Multimed Tools Appl (2018) 77:8695–8710
DOI 10.1007/s11042-017-4764-0

A hybrid architecture based on CNN for cross-modal semantic instance annotation

Yongzhe Zheng[1,2] **· Zhixin Li**[1,2] **· Canlong Zhang**[1,2]

Received: 14 October 2016 / Revised: 24 April 2017 / Accepted: 26 April 2017 /
Published online: 5 May 2017

Abstract With the rapid growth of various media data, how to effectively manage and retrieve multimedia data has become an urgent problem to be solved. Due to semantic gap, overcoming the semantic gap has become a difficult problem for image semantic annotation. In this paper, a hybrid approach is proposed to learn automatically semantic concepts of images, which is called CNN-ECC. It's divided into two processes generative feature learning and discriminative semantic learning. In feature learning phase, the redesigned convolutional neural network (CNN) is utilized for feature learning, instead of traditional methods of feature learning. Besides the reconstructed CNN model has the ability to learn multi-instance feature, which can enhance the image features' representation when extracting features from images containing multiple instances. In semantic learning phase, the ensembles of classifier chains (ECC) are trained based on obtained visual feature for semantic learning. In addition, the ensembles of classifier chains can learn semantic association between different labels, which can effectively avoid generating redundant labels when resolving multi-label classification task. Furthermore, the experimental results confirm that proposed approach performs more effectively and accurately than state-of-the-art for image semantic annotation.

Keywords Feature learning · Semantic learning · Cross-modal semantic annotation · Convolutional neural networks

✉ Zhixin Li
lizx@gxnu.edu.cn

Yongzhe Zheng
zhengyzms@163.com

Canlong Zhang
clzhang@gxnu.edu.cn

[1] Guangxi Key Lab of Multi-source Information Mining and Security, Guangxi Normal University, Guilin 541004, China

[2] Guangxi Experiment Center of Information Science, Guilin 541004, China

1 Introduction

With the rapid growth of various media data, how to effectively manage and retrieve multimedia data becomes an urgent problem to be solved. The previous image retrieval methods can be divided into two categories in general: text-based image retrieval methods and content-based image retrieval methods. Text-based image retrieval technology initially label images artificially, based on which subsequently using traditional text search engine query images. This method is intuitive, however, due to the high cost of manual annotation, this retrieval method is not adapt to massive image databases. Content-based image retrieval applies feature extraction and high-dimensional indexing techniques to image retrieval. It extracts several low level visual features of each image which is processed into the form of high dimensional visual vector after, and saves processed vectors in the database, eventually obtaining the search results by measuring the similarity between feature vectors. This method is well applied in some special fields, but images of similar visual characteristics are likely to be semantically irrelevant because of the notorious semantic gap [18, 28]. To obtain semantic-related retrieval results and avoid a large number of manual annotations, automatic image annotation has become a research hotspot.

The main goal of automatic image annotation is to determine the probability for certain semantic concept given by metadata. Automatic image annotation establishes the foundation for semantic retrieval of images and is closely related to these works such as automatic concept detection and language index, etc. At present, several approaches have been proposed to solve the problems of automatic image annotation and retrieval, which can be roughly categorized into two different models. The first one is based on generative model. In the beginning, the automatic annotation is defined as a traditional supervised classification problem [2, 16], which mainly depends on similarity between visual features and predefined tags to model the classifier, then an unknown image is annotated relevant tags based on computed similarity of visual level. The other is based on discriminative model, which regard image and text as equivalent data. The method try to mine the correlation between visual features and labels on an unsupervised basis by estimating the joint distributions of multi-instance features and words of each image [18, 28]. These approaches greatly reduces the ability of feature presentation by extracting various low-level visual features, therefore it makes the semantic gap become more narrow between images and semantic.

The performances of image annotation are highly dependent on the representation of visual feature and semantic mapping. In view of the fact that deep convolutional neural networks (CNNs) has been demonstrated an outstanding performance in computer vision in recent years. For example, many works [9, 14, 19, 20] have demonstrated that CNN has a better effect than existing methods of hand-crafted features in many computer vision applications. Inspired by these articles, this paper proposes a hybrid architecture based on CNN for image semantic annotation to improve the performances of image classification and annotation.

In this work, we propose a novel hybrid architecture for image semantic annotation, and name it CNN-ECC. Firstly, a redesigned CNN model is used to learn high-level visual features. Secondly, the ensembles of classifier chains (ECC) are exploited to train model based on visual features and predefined tags. Finally, a hybrid framework is put forward to learn semantic concepts of images combining CNNs. The experimental results show that our approach performs more effectively and accurately than previous approaches for image classification and annotation tasks.

2 Related work

Over the past decades, various approaches based on discriminative model have been proposed for semantic image annotation and retrieval. For example, the content-based soft annotation (CBSA) system [1] is based on binary classifiers used to train each word and it indexes a new image according to the output of each classifier. To improve the accuracy of class prediction, Goh K S et al. [8] annotate images by classification based on multi-class SVMs. Particle swarm model selection (PSMS) [6] uses a one-vs-all (OVA) strategy which divides a multi-class problem into a series of binary classification problems and each problem is applied to deal with whether a region belongs to a particular class or not. In addition, a nearest spanning chain method is proposed to construct the image-based graph. Recently, Zhang et al. [32] annotate images by incorporating word correlations into multi-class SVM, which employs optimal principle of minimum probability of word correlations and combines annotation as a multi-class classification problem, where each of the word or concept correlations are computed by a co-occurrence matrix, etc.

