Cross-Modal Retrieval With CNN Visual Features: A New Baseline

Yunchao Wei, Yao Zhao, *Senior Member, IEEE*, Canyi Lu, Shikui Wei, Luoqi Liu, Zhenfeng Zhu, and Shuicheng Yan, *Senior Member, IEEE*

***Abstract*—Recently, convolutional neural network (CNN) visual features have demonstrated their powerful ability as a universal representation for various recognition tasks. In this paper, cross-modal retrieval with CNN visual features is implemented with several classic methods. Specifically, off-the-shelf CNN visual features are extracted from the CNN model, which is pretrained on ImageNet with more than one million images from 1000 object categories, as a generic image representation to tackle cross-modal retrieval. To further enhance the representational ability of CNN visual features, based on the pretrained CNN model on ImageNet, a fine-tuning step is performed by using the open source Caffe CNN library for each target data set. Besides, we propose a deep semantic matching method to address the cross-modal retrieval problem with respect to samples which are annotated with one or multiple labels. Extensive experiments on five popular publicly available data sets well demonstrate the superiority of CNN visual features for cross-modal retrieval.**

***Index Terms*—Convolutional neural network (CNN) visual features, cross-media, cross-modal, deep learning, multimodal.**

I. Introduction

With rapid development of information technology, there has been an enormous amount of data with various modalities (e.g., image, text, audio, video, etc.) generated on the Internet. These data usually co-occur to describe the same objects or events and thus cross-modal retrieval is becoming imperative for many real-world applications, such as using image to search the relevant text documents or using text to search the relevant images. However, multimodal data usually span different feature spaces. This heterogeneous characteristic has been widely considered as a great challenge to cross-modal retrieval.

During the past few years, a great number of approaches have been proposed to address cross-modal retrieval. Some articles [13], [14], [38], [39], [42], [56] learn an optimal common representation of different modalities for cross-modal retrieval. This kind of approaches project representations of multiple modalities into a common (or an isomorphic) space, such that the distance between two objects with similar semantics could be minimized while the distance between two objects with dissimilar semantics could be maximized. To address the problem of prohibitively expensive nearest neighbor search, some hashing-based approaches [3], [26], [28], [43], [44], [59], [65], [67], [69], [70] to large scale similarity search have drawn much interest from the cross-modal retrieval community. Besides, ranking models [32], [58], [61], [63] and deep models [1], [11], [30], [34], [45], [53] have also been widely considered for multimodal problems in recent years. Despite their contributions to the solution of cross-modal retrieval, the performances of most of these techniques are still far from satisfactory. This may be the case because the performance of cross-modal retrieval is highly dependent on the visual feature representation and the traditional hand-crafted feature extraction techniques such as scale-invariant feature transform (SIFT) [31] and histogram of oriented gradients (HoG) [6], have limited the performance of cross-modal retrieval.

Recently, significant progress has been made in visual recognition, e.g., classification and detection, due to the development of convolutional neural network (CNN) [25], [27]. Especially, a big breakthrough in image classification was made by [25], which has achieved the state-of-the-art performance (with 10% gain over the method based on hand-crafted features) in large-scale object recognition, i.e., ImageNet large scale visual recognition challenge (ILSVRC) [7] with 1000 object categories and 1.2 million images. More recently, Donahue *et al.* [8], Razavian *et al.* [40], and Sermanet *et al.* [41] demonstrated that features extracted from the pretrained CNN can be considered as a generic image representation for diverse visual recognition tasks. To the best of our knowledge, few of the previous articles has applied CNN visual features to cross-modal retrieval. In this paper, we exhaustively compare several classic cross-modal retrieval methods based on CNN visual features and traditional visual features, e.g., SIFT bag-of-visual-words (BoVW).

Manuscript received October 13, 2015; revised December 22, 2015; accepted January 14, 2016. Date of publication March 8, 2016; date of current version January 13, 2017. This work was supported in part by the National Basic Research Program of China under Grant 2012CB316400, in part by the Fundamental Scientific Research under Project K15JB00360, in part by the National Natural Science Foundation of China under Grant 61210006 and Grant 61532005, and in part by the Program for Changjiang Scholars and Innovative Research Team in University under Grant IRT201206. This work was performed when Yunchao Wei was visiting the National University of Singapore. This paper was recommended by Associate Editor X. He. *(Corresponding author: Shikui Wei.)*

Y. Wei, Y. Zhao, S. Wei, and Z. Zhu are with Institute of Information Science, Beijing Jiaotong University, Beijing 100044, China, and also with Beijing Key Laboratory of Advanced Information Science and Network Technology, Beijing 100044, China (e-mail: 11112065@bjtu.edu.cn; yzhao@bjtu.edu.cn; shkwei@bjtu.edu.cn; zhfzhu@bjtu.edu.cn).

C. Lu, L. Liu, and S. Yan are with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore 2423525 (e-mail: canyilu@nus.edu.sg; liuluoqi@nus.edu.sg; eleyans@nus.edu.sg).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCYB.2016.2519449

Furthermore, we propose a simple but effective deep semantic matching (deep-SM) method to address cross-modal retrieval.

The main contributions of this paper are listed as follows.

1) We investigate using off-the-shelf CNN visual features to implement cross-modal retrieval between images and text. Specifically, the off-the-shelf CNN visual features are extracted from the CNN which is pretrained on a large-scale image data set, i.e., ImageNet. As far as we know, this is the first dedicated study to survey the cross-modal retrieval between images and text based on CNN visual features.
2) To better adapt the pretrained CNN model to specific data sets, we utilize the images from the target data set to fine-tune the pretrained model. We compare the off-the-shelf CNN visual features with the fine-tuned CNN visual features on cross-modal retrieval tasks, and experimental results demonstrate that further improvement can be made with CNN visual features fine-tuned by the images from the target data set.
3) We present a simple but effective deep-SM method to address the cross-modal retrieval problem with respect to samples which are annotated with one or multiple labels. In particular, two independent deep networks are learned to map image and text into a common semantic space with higher level abstraction. The correlation between two modalities can be built according to their shared ground truth label(s).
4) Extensive experiments on five public available data sets, including Wikipedia [38], Pascal sentence [37], INRIA-Websearch [23], Pascal VOC 2007 [9], and NUS-WIDE [4], well demonstrate the superiority of CNN visual features for cross-modal retrieval.

The remainder of this paper is organized as follows. We briefly review the related work on cross-modal retrieval in Section II. Section III details the CNN visual features extraction process and the proposed deep-SM method. Extensive experiments and conclusions are given in Section IV and Section V, respectively.

II. Related Work

A. CCA-Based Models

As one of the most popular cross-modal retrieval models, canonical correlation analysis (CCA) [14] is usually employed to find a pair of linear transformations to maximize the correlations between representations of two modalities. Recently, based on CCA, many extensions [5], [13], [38], [39], [42], [56] are applied to cross-modal retrieval. Rasiwasia *et al.* [38] proposed a semantic correlation matching (SCM) approach, where the multiclass logistic regression is applied to the maximally correlated feature representations obtained by CCA, to produce an isomorphic semantic space for cross-modal retrieval. As a supervised extension of CCA, Sharma *et al.* [42] proposed a generic framework called generalized multiview analysis to map data representations in different modality spaces to a common (non)linear subspace. More recently, Gong *et al.* [13] proposed a three-view CCA model by introducing a semantic view, which can be obtained by supervised information or clustering analysis, to achieve a better separation for multimodal data of different classes in the learned common subspace. Similarly, Rasiwasia *et al.* [39] presented a cluster CCA approach to learn discriminant common representations that maximize the correlation between the two modalities while segregating the different classes in the learned common space.

B. Hashing-Based Models

With the explosive growth of high-dimensional cross-modal data, the problem of nearest neighbor search becomes more expensive than ever before. To address this problem, hashing-based approaches [3], [26], [28], [43], [44], [59], [65], [67], [69], [70] for large scale similarity search have attracted considerable interest in the cross-modal retrieval community. Using hashing for multimodal problems was proposed by Bronstein *et al.* [3], named cross modal similarity sensitive hashing (CMSSH). However, CMSSH only considers the intermodality correlation and ignores the intramodality similarity. To address this problem, Kumar and Udupa [26] proposed cross view hashing to generate hash codes by minimizing the distance of hash codes for the similar data and maximizing the distance for the dissimilar data. Most recently, Wu *et al.* [59] proposed a sparse multimodal hashing method, which can obtain sparse code-sets for the data across different modalities via joint multimodal dictionary learning, to address cross-modal retrieval.

C. Ranking Models

In recent years, leaning to rank techniques [32], [58], [61], [63] have been attracted extensive attention for multimodal problems. In general, these methods are supervised but do not enforce the assumption that the trained multimodal data must be paired as needed for CCA-based models (e.g., one image is in pair-correspondence with one text description). Specifically, Yang *et al.* [61] proposed a semi-supervised algorithm called ranking with local regression and global alignment to learn a robust Laplacian matrix for multimodal data ranking. Lu *et al.* [32] proposed a latent semantic cross-modal ranking algorithm to optimize the listwise ranking loss with a low rank embedding for cross-modal retrieval. To take advantage of bi-directional ranking examples, which means that both text-query-image and image-query-text ranking examples are utilized during the training process, Wu *et al.* [58] presented a bi-directional cross-media semantic representation model to achieve a better performance for cross-modal retrieval.

