A Novel method for Target Detection

Xiangxiang Li[1, a], Songhao Zhu[2, b], Lingling Chen[2, c]

[1] *School of Automatic, Nanjing University of Posts and Telecommunications, Nanjing, 210046*

[2] *School of Automatic, Nanjing University of Posts and Telecommunications, Nanjing, 210046*

[3] *School of Automatic, Nanjing University of Posts and Telecommunications, Nanjing, 210046*

[a]*lixxnjupt@126.com*, [b]*zhush@njupt.edu.cn*, [c]*chenllnjupt@126.com*

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.

Keywords: Image Annotation, Deep Learning, Multi-Label, Multi-Modal

Introduction

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 [2-7]. 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] 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. [5] explores multimodal deep neural network to learn representations in image annotation and image retrieval tasks by fusing multiple sources with shared hidden representation. [6] completes the task of speech recognition by a deep belief network. [7] deals with the problem of assigning labels to images based on a multi-task deep neural network architecture. [8] 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.

NETWORKS LEARNING

● 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:

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

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:

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

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

- **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:

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

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

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

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.

- **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 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:

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

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 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 is performed to update the parameters of other two densely connected network layers and five convolutional layers.

- **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 is here adopted to find the combinational weights sequentially. Specifically, the optimization problem is formulated as follows:

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

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

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

and the formula of S_t is described as:

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

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:

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

Test results

The results of comparative experiments using different methods for labeling images with milti-annotations are shown in Table 1. 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 with respect to Hamming Loss.

Method	Natural scene image dataset[9]	NUS-WIDE image dataset[10]	IAPRTC-12 image dataset[11]
LL[9]	0.227	0.0364	0.0545
DRC[12]	0.176	0.0321	0.0493
Proposed	0.134	0.0219	0.0291

Acknowledgement

In this paper, the research is supported by Postdoctoral Foundation of China under No. 2014M550297, Postdoctoral Foundation of Jiangsu Province under No. 1302087B, Jiangsu College Graduate Research Innovation Projects under No. SJ22-0106 and KYLX_0820.

References

[1] G. Hinton and S. Osindero. A fast learning algorithm for deep belief nets. Neural Computation, 2006: 18(7): 1527-1554.

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

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

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

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

[8] 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.

[9] M. Zhang and Z. Zhou. ML-KNN: a lazy learning approach to multi-label learning. Pattern Recognition, 2007, 40 (6): 2038-2048.

[10] 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. ACM Conference on Image and Video Retrieval, 2009: 1-10.

[11] K. Yu, F. Lv, T. Huang, J. Wang, J. Yang, and Y. Gong. Locality-constrained linear coding for image classification. IEEE Conference on Computer Vision and Pattern Recognition, 2010: 3360-3367.

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