Most approaches based on generative model implement image to semantic mapping by learning the correlations between visual features and textual words. For example, Monay et al. [21] propose PLSA-WORDS to model multi-modal co-occurrences. This approach considers both semantic terms (words) and visual information (visual features) including color, texture information, and three discrete feature types that are blobs (region-based), Hue-Saturation-Value (HSV) and Scale-Invariant Feature Transform (SIFT) respectively. Jacobs et al. [11] propose a general multi-view feature extraction approach (GMA) for image annotation. GMA can obtain a single linear or nonlinear subspace over different feature spaces, which is useful for cross-view classification and retrieval. Mahendran et al. [19] propose a classical Haar and HoG features versus bag of words method for image annotation and retrieval, etc.

To sum up, these approaches all employed hand-craft features, even though these methods accomplish the annotation task based on different thoughts. For computer vision and multimedia analysis task, extracting useful features from target is essential in the processing of model, and the method of image feature extraction directly affects the performance of image annotation and retrieval. However, the traditional feature extraction methods reduce the representation ability of visual features, and these methods are not able to fully learn the semantic correlation between text labels. Therefore, we propose a multi-instance learning method based on deep learning to replace the traditional feature extraction methods.

Deep learning techniques aim to learn hierarchical feature representations from original images, where the higher level features are defined from lower level ones. Since convolutional neural network (CNN) [14] is proposed, deep learning has made outstanding achievements in the field of computer vision. In recent work, Mahendran et al. [19] have demonstrated CNNs is better performance than existing methods based on hand-crafted feature for many computer vision applications, such as object classification [9, 14], face recognition [22] and image retrieval [23]. Furthermore, Razavian et al. [25] have demonstrated the pre-trained CNN can be used as a generic image representation model to extract visual features for diverse visual recognition tasks.

By studying plenty of papers about image auto-annotation, we notice that most authors don't consider how to represent an object's feature better, and they just extract the low-level features of objects. Althought it's full of difficulties for automatically extracting the high-level features, it's a worthwhile work for the image model problem. Considering all above discussion and our previous work cPLSA [17], we have a nature choice to employ the CNNs model instead of cPLSA model. In generative learning step, CNN directly improves representation ability of visual features by automatic learning, which extract high-level visual features of each image on test data set by pre-trained CNN model on target data sets. In addition, this paper employ ensemble of classification chains to model extracted visual vectors and tags. Therefore, using multi-label classification to learn semantic concepts is able to overcome semantic gaps between image and text [33]. In Section 4, abundant experiments are conducted on two internationally data sets to compare the effect of CNN visual feature and traditional visual features for cross-modal image semantic annotation. The experimental results show good performance can be achieved by CNN visual features based on several classic cross-modal retrieval methods, such as PLSA-WORDS and GHM. Hybrid framework achieve inconceivably superior performance in image annotation and retrieval.

3 Hybrid framework for image semantic annotation

In this section, we present the two components of our framework. Combining deep model with ensembles of classifier chains, we propose a hybrid learning framework to address cross-modal semantic annotation problem between images and text labels. As shown in Fig. 1, the hybrid framework is divided into two steps, including generative feature learning process and discriminative semantic learning process.

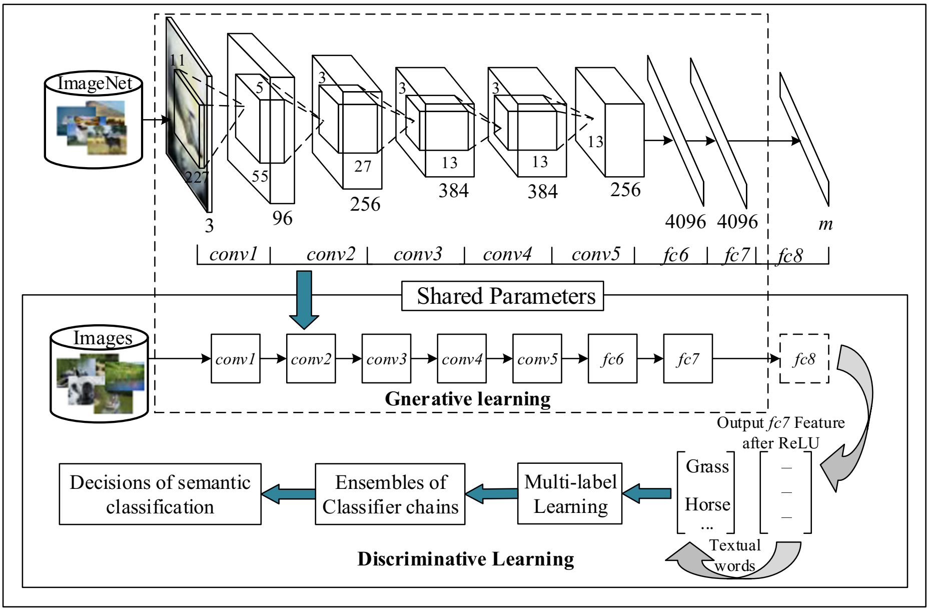

Fig. 1 Illustration of the CNN visual features and the proposed CNN-ECC image semantic annotation system. The high-level CNN visual features fc7, can be directly extracted from the pretrained CNN model. The fine-tuned CNN visual features, i.e., FT-fc6 and FT-fc7, are extracted from the CNN model, which is first pretrained on ImageNet and then fine-tuned on the target data set. For CNN-ECC, as shown in the lower part, the fc7 outputs after ReLU are employed for cross-modal annotation

3.1 CNN visual features extraction

The shared CNN contains five convolutional layers and three fully-connected layers with numerous parameters. Consequently, without enough training images, it is very difficult to obtain an effective deep model for multi-label classification. However, it is generally unaffordable to collect and annotate a large-scale multi-label data set. Fortunately, a large-scale single-label image dataset, i.e., ImageNet, can be used to pre-train the shared CNN for parameters initialization.

