\ &= \frac{\partial Tr\left(X_i - \hat{X}_i\right)\left(X_i - \hat{X}_i\right)'\right)}{\partial \hat{X}_i} \\ &= 2\left(\hat{X}_i - X_i\right) \end{aligned} \tag{7}$$

The loss function is minimized by using L-BFGS method, which is particularly suitable for the optimization of a large number of variables. Any mistake in the computation of gradient or loss will lead to the failure of line search in L-BFGS. This is very helpful for developers.

2.2 C2Net

To further improve the cross-modal retrieval performance, we build a fully-convolutional network to learn better representations. In this section, we describe details of our C2Net.

2.2.1 Model architecture

An illustration of the C2Net structure is shown in Fig. 4. Our network includes 3 fully-convolutional layers followed by batch normalization layers, taking the representations obtained from 3-view DCCA as the input. With the help of batch normalization layers,

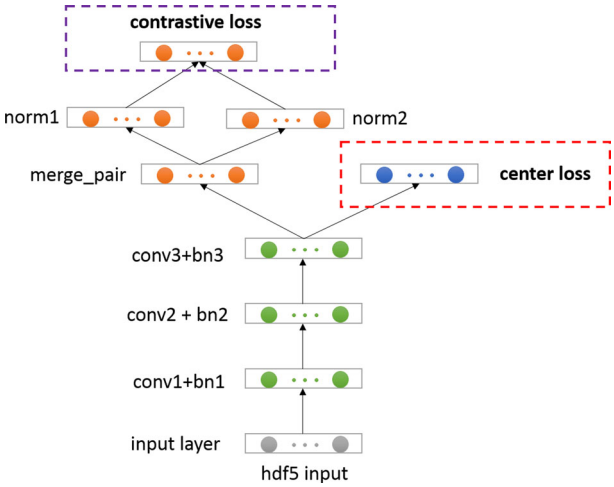


Fig. 4 The architecture of C2Net

our model is able to speed up the training from a higher learning rate, without ill side effects. Similar pairs and dissimilar pairs are obtained in the pair-merging layer. Features are $L2-$normalized before calculating the contrastive loss.

2.2.2 Formulation

Contrastive loss [13] is proposed to learn an invariant mapping for dimension reduction. Similar points are mapped to nearby points on the output manifold and dissimilar points to distant points. In cross-modal retrieval, paired text and image are similar if they are from the same semantic class. The contrastive loss [13] is defined to penalize large distance between intra-class pairs and distance smaller than a margin between inter-class pairs:

$$L_{con} = \frac{1}{2}\sum_{i=1}^{m/2} s_i d_i^2 + (1 - s_i) max(\gamma - d_i, 0)^2 \tag{8}$$

where $d_i = ||x_{2i-1} - x_{2i}||_2$, γ is the margin, m is the batch size, and binary s_i specifies whether x_{2i-1} and x_{2i} belong to the same class. The gradient of L_{con} with respect to x_{2i-1} is computed as:

$$\frac{\partial L_{con}}{\partial x_{2i-1}} = \begin{cases} x_{2i-1} - x_{2i} & s_i = 1 \\ -\frac{\gamma - d_i}{d_i}(x_{2i-1} - x_{2i}) & s_i = 0, \gamma - d_i > 0 \\ 0 & s_i = 0, \gamma - d_i \leq 0 \end{cases} \tag{9}$$

Center loss is proposed to minimize the intra-class distances of deep features and has been verified to be able to effectively enhance the discriminative power of features. Namely, the learned features are not only separable but also discriminative in terms of the compactness of intra-class features. The center loss function is defined as:

$$L_{cen} = \frac{1}{2}\sum_{i=1}^{m} ||x_i - c_{y_i}||_2^2 \tag{10}$$

where c_{y_i} denotes the center of the y_i-th class, and m is the batch size. The gradient of L_{cen} with respect to x_i is:

$$\frac{\partial L_{cen}}{\partial x_i} = x_i - c_{y_i} \tag{11}$$

c_{y_i} changes along with the changes of deep features. The error term of c_j is defined as:

$$\Delta c_j = \frac{\sum_{i=1}^{m} \delta\{y_i = j\}(c_j - x_i)}{\epsilon + \sum_{i=1}^{m} \delta\{y_i = j\}} \tag{12}$$

where ϵ is added to cope with the absence of samples belonging to the $j-$th class, and $\delta\{condition\}$ is defined as:

$$\delta\{condition\} = \begin{cases} 1 & if\ \ condition \\ 0 & otherwise \end{cases} \tag{13}$$

3 Experiments and results

3.1 Data sets

Wikipedia contains 2866 text-image pairs belonging to 10 semantic categories. We randomly split the data set into three subsets: 2,173 pairs for training, 231 pairs for validation and the last 462 pairs for testing. Each image is represented by three descriptors, including 1000-D pyramid histogram of dense SIFT, 512-D Gist, and 784-D MPEG-7. And each text is represented by 3000-D bag of high-frequency words.