D. Deep Models

With the development of deep learning, many deep models [1], [11], [30], [34], [45], [53] have been proposed to address multimodal problems. Specifically, Ngiam *et al.* [34] and Srivastava and Salakhutdinov [45] proposed to learn a shared representation between different modalities based on restricted Boltzmann machine [15]. Andrew *et al.* [1] introduced a deep CCA model, which can be viewed as a nonlinear extension of the linear CCA, to learn complex nonlinear transformations of two modalities of the data. Frome *et al.* [11] presented a deep visual-semantic embedding model to identify visual object using both labeled image data and semantic information obtained from unannotated text documents. Wang *et al.* [53] proposed an effective mapping mechanism, which can capture both intramodal and intermodal semantic relationships of multimodal data from heterogeneous sources, based on the stacked auto-encoders deep model. However, most of these articles focus on using traditional visual features (e.g., SIFT BoVW) as the input of the their proposed networks for cross-modal retrieval and little work have been conducted for cross-modal retrieval by employing CNN visual features.

Beyond the above mentioned models, other models [18], [20], [22], [29], [33], [47], [49]–[52], [54], [57], [60], [62], [64], [66], [68] are also proposed to address multimodal problems. Specially, Hwang and Grauman [18] proposed an unsupervised learning method based on kernel CCA to discover the relationship between human tags and the relative importance of objects in the image. Wang *et al.* [51] introduced a novel approach to facilitating image search based on a compact semantic embedding. Jia *et al.* [20] presented a topic model to learn cross-modality similarity for multimodal data. In [33], a parallel field alignment method, which integrates a manifold alignment framework from the perspective of vector fields, was proposed to address cross-modal retrieval problem. In [66], both the intramodal and the intermodal correlation are explored for cross-modal retrieval. Although these models have made great contributions to the solution of cross-modal retrieval, the performances of most of them are still far from satisfactory. The reason may be that the visual features extracted by traditional feature extraction techniques cannot effectively express the image semantics. Most of the existing cross-modal retrieval methods employ traditional global feature extraction techniques (e.g., color, GIST [35]) or local features (e.g., SIFT [31] and HoG [6]) extraction-coding-pooling pipeline to generate feature representation for images. However, these traditional feature extraction techniques have limited the performance of image recognition during the past few years.

Recently, significant progress has been made for visual recognition tasks due to the development of CNN. Specifically, Razavian *et al.* [40] have demonstrated that features extracted from the pretraind CNN can be utilized as a generic image representation to tackle diverse visual recognition tasks. However, as far as we know, there has been no work which surveys the effect of CNN visual features for cross-media retrieval. In this paper, extensive experiments are conducted on five publicly available data sets to compare the effectiveness of CNN visual features and traditional visual features for cross-modal retrieval. Inconceivably, good performance can be achieved by CNN visual features based on several classic cross-modal retrieval methods, such as CCA and three-view CCA. The results strongly suggest that visual features obtained from the pretrained or fine-tuned CNN model should be the primary candidates for cross-modal retrieval.

III. CNN Visual Features Extraction and Deep Semantic Matching

During the past few years, deep CNN has demonstrated a strong capability for image classification on some publicly available data sets, such as CIFAR-10/100 [24], Pascal VOC [9], and ImageNet [7]. Some recent articles [8], [12], [36], [40], [41], [55] demonstrated that the CNN models pretrained on large data sets with data diversity, e.g., ImageNet, can be directly transferred to extract CNN visual features for various visual recognition tasks such as image classification and object detection. Inspired by these articles, we propose to utilize CNN visual features to implement cross-modal retrieval.

The pretrained CNN model has a similar network structure to that of Krizhevsky *et al.* [25]. As shown in the upper part of Fig. 1, which contains five convolutional layers (short as cov) and three fully-connected layers (short as fc). The CNN model is pretrained by 1.2 million images of 1000 categories from ImageNet. Two kinds of CNN visual features (i.e., fc6 and fc7 as described in Table I) are utilized for cross-modal retrieval. To adapt the parameters pretrained on ImageNet to the target data set, we utilize the images from the target data set to fine-tune the CNN. Then, we extract the fine-tuned CNN visual features of the first two fully-connected layers (i.e., FT-fc6 and FT-fc7 as described in Table I) for cross-modal retrieval. Besides, motivated by Rasiwasia *et al.* [38], we propose a deep-SM approach to address the cross-modal retrieval problem between images and text with respect to the samples with one or multiple labels. Specifically, we employ the fine-tuned CNN and the trained fully-connected neural network to project image and text into an isomorphic semantic space with high level abstraction. The correlation between two modalities is built according to their shared ground truth label(s).

A. Extracting Visual Features From Pretrained CNN Model

Inspired by Donahue *et al*. [8], Razavian *et al*. [40], and Sermanet *et al.* [41], which demonstrated the outstanding performance of the off-the-shelf CNN visual features in various recognition tasks, we utilize the pretrained CNN model[1] to extract CNN visual features for cross-modal retrieval. In particular, each image is first resized to 256×256 and fed into the CNN model. We only utilize the center patch of the image to produce the CNN visual features. As shown in Fig. 1, we exploit two kinds of off-the-shelf CNN visual features in this paper. fc6 and fc7 denote the 4096 dimensional features of the first two fully-connected layers after the rectified linear units (ReLU) [25].

B. Extracting Visual Features From Fine-Tuned CNN Model

Since the categories (and the number of categories) between ImageNet and the target data set are usually different, directly using the pretrained CNN model to extract visual features

[1]The off-the-shelf CNN visual features used in this paper are extracted from DeCAF [8]. Our experiments based on off-the-shelf CNN visual features were conducted before the release of Caffe [21], which could also be utilized to extract CNN visual features. We did some comparative experiments by using the off-the-shelf CNN features from the Caffe model on the Pascal sentence data sets. The results were very similar with those based on DeCAF.

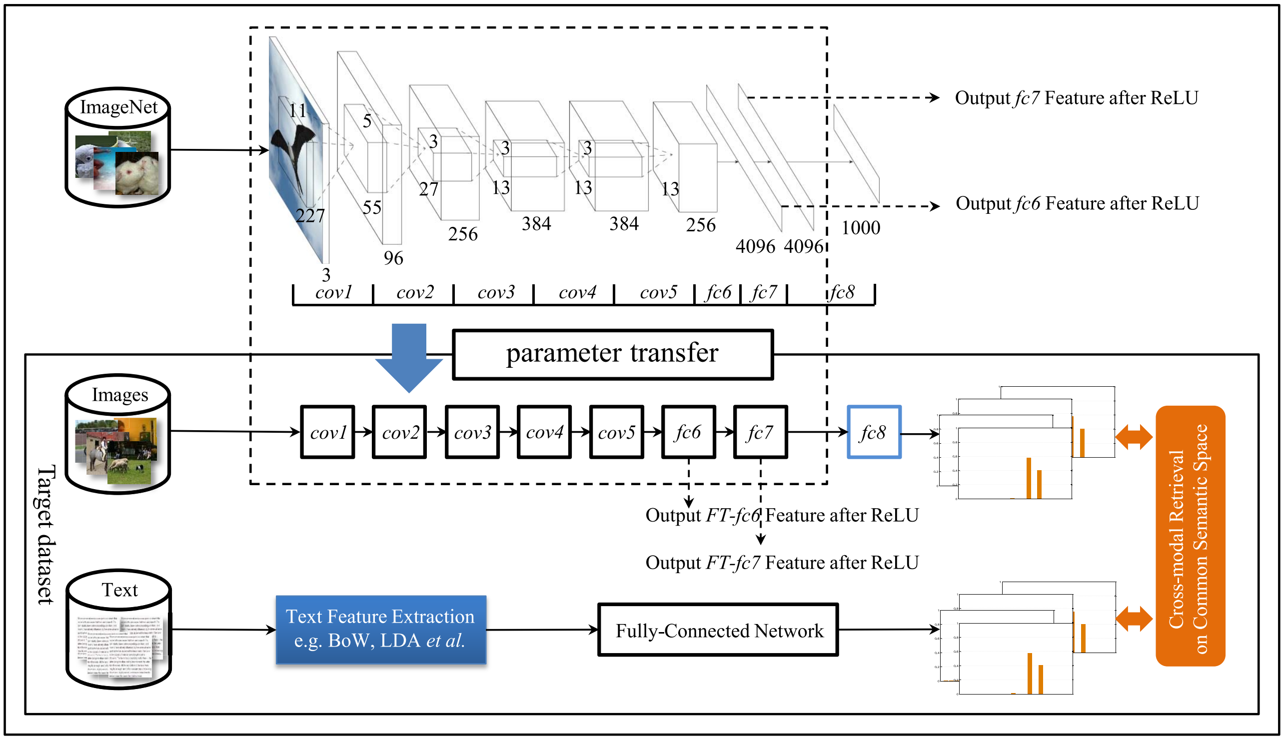


Fig. 1. Illustration of the CNN visual features and the proposed deep-SM approach. The off-the-shelf CNN visual features, i.e., fc6 and fc7, can be directly extracted from the pretrained CNN model. The fine-tuned CNN visual features, i.e., FT-fc6 and FT-fc7, are extracted from the CNN model, which is first pretrained on ImageNet and then fine-tuned on the target data set. For deep-SM, as shown in the lower part, the c dimensional outputs (c is the number of classes of the target data set) of the Softmax layer from the image fine-tuned net and the text fully-connected net are employed for cross-modal retrieval.