3.1.1 Extracting visual features from pre-trained CNN model

These works [15, 25] have demonstrated the outstanding performance of the off-the-shelf CNN visual features in various recognition tasks, so the pre-trained CNN model is utilized to extract visual features in this paper. CNN is a special form of neural network that consists of three different types of layers, such as convolutional layers, spatial pooling layers, and fully connected layers. Different network structures will show different ability of visual features representation. As shown in the top of the Fig. 1, reconstructed CNN model has the similar network structure to the AlexNet [14] in this paper, which contains five convolutional layers (short as conv) and three fully-connected layers (short as fc). Particularly, the reconstructed model is pre-trained in 1.2 million images of 1000 categories from ImageNet [3] in this paper, each image is resized to 227*227 and fed into the CNN model, then data dirve neural networks to learning parameters. However, the BP neural network has slow convergence speed and is easy to fall into local minimum problems in practical application. So using the activation function correctly can accelerate the convergence of the network. Rectified Linear Units (ReLUs) is a kind of activation function applied in CNN. Krizhev et al. [14] have proved that the Rectified Linear Units (ReLUs)not only saves the computing time, but also implements the features' sparse representation, and ReLU also increases the sample characteristic diversity. So to improve the generalization ability of the feature representation, the $fc7$ features are extracted from the secondly convolution layer after ReLU. The $fc7$ denote the 4096 dimensional features of the last two fully-connected layers after the rectified linear units (ReLU).

3.1.2 Exacting fusion visual features from redesigned CNN model

Taking into account the different categories between the target dataset and ImageNet, if we directly utilize the pre-trained model to exact image visual features on the ImageNet, it may not be the optimum strategy. To make the model fit the parameters better, the last hidden layer is redesigned for visual feature learning task, later CNN model is redesigned by fine-tuning parameters of each of images in the target dataset. As shown in the mid of Fig.1, the overall architecture of our CNN model still contains five conv layers including a pooling layer and three fully-connected layers. The last hidden layer is redesigned for feature learning task. Given the number of the target dataset's categories m, after the output of the last fully-connected layer is then fed into an m-way softmax and produces a probability distribution for m categories, the number of neural units of the last fully-connected layer is reduced from 1000 to m.

Given one training sample x, the network extracts layer-wise representations from the first $conv$ layer to the output of the last fully connected layer fc_8, which can be viewed as high level features of the input image $fc_8 \in \mathbb{R}^m$. Followed by a softmax layer, fc_8 is transformed into a probability distribution p for objects of m categories, $\boldsymbol{p} \in \mathbb{R}^m$. CNN model measures the prediction loss of the network by cross entropy, and the computational formula is shown as follows.

$$p_i = \frac{\exp(\overset{\wedge}{v_i})}{\sum_i \exp(\overset{\wedge}{v_i})}, \text{ and } L = -\sum_i t_i \log(p_i), i = 1, ..., m \tag{1}$$

where L is the loss function based on cross entropy, and p_i is probability of that object belongs to the ith class, t_i denotes the true label of the sample x_i, and v_i denotes the feature vector set of the ith column. After CNN model completes forward propagation and outputs probability distribution, it is necessary to calculate the loss value according to the loss function. To reduce the loss value, back propagation are utilized to compute gradient parameters. Gradient is computed as follows in the processing of back propagation.

$$\frac{\partial L}{\partial \overset{\wedge}{v_i}} = p_i - t_i \tag{2}$$

To learn multiple instances as a fusion features, we combine deep representation with multiple instance learning. Denote $\{x_j | j = 1, 2, ..., n\}$ as a bag of n instances and $t = \{t_i | t_i \in 0, 1, i = 1, ..., m\}$ as the label of the bag. Neural network extracts visual features of the bag $v = \{v_{ij}\} \in R^{m \times n}$. So an image can be viewed as multi-instance bag, in which each column is the representation of an instance. The merged representation of the bag for visual vectors are defined as:

$$\underset{i}{\hat{v}} = f(v_{i1}, v_{i2}, ..., v_{in}) \tag{3}$$

where function f represents the mapping function of feature set. Here we choice max pooling layer to merge the multi-instance bag.

In the training phase, stochastic gradient descent algorithm is used to optimize the loss function L. Suppose that we have a set of training images $I = \{M_i | i = 1, 2, ...n\}$. In the process of training network, training samples are regarded as bags I_i, and there are a number of t_i instances in each bag. The network extracts layer-wise representations from the first conv layer to the output of the last fully connected layer visual vectors v_i, which can be viewed as high level features of the input image. Fine-tuning by training with classes of particular objects, is known to improve classification accuracy. By fine-tuning the transferred parameters in CNN model, the better parameters can be obtained, and predicted value is closer to real value. In order to improve the effect of visual feature learning, we first employ existing model to fine-tune the parameters in the target dataset, then we apply the fine-tuned CNN model to learn image visual features. Similarly, the $FT - fc7$ denotes the 4096 dimensional features of the last two fully-connected layers after ReLU.