NUS-WIDE-10k. This data set is a subset of NUS-WIDE [3], which contains about 269,648 images with tag annotations from 81 categories. Because some categories are scare, we only choose 10 most common categories. We have 8000 image-text pairs for training, 1000 for parameter validation, and 1000 for test. For image representation, six types of low-level features are extracted from these images, including 64-D color histogram, 144-D color correlogram, 73-D edge direction histogram, 128-D wavelet texture, 225-D block-wise color moments and 500-D bag of words based on SIFT descriptors. For text representation, we use 1000-D bag of words.

3.2 Evaluation metric

We use mean average precision (mAP) as the evaluation criterion. Given one query and top-R retrieved data, the average precision is defined as:

$$\frac{1}{N} \sum_{i=1}^{R} prec(i) * rel(i) \tag{14}$$

where, N is the number of the relevant documents in the retrieved set, prec(i) is the percentage of the relevant text documents (images) in the top-i text documents (images). rel(i) is an indicator function, when the i-th result is relevant to the query it equals 1, otherwise 0.

3.3 Implement details

For 3-view DCCA, feature whitening and dimension reduction are applied. For Wikipedia(NUS-WIDE-10k), image features and text features are reduced to 128-D(256-D). We perform grid search for the number of units in the hidden layers with the setting

128, 256, 384, 512, 768, 1024. In the output layer, the number of units is set to 9 for both data sets. The reconstruct layer has the same number of units as the input layer. After setting up the network architecture, the parameters of each layer are pretrained with a Denoising AutoEncoder(DAE), which reconstructs the original input from a distorted status, to ensure the representations of the input are robust to small irrelevant changes. For the tolerance of L-BFGS algorithm, we use the default setting of DCCA. The implementation of 3-view DCCA is built on Galen Andrew's DCCA library [1].

As for C2Net, the implementation is built based on Caffe [17]. Features obtained from 3-view DCCA are stored in hdf5 format. C2Net consists of 3 fully-convolutional layers with 256 filters of size 1×1, followed by batch normalization layers. The network is trained with a base learning rate 0.1 by stochastic gradient descent with 0.9 momentum, and the weight decay parameter is 0.005. C2Net is trained with batch size of 1024.

Experiments for our method are conducted on a computer which has Intel i7 4.0 GHZ 8 processors, 8 GB RAM,6GB Nvidia Gefore GTX TITAN GPU, Intel MKL 11.2 and Windows7. The training time increases with the number of training instances increasing. The training time of our method on NUS-WIDE-10k is less than 20 minutes.

3.4 Performance comparison

In this subsection, we compare our approach with several typical cross-modal retrieval methods, including both supervised and unsupervised algorithms. For all baselines listed below, two cross-modal tasks are investigated: text-to-image retrieval and image-to-text retrieval.

CCA [14]: CCA learns a common latent space by maximizing the correlation between original features of image view and text view.

PCA-CCA : CCA is performed on features reduced by PCA.

KCCA [4]: KCCA is a nonlinear extension of CCA. And we choose Gaussian kernel RBF as the kernel function of KCCA.

DCCA [1]: DCCA builds two separate deep networks to learn nonlinear correlation between text and image. Input features of text view and image view are whitened and reduced. This preprocessing is the same for all DCCA-based methods.

CFA [19]: CFA proposes to learn orthogonal transformation for each modal by minimizing the pair-wise distances in the latent space.

Multimodal DBN [34]: Multimodal DBN learns a shared representation for image view and text view with a separate two-layer DBN.

Bimodal AE [20]: Bimodal AE connects image features and text features with a joint layer followed by two intra-modal autoencoders.

