Image Annotation Via Deep Neural Network

Sun Chengjian, Songhao Zhu, Zhe Shi
School of Automation, Nanjing University of Posts and Telecommunications
Nanjing 210046, China
{suncjnjupt, njuptzsl, shizenjupt}@yeah.net

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 fune-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 the NUS-WIDE dataset 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 hand-held mobile devices. There is an urgent need to developing 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 organize and large scale image retrieval.

From a 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, 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 multi-layer 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. [2] presents a convolutional DBN to achieve better performance in image classification and speaker identification tasks by unsupervised learning of hierarchical feature representation. [3] proposes an unsupervised framework to derive hierarchical image representations to deal with the image denoising or object recognition tasks. [4] develops a generative deep learning model to achieve high-resolution images by merging a deep belief network with the gated Markov random field. [5] employs a bilinear deep belief network framework to deal with the image classification task by utilizing a bilinear discriminant strategy to simulatie the "initial guess" in human object recognition and effectively avoid falling into a bad local optimum simultaneously. [6] explores multimodal deep neural network to learn representations in image annotation and image retrieval tasks by fusing multiple sources with shared hidden representation. [7] completes the task of speech recognition by a deep belief network. [8] deals with the problem of assigning labels to images based on a multi-task deep neural network architecture, and [9] performs image annotation by combining convolutional architectures with approximate top-k ranking objectives. [10] presents an unsupervised deep learning framework to derive spatio-temporal features for human-robot interaction. [11] tackles 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 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 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.

2. Networks Architecture

The overall architecture of the proposed convolutional neural networks model is shown in Figure 1. 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.

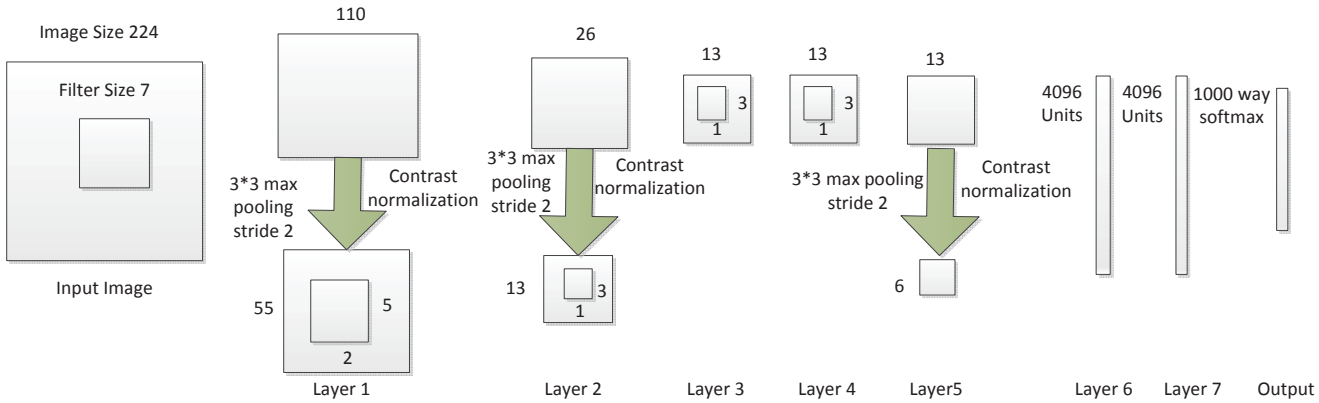

Figure 1: The overall architecture of the convolutional neural networks model with 1000-way softmax layer.

Before feeding the images to 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 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 output sizes 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-tune the convolutional neural network. Most features patterns of the training images can be obtained through the convolutional layers and pooling layers with respect to the training set, and 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 varies of the new dataset work and well on the new dataset.

3. Networks Learning

3.1. Multimodal

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

$$S(x, \Gamma_1) > S(x, \Gamma_2) \tag{1}$$

Where Γ_1 and Γ_2 are both annotations, and the location of Γ_1 is on the top of the location of Γ_2 in the ranking list with respect to the image content.