TABLE I
DESCRIPTION OF CNN VISUAL FEATURES USED IN THIS PAPER

CNN Feature	Description
fc6	The 4,096 dimensional feature of the first fully-connected layer after ReLU from the pre-trained CNN model.
fc7	The 4,096 dimensional feature of the second fully-connected layer after ReLU from the pre-trained CNN model.
FT-fc6	The 4,096 dimensional feature of the first fully-connected layer after ReLU from the fine-tuning network.
FT-fc7	The 4,096 dimensional feature of the second fully-connected layer after ReLU from the fine-tuning network.

may not be the best strategy. To better adapt the pretrained model on ImageNet to the target data set, we employ the images from the target data set (e.g., Wikipedia, Pascal sentence, INRIA-Websearch, Pascal VOC 2007, and NUS-WIDE) to fine-tune the pretrained parameters.

Each image from the target data set is resized into 256×256 without cropping. We randomly extract 227×227 patches (and their horizontal reflections) from the given image and fine-tune the pretrained CNN model based on these extracted patches. The number of neural units of the last fully-connected layer is modified from 1000 to c, where c is the number of classes of the target data set. The output of the last fully-connected layer is then fed into a c-way soft max which produces a probability distribution over c classes.

In this paper, we adopt different loss functions for different target data sets. We note that squared loss can achieve a similar or even better classification accuracy when the number of classes of the target data set is small. However, with the growth of the number of classes, cross entropy loss function [25] can reach a better classification result. Therefore, we employ the squared loss to fine-tune the pretrained parameters for Wikipedia, Pascal sentence, Pascal VOC 2007, and NUS-WIDE. The number of classes of these four data sets are no more than 21. For INRIA-Websearch,which includes 100 classes, we utilize cross entropy as the loss function during fine-tuning. In this paper, we mainly make a detailed introduction of the squared loss function.

Suppose there are N images in the target data set, and $\boldsymbol{y}_i = [y_{i1}, y_{i2}, \ldots, y_{ic}]$ is the label vector of the ith image. $y_{ij} = 1$ $(j = 1, \ldots, c)$ if the image is annotated with class j, and otherwise $y_{ij} = 0$. The ground-truth probability vector of the ith image is defined as $\hat{\boldsymbol{p}}_i = \boldsymbol{y}_i / ||\boldsymbol{y}_i||_1$ ($||\cdot||_1$ denotes the ℓ_1 norm) and the predictive probability vector is $\boldsymbol{p}_i = [p_{i1}, p_{i2}, \ldots, p_{ic}]$. Then the cost function to be minimized is defined as

$$J = \frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{c} (p_{ik} - \hat{p}_{ik})^2. \quad (1)$$

As shown in Fig. 1, the parameters of the first seven layers are initialized by the parameters pretrained on ImageNet and the parameters of the last fully-connected layer are randomly initialized with a Gaussian distribution $G(\mu, \sigma)$ ($\mu = 0$, $\sigma = 0.01$). During the fine-tuning process, we adopt a discriminating learning rate scheme for different layers. Inspired by Girshick *et al.* [12] and Wei *et al.* [55], we experimentally set the learning rates of the convolutional layers, the first two fully-connected layers and the last fully-connected layers as 0.001, 0.002, and 0.01 at the beginning, respectively. By setting the different learning rates for different layers, the updating rates for parameters of different layers are also different. The first five convolutional layers mainly extract some low-level invariant representations, thus the parameters are quite consistent from the pretrained data set to the target data set, which can be achieved by a low learning rate (i.e., 0.001 at the beginning). However, for the fully-connected layers, especially the last fully-connected layer, which are specifically adapted to the target data set, a much higher learning rate is required to guarantee its fast convergence to the new optimum. By fine-tuning like this, the parameters can better adapt to the target data set without clobbering the transferred parameters.

Fine-tuning is performed using the open source Caffe CNN library [21]. The pretrained model provided by [21] is used to initialize the first seven layers of the fine-tuned CNN. We fine-tune the CNN by stochastic gradient descent (SGD) with a momentum of 0.9 and weight decay of 0.0005. Besides, each layer is followed by a drop-out [16] operation with a drop-out ratio of 0.5 to combat over fitting. Specifically, momentum indicates the weight of the previous update, weight decay is the weight of a regularizer to reduce the training error and drop-out is to set the output of each hidden neuron to zero with a setting probability. For more details of these parameters please refer to [25]. We carry out 60 epoches for fine-tuning and the learning rate of each layer is reduced to one tenth of the current rate after every 20 epoches. After fine-tuning, we utilize the fine-tuned model to extract the output of the first two fully-connected layers after ReLU[2] as the CNN visual features for cross-modal retrieval. The feature extraction process is the same with the process described in Section III-A.

C. Deep Semantic Matching

Rasiwasia *et al.* [38] proposed an SM approach to address the cross-modal retrieval problem. In particular, SM is to represent data of different modalities at a higher level of abstraction, so that there are natural correspondences between the text and image spaces. Inspired by SM, we propose a deep-SM method to address the case where the image (or text) is labeled with one or multiple class labels.

There are some differences between SM and deep-SM. SM tries to learn a shallow (or surface) linear classifier with a probabilistic interpretation to produce a probability distribution over classes as the semantic features. Different from SM, deep-SM learns a deep neural network composed of multiple no-linear transformations to produce a probability distribution over classes as the semantic features. For deep-SM, the outputs of the neural network are the intrinsic probability distribution over the class labels for image or text. We simply use these probabilistic scores as the learned features on the common semantic space for cross-modal retrieval.

For the image, during the fine-tuning process, the neural unit number of the last fully-connected layer (i.e., fc8 with blue bounding box as indicated in Fig. 1) is modified as c, where c is the number of classes of the target data set. We directly employ the c dimensional output of the Softmax layer as the semantic representation for the image. Actually, soft max produces a probability distribution over c classes, which is essentially the same as SM.

For the text, since the representation of text is usually much more discriminative than the image, the relationship between text features and their ground-truth labels can be more easily built. Therefore, we directly build a TextNet with three fully-connected layers to map text features from the original feature space to the semantic space. Specifically, many text feature extraction techniques, such as tf-idf and latent Dirichlet allocation (LDA) [2], can be employed to extract the input text features for TextNet. Similar as the fully-connected layers in CNN, we utilize ReLU as the nonlinear activation function for each fully-connected layer in TextNet and the output of the last fully-connected layer is fed into a c-way soft max, which generates predictive scores (i.e., semantic feature) over c classes. The TextNet is trained by SGD, and the learning rate for each layer is set as 0.01 at the beginning and dynamically changed according to the squared loss (or cross entropy loss) as mentioned in Section III-B.

IV. EXPERIMENTS

A. Data Set and Metric

1) Wikipedia [38]: This data set contains 2866 image-text pairs from ten categories in total. Each image accompanies a text document. The whole data set is randomly split into a training set and a testing set with 2173 and 693 pairs, respectively. We employ the hand-crafted visual feature, i.e., 128 dimensional SIFT BoVW feature provided by [38], to compare with CNN visual features. For text representation, we first obtain the feature vector based on 500 tokens (with stop words removed) and then the LDA model is used to compute the probability of each document under 100 topics. The probability vector is used for text representation.

2) Pascal Sentence [37]: This data set, which is a subset of Pascal VOC, contains 1000 pairs of image and text description (several sentences) from 20 categories (50 for each category). We randomly select 30 pairs from each category for training and the rest for testing. We extract 1024 dimensional SIFT BoVW feature for the image to compare with CNN visual features. For text features, we first extract the feature vector based on the 300 most frequent tokens (with stop words removed) and then utilize the LDA to compute the probability of each document under 100 topics. The 100 dimensional probability vector is used for text representation.

3) INRIA-Websearch [23]: This data set contains 71 478 pairs of image and text description (tags or sentences) from 353 categories. We remove those pairs which are marked as irrelevant, and select those pairs that belong to any one of the 100 largest categories. Then, we get a subset of 14 698

[2]ReLU [25] is a nonlinear transformation $f(x) = \max(0, x)$.

pairs for evaluation. We randomly select 70% of the pairs from each category as the training set (10 332 pairs), and the rest are treated as the testing set (4366 pairs). We employ locality-constrained linear coding (LLC) [48] to extract 2560 dimensional features (with a codebook of size 512 and a two level spatial pyramid) for image representation. For text representation, we first obtain the feature vector based on 25 000 most frequent tokens (with stop words removed) and then use the LDA to compute the probability of each document under 1000 topics. The 1000 dimensional probability vector is used for text representation.