Corr-AE [6]: Corr-AE builds two separate uni-modal autoencoders and try to learn the correlation between two modalities by correlating hidden representations of two autoencoders. Corr-AE performs grid search for the number of hidden units of each layer with the setting 32, 64, 128, 256, 512, 1024. The number of units for all hidden layers in Corr-AE is restricted to the same. The code with parameter specifications of Multimodal DBN, Bimodal AE and Corr-AE is available online.[1]

DCCA-PHS [26]: DCCA-PHS, our previous method, proposes to extract semantic features from text features with hypergraph learning. We embed image view, text view and

[1] https://github.com/fangxiangfeng/deepnet

Table 1 mAP@50 results of different methods for image-to-text(Img2Txt) and text-to-image(Txt2Img) retrieval

Methods	Wikipedia			NUS-WIDE		
	Img2Txt	Txt2Img	average	Img2Txt	Txt2Img	average
CCA	0.173	0.179	0.176	0.286	0.297	0.292
PCA-CCA	0.295	0.313	0.304	0.340	0.341	0.341
KCCA(RBF)	0.293	0.262	0.278	0.333	0.342	0.338
DCCA	0.235	0.226	0.231	0.284	0.290	0.287
CFA [19]	0.315	0.328	0.322	0.324	0.332	0.328
Bimodal AE [20]	0.282	0.327	0.305	0.250	0.297	0.274
Bimodal DBN [34]	0.189	0.222	0.206	0.173	0.203	0.188
Corr-AE [6]	0.336	0.368	0.352	0.331	0.379	0.355
DCCA-PHS [26]	0.341	0.379	0.360	0.395	0.408	0.387
3-view CCA*	0.337	0.401	0.369	0.384	0.434	0.409
SCM*	0.347	0.382	0.365	0.412	0.501	0.457
JRL* [46]	0.310	0.386	0.348	0.348	0.458	0.403
CMDN* [22]	0.360	0.487	0.424	0.432	0.497	0.465
CDMLMR* [16]	0.388	0.517	0.453	0.487	0.553	0.520
Our method*	0.380	0.474	0.427	0.447	0.511	0.479

Methods marked with "*" are supervised

unsupervised semantic into a progressive framework to learn complex nonlinear correlation. The numbers of layers and numbers of nodes per layer in DCCA-PHS are described in the experiment section of [26].

3-view CCA : For 3-view CCA, an Eigenvector implementation is available at http://www.unc.edu/~yunchao/crossmodal.htm. We implement 3-view CCA based on DCCA. It's clear that 3-view CCA can be viewed as a single layer DCCA with linear activation function.

SCM [23]: SCM combines correlation matching(CM), an unsupervised method which models cross-modal correlation, and semantic matching(SM), an supervised technique that relies on semantic representation. Our method is inspired by SCM. The source code of CCA and SCM is available at http://www.svcl.ucsd.edu/projects/crossmodal/.

JRL [46]: JRL simultaneously learns linear projections for different modalities with semi-supervised regularization and sparse regularization.

CMDN [22]: CMDN proposes to learn intra-modal representations and inter-modal representations with Stacked AutoEncoders(SAE) and Multimodal DBN [34] respectively. Both representations are combined to generate the shared representations with Bimodal Autoencoder [20] in a stacked style. The implementation of CMDN is based on deepnet.[2]

CDMLMR [16]: CDMLMR integrates quadruplet ranking loss and semi-supervised contrastive loss for modeling cross-modal semantic similarity in a unified multi-task learning architecture.

Table 1 summarizes the mAP@50 scores of different methods for cross-modal retrieval. PCA-CCA outperforms CCA on both data sets. Dimension reduction is an effective way

[2] https://github.com/nitishsrivastava/deepnet

to boost the performance of CCA. So we apply dimension reduction on DCCA and 3-view DCCA. The performance of DCCA is worse than PCA-CCA due to over-fitting. 3-view CCA outperforms CCA, which indicates that the semantic information is very useful for improving the semantic consistency in the latent space. Overall, supervised methods achieve better performance than unsupervised ones. Our method can be viewed as a nonlinear extension of SCM. It benefits from the better representation power of deep models. CDMLMR achieves state-of-the-art performance in terms of mAP@50. However, CDMLMR is trained with the help of data from the test set in the semi-supervised learning stage as mentioned in [16], which is not fair for other methods. So in the following section, we mainly compare our method with another state-of-the-art method, CMDN. Although our algorithm and CMDN build two cascaded networks, the network structures are quite different: CMDN is optimized with multiple reconstruct loss, L2 loss and softmax loss, while our method is trained by multiple reconstruct loss, CCA loss, contrastive loss and center loss. The reconstruct loss in representation learning is very beneficial for correlation learning in cross-modal retrieval. CCA loss and L2 loss are common losses for building correlation. In our experiments, CCA loss could achieve the same performance as L2 loss with much fewer dimensions. We have also tested single softmax loss, but did not achieve satisfactory results. The advantages of CMDN are the combination of two representations in the first stage and the stacked architecture in the second stage.