3.2 Ensembles of classifier chains for semantic learning

The Classifier Chains [26] are used to accomplish the task of multi-label classification. Taking into account the semantic correlations between tags, Classifiers Chain can't classify images into multiple semantic classes, with a high degree of confidence and acceptable computational complexity. Based on this research, we propose the the Ensembles of Classifier Chains (ECC) to improve the accuracy of the annotation system. In the discriminative learning phase, the ensembles of classifier chain model consists of m binary classifiers, and each of the binary classifier is implemented by SVM [13]. Furthermore, classifier chain can effectively overcome the problems of label independence in image binary classification by learning the semantic relevance between labels.

The ensembles of classifier chains model consist of m binary classifiers, where m denotes real classes of label sets and target label sets is denoted as T. Classifiers are linked along a chain where each classifier deals with the binary relevance problem associated with label $l_j \in \{T_j | j = 1, 2, ...n\}$. The feature space of each linked in the chain is extended with the $0-1$ label associations of all previous links. The training procedure is outlined in Algorithm 1 in the left of Table 1. Lastly, the notation is noted for a training example (x, S), where $S \subseteq T$ and x is an instance feature vector.

Stated thus a chain $C_1, C_2, ..., C_i$ of binary classifier is formed. Each classifier C_j in the chain is responsible for learning and predicting the binary association of label l_j, which is given in the feature space and is augmented by all prior binary relevance predictions in the chain $l_1, l_2, ..., l_{j-1}$. The classification procedure begins at and propagates along the chain C_1 to determine $Pr(l_1|x)$ and every following classifier $C_2, ..., C_j$ predicts $Pr(l_j|x_i, l_1, l_2, ..., l_{j-1})$. This classification procedure is described in Algorithm 2 in the right of Table 1.

This training method takes into account label semantic correlations in classifier chains, which overcomes the label independence problem of binary relevance method. Although $|T|/2$ features are added to each instance on an average, this item is negligible in computational complexity because $|T|$ is invariably limited in practices, therefore classifier chain still remains advantages of binary relevance method including low memory and runtime complexity. Different order of the chain clearly has a different effect on accuracy. This problem can be solved by using an ensemble framework with a different random train ordering for each iteration. Ensembles of classifier chains train m classifier chains including $C_1, C_2, ..., C_m$. Each C_k model is trained with a random chain which can order the L outputs and get a random subset of D. Hence each C_k model is likely to be unique and able to give different multi-label predictions. These predictions are summed by label so that each label receives a number of votes. A threshold is used to select the most popular labels which form the final predicted multi-label set.

Table 1 Training procedures of ensembles of classifier chains for multi-label learning

	Algorithm 1	Algorithm 2
Processing	Training steps of classifier chain	Classifying procedure ECC
Input	Training set	Test example x
	$I = (x_1, S_1), (x_2, S_2), ..., (x_n, S_n)$	
Output	Classifier chains $C_1, C_2, ..., C_m$	$Y = l_1, l_2, ..., l_m$
procedures		
1	For $i \in 1, 2, ...m$	Y
2	Semantic learning	For $i \in 1, 2, ..., m$
3	$I' \leftarrow \{\}$	Do $Y \leftarrow \cup$
		$(l_i \leftarrow C_i : (x_i, l_1, l_2, ...l_{j-1}))$
4	For $(x, S) \in I$	Return (x, Y)
5	Do $I' \leftarrow I' \cup ((x, l_1, l_2, ..., l_{i-1}), l_i)$	
6	Train C_i	
	to predict binary relevance of l_i	
7	$C_i : I \to l_i \in 0, 1$	

Given the kth individual model predicts vector $y_k = (l_1, l_2, ..., l_{|T|}) \in \{0, 1\}^{|T|}$. The sums are stored in a vector $W = (\lambda_1, \lambda_2, ..., \lambda_{|L|}) \in \mathbb{R}^m$, where λ_j is defined as $\lambda_j = \sum_{k=1}^{m} l_j \in y_k$. Hence each $\lambda_j \in W$ represents the sum of the votes for label $l_j \in T$. Then, we normalize W to W_{norm}, which represents a distribution of scores for each label in [0, 1]. A threshold is used to choose the final multi-label set Y such that $l_j \in Y$ where $\lambda_j \geq t$ for threshold t. Hence the relevant labels in Y represent the final multi-label prediction.

3.3 Image semantic annotation

We now explain our method for semantic image annotation. As shown in Fig. 1, the training process of CNN-ECC is divided into two steps.