4) Pascal VOC 2007 [9]: There are 9963 images of 20 categories in this data set. Each image accompanies 399 tags annotated by [17]. This data set is divided into train, val, and test subsets. We conduct experiments on trainval and test splits, which contain 5011 and 4952 pairs, respectively. We employ the 776 dimensional visual feature (GHB for short), which contains a 512-D GIST feature, a 64 dimensional Hue-saturation-value (HSV) feature and a 200 dimensional SIFT BoVW feature, provided by [17] to compare with CNN visual features. For text representation, the 798 dimensional tag ranking feature (relative and absolute) provided by [17] is employed as the text feature.

5) NUS-WIDE [4]: This data set contains 269 648 images. Each image is accompanied with 81 ground truth labels and 1000 text tags. We drop those pairs containing images without any ground truth label or text annotation, and only select those pairs belonging to any one of the 21 largest categories. Then, based on the division provided by [4], a subset of 114 117 pairs for training and 76 303 pairs for testing can be obtained for evaluation. We employ the 500 dimensional SIFT BoVW feature provided by [4] to compare with CNN visual features and use the 1000 dimensional text annotations provided by [4] as the text features.

Specifically, Wikipedia, Pascal sentence, and INRIA-Websearch are single-label (each pair of image and text is annotated with one label) data sets, and Pascal VOC 2007 and NUS-WIDE are two multilabel (each pair of image and text is annotated with one or more labels) data sets. Retrieval performance is evaluated by mean average precision (mAP), which is one of the standard information retrieval metrics. Given a set of queries, the average precision (AP) of each query is defined as

$$\mathrm{AP} = \frac{\sum_{k=1}^{R} P(k)\mathrm{rel}(k)}{\sum_{i=1}^{R} \mathrm{rel}(k)}$$

where R denotes the number of the retrieved results. $\mathrm{rel}(k) = 1$ if the item at rank k is relevant, $\mathrm{rel}(k) = 0$ otherwise. $P(k)$ is the precision of the retrieved results ranked at k. We can get the mAP score by averaging AP for all queries. For each data set, the TextNet is composed of three fully-connected layers which are denoted as T-fc1, T-fc2, and T-fc3. In this paper, deep learning for text data is not the key point and it has not been will studied. We experimentally change the number of neural units of each layer so that the TextNet could well converge on the training set. Details of neural unit settings can be found in Table II. Since Pascal VOC 2007 and NUS-WIDE are two multilabel data sets, it is regarded as a relevant result if the retrieved result shares at least one class label with the query.

TABLE II
NEURAL UNIT NUMBER SETTING OF TEXTNET FOR EACH DATA SET

Dataset	*T-fc1*	*T-fc2*	*T-fc3*
Wikipedia	50	20	10
Pascal Sentence	50	20	20
INRIA-Websearch	100	100	100
Pascal VOC 2007	512	256	20
NUS-WIDE	512	256	21

We compare the CNN visual features and traditional visual features for cross-modal retrieval over three common subspace learning approaches.

1) *Canonical Correlation Analysis [14]:* CCA attempts to find a pair of linear transformations (i.e., matrices) to project features of different feature spaces into a common subspace, so that the correlations between these two variables can be maximized.
2) *Three View CCA (T-V CCA) [13]:* Different from CCA, which attempts to model the relationship between two modalities (views), Gong *et al.* [13][3] introduced a semantic view, which can be obtained by supervised information or clustering analysis, to achieve a better separation for multimodal data of different classes in the learned common subspace.
3) *Semantic Matching (SM) [38]:* SM represents the image as well as text at a higher level of abstraction, so that there are natural correspondences between the text and image spaces. Rasiwasia *et al.* [38][4] adopted a multiclass logistic regression [10] operation to generate common semantic representations of multimodal data for cross-modal retrieval.

B. Cross-Modal Retrieval on Wikipedia

Table III reports our experimental results on Wikipedia data set over CCA, T-V CCA, and SM. We can see that off-the-shelf CNN visual features (e.g., fc6 and fc7) yield a great improvement (6.6%–9.7% based on CCA, 6.8%–11.0% based on T-V CCA, and 14.7%–15.3% based on SM) compared with traditional SIFT BoVW feature. We notice that fc6 makes a better improvement than fc7, which is consistent with the conclusion in [8]. After fine-tuning by images from Wikipedia, the CNN visual features can further improve the performance of cross-modal retrieval (0.8% based on CCA, 0.4% based on T-V CCA, and 5.3% based on SM).

In addition, based on CNN visual features, the overall performance of SM is better than that of CCA and T-V CCA. To explain this observation, we give the uni-modal classification (logistic regression is utilized as the classifier [10]) confusion matrices for image and text as shown in Fig. 2. We observe that the text feature possesses a greater discriminative ability (with a text classification accuracy of 75.6%) than the traditional SIFT BoVW feature (with an image classification accuracy of 26.0%). However, if we replace the

[3]http://www.unc.edu/~yunchao/crossmodal.htm
[4]http://www.svcl.ucsd.edu/projects/crossmodal/

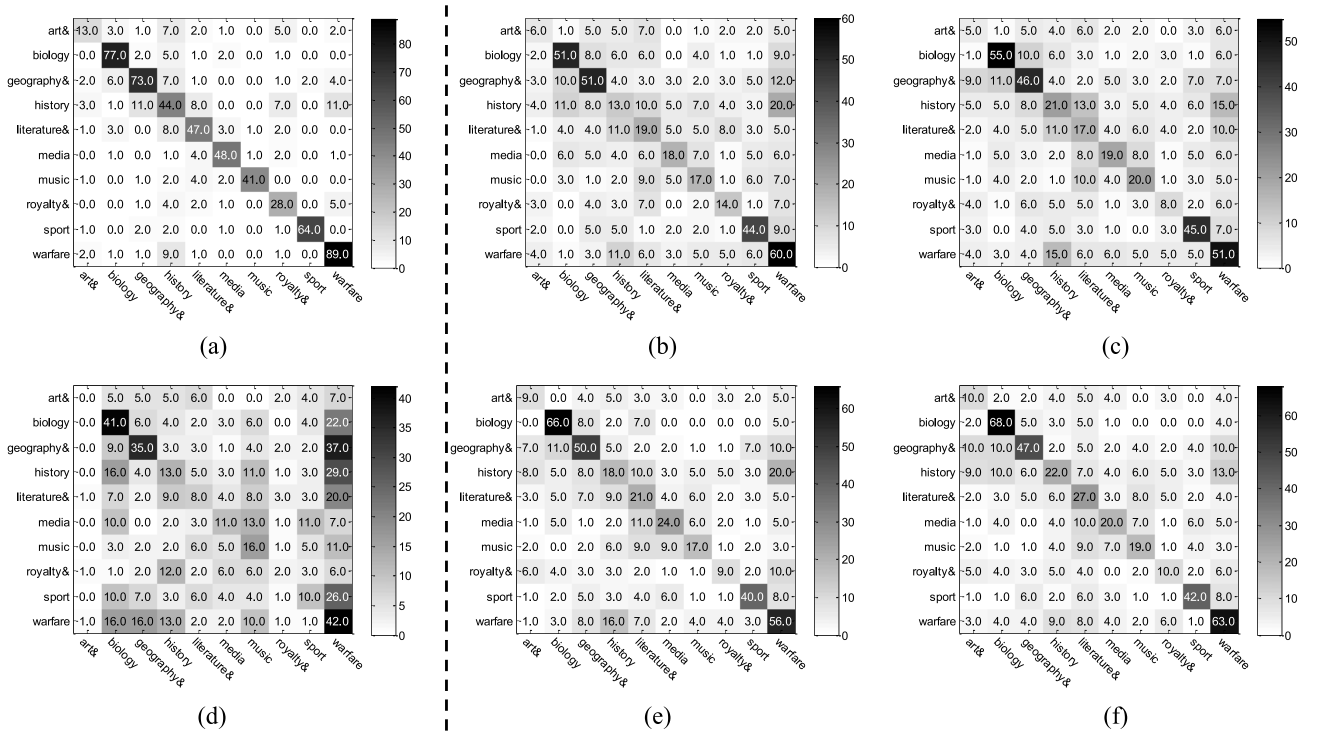


Fig. 2. Confusion matrices of classification results of text and image on Wikipedia data set. Specifically, we compare the classification results of CNN visual features with that of SIFT BoVW feature, which demonstrates that CNN visual features are more discriminative than traditional SIFT BoVW feature. (a) Text. (b) Image-fc6. (c) Image-fc7. (d) Image-BoVW. (e) Image-FT-fc6. (f) Image-FT-fc7.