The above discussion generally assumes similarity learning is performed on uni-modal data. This paper aims to generalize it for multi-modal data, where each image is represented by different kinds of low-level features including color, shape, or texture, and the similarity 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 feature.

The proposed multimodal similarity learning scheme 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:

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

where α_j is the combination weight for the j^{th} modality, and x_j and Γ_j are the feature space within the j^{th} modality.

3.2. Pre-Training

Unlabeled data are utilized to learn abstract and discriminative intermediate representation for the objects in the images, and also provide a good initialization for the 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:

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

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

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

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 back-propagating the derivatives of label assignment error. From the of 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.

Take the l^{th} annotation assignment error as an example. The posterior probability of an image x with the j^{th} feature x^j and the l^{th} annotation Γ_l, namely the probability an image x with the j^{th} feature x^j owns the l^{th} annotation Γ_l, can be expressed using the following equation:

$$p_{jl}=\frac{\exp\left(p\left(\Gamma_l\left|x^j\right.\right)\right)}{\sum_{k=1}^{L}p\left(\Gamma_k\left|x^j\right.\right)} \tag{5}$$

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 that there is an annotation vector $y\in R^{1\times c}$ where y_l=1 denotes the presence of the l^{th} annotation and y_l=0 denotes the absence of the l^{th} annotation for an image, the ground-truth probability can be achieved by normalizing y as $y/\|y\|_1$. If the ground truth probability for an image x_i and annotation l is defined as q_{il}, the cost function for the l^{th} annotation assignment to be minimized is formulated as follows:

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

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 back-propagation algorithm[12] is performed to update the parameters of other two densely connected network layers and five convolutional layers.

3.4. Fine-Tuning of Multi Modality

For the proposed multi modality deep networks, another key task is to learn the optimal combinational weights α=(α_1, α_2, …, α_n, …, α_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 problem is formulated as:

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

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

$$\begin{cases}D_{KL}\left(u\left|v\right.\right)=\sum_i u_i\ln\left(\frac{u_i}{v_i}\right)\\ h_t\left(\alpha\right)=\max\left(0,\ \psi\text{-}\alpha^T S_t\right)\end{cases} \tag{9}$$

and the formula of S_t is described as:

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

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

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

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

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

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

Algorithm 1: Multimodal Deep Learning Algorithm

INPUT unlabelled data: U
Initialize weights: $\alpha_{1,j}$=1/N, j = 1, 2, . . . , N
Pretrain N eight-layer deep networks with unlabelled data for each feature space by utilizing as the convolutional neural networks shown in Figure 2
for t=1, 2, . . ., M **do**
  Receive: (x_t, Γ^+, Γ^-)
  for j =1, 2, . . . , N **do**
    Update the deep network parameters W_8 of last layer by utilizing the equation (8)
    Adopt Backpropagation to finetune the parameters of other deep network layers
  end for
  Compute: $S_{t,j}=S_j(x_t, \Gamma^+)- S_j(x_t, \Gamma^-), j$ = 1, 2, . . . , N
  Compute: $h_t(\alpha_t)=\max(0, \alpha_t^T S_t)$
  if $h_t(\alpha_t)>0$ **then**
    $\alpha_{t+1,j}=\frac{\alpha_{t,j}e^{-\mu N h_t(\alpha_t)_j}}{\sum_{k=1}^{N}\alpha_{t,k}e^{-\mu N h_t(\alpha_t)k}}$, j=1, 2, ..., 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 algorithm for labelling image with multi annotations. Specifically, the dataset chosen to evaluate the proposed algorithm is first described; then, typical visual features for representing images and optimal parameters for achieving good performance are investigate; 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[14], NUS-WIDE image dataset[15], and IAPRTC-12 image dataset[16].

In the current implementation, the following global features are extracted as the visual descriptors: (1) 128-dimension HSV color histogram and 225- imension 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, and the following local features are extracted as the visual descriptors: the dense sampling method and a Harris corner detector are first adopted as the patch-sampling methods; then, SIFT feature[17], CSIFT feature[18], and

RGBSIFT feature[18] are extracted to form a codebook of size 1000 using kmeans clustering; next, a two-level spatial pyramid[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 the range of [0, 1].