Step1: feature learning based on resigned CNN model from outside training data As many efficient and open source implementations of CNNs are available, we will not go into the full details of implementing convolutional, max polling or fully connected layers. For that, we relied on the sources provided by the Caffe library [12], itself based on the Nvidia CuDNN library. We utilize the ImageNet [3] to pre-train the shared CNN model. In experiments, we handle the training images with a pre-processing technology. Given a training sample, we first resize it into 256 × 256 pixels, after that we extract random 227 × 227 patches from the given image and train our network based on these extracted patches. Each extracted patch is pre-processed by subtracting the image mean. We train the network by using stochastic gradient descent with a momentum of 0.9 and weight decay of 0.0005. To overcome over-fitting, each of the first two fully-connected layers is followed by a drop-out operation with a drop-out [30] ratio of 0.5. The learning rate is initialized as 0.01 for all layers and reduced to one tenth of the current rate after every 20 epochs (70 epochs in all). At last, the trained CNN model is utilized to extracted visual features. Suppose that we have a test image I, CNN model extracts visual vectors by pre-trained CNN model and we denote the space of visual vectors as $v = \{v_1, v_2, ..., v_i\}$, where v_i denotes the visual vector of image I. Noting the notation for a training example (v_i, S), where $S \in T$, T denotes the label sets and v is a feature vector extracted from CNN model. Then, by making use of the aspect distribution and original labels of each training image, we build a series of classifiers in which every word in the vocabulary is treated as an independent class. The classifier chain model implements the feature classification task and it can effectively learn the semantic correlation between labels in discriminative step. Finally, given a test image, the CNN-ECC system will return a correlative label subset $l \in T$. Specifically, we combine the outputs of image and text understanding systems in the final fully connected layer, as illustrated in Fig. 1.

Step2: semantic learning based on ensembles of classifier chains In discriminative semantic learning phase, we utilize visual vectors extracted from pre-trained CNN model and corresponding text labels to fit the ensembles of classifier chains. This training method passes label information between classifiers, allowing classifier chain considers label correlations and thus overcoming the label independence problem of binary relevance method. Then, we classify the aspect distribution of each test image with the trained multi-class classifier. Following by [21], multi-class classifier model outputs 5 words with highest confidence as semantic labels of the test image. After each image in the database is annotated, the retrieval algorithm can rank the images labeled with the query word by decreasing confidence.

Based on the trained CNN and multi-class classifier model, the multi-label classification of a given image can be summarized as follows. We firstly generate the multi-instance fusion feature of the given image based on the redesigned CNN method. Then, for each test image, the top 5 predictive results can be obtained by the trained ensembles of classifier chains. We integrate deep features and semantic learning to truly find out discriminative and relevant labels for each image.

4 Experiments and results

In this section, we discuss implementation details of our training, and evaluate different components of our method. We conduct experiments of our CNN-ECC learning framework on both image classification and auto-annotation. We choose two image datasets Corel5K and Pascal VOC 2007, which are widely used in image classification and annotation. In order to make the experimental result more convinced, we simultaneously compare the experimental results with the existing traditional model and deep model.

4.1 Datasets and evaluation measures

To test the effectiveness and accuracy of the proposed approach, our experiments are conducted on a baseline annotated image datasets Corel5K [5] and Pascal VOC 2007 [7]
.

- Corel5k: it's a basic comparative dataset which contains 5000 images from 50 Corel Stock Photo cds for recent research works on image annotation. The training set of 4500 images and the test set of 500 images are obtained by dividing this dataset into 3 parts: a training set of 4000 images, a validation set of 500 images and a test set of 500 images. Like the Duygulu et al. [5].
- Pascal VOC 2007 [7]: There are 9963 images of 20 categories in this data set. Each image accompanies 399 tags annotated by methods in [10]. First, the data set is divided into three subsets including train, val, and test, and the total number of images contained in train and validation is 5011, the number of images contained in test is 4952. Second, experiments are conducted on train (including validation) and test respectively. Eventually, the obtained visual features by using methods in [10], which contains a 180 dimensional SIFT BoVW features, are compared with CNN visual features.

Specifically, image annotation performance is evaluated by comparing the automatically generated results on the test set with the human-produced ground truth. It's essential to use several evaluation measures in multi-label evaluation. Similar to Monay et al. [21], we use mAP as evaluation measures. Naturally, we define the automatic annotation as the top 5 semantic words of largest posterior probability, and compute the recall and precision of every word in the test set.

4.2 Image annotation on Corel5K

In this section, the performance of our model on the corel5k data set for image multi-label annotation is demonstrated, and the results are compared with some existing image annotation methods, such as PLSA-WORDS [21], HGMD [17] and DNN [27]. After evaluating the returned keywords in a class-wise manner, the performance of image annotation is

Table 2 Comparing 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

Method	Visual	Result on all words		mAP
	features	Precision	Recall	
PLSA-WORDs	BoVW	22.1	12.1	19.1
	fc7	27.5	21.7	26.9
	FT-fc7	29.3	22.6	27.3
HGDM	BoVW	32.1	29.3	26.3
	fc7	36.4	30.5	29.7
	FT-fc7	37.6	32.9	30.9
DNN	FT-fc7	42.5	40.5	40.7
CNN-ECC(our)	FT-fc7	47.7	43.6	44.9

evaluated by comparing the automatically generated results with the original manual annotations. Similar to Monay F et al. [21], the recall and precision of every word in the test set is computed, and their mean is used to summarize the system performance.

Table 2 reports results of several models on the set of all 260 words which occur in the training set. Data in precision and recall columns denotes mean precision and mean recall of each word. The off-the-shelf CNN features (i.e. fc7 and FT-fc7) obtain significant improvements (7.8% based on PLSA-WORDS, 3.4% based on HGDM) compared with these traditional feature learning methods. After fine-tuning, a further improvement (8.2% based on PLSA-WORDS, 4.6% based on HGDM) can be achieved with the best performance of the CNN visual features FT-fc7. Annotations of several images obtained by proposed method annotation system are show in Fig. 2. We can see that annotations generated by CNN-ECC are more accurate than HGDM in most cases. To intuitively compare with precision and recall of various methods, the Fig. 3 presents the precision-recall curves of several annotation models on the Corel5k data set. As is shown in Fig. 3, CNN-ECC performs consistently better than other models, where the precision and recall values are the mean values calculated based on all words.