TABLE III
PERFORMANCE (mAP IN %) COMPARISON IN TERMS OF DIFFERENT METHODS AND VISUAL FEATURES ON WIKIPEDIA DATA SET

Method	visual features	Image Query	Text Query	Average
CCA	BoVW	18.8	17.8	18.3
	fc6	27.2	28.7	28.0
	fc7	25.4	24.4	24.9
	FT-fc6	28.0	29.1	28.6
	FT-fc7	28.8	28.8	28.8
T-V CCA	BoVW	19.6	21.2	20.4
	fc6	31.1	31.6	31.4
	fc7	28.7	25.8	27.2
	FT-fc6	32.0	31.5	31.8
	FT-fc7	31.6	30.8	31.2
SM	BoVW	16.3	22.5	19.4
	fc6	41.6	27.8	34.7
	fc7	40.9	27.2	34.1
	FT-fc6	41.1	36.1	38.6
	FT-fc7	**43.0**	**37.0**	**40.0**
deep-SM	-	39.8	35.4	37.6

SIFT BoVW feature by the CNN visual features, the mean image classification accuracy can reach 43.9% (fc6: 42.3%, fc7: 41.4%, FT-fc6: 44.7%, and FT-fc7: 47.3%). Therefore, based on CNN visual features, better semantic representations of images can be obtained at a higher level of abstraction.

Besides, it is worth noting that SM with the fine-tuned CNN visual features can achieve a better performance than deep-SM. Since the image classification accuracy of the fine-tuned CNN is 48.5%, the main reason may be that the text semantic feature representations generated by logistic regression is better than that from TextNet (the classification accuracy of TextNet is 72.9%).

TABLE IV
STATE-OF-THE-ART CROSS-MODAL RETRIEVAL PERFORMANCES (mAP IN %) WITH TRADITIONAL VISUAL FEATURES ON WIKIPEDIA DATA SET

Wikipedia	Image Query	Text Query	Average
CMCP-2012 [66]	32.6	25.1	28.9
GMMFA-2012 [42]	26.4	23.1	24.8
GMLDA-2012 [42]	27.2	23.2	25.3
PFAR-2013 [33]	29.8	27.3	28.6
SCM-2014 [5]	36.2	27.3	31.8
clusterCCA-2014 [39]	33.4	25.0	29.2
clusterKCCA-2014 [39]	36.5	28.8	32.7

Wikipedia is a very popular data set for cross-modal retrieval evaluation, and many articles have utilized this data set to evaluate their proposed methods. To further validate the effectiveness of CNN visual features for cross-modal retrieval, some performances of the state-of-the-art methods which use the same train/test division as ours are shown in Table IV. We can observe that, based on CNN visual features, the best performance can reach 40.0% by employing FT-fc7 with SM, which significantly outperforms the state-of-the-art methods with a large margin of more than 7%.

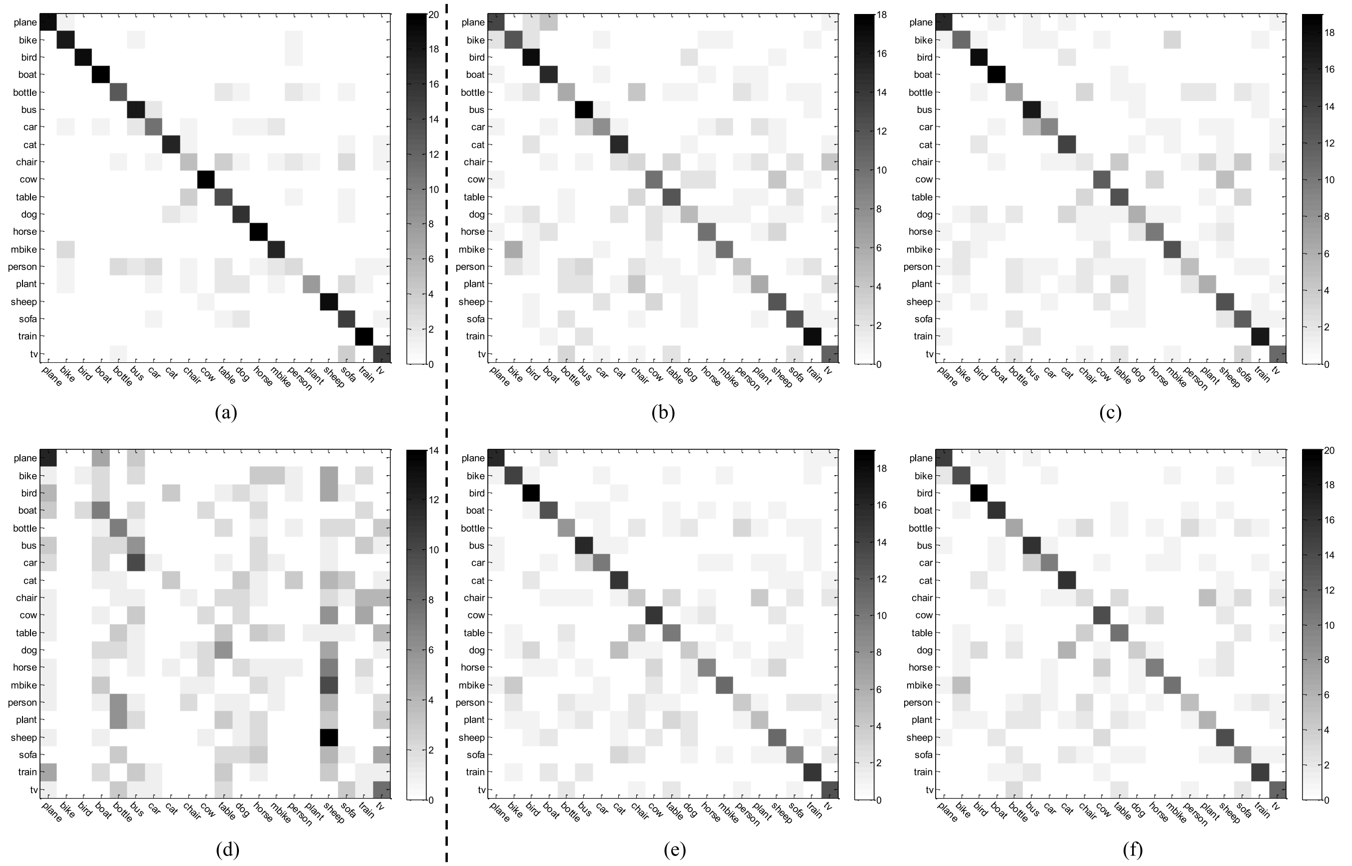


Fig. 3. Confusion matrices of classification results of text and image on Pascal sentence data set. In particular, we compare the classification results of CNN visual features with that of SIFT BoVW feature, which demonstrates that CNN visual features are more discriminative than traditional SIFT BoVW feature. (a) Text. (b) Image-fc6. (c) Image-fc7. (d) Image-BoVW. (e) Image-FT-fc6. (f) Image-FT-fc7.

C. Cross-Modal Retrieval on Pascal Sentence

Table V reports our experimental results on Pascal sentence data set over CCA, T-V CCA, and SM. Similar as on Wikipedia, the off-the-shelf CNN visual features (e.g., fc6 and fc7) also obtain significant improvements (22.4%–24.1% based on CCA, 24.3%–27.4% based on T-V CCA, and 33.3%–35.8% based on SM) compared with the traditional SIFT BoVW feature. Different from on Wikipedia, where fc6 performs better than fc7, fc7 achieves a greater improvement than fc6. This may because the classes in Pascal sentence are all included in the 1000 classes of ImageNet. Therefore, images in Pascal sentence are very similar to those in ImageNet, which results in features from the later fully-connected layer with more discriminative power. After fine-tuning, a further improvement (2.0% based on CCA, 2.2% based on T-V CCA, and 0.6% based on SM) can be made compared with the best performance of the off-the-shelf CNN visual features for cross-media retrieval.

Similar as Wikipedia, based on CNN visual features, the overall performance of SM is better than that of CCA and T-V CCA, and the deep-SM cannot compare with SM on fc7 and FT-fc7. Fig. 3 shows the uni-modal classification confusion matrices for the image and text. In particular, for the image, classification accuracies of SIFT BoVW, fc6, fc7, FT-fc6, FT-fc7, and the fine-tuned CNN are 17%, 54%, 57.8%, 55.3%, 57.0%, and 56.0%, respectively. For the text, the classification accuracies of logistic regression and TextNet are 76.8% and 70.8%. Based on CNN visual features, semantic representations of images can be more consistent with those of text.

TABLE V
PERFORMANCE (mAP IN %) COMPARISON IN TERMS OF DIFFERENT METHODS AND VISUAL FEATURES ON PASCAL SENTENCE DATA SET

Method	visual features	Image Query	Text Query	Average
CCA	BoVW	11.1	11.8	11.5
	fc6	30.7	37.2	33.9
	fc7	34.3	36.9	35.6
	FT-fc6	32.3	35.7	34.0
	FT-fc7	36.4	38.7	37.6
T-V CCA	BoVW	13.1	15.9	14.5
	fc6	33.8	43.8	38.8
	fc7	39.5	44.3	41.9
	FT-fc6	35.7	44.0	39.9
	FT-fc7	41.7	46.5	44.1
SM	BoVW	8.0	14.8	11.4
	fc6	42.6	46.7	44.7
	fc7	48.3	46.0	47.2
	FT-fc6	45.0	45.8	45.4
	FT-fc7	**49.6**	46.0	**47.8**
deep-SM	-	44.6	**47.8**	46.2

TABLE VI
PERFORMANCE (mAP IN %) COMPARISON IN TERMS OF DIFFERENT METHODS AND OTHER HAND-CRAFTED VISUAL FEATURES ON PASCAL SENTENCE DATA SET

Method	visual features	Image Query	Text Query	Average
CCA	LLC	9.4	9.5	9.5
	LLC-P	10.3	10.4	10.4
	VLAD	12.1	13.9	13.0
	VLAD-P	11.3	14.7	13.0
T-V CCA	LLC	10.3	10.4	10.4
	LLC-P	11.2	10.6	10.9
	VLAD	14.0	15.9	15.2
	VLAD-P	13.3	17.5	15.4
SM	LLC	15.4	19.7	17.6
	LLC-P	15.2	20.5	17.9
	VLAD	18.7	20.0	19.4
	VLAD-P	18.4	21.7	20.1

Due to different training/testing division or different problems, we do not compare this paper with other articles such as [46] on this data set.

