Pattern Recognition Letters 65 (2015) 103–108

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Deep neural network based image annotation☆

Songhao Zhu[a,*], Zhe Shi[a], Chengjian Sun[a], Shuhan Shen[b]

[a] School of Automation, Nanjing University of Posts and Telecommunications, Nanjing, 210046, China
[b] Institute of Automation, Chinese Academy of Sciences, Beijing, 110093, China

ARTICLE INFO

Article history:
Received 4 February 2015
Available online 7 August 2015

Keywords:
Deep learning
Multi-label
Multi-modal
Image annotation

ABSTRACT

Multilabel image annotation is one of the most important open problems in computer vision field. Unlike existing works that usually use conventional visual features to annotate images, features based on deep learning have shown potential to achieve outstanding performance. In this work, we propose a multimodal deep learning framework, which aims to optimally integrate multiple deep neural networks pretrained with convolutional neural networks. In particular, the proposed framework explores a unified two-stage learning scheme that consists of (i) learning to fine-tune the parameters of deep neural network with respect to each individual modality, and (ii) learning to find the optimal combination of diverse modalities simultaneously in a coherent process. Experiments conducted on a variety of public datasets evaluate the performance of the proposed framework for multilabel image annotation, in which the encouraging results validate the effectiveness of the proposed algorithms.

1. Introduction

Recent years have witnessed an explosive growth of digital images, and most of them are captured by handheld mobile devices. There is an urgent need to develop effective techniques to annotate images with several labels according to the semantic contents, which can be deployed in many applications, such as personal image collection organization and large-scale image retrieval.

From the point of view of pattern recognition, the issue of image annotation can be considered as an issue of assigning a set of relevant tags to an image according to the contents inside it, in which learning good features is a very important task and will significantly improve the overall system performance. Many efforts have been put forward to train hierarchical models which contain multiple levels of feature extractors, such as Gabor-like edges, object contour, shape, and texture. Recently, deep neural network (DNN), a typical hierarchical model has received more and more attention again since Hinton et al. introduce deep belief networks (DBNs) to efficiently train multilayer to learn features from unlabeled data [1]. The variants of DBN have been successfully applied to a variety of language and information retrieval applications [2–11]. By exploiting deep architectures, deep learning technologies can discover from training data the hidden structures and effective features to help improve performance. A convolutional DBN is presented in [2], aiming to achieve better performance in image classification and speaker identification tasks by unsupervised learning of hierarchical feature representation. Zeiler et al. [3] proposed an unsupervised framework to derive hierarchical image representations to deal with the image denoising or object recognition tasks. A generative deep learning model is presented in [4], aiming to achieve high-resolution images by merging a deep belief network with the gated Markov random field. Zhong et al. [5] employ a bilinear deep belief network framework to deal with the image classification task by utilizing a bilinear discriminant strategy to simulate the "initial guess" in human object recognition and effectively avoid falling into a bad local optimum simultaneously. Srivastava et al. [6] explore multimodal deep neural network to learn representations in image annotation and image retrieval tasks by fusing multiple sources with shared hidden representation. Mohamed et al. [7] complete the task of speech recognition by a deep belief network. Huang et al. [8] deal with the problem of assigning labels to images based on a multi-task deep neural network architecture. Gong et al. [9] perform image annotation by combining convolutional architectures with approximate top-k ranking objectives. An unsupervised deep learning framework is presented in [10], aiming to derive spatio-temporal features for human–robot interaction. Dong et al. [11] tackle the task of image super-resolution by learning a deep convolutional neural network.

Inspired by a variety of image annotation algorithms based on the idea of deep neural networks, this paper proposes a novel framework of multimodal deep learning, as shown in the following Fig. 1. Specifically, the convolutional neural networks with unlabeled data is utilized to pre-train the multimodal deep neural network to learn intermediate representations and provide a good initialization for the

☆ This paper has been recommended for acceptance by R. Davies.