Figure 5 presents a more detailed analysis of the retrieval performance, in the form of 11-point interpolated precision-recall(PR) curves. These PR curves show that our method attains higher precision at lower levels of recall, which indicate that the top retrieved results of our method are more relevant to the queries. However, our method fails to obtain higher

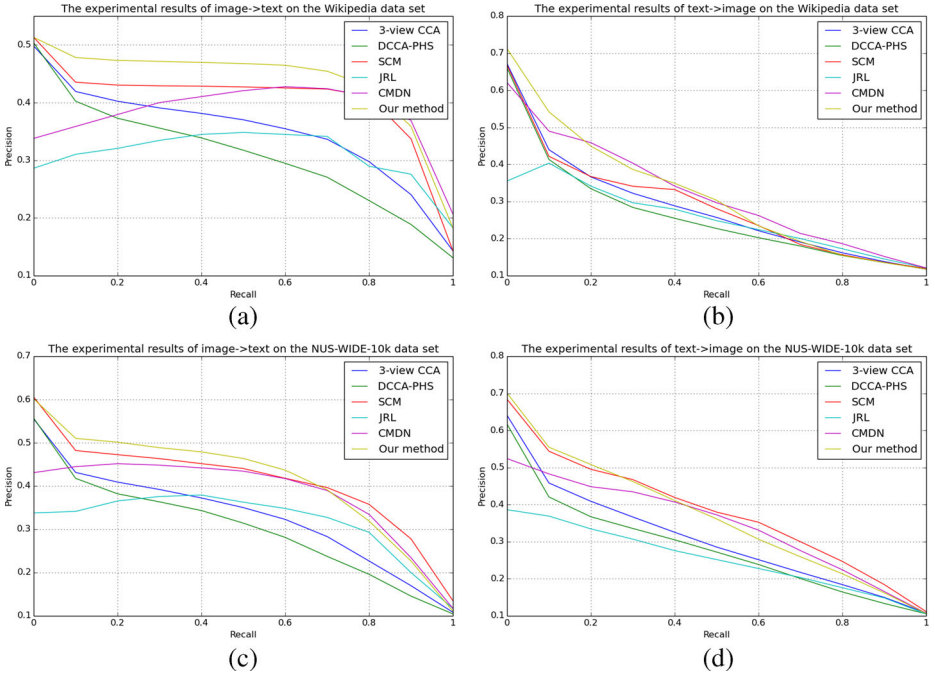
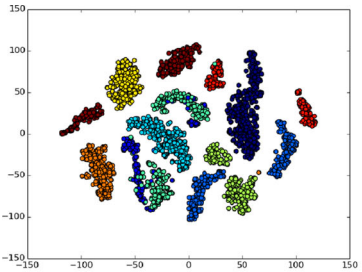
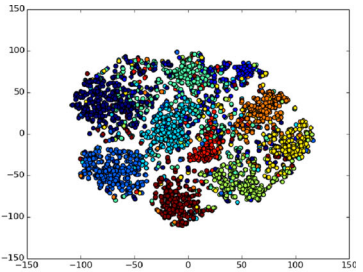


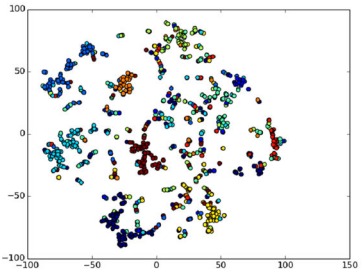
Fig. 5 The PR curves on Wikipedia and NUS-WIDE-10k