To further validate the effectiveness of CNN visual features for the cross-modal retrieval task, we experimentally compare them with some more powerful hand-crafted features, i.e., LLC [48] and vector of locally aggregated descriptors (VLAD) [19]. We first extract SIFT interest points for each image and then encode them with LLC or VLAD to generate the feature representation. As shown in Table VI, LLC, LLC-P, VLAD, and VLAD-P are encoded with the codebook size of 512, 1024, 64, and 128, respectively. It can be observed that their performance is not satisfactory even with more powerful hand-crafted features. If we continue to enlarge the codebook size, the performance for both LLC and VLAD will improve but may still be limited. In addition, with 4096 dimensional CNN visual feature, the performance can reach 47.8%, which is much better than that of 16 384 denominational VLAD-P feature, i.e., 20.1%.

D. Cross-Modal Retrieval on INRIA-Websearch

Table VII reports our experimental results on the INRIA-Websearch data set over CCA, T-V CCA, SM, and deep-SM. The off-the-shelf CNN visual features (i.e., fc6 and fc7) obtain significant improvements (20.5%–21.1% based on CCA, 24.6%–24.7% based on T-V CCA, and 16.6%–17.2% based on SM) compared with LLC. After fine-tuning, a further improvement (5.7% based on CCA, 4.7% based on T-V CCA, and 1.9% based on SM) can be made compared with the best performance of the off-the-shelf CNN visual features.

Based on CNN visual features, the overall performance of SM is better than that of CCA and T-V CCA, and the deep-SM cannot compare with SM on fc6, FT-fc6, and FT-fc7. For the image, classification accuracies of LLC, fc6, fc7, FT-fc6, FT-fc7, and the fine-tuned CNN are 52.0%, 68.1%, 67.2%, 70.0%, 69.8%, and 69.0%, respectively. For the text, the classification accuracies of logistic regression and TextNet are 73.0% and 72.0%. Therefore, with the CNN visual features, semantic representations of images can be more consistent with those of text. Since this data set is constructed by ourselves, we do not compare this paper with other articles.

TABLE VII
PERFORMANCE (mAP IN %) COMPARISON IN TERMS OF DIFFERENT METHODS AND VISUAL FEATURES ON INRIA-WEBSEARCH DATA SET

Method	visual features	Image Query	Text Query	Average
CCA	LLC	10.4	15.2	12.8
	fc6	27.4	39.2	33.3
	fc7	29.2	38.6	33.9
	FT-fc6	30.5	42.5	36.5
	FT-fc7	34.5	44.6	39.6
T-V CCA	LLC	11.1	22.8	16.9
	fc6	32.9	50.0	41.5
	fc7	34.1	49.2	41.6
	FT-fc6	35.8	52.1	43.9
	FT-fc7	39.7	52.9	46.3
SM	LLC	22.6	38.7	30.6
	fc6	43.9	51.7	47.8
	fc7	43.0	51.4	47.2
	FT-fc6	46.5	**52.7**	49.6
	FT-fc7	**47.7**	51.6	**49.7**
deep-SM	-	43.9	51.5	47.7

TABLE VIII
PERFORMANCE (mAP IN %) COMPARISON IN TERMS OF DIFFERENT METHODS AND VISUAL FEATURES ON PASCAL VOC 2007 DATA SET

Method	visual features	Image Query	Text Query	Average
CCA	GHB	36.8	34.5	35.7
	fc6	63.5	64.3	63.9
	fc7	66.6	67.2	66.9
	FT-fc6	63.7	64.2	64.0
	FT-fc7	66.1	66.8	66.5
T-V CCA	GHB	47.5	43.6	45.6
	fc6	68.9	71.4	70.2
	fc7	71.7	73.7	72.7
	FT-fc6	68.8	71.2	70.0
	FT-fc7	71.5	73.4	72.5
deep-SM	-	**82.3**	**77.6**	**80.0**

E. Cross-Modal Retrieval on Pascal VOC 2007

Table VIII reports our experimental results on Pascal VOC 2007 data set over CCA, T-V CCA, and deep-SM. Since Pascal VOC 2007 is a multilabel data set, we implement the cross-modal retrieval based on the criterion that it is regarded as a relevant result if the retrieved result shares as least one class label with the query.

From Table VIII, we can see that CNN visual features outperform the traditional visual feature,

TABLE IX
PERFORMANCE (mAP IN %) COMPARISON IN TERMS OF DIFFERENT METHODS AND VISUAL FEATURES ON NUS-WIDE DATA SET

Method	visual features	Image Query	Text Query	Average
CCA	BoVW	38.0	38.6	38.3
	fc6	45.2	47.1	46.1
	fc7	46.2	47.6	46.9
	FT-fc6	46.5	49.4	48.0
	FT-fc7	48.4	50.9	49.7
T-V CCA	BoVW	46.4	47.8	47.1
	fc6	58.7	60.7	59.7
	fc7	59.3	60.8	60.1
	FT-fc6	60.6	62.6	61.6
	FT-fc7	66.3	63.1	64.7
deep-SM	-	**67.9**	**69.3**	**68.6**

i.e., GIST-HSV-BoVW (GHB), with a large margin (27.1%–31.2% based on CCA and 22.4%–27.1% based on T-V CCA). Besides, with the proposed deep-SM, a significant improvement could be achieved (from 72.7% obtained using T-V CCA to 80.0%). We may note that the results on this data set does not show consistent improvements by using CNN visual features after fine-tuning. The reason may be explained as follows. On one hand, the 20 classes are all included in the ImageNet and many images from the training set of ILSVRC 2012 are very similar as those from Pascal VOC 2007. Therefore, CNN features directly extracted from the pretrained CNN model are still with powerful discriminative ability. On the other hand, Pascal VOC 2007 is a multilabel data set, we define that it is regarded as a relevant result if the retrieved result shares at least one class label with the query. Therefore, the results of CNN features (without fine-tuning) may outperform those of fine-tuned CNN features with a certain probability.

In addition, Sharma *et al.* [42] utilized single-label pairs of image and text from VOC 2007 for cross-modal retrieval evaluation and achieved an average mAP score of 38.3% (image query: 42.7% and text query: 33.9%). With the same train/test setting, Rasiwasia *et al.* [39] achieved an average mAP score of 44.0% (image query: 44.5% and text query: 43.6%) on this data set.

F. Cross-Modal Retrieval on NUS-WIDE

Table IX reports our experimental results on NUS-WIDE data set over CCA, T-V CCA, and deep-SM. Similar as on Pascal VOC 2007, CNN visual features also outperform the traditional SIFT BoVW feature, with a large margin (7.8%–11.4% based on CCA and 9.0%–17.6% based on T-V CCA). Besides, with the proposed deep-SM, the performance can be further improved from 64.7% to 68.6%.

As far as we know, NUS-WIDE is one of the largest publicly available multilabel data set for cross-modal retrieval. Many articles have utilized this data set to evaluate their algorithms. However, we cannot directly compare our method with previous articles due to the different ways of using this data set (e.g., different train/test split). Similar with our criterion, which only selects those pairs belonging to one of the 21 most frequent categories, MASE-2014 [53] achieved an average mAP of 44.0% (image query: 44.7% and text query: 43.2%) based on its division. As one of the state-of-the-art hashing based methods, SM^2H-2013 [59] achieved an average mAP of 48.4% (image query: 48.0% and text query: 48.8%) by using those pairs belonging to the ten most frequent categories.

To sum up, based on the above reported experimental results, we can see that CNN visual features are very effective for cross-modal retrieval.

V. CONCLUSION

In this paper, cross-modal retrieval with CNN visual features is implemented and compared with several classic methods based on five publicly available data sets. From experimental results, we can see that cross-modal retrieval with images represented by CNN visual features can easily achieve superior results compared with using traditional visual features, e.g., SIFT BoVW or LLC. The experimental results strongly suggest that the visual feature obtained from the pretrained or fine-tuned CNN model should be the primary candidate for cross-modal retrieval. Based on CNN visual features, some more effective approaches may be designed for cross-modal retrieval. However, deep learning for text data has not been well studied in this paper. We just employ a fully-connected neural network for semantic features extraction. In the future, some more appropriate neural networks such as recurrent neural network will be explored to build the relationship between low-level features and high-level semantics for text data.