* Corresponding author. Tel.: +86 15952021713; fax: +86 25 85866512.
E-mail address: zhush@njupt.edu.cn (S. Zhu).

http://dx.doi.org/10.1016/j.patrec.2015.07.037
0167-8655/

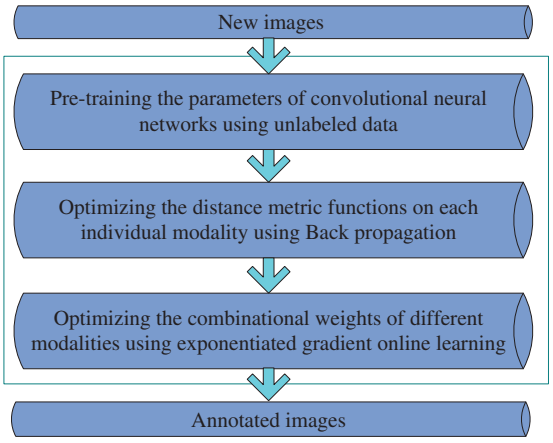


Fig. 1. The proposed multilabel deep learning framework.

network; then, backpropagation is adopted to optimize the distance metric functions on each individual modality; finally, the exponentiated gradient online learning algorithm is applied to optimize the combinational weights of different modalities.

The rest of this paper is organized as follows. Section 2 discusses the Networks Architecture. Section 3 presents Network Learning. Section 4 presents experimental results and Section 5 concludes this work.

2. Networks architecture

The overall architecture of the proposed convolutional neural networks model is shown in Fig. 2. The network contains eight layers with weights, where the first five are convolutional layers and the remaining three are densely connected layers. The outputs of the densely connected layer are fed into a 1000-way softmax classifier which produces a distribution over 1000 labels. For both the pre-training and fine-tuning phases, a multinomial logistic regression objective function is used.

Within the constructed convolutional neural networks, normalization layers are utilized in the first, second and last convolutional layers and max-pooling layers are utilized in all the normalization layers to introduce invariance. Furthermore, rectified linear unit is utilized as the nonlinear activation function for every convolutional layer and every densely connected layer.

Before feeding images into the convolutional layers, each image is resized to 256×256. Next, the first two convolutional filter sizes are set as 7×7 with a stride of 2 pixels and 5×5 with a stride of 2 pixels respectively, and the filter number are set as 96 and 256 respectively. Such a size of filter is utilized to obtain the mid-frequency information as well as the extremely low and high frequencies, and smaller stride is utilized to avoid the "dead features" which is harmful to the next layers. Then, the last three convolutional layers are connected to each another without any inter-value pooling or normalization layer. The last three convolutional filter sizes are all set as the critical 3×3 with a stride of 1 pixel, and the filter number are set as 384, 384 and 256 respectively. Each densely connected layer has the output size of 4096. Dropout in the first two densely connected layers is set as 0.6 during the pre-training phase.

The networks architecture remains the same during the pre-training and fine-tuning phases. Only the last densely connected layer and the classifier will be changed when fine-tuning the convolutional neural network. Most feature patterns of training images can be obtained through the convolutional layers and pooling layers with respect to the training set. The densely connected layers combine these features together and feed them into a softmax classifier. At the fine-tuning phase, the detectors in the convolutional layers are fine-tuned to cover the variation of new dataset.

3. Networks learning

3.1. Multimodal

To formulate the annotation learning task, the similarity function between an image annotation K and an input image x is denoted as $S(x, K)$. The learning goal is to learn a similarity function $S(\cdot, \cdot)$ that can always produce the similarity values satisfying the following inequality:

$$S(x, K_1) > S(x, K_2) \tag{1}$$

where K_1 and K_2 are both annotations, and the location of K_1 is on the top of the location of K_2 in the ranking list with respect to the image content.

The above discussion generally assumes that similarity learning is performed on uni-modal data. This paper aims to generalize it for multimodal data, where each image is represented by different kinds of low-level features including color, shape, or texture, and the similarity between an image annotation and an input image is computed by defining different kinds of distance measures including linear similarity, cosine similarity, and radial distance. Suppose n_f kinds of feature descriptors and n_s types of similarity measures construct $N = n_f \times n_s$ modalities, where each of which applies one kind of distance measure to compute the similarity between an image annotation and an input image with respect to one kind of low-level feature.