For each query-annotation pair, three similarity measures are investigated as shown in equation (4), where the parameter μ is chosen 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 36 modalities investigated to measure the similarity for image annotation.

an experimental comparison is performed between three image classification methods: Lazy learning based approach (LL)[14], Deep representations and codes based approach (DRC)[20], the proposed approach.

4.2. Performance Comparison

The results of comparative experiments using different methods for labeling images with milti-annotations are shown in Table 1, where the evaluation metric is the Hamming Loss . It can be seen from the results that the proposed deep structured semantic model considerably surpasses the other two approaches for all cases. That is, the proposed model is the best performer, beating other approaches by a statistically significant margin in Hamming Loss and validating the efficacy of learning effective similarity functions on multi-modal data.

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
Ours	0.134	0.0219	0.0291

5. Conclusions

In this paper, we propose a novel image annotation method 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 fune-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 scheme compared with other exsiting state-of-the-art algorithms.

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.

References

[1] G. Hinton and S. Osindero, “A fast learning algorithm for deep belief nets,” Neural Computation, 2006: .

[2] H. Lee, R. Grosse, R. Ranganath, and A. Ng, “Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations,” International Conference on Machine Learning, 2009: 609-616.

[3] M Zeiler, D. Krishnan, G. Taylor, and R. Fergus, “Deconvolutional networks,” IEEE Conference on Computer Vision and Pattern Recognition, 2010: 2528-2535.

[4] M. Ranzato, J. Susskind, V. Mnih, and G. Hinton, “On deep generative models with applications to recognition,” IEEE Conference on Computer Vision and Pattern Recognition, 2011: 2857-2864.

[5] S. Zhong, Y. Liu, and Yang Liu, “Bilinear deep learning for image classification,” ACM Conference on Multimedia, 2011: 343-352.

[6] N. Srivastava and R. Salakhutdinov, “Learning representations for multimodal data with deep belief nets,” International Conference on Machine Learning, 2012: 1-8.

[7] A. Mohamed, G. E. Dahl, and G. E. Hinton, “Acoustic modeling using deep belief networks,” IEEE Transactions on Audio, Speech, and Language Processing, 2012, 20(1): 14-22.

[8] Y. Huang, W. Wang, L. Wang, T. Tan, “Multi-Task Deep Neural Network For Multi-Label Learning,” International Conference on Image Processing, 2013: 2897-2900.

[9] Y. Gong, Y. Jia, Thomas Leung, et al, “Deep Convolutional Ranking for Multilabel Image Annotation, ” Computing Research Repository, 2013, 21(12): 1-9.

[10] K. Charalampous and A. Gasteratos, “A tensor-based deep learning framework,” Image Vision Comput, 2014, 32(11): 916-929.

[11] C. Dong, C. Loy, K. He, and X. Tang, “Learning a Deep Convolutional Network for Image Super-Resolution,” European Conference on Computer Vision , 2014:184-199.

[12] D. Rumelhart, G. Hinton, and R. J.Williams. Neurocomputing: foundations of research. Massachusetts Institute of Technology Press, USA, 1988.

[13] N. Cesa-Bianchi and G. Lugosi. Prediction, Learning, and Games. Cambridge University Press, USA, 2006.

[14] M. Zhang and Z. Zhou. ML-KNN: a lazy learning approach to multi-label learning. Pattern Recognition, 2007.

[15] T. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y. Zheng. Nus-wide: A real-world web image database from national university of singapore. In Proc. of ACM Conf. on Image and Video Retrieval, 2009.

[16] K. Yu, F. Lv, T. Huang, J. Wang, J. Yang, and Y. Gong. Locality-constrained linear coding for image classification. In CVPR, 2010: 3360-3367.

[17] D. Lowe. Distinctive image features from scale-invariant keypoints. IJCV, 2004.

[18] K. Sande, T. Gevers, and C. Snoek. Evaluating color descriptors for object and scene recognition. PAMI, 2010.

[19] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. CVPR, 2006.

[20] R. Kiros and C. Szepesvári. Deep Representations and Codes for Image Auto-Annotation. IEEE Conference on Neural Information Processing Systems, 2012: 917-925.