REFERENCES

[1] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, "Deep canonical correlation analysis," in *Proc. ICML*, Atlanta, GA, USA, 2013, pp. 1247–1255.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Mar. 2003.

[3] M. M. Bronstein, A. M. Bronstein, F. Michel, and N. Paragios, "Data fusion through cross-modality metric learning using similarity-sensitive hashing," in *Proc. CVPR*, San Francisco, CA, USA, Jun. 2010, pp. 3594–3601.

[4] T.-S. Chua *et al.*, "NUS-WIDE: A real-world Web image database from National University of Singapore," in *Proc. CIVR*, 2009, Art. ID 48.

[5] J. C. Pereira *et al.*, "On the role of correlation and abstraction in cross-modal multimedia retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 3, pp. 521–535, Mar. 2014.

[6] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. CVPR*, vol. 1. San Diego, CA, USA, Jun. 2005, pp. 886–893.

[7] J. Deng *et al.*, "ImageNet: A large-scale hierarchical image database," in *Proc. CVPR*, Miami, FL, USA, Jun. 2009, pp. 248–255.

[8] J. Donahue *et al.*, "DeCAF: A deep convolutional activation feature for generic visual recognition," *arXiv preprint arXiv:1310.1531*, 2013.

[9] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, 2010.

[10] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "LIBLINEAR: A library for large linear classification," *J. Mach. Learn. Res.*, vol. 9, pp. 1871–1874, Aug. 2008.

[11] A. Frome *et al.*, "DeViSE: A deep visual-semantic embedding model," in *Proc. NIPS*, 2013, pp. 2121–2129.

[12] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," *arXiv preprint arXiv:1311.2524*, 2013.

[13] Y. Gong, Q. Ke, M. Isard, and S. Lazebnik, "A multi-view embedding space for modeling Internet images, tags, and their semantics," *Int. J. Comput. Vis.*, vol. 106, no. 2, pp. 210–233, 2014.

[14] D. R. Hardoon, S. R. Szedmak, and J. R. Shawe-Taylor, "Canonical correlation analysis: An overview with application to learning methods," *Neural Comput.*, vol. 16, no. 12, pp. 2639–2664, Dec. 2004.

[15] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[16] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," *arXiv preprint arXiv:1207.0580*, 2012.

[17] S. J. Hwang and K. Grauman, "Accounting for the relative importance of objects in image retrieval," in *Proc. BMVC*, Aberystwyth, U.K., 2010, pp. 1–12.

[18] S. J. Hwang and K. Grauman, "Learning the relative importance of objects from tagged images for retrieval and cross-modal search," *Int. J. Comput. Vis.*, vol. 100, no. 2, pp. 134–153, 2012.

[19] H. Jégou, M. Douze, C. Schmid, and P. Pérez, "Aggregating local descriptors into a compact image representation," in *Proc. CVPR*, San Francisco, CA, USA, 2010, pp. 3304–3311.

[20] Y. Jia, M. Salzmann, and T. Darrell, "Learning cross-modality similarity for multinomial data," in *Proc. ICCV*, Barcelona, Spain, 2011, pp. 2407–2414.

[21] Y. Jia *et al.*, "Caffe: Convolutional architecture for fast feature embedding," in *Proc. ACM MM*, Orlando, FL, USA, 2014, pp. 675–678.

[22] M. Katsurai, T. Ogawa, and M. Haseyama, "A cross-modal approach for extracting semantic relationships between concepts using tagged images," *IEEE Trans. Multimedia*, vol. 16, no. 4, pp. 1059–1074, Jun. 2014.

[23] J. Krapac, M. Allan, J. Verbeek, and F. Jurie, "Improving Web image search results using query-relative classifiers," in *Proc. CVPR*, San Francisco, CA, USA, 2010, pp. 1094–1101.

[24] A. Krizhevsky, "Learning multiple layers of features from tiny images," M.S. thesis, Dept. Comput. Sci., Univ. Toronto, Toronto, ON, Canada, 2009.

[25] A. Krizhevsky, I. Sutskever, and G. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. NIPS*, 2012, pp. 1106–1114.

[26] S. Kumar and R. Udupa, "Learning hash functions for cross-view similarity search," in *Proc. IJCAI*, vol. 22. Barcelona, Spain, 2011, pp. 1360–1365.

[27] Y. LeCun, K. Kavukcuoglu, and C. Farabet, "Convolutional networks and applications in vision," in *Proc. ISCAS*, Paris, France, 2010, pp. 253–256.

[28] X. Liu, Y. Mu, D. Zhang, B. Lang, and X. Li, "Large-scale unsupervised hashing with shared structure learning," *IEEE Trans. Cybern.*, vol. 45, no. 9, pp. 1811–1822, Sep. 2015.

[29] X. Liu, M. Wang, B.-C. Yin, B. Huet, and X. Li, "Event-based media enrichment using an adaptive probabilistic hypergraph model," *IEEE Trans. Cybern.*, vol. 45, no. 11, pp. 2461–2471, Nov. 2015.

[30] Y. Liu, S.-H. Zhong, and W. J. Li, "Query-oriented multi-document summarization via unsupervised deep learning," in *Proc. AAAI*, Toronto, ON, Canada, 2012, pp. 1699–1705.

[31] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[32] X. Lu *et al.*, "A low rank structural large margin method for cross-modal ranking," in *Proc. SIGIR*, Dublin, Ireland, 2013, pp. 433–442.

[33] X. Mao, B. Lin, D. Cai, X. He, and J. Pei, "Parallel field alignment for cross media retrieval," in *Proc. ACM Multimedia*, Barcelona, Spain, 2013, pp. 897–906.

[34] J. Ngiam *et al.*, "Multimodal deep learning," in *Proc. ICML*, Bellevue, WA, USA, 2011, pp. 689–696.

[35] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *Int. J. Comput. Vis.*, vol. 42, no. 3, pp. 145–175, 2001.

[36] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Learning and transferring mid-level image representations using convolutional neural networks," in *Proc. CVPR*, Columbus, OH, USA, 2014, pp. 1717–1724.

[37] C. Rashtchian, P. Young, M. Hodosh, and J. Hockenmaier, "Collecting image annotations using Amazon's mechanical turk," in *Proc. Workshop Creating Speech Lang. Data Amazon's Mech. Turk*, Los Angeles, CA, USA, 2010, pp. 139–147.

[38] N. Rasiwasia *et al.*, "A new approach to cross-modal multimedia retrieval," in *Proc. ACM Multimedia*, Florence, Italy, 2010, pp. 251–260.

[39] N. Rasiwasia, D. Mahajan, V. Mahadevan, and G. Aggarwal, "Cluster canonical correlation analysis," in *Proc. AISTATS*, Reykjavik, Iceland, 2014, pp. 823–831.

[40] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "CNN features off-the-shelf: An astounding baseline for recognition," *arXiv preprint arXiv:1403.6382*, 2014.

[41] P. Sermanet *et al.*, "Overfeat: Integrated recognition, localization and detection using convolutional networks," *arXiv preprint arXiv:1312.6229*, 2013.

[42] A. Sharma, A. Kumar, H. Daume, and D. W. Jacobs, "Generalized multiview analysis: A discriminative latent space," in *Proc. CVPR*, Providence, RI, USA, 2012, pp. 2160–2167.

[43] J. Song, Y. Yang, X. Li, Y. Yang, and Z. Huang, "Robust hashing with local models for approximate similarity search," *IEEE Trans. Cybern.*, vol. 44, no. 7, pp. 1225–1236, Jul. 2014.

[44] J. Song, Y. Yang, Y. Yang, Z. Huang, and H. T. Shen, "Inter-media hashing for large-scale retrieval from heterogeneous data sources," in *Proc. ACM SIGMOD*, New York, NY, USA, 2013, pp. 785–796.

[45] N. Srivastava and R. Salakhutdinov, "Multimodal learning with deep Boltzmann machines," in *Proc. NIPS*, 2012, pp. 2231–2239.

[46] Y. Verma and C. V. Jawahar, "Im2text and text2im: Associating images and texts for cross-modal retrieval," in *Proc. BMVC*, Nottingham, U.K., 2014, pp. 1–13.

[47] A. Wang, J. Lu, J. Cai, T.-J. Cham, and G. Wang, "Large-margin multi-modal deep learning for RGB-D object recognition," *IEEE Trans. Multimedia*, vol. 17, no. 11, pp. 1887–1898, Nov. 2015.

[48] J. Wang *et al.*, "Locality-constrained linear coding for image classification," in *Proc. CVPR*, San Francisco, CA, USA, 2010, pp. 3360–3367.

[49] M. Wang *et al.*, "Unified video annotation via multigraph learning," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 19, no. 5, pp. 733–746, May 2009.

[50] M. Wang, H. Li, D. Tao, K. Lu, and X. Wu, "Multimodal graph-based reranking for Web image search," *IEEE Trans. Image Process.*, vol. 21, no. 11, pp. 4649–4661, Nov. 2012.