Image					
Ground Truth	temple,sky, buddhist, mountains	elephant, trees,planes, sky	cabin,trees, utumn,field	trees,sky, road,park	people, water,trees, sand
HGDM annotations	house,sky, water,clound, mountains	africa,sky, animal,land, beach	field,grass, land,trees, mountains	trees,road, sky,pant, sea	sand,water, people,tress, clound
CNN-ECC annotations	temple,sky, palace,land, mountains	elephant, land,planes, trees,sky	chair,trees, mountains, land,field	park,road trees,sky, mountains	sand,coast, trees,water, people,

Fig. 2 Comparison of annotations made by HGDM and CNN-ECC on Corel5k

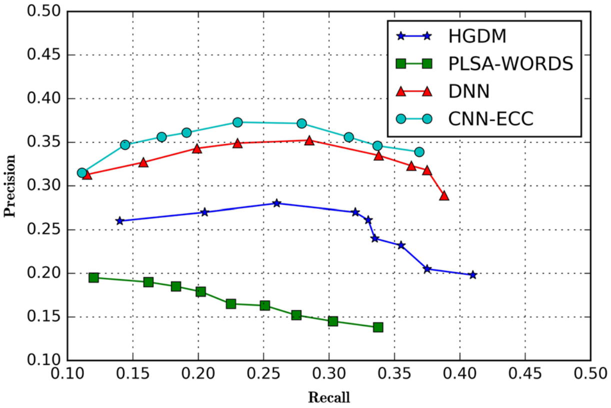

Fig. 3 Precision–recall curves of several models for image annotation on Corel5K

4.3 Cross-modal annotation on pascal VOC 2007

In this experiment, we compare our method with several different methods in image classification and annotation tasks. The results of the experiment demonstrates redesigned CNN powerful ability as a universal representation for various recognition tasks. Particularly, the extracted CNN features have strong capacity of the image features' representation based on the good experimental results on Corel 5K. So the results on Pascal 07 only compare with deep features (i.e. fc7 and FT-fc7). Table 3 reports our experimental results of state-of-the-arts and CNN-ECC on Pascal VOC2007 data set. Because Pascal VOC 2007 is a multi-label data set, the cross-modal retrieval based on the criterion that it's regarded as a relevant result, if the retrieved result shares as least one class label with the query is implemented. We compare our approach with HGDM [17], GHM [31], AGS [4], NUS [29] and DNN [27] for image classification and annotation tasks. It reported the classification results on Pascal 2007, which achieved the state-of-the-art performance. As shown in Table 3, compared with HGDM, the proposed CNN-ECC has an improvement of 6.1%. Both pre-trained on the ImageNet dataset with 1,000 classes, CNN-ECC gives a more competitive result compared with DNN (79.1% vs. 73.0%).

After evaluating the off-the-shelf CNN and our fine-tuned ones in different number of training epochs, our different strategies for different feature selection are evaluated independently in order to decompose the benefit of each ingredient. Finally, a comparison with the traditional method is performed, and the same training data is trained exactly as the ones used in our architecture. The results in Fig. 4 show that CNN features with robust feature representation ability, both acquire a consistent increase in the performance. However, we notice that oscillation of the error in the validation set from early epochs, which maybe imply over-fitting. Considering this situation, we draw on the experience of Prechelt [24], which employs early stopping when training model. It indicates we will stop training model, if the predict results of the model in a certain number of iterations do not improve. As shown in Fig. 4, by training on target dataset, the error in the validation set gradually level off.

Table 4 shows our experimental results compared with the state-of-the-arts on Pascal VOC 2007. The results imply a comprehensive measure of annotation and retrieval accuracy. Obviously, CNN-ECC similarly obtains significant improvements based on extracted CNN visual features (e.g., fc7and FT-fc7).

Table 3 Image classification results on Pascal VOC 2007

	HGDM	GHM	AGS*	NUS*	DNN*	CNN-ECC*
aero	61.3	76.7	82.2	82.5	91.2	92.7
bike	57.6	74.7	83.0	79.6	81.4	83.1
bird	51.1	53.8	58.4	64.8	82.1	86.7
blt	39.8	40.4	76.1	73.4	51.6	53.7
boat	63.9	72.1	56.4	54.2	81.1	83.1
bus	58.2	71.7	77.5	75.0	84.4	86.6
car	63.5	83.6	88.8	77.5	83.9	85.1
cat	44.7	66.5	69.1	79.2	54.5	54.5
chair	41.6	52.5	62.2	46.2	61.0	67.6
cow	43.9	57.5	61.8	62.7	61.0	67.6
dog	36.8	51.1	64.2	41.4	72.3	73.2
hrs	59.7	81.4	51.3	74.6	74.9	78.7
mbk	63.7	71.5	85.4	85.0	75.6	79.1
per	71.2	86.5	80.2	76.8	83.7	86.6
plant	62.1	36.4	91.1	91.1	47.4	51.7
shp	46.9	55.3	48.1	53.9	71.7	74.3
sofa	38.2	60.6	61.7	61.0	60.0	63.2
tabel	51.8	62.8	67.7	67.5	53.8	67.2
train	62.1	80.6	86.3	83.6	88.3	91.5
tv	51.2	57.8	70.9	70.6	79.4	80.4
mAP	53.5	64.7	71.1	70.5	73.0	79.1