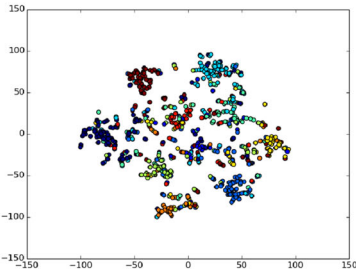
(a) Wikipedia train α=0



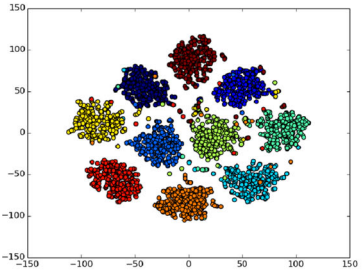
(b) Wikipedia train α=1



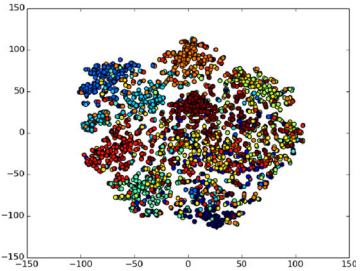
(c) Wikipedia test α=0



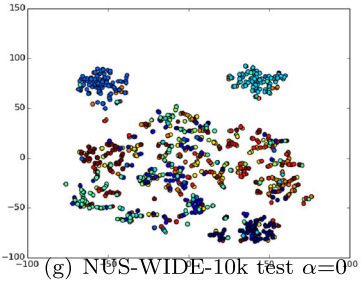
(d) Wikipedia test α=1



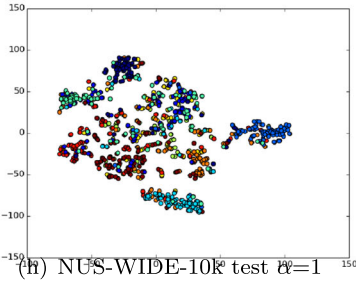
(e) NUS-WIDE-10k train α=0



(f) NUS-WIDE-10k train α=1



(g) NUS-WIDE-10k test α=0



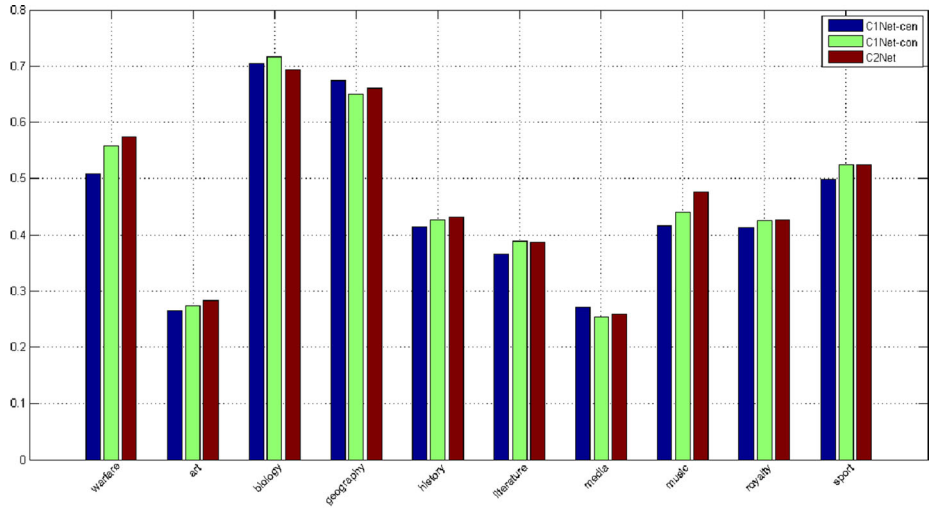
(h) NUS-WIDE-10k test α=1

Fig. 6 Visualizations of the representations in the output layer learned by 3-view DCCA when α=0 and α=1. The points in the same color belong to the same category