[51] M. Wang *et al.*, "Facilitating image search with a scalable and compact semantic mapping," *IEEE Trans. Cybern.*, vol. 45, no. 8, pp. 1561–1574, Aug. 2014.

[52] M. Wang, K. Yang, X.-S. Hua, and H.-J. Zhang, "Towards a relevant and diverse search of social images," *IEEE Trans. Multimedia*, vol. 12, no. 8, pp. 829–842, Dec. 2010.

[53] W. Wang, B. C. Ooi, X. Yang, D. Zhang, and Y. Zhuang, "Effective multi-modal retrieval based on stacked auto-encoders," *Proc. VLDB Endowment*, vol. 7, no. 8, pp. 649–660, 2014.

[54] S. Wei, Y. Wei, L. Zhang, Z. Zhu, and Y. Zhao, "Heterogeneous data alignment for cross-media computing," in *Proc. ICIMCS*, Zhangjiajie, China, 2015, Art. ID 84.

[55] Y. Wei *et al.*, "CNN: Single-label to multi-label," *arXiv preprint arXiv:1406.5726*, 2014.

[56] Y. Wei *et al.*, "Modality-dependent cross-media retrieval," *arXiv preprint arXiv:1506.06628*, 2015.

[57] Y. Wei, Y. Zhao, Z. Zhu, Y. Xiao, and S. Wei, "Learning a mid-level feature space for cross-media regularization," in *Proc. ICME*, Chengdu, China, 2014, pp. 1–6.

[58] F. Wu *et al.*, "Cross-media semantic representation via bi-directional learning to rank," in *Proc. ACM Int. Conf. Multimedia*, Barcelona, Spain, 2013, pp. 877–886.

[59] F. Wu *et al.*, "Sparse multi-modal hashing," *IEEE Trans. Multimedia*, vol. 16, no. 2, pp. 427–439, Feb. 2014.

[60] Y. Yang, Z. Ma, A. G. Hauptmann, and N. Sebe, "Feature selection for multimedia analysis by sharing information among multiple tasks," *IEEE Trans. Multimedia*, vol. 15, no. 3, pp. 661–669, Apr. 2013.

[61] Y. Yang *et al.*, "A multimedia retrieval framework based on semi-supervised ranking and relevance feedback," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 4, pp. 723–742, Apr. 2012.

[62] Y. Yang *et al.*, "Multi-feature fusion via hierarchical regression for multimedia analysis," *IEEE Trans. Multimedia*, vol. 15, no. 3, pp. 572–581, Apr. 2013.

[63] Y. Yang, D. Xu, F. Nie, J. Luo, and Y. Zhuang, "Ranking with local regression and global alignment for cross media retrieval," in *Proc. ACM Int. Conf. Multimedia*, Beijing, China, 2009, pp. 175–184.

[64] Y. Yang, Y.-T. Zhuang, F. Wu, and Y.-H. Pan, "Harmonizing hierarchical manifolds for multimedia document semantics understanding and cross-media retrieval," *IEEE Trans. Multimedia*, vol. 10, no. 3, pp. 437–446, Apr. 2008.

[65] Z. Yu *et al.*, "Discriminative coupled dictionary hashing for fast cross-media retrieval," in *Proc. SIGIR*, Gold Coast, QLD, Australia, 2014, pp. 395–404.

[66] X. Zhai, Y. Peng, and J. Xiao, "Cross-modality correlation propagation for cross-media retrieval," in *Proc. ICASSP*, Kyoto, Japan, 2012, pp. 2337–2340.

[67] D. Zhang and W.-J. Li, "Large-scale supervised multimodal hashing with semantic correlation maximization," in *Proc. AAAI*, Quebec City, QC, Canada, 2014, pp. 2177–2183.

[68] L. Zhang *et al.*, "Feature correlation hypergraph: Exploiting high-order potentials for multimodal recognition," *IEEE Trans. Cybern.*, vol. 44, no. 8, pp. 1408–1419, Aug. 2014.

[69] X. Zhu, Z. Huang, H. T. Shen, and X. Zhao, "Linear cross-modal hashing for efficient multimedia search," in *Proc. ACM 21st Int. Conf. Multimedia*, Barcelona, Spain, 2013, pp. 143–152.

[70] Y. T. Zhuang, Y. F. Wang, F. Wu, Y. Zhang, and W. M. Lu, "Supervised coupled dictionary learning with group structures for multi-modal retrieval," in *Proc. AAAI*, Bellevue, WA, USA, 2013, pp. 1070–1076.

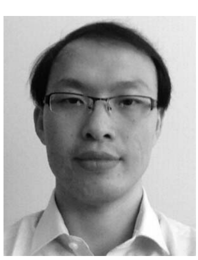
Yunchao Wei is currently pursuing the Ph.D. degree with the Institute of Information Science, Beijing Jiaotong University, Beijing, China.

He is currently a Visiting Student with the National University of Singapore, Singapore. His current research interests include object classification in computer vision and multimodal analysis in multimedia.

Yao Zhao (M'06–SM'12) received the B.S. degree from Fuzhou University, Fuzhou, China, in 1989, and the M.E. degree from Southeast University, Nanjing, China, in 1992, both from the Radio Engineering Department, and the Ph.D. degree from the Institute of Information Science, Beijing Jiaotong University (BJTU), Beijing, China, in 1996.

He became an Associate Professor with BJTU in 1998 and became a Professor in 2001. From 2001 to 2002, he was a Senior Research Fellow with the Information and Communication Theory Group, Faculty of Information Technology and Systems, Delft University of Technology, Delft, The Netherlands. He is currently the Director of the Institute of Information Science, BJTU. His current research interests include image/video coding, digital watermarking and forensics, and video analysis and understanding.

Prof. Zhao is currently leading several national research projects from the 973 Program, 863 Program, and the National Science Foundation of China. He serves on the Editorial Board of several international journals, including as an Associate Editor of the IEEE TRANSACTIONS ON CYBERNETICS, the IEEE SIGNAL PROCESSING LETTERS, an Area Editor of *Signal Processing: Image Communication* (Elsevier), and an Associate Editor of *Circuits, System, and Signal Processing* (Springer). He was named a Distinguished Young Scholar by the National Science Foundation of China in 2010, and was elected as a Chang Jiang Scholar of Ministry of Education of China in 2013. He is a fellow of IET.



Canyi Lu is currently pursuing the Ph.D. degree with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore.

His current research interests include computer vision, machine learning, pattern recognition, and optimization.

Mr. Lu was a recipient of the Microsoft Research Asia Fellowship 2014.

Shikui Wei received the Ph.D. degree in signal and information processing from Beijing Jiaotong University (BJTU), Beijing, China, in 2010.

From 2010 to 2011, he was a Research Fellow with the School of Computer Engineering, Nanyang Technological University, Singapore. He is currently an Associate Professor with the Institute of Information Science, BJTU. His current research interests include computer vision, image/video analysis and retrieval, and copy detection.

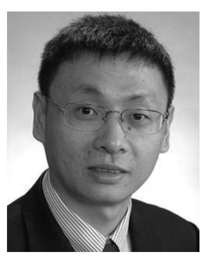
Luoqi Liu is currently pursuing the Ph.D. degree with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore.

His current research interests include computer vision and multimedia.

Mr. Liu was a recipient of the Best Paper Award in ACM MM 2013 and the Best Student Paper Award (Gold Prize) in PREMIA'14.

Zhenfeng Zhu received the M.S. degree from the Harbin Institute of Technology, Harbin, China, in 2001, and the Ph.D. degree from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2005.

He is currently a Professor with the Institute of Information Sciences, Beijing Jiaotong University, Beijing. His current research interests include image and video understanding, computer vision, and machine learning. He has authored or co-authored over 80 academic papers in the above areas.



Shuicheng Yan received the Ph.D. degree from Peking University, Beijing, China, in 2004.

He is currently an Associate Professor with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore, and the Founding Lead of the Learning and Vision Research Group (http://www.lv-nus.org). He has authored/co-authored nearly 400 technical papers over a wide range of research topics, with over 20 000 Google Scholar citations. His current research interests include machine learning, computer vision, and multimedia.

Dr. Yan was a recipient of the Best Paper Awards from ACM MM'13 (Best Paper and Best Student Paper), ACM MM12 (Best Demo), Pacific-Rim Conference on Multimedia'11, ACM MM10, IEEE International Conference on Multimedia & Expo10, and International Conference on Internet Multimedia Computing and Service'09, the Runner-Up Prize of ILSVRC'13, the Winner Prizes of the Classification Task in PASCAL Visual Object Classes Challenge (VOC) 2010–2012, the Winner Prize of the Segmentation Task in PASCAL VOC 2012, the Honorable Mention Prize of the Detection Task in PASCAL VOC'10, the 2010 TCSVT Best Associate Editor Award, the 2010 Young Faculty Research Award, the 2011 Singapore Young Scientist Award, and the 2012 National University of Singapore Young Researcher Award. He is an ISI Highly-Cited Researcher 2014, and an International Association of Pattern Recognition Fellow 2014. He has been serving as an Associate Editor of the IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, Computer Vision and Image Understanding, and the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY.