Multi-label Image Ranking based on Deep Convolutional Features

Guanghui Song[1,2,a], Xiaogang Jin[1,b,*], Genlang Chen[2,c] and Yan Nie[3,d]

[1]College of Computer Science, Zhejiang University, Hangzhou, China

[2]Ningbo Institute of Technology, Zhejiang University, Ningbo, China

[3]College of Science and Technology, Ningbo University, Ningbo, China

[a]songgh@nit.zju.edu.cn, [b]xiaogangj@cise.zju.edu.cn, [c]cgl@nit.zju.edu.cn, [d]nieyan@nbu.edu.cn

Keywords: feature learning, deep convolutional neural network, multi-label ranking

Abstract. Multi-label image ranking has many important applications in the real world, and it includes two core issues: image feature extraction approach and multi-label ranking algorithm. The existing works are mainly focused on the improvement of multi-label ranking algorithm based on the conventional visual features. Recently, image features extracted from the deep convolutional neural network have achieved impressive performance for a variety of vision tasks. Using these deep features as image representations have gained more and more attention on multi-label ranking problem. In this study, we evaluate the performance of the deep features using two baseline multi-label ranking algorithms. First, the deep convolutional neural network model pre-trained on ImageNet is fine-tuned to the target dataset. Second, the global deep features of raw image are extracted from the fine-tuned model and serve as the input data of ranking algorithms. Finally, experiments using the Tasmania Coral Point Count dataset demonstrate that the deep features enhance the expression ability in comparison with that of conventional visual features, and they can effectively improve multi-label ranking performance.

Introduction

Multi-label images have been widely used in many applications, such as image retrieval, semantic annotation, and other fields, because of the important practical significance [9]. Most real-world images contain more than one object of different categories. Using multi-label method to annotate the images can fully describe the original image content in comparison with that of single-label method. And on this basis, label ranking can further reflect the semantic information of multi-label images [3]. Multi-label image ranking problem is a very challenging task, and it has received considerable attention in computer vision recently. This problem consists of two parts: on the one hand, the relevant labels are assigned to each image automatically, namely multi-label classification; on the other hand, a proper ranking is predicted for the relevant labels, namely label ranking [2]. The goal of multi-label ranking is to learn a mapping from multiple instances of each image to the ranking of the corresponding labels. Figure 1 shows the single-label and multi-label images from different datasets. We can see that the description of image content is incomplete using single-label method. However, the important degree of multiple objects in an image can be obtained using multi-label ranking method.

To solve multi-label image ranking problem, image features extraction approach and multi-label ranking algorithm are two important steps. Both of them have great influence on the performance of multi-label ranking [1]. In previous studies, many methods are proposed to address this challenging task from above two aspects. Most of them are mainly focused on the improvement of multi-label learning algorithm based on the conventional visual features that serve as image representation [6,7]. Recently, the image features extracted from the deep convolutional neural network (CNN) have achieved impressive performance on single-label image classification, which is also known as the deep features [12]. These deep features can produce a rich representation of the raw image by embedding them to a fixed-length vector, such that this representation can be used for a variety of vision tasks [10,11]. Especially in some applications for generating image description, the deep

© 2016. The authors - Published by Atlantis Press

features based on object bounding boxes and multiple instance learning approach are adopted, and they have achieved good results. But, in essence, they still are single-label classification methods based on multiple local region of each image [13,14]. It is not yet clear whether the global deep features extracted from the whole raw image can improve multi-label ranking performance.

We evaluate in this study the performance of the deep features on multi-label ranking using two multi-label ranking algorithms. The deep CNN model pre-trained on ImageNet is fine-tuned toward our target dataset, and the global deep features are extracted from the fine-tuned model. Then, the deep features serve as the raw image representation and are used as the input data of label ranking algorithms. The experiments using Tasmania Coral Point Count (CPC) dataset demonstrate that the deep features enhance the expression ability in comparison with that of conventional visual features, and they can effectively improve multi-label ranking performance on some particular dataset.