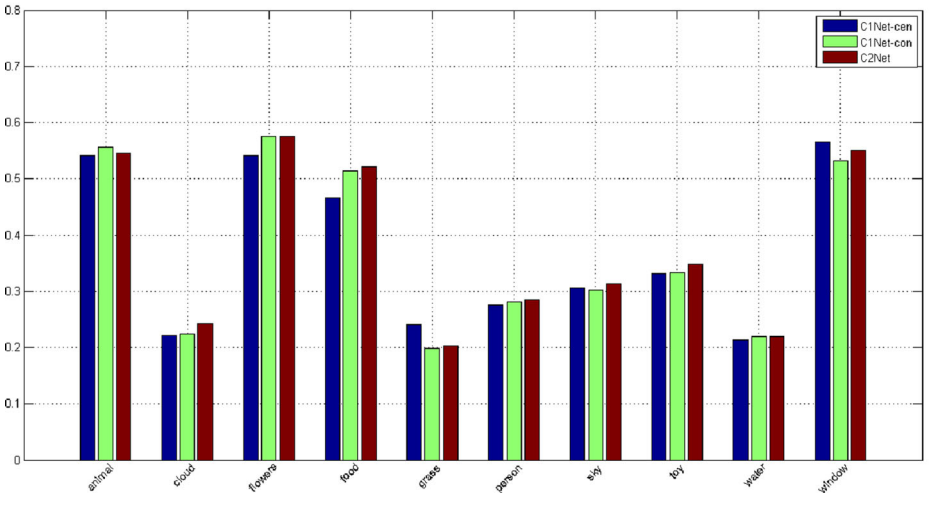
precision at all levels of recall. Therefore, better networks would be explored in the next work.

3.5 Effectiveness of reconstruct loss in the first stage

We use t-SNE to visualize the image and text representations, which are learned by 3-view DCCA when α=0 and α=1 in Fig. 6. When α is set to 0, the effect of the reconstruct loss



(a) Wikipedia.



(b) NUS-WIDE-10k.

Fig. 7 Average per-class mAP@50 of image-to-text retrieval and text-to-image retrieval for each category

is ignored. As discussed above, the similarity metric we use for retrieval is normalized correlation(NC) [23] which is not suitable for t-SNE. So we apply L2 normalization on the features in the latent space. In the normalized space, NC distance is equal to Euclidean distance. As shown in Fig. 6a and (b), 3-view DCCA gets a good latent space when α is set to 0, since a lot of relevant pairs are clustered. However, the results on the test set (c) and (d) are opposite. The result on NUS-WIDE-10k is the same as that on Wikipedia. Like most deep learning methods, 3-view DCCA easily gets into over-fitting without the reconstruct loss. The reconstruct loss back-propagated from the last layer can be viewed as a regularization term for the correlation loss in the output layer and prevents overfitting to some extent.

3.6 Effectiveness of joint loss in the second stage

In this subsection we demonstrate the effectiveness of the combination of contrastive loss and center loss. We check the performance of our method in 3 cases: (1) C1Net-con, the center loss layer is removed. (2) C1Net-cen, the contrastive loss layer and relevant normalization layers are removed. (3) C2Net, i.e., the proposed method. Figure 7 shows the mAP@50 scores achieved per category by above cases. C2Net is better than C1Net-con and C1Net-cen for most categories on both data sets. The contributions of contrastive loss and center loss are complementary and the best performance is achieved when the two are combined.

4 Conclusions and future work

In this paper, we propose a two-stage deep learning method for supervised cross-modal retrieval. We expand CCA from 2-view to 3-view to conduct supervised learning in both stages, which could maximize the value of semantics. Multiple supervisory signals are successfully combined in both stages to overcome over-fitting and learn better feature representations as well. Experiments on two public data sets show that our method is competitive with state-of-the-art performance. In the future, we will build an end-to-end network instead of cascaded two-stage network so as to easily adjust network structures and finetune parameters. Meanwhile, better network architecture will be explored to improve cross-modal retrieval performance.

Acknowledgments This work is supported by Chinese National Natural Science Foundation(61532018, 61471049), and Key Laboratory of Forensic Marks, Ministry of Public Security of China.

References

1. Andrew G, Arora R, Bilmes J, Livescu K (2013) Deep canonical correlation analysis. In: Proceedings of the 30th international conference on machine learning, pp 1247–1255
2. Cai J, Tang Y, Wang J (2016) Kernel canonical correlation analysis via gradient descent. Neurocomputing 182:322–331
3. Chua TS, Tang J, Hong R, Li H, Luo Z, Zheng Y (2009) Nus-wide: a real-world web image database from national university of singapore. In: Proceedings of the ACM international conference on image and video retrieval. ACM, p 48