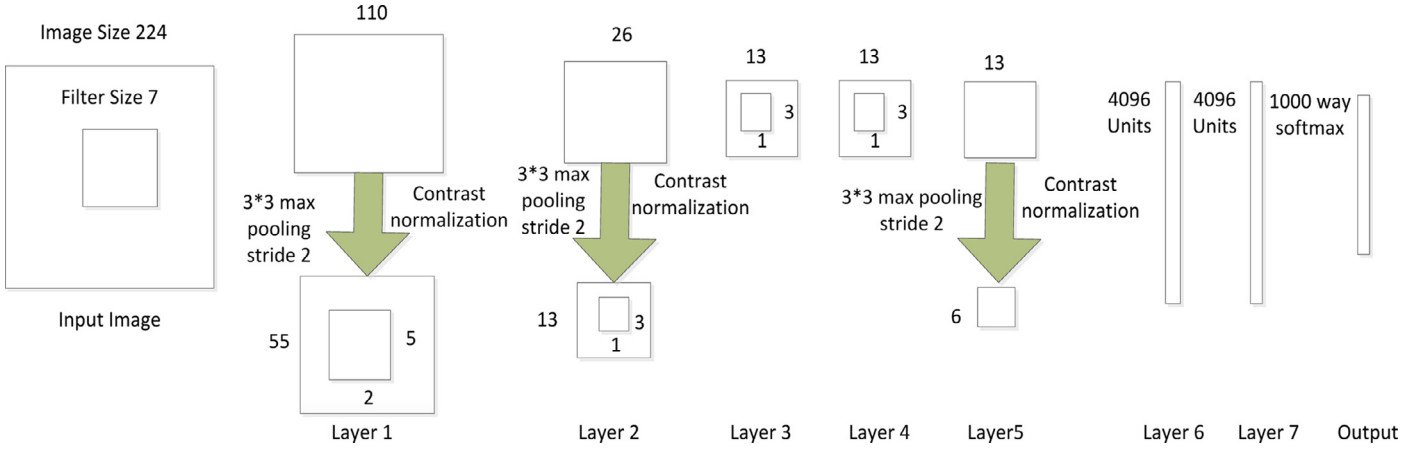


Fig. 2. The overall architecture of the convolutional neural networks model with 1000-way softmax layer.

The proposed multimodal similarity learning framework aims to deal with the following two issues: on the one hand, learning each optimal modality, namely learning each optimal similarity function $S(\cdot,\cdot)$ with respect to one specific low-level feature; on the other hand, identifying an optimal combination of these modalities to achieve the final optimal multimodal:

$$\begin{cases} S(x,K)=\sum_{j=1}^{N}\alpha_j S_j\left(x^j,K^j\right) \\ s.t.\ \sum\alpha_j=1 \text{and} \alpha_j\in[0,1] \end{cases} \tag{2}$$

where α_j is the combination weight for the jth modality, and x_j and K_j are the feature space within the jth modality.

3.2. Pre-training

Unlabeled data are utilized to learn abstract and discriminative intermediate representation for the image contents, and also provide a good initialization for the deep network. Specifically, the input layer and the first convolutional layer are combined to train the node weights W_1 with contrastive divergence. The conditional probability of the first convolutional layer nodes will be used as the input of the second convolutional layer:

$$p(K|x^j)=S(W_1,x^j) \tag{3}$$

where x^j is the jth feature vector and K is the label information. $S(\cdot,\cdot)$ is the similarity function, such as:

$$\begin{cases} S\left(W_1,x^j\right)=\frac{W_1^T x^j}{\|W_1\|\left\|x^j\right\|} & \text{Cosine function} \\ S\left(W_1,x^j\right)=W_1^T x^j & \text{Linear function} \\ S\left(W_1,x^j\right)=e^{-\frac{\left\|W_1-x^j\right\|^2}{2\sigma}} & \text{RBF function} \end{cases} \tag{4}$$

Then, the first convolutional layer and the second convolutional layer are combined to combine train the node weights W_2 in the similar way. This process is repeated for the remaining three convolutional layers and three densely connected layers.

3.3. Fine-tuning of individual modality

At the phase of fine-tuning of individual modality, the node weights are optimized with labeled data by backpropagating the derivatives of label assignment error. From the point of view of pattern recognition, the multi-label learning can be considered as a multi-task learning problem. Therefore, the whole assignment error of the proposed convolutional neural networks can be defined as the summation of each label assignment error.

The assignment error of the lth annotation is here taken as an example. The posterior probability of an image x with the jth feature x^j and the lth annotation K_l, namely the probability of an image x with the jth feature x^j owns the lth annotation K_l, can be expressed as:

$$p_{jl}=\frac{\exp\left(p\left(K_l|x^j\right)\right)}{\sum_{l=1}^{L}p\left(K_l|x^j\right)} \tag{5}$$

where L is the number of annotations.