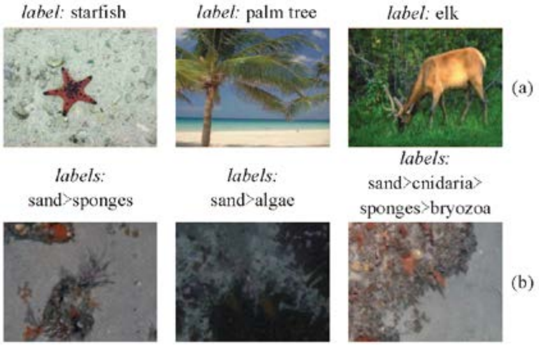


Fig.1: (a) Some examples of single-label images from the Caltech-256 dataset. (b) Some examples of multi-label images from the Tasmania CPC dataset. Multi-label ranking can adequately describe the semantic information of image content.

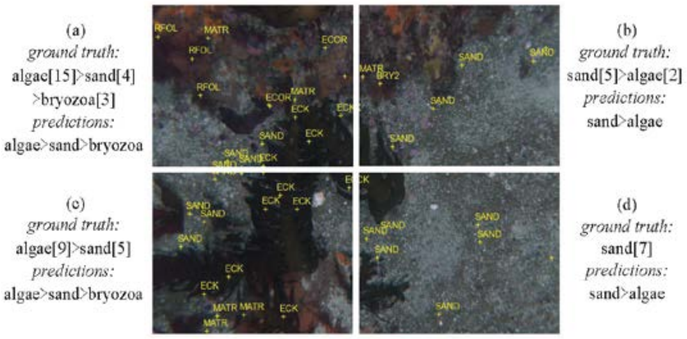


Fig.2: The numbers in square brackets represent the number of sampling point per category, and they are used as a basis for multi-label ranking. The comparison of prediction labels with ground truth labels based on deep features and CLR algorithm using the Tasmania CPC dataset.

Related Work

Related works can be broadly divided into the following two aspects: (i) deep convolution features, (ii) multi-label ranking algorithms. The network structure of the deep CNN proposed by Krizhevsky et al. is widely used because of its better performance [12]. The deep CNN model is composed of five convolutional layers (conv1 to conv5), two fully connected layers (fc6 and fc7), and a softmax output layer. The generalization ability of the features extracted from the deep CNN is very excellent [10,11]. These feature representations can deal with the different tasks and datasets that have a degree of bias with respect to ImageNet. In the most existing methods, the deep features extracted from the layer fc6 in a deep network serve as image representations [10]. As the expression ability of the layer fc6 is the strongest, the performance is the best compared with that of other layers. Recently, some studies have examined multi-label classification methods based on the deep features [13,14,15], which can be summarized as two ideas. One way is to convert multi-label problem into a single-label problem using multiple local bounding boxes and multiple instance learning method. Another way is to extract the global features of raw image and perform classification task by combining with multi-label learning algorithm. For example, a deep CNN model using multi-label loss function is proposed to perform the top-k ranking [8].

The existing multi-label ranking algorithms can be roughly divided into two categories: problem transformation methods (PTMs) and algorithm adaptation methods (AAMs) [1]. The PTMs are algorithm independent, and they transform the multi-label learning task into one or more single-label learning tasks. The calibrated label ranking (CLR) algorithm is a typical representative of the PTMs. It is based on the standard label ranking algorithm and adds an additional virtual label to the relevant category label set for each sample image [5,6]. This virtual label is used as a natural partition between the relevant and irrelevant labels. The AAMs extend specific single-label learning algorithms in order to handle multi-label data directly. In the AAMs, multi-label k nearest neighbor (ML-kNN) is usually used as a baseline algorithm, and it is derived from the conventional k nearest neighbor algorithm [4,7]. The ML-kNN method is simple, the time complexity is low and the performance is better. In this paper, we use the above two representative algorithms to evaluate the performance of the deep features.

Feature Extraction and Multi-label Ranking

The performance of multi-label ranking is determined by image features representation to a large extent. We compare the performance of the deep features with that of the conventional visual features using two representative multi-label ranking algorithms in our experiments. The details of image features and typical algorithm for experiments from each category are described as follows.

CIELUV Features. The CIELUV features are common benchmark features in multi-label scene classification [9]. In detail, a color image is first converted into the CIE L*U*V* space. Then, the image is divided into 80 blocks using a 10×8 grid. In each block, the mean and variance of each band are computed, corresponding to a low-resolution image and to computationally inexpensive texture features, respectively. Finally, the image is represented by a feature vector of 480 dimensions.