4. Costa PJ, Coviello E, Doyle G, Rasiwasia N, Lanckriet GR, Levy R, Vasconcelos N (2013) On the role of correlation and abstraction in cross-modal multimedia retrieval. IEEE Trans Pattern Anal Mach Intell 36(3):521–35
5. Costa Pereira J, Coviello E, Doyle G, Rasiwasia N, Lanckriet GR, Levy R, Vasconcelos N (2014) On the role of correlation and abstraction in cross-modal multimedia retrieval. IEEE Trans Pattern Anal Mach Intell 36(3):521–535
6. Feng F, Wang X, Li R (2014) Cross-modal retrieval with correspondence autoencoder. In: Proceedings of the ACM international conference on multimedia. ACM, pp 7–16
7. Feng F, Li R, Wang X (2015) Deep correspondence restricted boltzmann machine for cross-modal retrieval. Neurocomputing 154:50–60
8. Frome A, Corrado GS, Shlens J, Bengio S, Dean J, Mikolov T et al (2013) Devise: a deep visual-semantic embedding model. In: Advances in neural information processing systems, pp 2121–2129
9. Gao Z, Zhang H, Xu G, Xue Y, Hauptmann AG (2015) Multi-view discriminative and structured dictionary learning with group sparsity for human action recognition. Signal Process 112:83–97
10. Gao Z, Wang D, He X, Zhang H (2018) Group-pair convolutional neural networks for multi-view based 3d object retrieval
11. Gong Y, Ke Q, Isard M, Lazebnik S (2014) A multi-view embedding space for modeling internet images, tags, and their semantics. Int J Comput Vis 106(2):210–233
12. Grangier D, Bengio S (2008) A discriminative kernel-based approach to rank images from text queries. IEEE Trans Pattern Anal Mach Intell 30(8):1371–1384
13. Hadsell R, Chopra S, Lecun Y (2006) Dimensionality reduction by learning an invariant mapping. In: 2006 IEEE computer society conference on computer vision and pattern recognition, pp 1735–1742
14. Hardoon DR, Szedmak S, Shawetaylor J (2004) Canonical correlation analysis: an overview with application to learning methods. Neural Comput 16(12):2639
15. Hinton GE, Salakhutdinov R (2009) Replicated softmax: an undirected topic model. In: Advances in neural information processing systems, pp 1607–1614
16. Huang X, Peng Y (2017) Cross-modal deep metric learning with multi-task regularization. arXiv:1703.07026
17. Jia Y, Shelhamer E, Donahue J, Karayev S, Long J, Girshick R, Guadarrama S, Darrell T (2014) Caffe: Convolutional architecture for fast feature embedding. arXiv:1408.5093
18. Kang C, Liao S, He Y, Wang J, Xiang S, Pan C (2014) Cross-modal similarity learning: a low rank bilinear formulation. arXiv:1411.4738
19. Li D, Dimitrova N, Li M, Sethi IK (2003) Multimedia content processing through cross-modal association. In: Proceedings of the eleventh ACM international conference on multimedia. ACM, pp 604–611
20. Ngiam J, Khosla A, Kim M, Nam J, Lee H, Ng AY (2011) Multimodal deep learning. In: Proceedings of the 28th international conference on machine learning (ICML-11), pp 689–696
21. Nie W, Liu A, Su Y (2016) Cross-domain semantic transfer from large-scale social media. Multimed Syst 22(1):75–85
22. Peng Y, Huang X, Qi J (2016) Cross-media shared representation by hierarchical learning with multiple deep networks. In: International joint conference on artificial intelligence (IJCAI), pp 3846–3853
23. Rasiwasia N, Costa Pereira J, Coviello E, Doyle G, Lanckriet GR, Levy R, Vasconcelos N (2010) A new approach to cross-modal multimedia retrieval. In: Proceedings of the international conference on multimedia. ACM, pp 251–260
24. Rosipal R, Krämer N (2006) Overview and recent advances in partial least squares. In: Subspace, latent structure and feature selection. Springer, pp 34–51
25. Shao J, Zhao Z, Su F, Yue T (2015) 3view deep canonical correlation analysis for cross-modal retrieval. In: Visual communications and image processing (VCIP), 2015. IEEE, pp 1–4
26. Shao J, Wang L, Zhao Z, Cai A et al (2016) Deep canonical correlation analysis with progressive and hypergraph learning for cross-modal retrieval. Neurocomputing 214:618–628
27. Sharma A, Kumar A, Daume H III, Jacobs DW (2012) Generalized multiview analysis: a discriminative latent space. In: 2012 IEEE conference on computer vision and pattern recognition (CVPR). IEEE, pp 2160–2167
28. Smolensky P (1986) Parallel distributed processing: explorations in the microstructure of cognition, vol. 1. chapter information processing in dynamical systems: foundations of harmony theory. MIT Press, Cambridge. **15**, 18
29. Socher R, Ganjoo M, Manning CD, Ng A (2013) Zero-shot learning through cross-modal transfer. In: Advances in neural information processing systems, pp 935–943
30. Song J, Yang Y, Yang Y, Huang Z, Shen HT (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. ACM, pp 785–796