DNN is the popular OverFeat representation. * indicates methods using additional data (i.e., ImageNet) for training

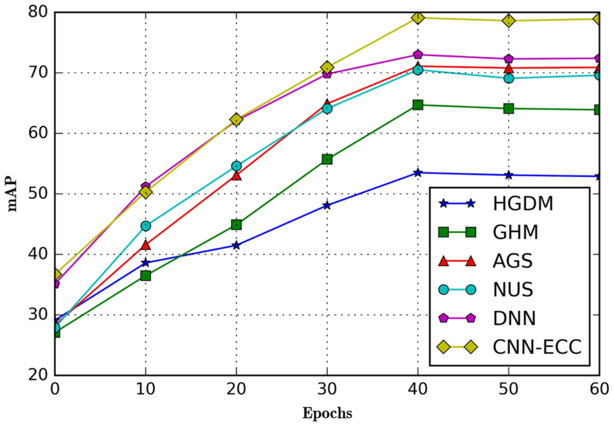

Fig. 4 Influence of the number of three methods used for CNN fine-tuning. Performance is evaluated on Pascal 07 dataset

Table 4 Image annotation results on Pascal VOC 2007

	Visual features	mAP
HGDM [17]	–	26.3
GHM [31]	–	32.7
AGS* [4]	–	49.6
NUS* [29]	–	50.4
DNN* [27]	–	53.7
CNN-ECC(our)*	FT-fc7	61.3

* indicates methods using additional data (i.e., ImageNet) for training. (The "-" means to use their method)

On the one hand, our feature learning strategy directly optimizes visual features when extracting features from image, and applying the fine-tuned networks to enhance feature representation. By integrating multi-instance learning in CNN, that is, first regarding each object as a region vector and then aggregating, performance is significantly enhanced. On the other hand, the ensembles of classifier chains can learn semantic association between different labels, which can effectively avoid generating redundant labels when resolving multi-label classification task. In summary, the advanced performances of our methods not only are due to the feature representation, also come from feature learning and semantic discrimination learning. By comparing results with other methods, the CNN-ECC image semantic annotation system outperforms many state-of-the-art approaches, which proves that the redesigned CNN and the ensembles of classification classifiers are separately effective in learning visual features and semantic concepts of images. By comparing with the other state-of-the-art for cross-media image annotation and retrieval, Tables 2 and 4 separately show the comparison in terms of rigid and articulated visual features among Corel5k and Pascal 2007. It proves that the extracted features from redesigned CNNs outperform almost all the original hand-crafted features. For image annotation, the ECC shows a strong learning ability of semantic association. Figures 2 and 4 show CNN-ECC system automatically generate semantic annotation. When it annotates image by multi-label, it is more reliable than other methods.

5 Conclusion

This paper proposes a hybrid method based on CNN for cross modal semantic instance annotation. Firstly, we utilize the trained reconstructed convolution neural networks to extract visual features. Secondly, ensembles of classier chains are trained based on obtained visual feature and corresponding text labels for semantic learning. At last, based on the whole model, the semantic annotation task is completed. In comparison to many state-of-the-art approaches, experimental results show that our method achieves superior results in the tasks of image classification and annotation on Corel5K and Pascal VOC 2007, therefore re-designed CNN model and ensembles of classifier chains can effectively improve image annotation accuracy.

However, in the process of learning visual features, CNN-ECC only employs single convolution neural network but not fully understanding multi-instances in the image. Furthermore, owing to the semantic gap between cross-modal data, how to mine the high-level semantic relevance between the tags is a wholly worthwhile task. In future research, we aim to take semi-supervised learning based on a large number of unlabeled data to improve its effectiveness.

Acknowledgements This work is supported by the National Natural Science Foundation of China (Nos. 61663004, 61363035, 61365009), the Guangxi Natural Science Foundation (Nos. 2016GXNSFAA380146, 2014GXNSFAA118368), the Director Fund of Guangxi Key Lab of Multi-source Information Mining and Security (16-A-03-02), the Guangxi "Bagui Scholar" Teams for Innovation and Research Project, Guangxi Collaborative Innovation Center of Multi-source Information Integration and Intelligent Processing.

References

1. Chang E, Goh K, Sychay G, Wu G (2003) Cbsa: content-based soft annotation for multimodal image retrieval using bayes point machines. IEEE Trans Circuits Syst Video Technol 13(1):26–38
2. Cusano C, Ciocca G, Schettini R (2003) Image annotation using svm. In: Proceedings of SPIE - the international society for optical engineering, vol 5304, pp 330–338
3. Deng J, Dong W, Socher R, Li LJ (2009) Imagenet: a large-scale hierarchical image database. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 248–255
4. Dong J, Xia W, Chen Q, Feng J, Huang Z, Yan S (2013) Subcategory-aware object classification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 827–834
5. Duygulu P, Barnard K, de Freitas JFG, Forsyth DA (2002) Object recognition as machine translation: learning a lexicon for a fixed image vocabulary. In: Proceedings of the european conference on computer vision (ECCV), pp 97–112
6. Escalante HJ, Montes M, Sucar LE (2012) Multi-class particle swarm model selection for automatic image annotation. Expert Syst Appl 39(12):11011–11021
7. Everingham M, Gool L, Williams CK, Winn J, Zisserman A (2010) The pascal visual object classes (voc) challenge. Int J Comput Vis 88(2):303–338
8. Goh KS, Chang EY, Li B (2005) Using one-class and two-class svms for multiclass image annotation. IEEE Trans Knowl Data Eng 17(10):1333–1346
9. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778
10. Hwang SJ, Grauman K (2010) Accounting for the relative importance of objects in image retrieval. In: Proceedings of the British machine vision conference, pp 1–12
11. Jacobs DW, Daume H, Kumar A, Sharma A (2012) Generalized multiview analysis: a discriminative latent space. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2160–2167
12. Jia Y, Shelhamer E, Donahue J, Karayev S, Long J, Girshick R, Guadarrama S, Darrell T (2014) Caffe: convolutional architecture for fast feature embedding. In: Proceedings of the 22nd ACM international conference on multimedia, pp 675–678
13. Joachims T (1998) Making large-scale svm learning practical. Technical report, Universitat Dortmund
14. Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. In: Proceedings of the advances in neural information processing systems (NIPS), pp 1106–1114
15. Li G, Yu Y (2015) Visual saliency based on multiscale deep features. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 5455–5463
16. Li J, Wang JZ (2003) Automatic linguistic indexing of pictures by a statistical modeling approach. IEEE Trans Pattern Anal Mach Intell 25(9):1075–1088
17. Li Z, Shi Z, Zhao W, Li Z, Tang Z (2013) Learning semantic concepts from image database with hybrid generative/discriminative approach. Eng Appl Artif Intell 26(9):2143–2152
18. Liu Y, Zhang D, Lu G, Ma WY (2007) A survey of content-based image retrieval with high-level semantics. Pattern Recogn 40(1):262–282
19. Mahendran A, Vedaldi A (2015) Understanding deep image representations by inverting them. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 5188–5196

20. Mnih V, Kavukcuoglu K, Silver D, Rusu AA, Veness J, Bellemare MG, Graves A, Riedmiller M, Fidjeland AK, Ostrovski G,, et al. (2015) Human-level control through deep reinforcement learning. Nature 518(7540):529–533
21. Monay F, Gatica-Perez D (2007) Modeling semantic aspects for cross-media image indexing. IEEE Trans Pattern Anal Mach Intell 29(10):1802–1817
22. Parkhi OM, Vedaldi A, Zisserman A (2015) Deep face recognition. In: Proceedings of the British machine vision conference, pp 6–18
23. Paulin M, Mairal J, Douze M, Harchaoui Z, Perronnin F, Schmid C (2017) Convolutional patch representations for image retrieval: an unsupervised approach. Int J Comput Vis 121(1):149–168
24. Prechelt L (1998) Early stopping—but when? In Lecture Notes in Computer Science 1524:55–69
25. Razavian AS, Azizpour H, Sullivan J, Carlsson S (2014) Cnn features off-the-shelf: an astounding baseline for recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 806–813
26. Read J, Pfahringer B, Holmes G, Frank E (2009) Classifier chains for multi-label classification. Mach Learn 85(3):254–269
27. Sermanet P, Eigen D, Zhang X, Mathieu M, Fergus R, LeCun Y (2014) Overfeat: integrated recognition, localization and detection using convolutional networks. In: Proceedings of international conference on learning representations, pp 1–16
28. Smeulders AWM, Worring M, Santini S, Gupta A, Jain R (2000) Content-based image retrieval at the end of the early years. IEEE Trans Pattern Anal Mach Intell 22(12):1349–1380
29. Song Z, Chen Q, Huang Z, Hua Y, Yan S (2011) Contextualizing object detection and classification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1585–1592
30. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R (2014) Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res 15(1):1929–1958
31. Yan S, Huang Z, Hua Y, Song Z, Chen Q (2012) Hierarchical matching with side information for image classification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3426–3433
32. Zhang L, Ma J (2011) Image annotation by incorporating word correlations into multi-class svm. Soft Comput 15(5):917–927
33. Zhang ML, Zhou ZH (2014) A review on multi-label learning algorithms. IEEE Trans Knowl Data Eng 26(8):1819–1837

Yongzhe Zheng is a postgraduate student at the College of Computer Science and Information Technology, Guangxi Normal University. His research interests include image understanding and deep learning.

Zhixin Li is a professor at the College of Computer Science and Information Technology, Guangxi Normal University. He obtained his Ph.D. degree in computer software and theory from Institute of Computing Technology, Chinese Academy of Sciences in 2010. He obtained his B.S. degree and M.S. degree at the Huazhong University of Science and Technology in 1992 and 2004 respectively. His research interests include image understanding, machine learning and multimedia information retrieval. His doctoral dissertation has won the best doctoral dissertation award of Chinese Association of Artificial Intelligence in 2011.

Canlong Zhang is an associate professor at the College of Computer Science and Information Technology Guangxi Normal University. In 2014, he obtained his Ph.D. degree in control technology and control engineering from Shanghai Jiao Tong University in China. He was engaged as an evaluation expert of science and technology project of Guangxi in 2011. His research interests include target tracking, pattern recognition and multi-sensor data fusion.