Then, the KL-divergence between the predictions and the ground-truth probabilities is minimized. Suppose that there are multiple labels for each image and an annotation vector $y\in R^{1\times c}$ where $y_l=1$ denotes the presence of the lth annotation and $y_l=0$ denotes the absence of the lth annotation, the ground-truth probability can be achieved by normalizing y as $y/||y||_1$. If the ground-truth probability for the ith image x_i and the lth annotation K_l is defined as q_{il}, the cost function for the lth annotation assignment to be minimized is formulated as:

$$J_l=-\sum_{i=1}^{M}\sum_{l=1}^{L}q_{il}\log\left(p_{il}\right)-\sum_{i=1}^{M}\sum_{l=1}^{L}\left(1-q_{il}\right)\log\left(1-p_{il}\right) \tag{6}$$

The whole assignment error over all the annotations errors can be achieved as follows:

$$J=\sum_{l=1}^{L}J_l \tag{7}$$

Finally, the derivatives of J over the third densely connected parameters are computed and the backpropagation algorithm as in [12] is performed to update the parameters of other two densely connected network layers and five convolutional layers.

3.4. Fine-tuning of multimodal

For the proposed multimodal deep networks, another key task is to learn the optimal combinational weights $\alpha=(\alpha_1,\alpha_2,\ldots,\alpha_n,\ldots,\alpha_N)$ where α_n is set to be $1/N$ at the beginning of the learning task. The exponentiated gradient online learning algorithm [13] is here adopted to find the combinational weights sequentially. Specifically, the optimization of combinational weight α at the $(t+1)$th running, α_{t+1} can be formulated as:

$$\alpha_{t+1}=\underset{\alpha}{\operatorname{argmin}}\,KL(\alpha|\alpha_t)+\mu h_t(\alpha) \tag{8}$$

where $KL()$ is the KL-divergence and $h(\alpha)$ is a hinge loss:

$$\begin{cases} KL(\alpha|\alpha_t)=\sum_i\alpha_i\ln\left[\frac{\alpha_i}{(\alpha_t)_i}\right] \\ h_t(\alpha)=\max\left(0,\psi-\alpha^T S_t\right) \end{cases} \tag{9}$$

and the formula of S_t is described as:

$$S_t=\left[S_1(x,K^+)-S_1(x,K^-),\ldots,S_N(x,K^+)-S_N(x,K^-)\right]^T \tag{10}$$

where the annotation K^+ reveals the more content of the image x in contrast to the annotation K^-.

The first-order Taylor expansion of $h_t(\alpha)$ at α_t is performed to simplify the optimization, and thus the optimization formula (8) is formulated as:

$$\alpha_{t+1}=\underset{\alpha}{\operatorname{argmin}}\,KL(\alpha|\alpha_t)+\mu[h_t(\alpha_t)+\nabla h_t(\alpha_t)(\alpha-\alpha_t)] \tag{11}$$

It can be seen from the above formula that α will be updated whenever the current α fails to rank the order of K^+ and K^- with respect to the input unlabeled image x correctly at a sufficiently large margin.

The details of the proposed multimodal deep learning algorithm are summarized in the following algorithm 1.

Algorithm 1: Multimodal Deep Learning Algorithm

1: **INPUT** unlabelled data: U
2: **Initialize** weights: $\alpha_{1,j}=1/N, j=1,2,\ldots,N$
3: **Pretrain** N eight-layer deep networks with unlabelled data for each feature space by utilizing the convolutional neural networks as shown in Fig. 2
4: **for** $t=1,2,\ldots,$ M **do**
  Receive: (x_t,K^+,K^-)
  for $j=1,2,\ldots,N$ **do**
    Update the deep network parameters W_8 of the last layer by utilizing the formula (8)
    Adopt backpropagation to fine-tune the parameters of other deep network layers
  end for
  Compute: $S_{t,j}=S_j(x_t,K^+)-S_j(x_t,K^-), j=1,2,\ldots,N$
  Compute: $h_t(\alpha_t)=\max(0,\alpha_t{}^T S_t)$
  if $h_t(\alpha_t)>0$ **then**
    $\alpha_{t+1,j}=\frac{\alpha_{t,j}e^{-\mu\nabla h_t(\alpha_t)_j}}{\sum_{k=1}^{N}\alpha_{t,k}e^{-\mu\nabla h_t(\alpha_t)k}}, j=1,2,\ldots,N$
  end if
end for

4. Experiments

In this section, an extensive set of experiments will be conducted to evaluate the efficacy of the proposed multimodal deep learning

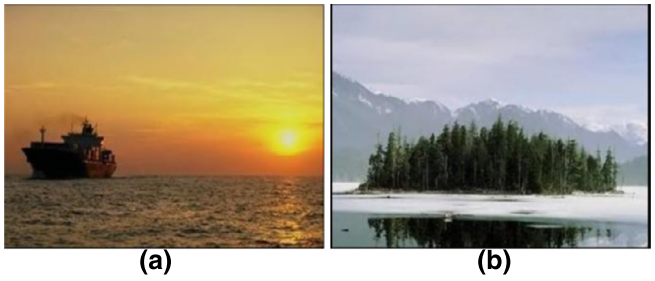


Fig. 3. Two images from the natural scene image dataset.

Fig. 4. Two images from the NUS-WIDE image dataset.

algorithm for labeling image with multi annotations. Specifically, the dataset chosen to evaluate the proposed algorithm is first described; then, typical visual features chosen to represent images and optimal parameters for achieving good performance are investigated; finally, the comparison experiments are performed between the proposed algorithm and other state-of-the-art algorithms.

4.1. Experimental settings

Three publicly available image datasets are adopted in our experiments, including natural scene image dataset as in [14], NUS-WIDE image dataset as in [15], and IAPRTC-12 image dataset as in [16]. The detail information of these three image datasets is described as follows:

Natural scene image dataset contains 2000 natural scene images. All of these images include the following 5 labels, such as desert, mountains, sea, sunset, and trees. More than 20% images have more than one label, and the average number of labels for each image is 1.3. Fig. 3 shows two images of this dataset. It can be seen that sunset and sea can be assigned to the Fig. 3(a), and mountains and trees can be assigned to the Fig. 3(b).

NUS-WIDE image dataset is crawled from the Flickr image dataset, which contains 30,000 images and 31 concepts. The images in this image dataset are exacted from the photo sharing web site Flickr.com. The concepts of these images are various, such as boats, cars, flags, horses, sky, sun, tower, plane and zebra. Fig. 4 shows two images of this dataset. It can be seen that sky and plane can be assigned to the Fig. 4(a), and sea and sunset can be assigned to the Fig. 4(b).

IAPRTC-12 image dataset is a collection of 20,000 images, where the total number of annotations is 291 and the average number of annotations for each image is 5.7. Fig. 5 shows two images of this dataset. It can be seen that brown, face, hair, man, and woman can be assigned to the Fig. 5(a), and boat, lake, sky, mountain, and tree can be assigned to the Fig. 5(b).

4.2. Visual features and similarity measures

The problem of feature selection has been an active research topic for many decades due to the fact that feature selection might have a great impact on final annotation results. In the current implementation, the following global and local features are extracted as the visual descriptors:

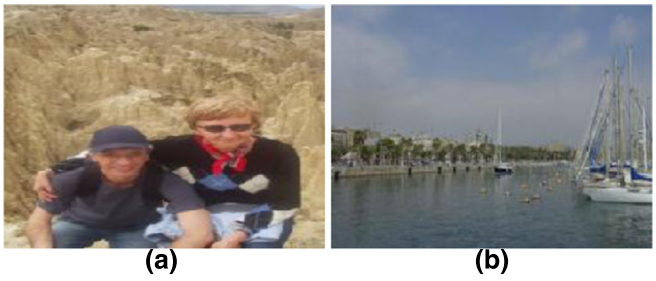


Fig. 5. Two images from the IAPRTC-12 image dataset.

- **Global features**: (1) 128-dimension HSV color histogram and 225-dimension LAB color moments, (2) 37-dimension edge direction histogram, (3) 36-dimension pyramid wavelet texture, (4) 59-dimension local binary pattern feature descriptor, and (5) 960-dimension GIST feature descriptor.
- **Local features:** Two different sampling methods and three different local descriptors are utilized to extract local texture features. Specifically, the dense sampling method and a Harris corner detector are first adopted as the patch-sampling methods; then, SIFT feature as in [17], CSIFT feature and RGBSIFT feature as in [18] are extracted to form a codebook of size 1000 using k-means clustering; next, a two-level spatial pyramid as in [19] is adopted to construct a 5000-dimensional vector for each image; finally, the TF-IDF weighing scheme is utilized to generate the final bag-of-visual-words. For all experiments, the feature vectors are all normalized to [0, 1].