31. Song J, Gao L, Nie F, Shen HT, Yan Y, Sebe N (2016) Optimized graph learning using partial tags and multiple features for image and video annotation. IEEE Trans Image Process 25(11):4999–5011
32. Song J, Guo Y, Gao L, Li X, Hanjalic A, Shen HT (2018) From deterministic to generative: multimodal stochastic rnns for video captioning. IEEE Trans Neural Netw Learn Syst PP(99):1–12. https://doi.org/10.1109/TNNLS.2018.2851077
33. Song J, Zhang H, Li X, Gao L, Wang M, Hong R (2018) Self-supervised video hashing with hierarchical binary auto-encoder. IEEE Trans Image Process 27(7):3210–3221
34. Srivastava N, Salakhutdinov R (2012) Learning representations for multimodal data with deep belief nets. In: International conference on machine learning workshop
35. Sun S (2013) A survey of multi-view machine learning. Neural Comput & Appl 23(7-8):2031–2038
36. Sun S, Hardoon DR (2010) Active learning with extremely sparse labeled examples. Neurocomputing 73(16):2980–2988
37. Sun Y, Chen Y, Wang X, Tang X (2014) Deep learning face representation by joint identification-verification. In: Advances in neural information processing systems, pp 1988–1996
38. Tenenbaum JB, Freeman WT (2000) Separating style and content with bilinear models. Neural computation 12(6):1247–1283
39. Wang X, Gao L, Song J, Shen H (2017) Beyond frame-level cnn: saliency-aware 3-d cnn with lstm for video action recognition. IEEE Signal Process Lett 24(4):510–514
40. Wang X, Gao L, Wang P, Sun X, Liu X (2018) Two-stream 3-d convnet fusion for action recognition in videos with arbitrary size and length. IEEE Trans Multimed 20(3):634–644
41. Wei Y, Zhao Y, Zhu Z, Wei S, Xiao Y, Feng J, Yan S (2016) Modality-dependent cross-media retrieval. ACM Trans Intell Syst Technol (TIST) 7(4):57
42. Welling M, Rosen-Zvi M, Hinton GE (2004) Exponential family harmoniums with an application to information retrieval. In: Advances in neural information processing systems, pp 1481–1488
43. Wen Y, Zhang K, Li Z, Qiao Y (2016) A discriminative feature learning approach for deep face recognition. In: European conference on computer vision. Springer, pp 499–515
44. Weston J, Bengio S, Usunier N (2010) Large scale image annotation: learning to rank with joint word-image embeddings. Mach Learn 81(1):21–35
45. Wu F, Lu X, Zhang Z, Yan S, Rui Y, Zhuang Y (2013) Cross-media semantic representation via bi-directional learning to rank. In: Proceedings of the 21st ACM international conference on multimedia. ACM, pp 877–886
46. Zhai X, Peng Y, Xiao J (2014) Learning cross-media joint representation with sparse and semisupervised regularization. IEEE Trans Circuits Syst Video Technol 24(6):965–978
47. Zhu X, Li X, Zhang S, Xu Z, Yu L, Wang C (2017) Graph pca hashing for similarity search. IEEE Trans Multimed 19(9):2033–2044
48. Zu C, Zhang D (2016) Canonical sparse cross-view correlation analysis. Neurocomputing 191:263–272

Jie Shao is a Ph.D. candidate in School of Information and Communication Engineering, Beijing University of Posts Telecommunications. His current research interests include cross-modal retrieval and deep learning.

Dr. Zhicheng Zhao now is a lecture of BUPT. His research interests are computer vision, image and video semantic understanding and retrieval.

Fei Su is a female professor in the multimedia communication and pattern recognition lab, school of information and telecommunication, Beijing university of posts and telecommunications. She received the Ph.D. degree majoring in Communication and Electrical Systems from BUPT in 2000. She was a visiting scholar at electrical computer engineering department, Carnegie Mellon University from 2008 to 2009, Her current interests include pattern recognition, image and video processing and biometrics. She has authored and co-authored more than 70 journal and conference papers and some textbooks.