For each query-annotation pair, three similarity measures are investigated as shown in the formula (4), where the margin parameter μ is chosen by using the cross validation scheme. Specifically, μ is set to be 0.18 for Cosine similarity measure, μ is set to be 1 for linear similarity measure, and σ is set to be 2, μ is set to be 0.18 for RBF similarity measure. Finally, there are a total of 18 modalities investigated to measure the similarity.

4.3. Compared algorithms

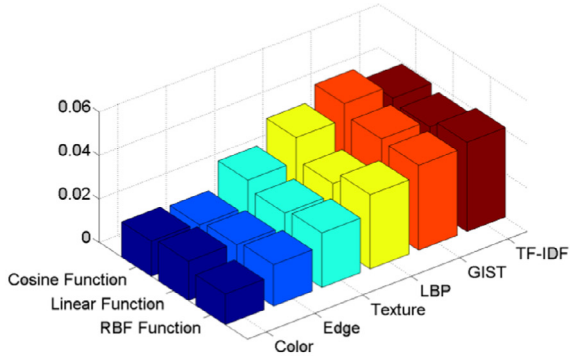
An experimental comparison is performed between four different image classification methods:

- Lazy learning based approach (LL) [14] where K-nearest neighbors in training set for each testing image are identified, and then maximum a posteriori principle is utilized to assign labels to unlabeled images based on the statistical information from the label sets of neighboring instances.
- Deep representations and codes based approach (DRC)[20] where a hierarchical model is utilized to learn the representations of images from the pixel level.
- Multi-task deep neural network based approach (MTDNN) [8] where the issue of image annotation is defined as a binary classification task.
- The proposed approach labels images by optimally integrating multiple deep neural networks pretrained with convolutional neural networks.

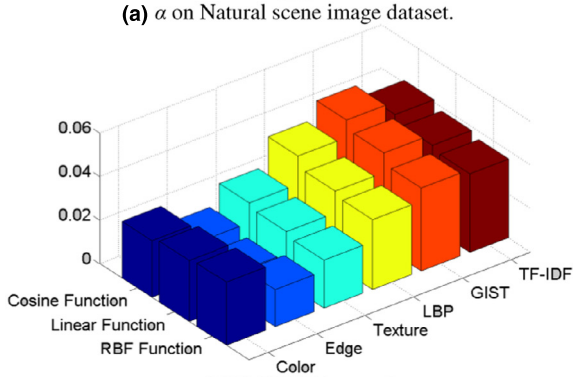
4.4. Evaluation measure

Hamming loss is here adopted as the evaluation criterion, which computes how many times an image-annotation pair is misclassified:

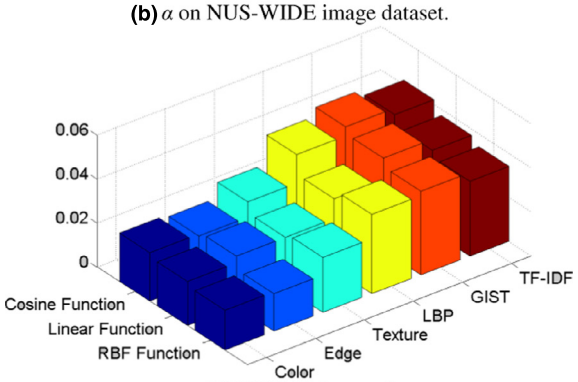
$$h_{\text{loss}}(h)=\frac{1}{V}\sum_{i=1}^{V}\frac{1}{B}|h(x_i)\Delta Y_i| \tag{12}$$



(a) α on Natural scene image dataset.



(b) α on NUS-WIDE image dataset.



(c) α on IAPRTC-12 image dataset.

Fig. 6. The combination weights of different modalities α learned by the proposed approach.

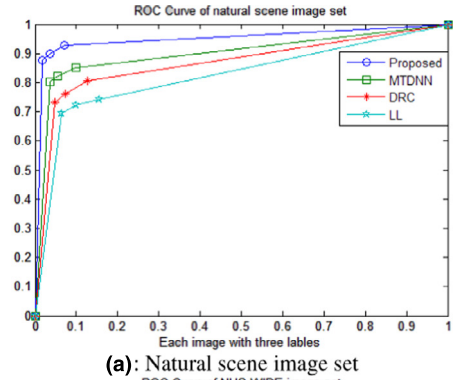
Table 1
Comparative results.

Method	Natural scene	NUS-WIDE	IAPRTC-12
LL [14]	0.227	0.0364	0.0545
DRC [20]	0.176	0.0321	0.0493
MTDNN [8]	0.147	0.0246	0.0342
Proposed	0.134	0.0219	0.0291

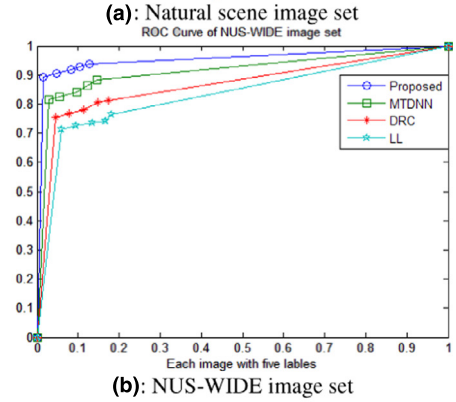
where V is the total number of tested images, and B is the total number of labels. Δ is the symmetric difference between two sets. It can be seen that the smaller the hamming loss of a method is, the better the performance of the method is.

4.5. Modality weights

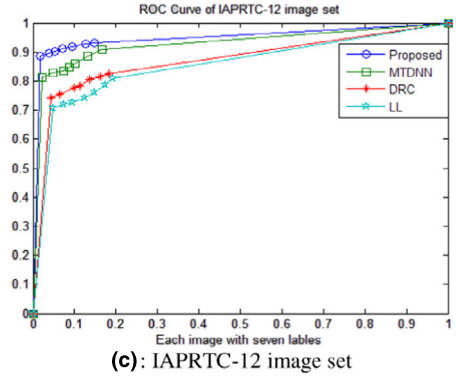
For the proposed image annotation algorithm, the combination weight of different modalities α has a very big impact on final system performance. Fig. 6 visualizes the results of weights α for diverse modalities learned on three different datasets.



(a): Natural scene image set



(b): NUS-WIDE image set



(c): IAPRTC-12 image set

Fig. 7. ROC curves of different methods.

It can be easily seen from the results that there is no significant difference between the ratios of different modalities, which means each modality makes more or less contributions for different scenarios. This is primarily because these three image datasets contain many pictures about natural scenes and diverse categories, which further validates the importance of finding the optimal combination of different modalities.

4.6. Performance comparison

The results of comparative experiments are illustrated in Table 1 in terms of hamming loss and Fig. 7 in terms of receiver operating characteristic curve (ROC curve).

It can be seen from the comparison results that the proposed deep structured semantic model considerably surpasses the other three approaches for all cases, which confirms the method from the following two aspects: (1) compared with other image annotation approaches, the proposed approach based on the deep learning algorithm can significantly improve the system performance by extracting useful feature information for building classifiers; (2) compared with other deep learning algorithms, the proposed deep neural network can further improve the system performance by learning parameters in feature representations.

Table 2
Future work of the proposed algorithm.

	Shortcomings of the proposed method	Future work of the proposed method
1	The number of feature dimensions achieving satisfactory system performance needs to be determinated by long training.	Choose optimally the number of feature dimensions to achieve satisfactory system performance of a given depth learning architecture.
2	Which mechanism is to be utilized to achieve satisfactory system performance needs to be determinated by long training.	Determinate optimally which mechanism can be utilized to enhance the robustness of a given depth learning architecture.

5. Conclusions

In this paper, we propose a novel image annotation framework which aims to optimally integrate multiple deep neural networks pretrained with convolutional neural networks. In particular, the proposed framework explores a unified two-stage learning scheme by (i) learning to fine-tune the parameters of deep neural network with respect to each individual modality, and (ii) learning to find the optimal combination of diverse modalities simultaneously in a coherent process. Experiments conducted on a variety of public datasets demonstrate the most competitive performance of the proposed algorithm compared with existing state-of-the-art algorithms.

Although more emphasis is put on Natural scene dataset, NUS-WIDE dataset and IAPRTC-12 dataset in this work, the proposed approach can be easily extended to deal with a variety of online media repositories, such as Flickr and Zooomr, as well as any other media databases in image annotation. Furthermore, further deep learning research is required to focus on the following aspects: one aspect is to determinate the number of feature dimensions to achieve satisfactory system performance for a particular neural network framework, the other aspect is which mechanism can be utilized to enhance a given depth learning architecture to improve its robustness, as shown in the following Table 2.

Acknowledgments

This work is supported by Postdoctoral Foundation of China under No. 2014M550297, Postdoctoral Foundation of Jiangsu Province under No. 1302087B, Education Reform Research and Practice Program of Jiangsu Province under No. JGZZ13_041, and Graduate Research and Innovation Program of Jiangsu under No. KYLX15_0854 and SJZZ15_0105.

References

[1] G. Hinton, S. Osindero, A fast learning algorithm for deep belief nets, Neural Comput. 18 (7) (2006) 1527–1554.
[2] H. Lee, R. Grosse, R. Ranganath, A. Ng, Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations, in: International Conference on Machine Learning, 2009, pp. 609–616.
[3] M Zeiler, D. Krishnan, G. Taylor, R. Fergus, Deconvolutional networks, in: IEEE Conference on Computer Vision and Pattern Recognition, 2010, pp. 2528–2535.
[4] M. Ranzato, J. Susskind, V. Mnih, G. Hinton, On deep generative models with applications to recognition, in: IEEE Conference on Computer Vision and Pattern Recognition, 2011, pp. 2857–2864.
[5] S. Zhong, Y. Liu, Y. Liu, Bilinear deep learning for image classification, in: ACM Conference on Multimedia, 2011, pp. 343–352.
[6] N. Srivastava, R. Salakhutdinov, Learning representations for multimodal data with deep belief nets, in: International Conference on Machine Learning, 2012, pp. 1–8.
[7] A. Mohamed, G. Dahl, G. Hinton, Acoustic modeling using deep belief networks, IEEE Trans Audio, Speech, and Language Process. 20 (1) (2012) 14–22.
[8] Y. Huang, W. Wang, L. Wang, T. Tan, Multi-task deep neural network for multi-label learning, in: IEEE Conference on Image Processing, 2013, pp. 2897–2900.
[9] Y. Gong, Y. Jia, T Leung, Deep convolutional ranking for multilabel image annotation, Comput. Res. Reposit. 21 (12) (2013) 1–9.
[10] K. Charalampous, A. Gasteratos, A tensor-based deep learning framework, Image Vision Comput. 32 (11) (2014) 916–929.
[11] C. Dong, C. Loy, K. He, X. Tang, Learning a deep convolutional network for image super-resolution, in: European Conference on Computer Vision, 2014, pp. 184–199.
[12] D. Rumelhart, G. Hinton, R.J. Williams, Neurocomputing: foundations of research, Massachusetts Institute of Technology Press, USA, 1988.
[13] N. Cesa-Bianchi, G. Lugosi, Prediction, learning, and games, Cambridge University Press, USA, 2006.
[14] M. Zhang, Z. Zhou, ML-KNN: a lazy learning approach to multi-label learning, Pattern Recognit. 40 (6) (2007) 2038–2048.
[15] T. Chua, J. Tang, R. Hong, H. Li, Z. Luo, Y. Zheng, NUS-WIDE: a real-world web image database from National University of Singapore, in: ACM Conference on Image and Video Retrieval, 2009, pp. 1–10.
[16] K. Yu, F. Lv, T. Huang, J. Wang, J. Yang, Y. Gong, Locality-constrained linear coding for image classification, in: IEEE Conference on Computer Vision and Pattern Recognition, 2010, pp. 3360–3367.
[17] D. Lowe, Distinctive image features from scale-invariant keypoints, Int. J. Comput. Vis. 60 (2) (2004) 91–110.
[18] K. Sande, T. Gevers, C. Snoek, Evaluating color descriptors for object and scene recognition, IEEE Trans. Pattern Anal. Mach. Intell. 32 (9) (2010) 1582–1596.
[19] S. Lazebnik, C. Schmid, J. Ponce, Beyond bags of features: spatial pyramid matching for recognizing natural scene categories, in: IEEE Conference on Computer Vision and Pattern Recognition, 2006, pp. 2169–2178.
[20] R. Kiros, C. Szepesvári, Deep representations and codes for image auto-annotation, in: IEEE Conference on Neural Information Processing Systems, 2012, pp. 917–925.