Multi-feature Fusion. Multi-feature fusion approach is used in [16] and achieves state-of-the-art performance for image retrieval and annotation firstly. Therefore, we use it as a contrast feature in our experiments. It consists of four parts as follows: GIST, SIFT, HOG and Color. Using the above four approaches, a total of 36,472-dimensional features are extracted and fused from each image. Learning is very expensive using such a large dimension. Therefore, a kernel PCA (KPCA) separately is performed on each feature extraction approach to reduce the dimensionality to 500, and then all of the feature vectors are concatenated to form a 4500-dimensional global image feature vector and perform different multi-label ranking algorithms on it.

Deep Convolutional Features. To obtain the deep convolutional features of raw images, the deep CNN model is used as a feature extractor not a classifier. Each image is pre-processed by subtracting the image mean of ImageNet dataset, and fed into the first convolutional layer of the CNN. A base deep CNN model is pre-trained on the ImageNet dataset, and then we take a pre-trained deep CNN, modify the output category numbers, initialize the weights with random values on the last softmax layer, keep the remaining layers with no change, fine-tune the network toward the target dataset. The deep features extracted from the layer fc6 in the fine-tuned network model serve as image representation, which is a 4096-dimensional feature vector.

Calibrated Label Ranking Algorithm. Calibrated label ranking algorithm is a problem transformation method. For the formal description of CLR algorithm, $L = \{\lambda_j : j = 1 \cdots m\}$ is used to denote the finite set of labels in a multi-label ranking task and $D = \{(x_i, Y_i), i = 1 \cdots n\}$ to denote a set of multi-label training samples, where x_i is the feature vector of raw image and $Y_i \subseteq L$ is the set of labels of the i-th sample.

The ranking by pairwise comparison (RPC) algorithm transforms a multi-label dataset into $m(m-1)/2$ binary-label dataset, each pair of labels is expressed as (λ_i, λ_j),$1 \le i < j \le m$ [17]. The samples of D are annotated by at least one of the two corresponding labels, but not both. A binary classifier learns to discriminate between the two labels, and it is trained on each of these binary-label dataset. Given a new instance, all binary classifiers are invoked, and a ranking is obtained by counting the votes received by each label. The CLR algorithm [5] extends RPC by introducing an additional virtual label λ_0, which acts as a natural breaking point between relevant and irrelevant labels. All relevant labels are preferred to λ_0, which in turn is preferred to all irrelevant labels. Thus, a calibrated ranking

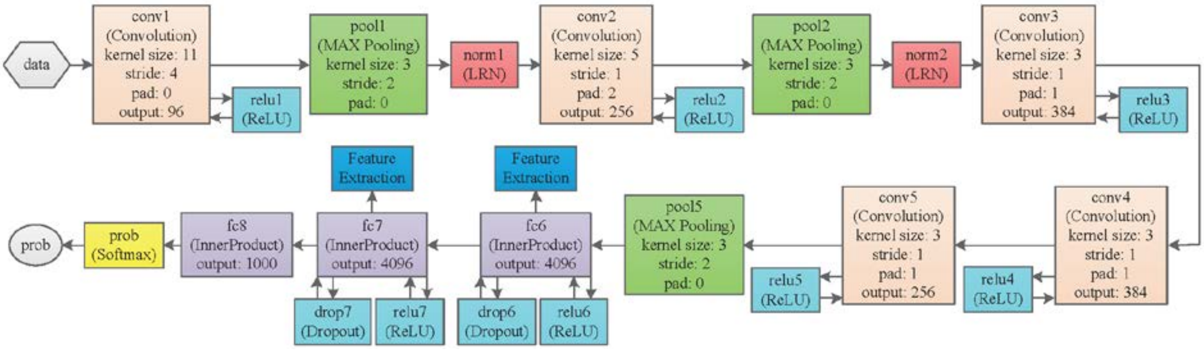
$$\lambda_1 \succ \cdots \succ \lambda_j \succ \lambda_0 \succ \lambda_{j+1} \succ \cdots \succ \lambda_m \tag{1}$$

generates a bipartite partition (P expresses relevant labels and N expresses irrelevant labels) into

$$P = \{\lambda_1 \cdots \lambda_j\} \; and \; N = \{\lambda_{j+1} \cdots \lambda_m\} \tag{2}$$

in a straightforward way. In this way, the CLR algorithm solves the multi-label ranking problem.

Multi-label k Nearest Neighbor. The ML-kNN algorithm is an algorithm adaptation method, and it is also a lazy learning algorithm. ML- kNN is derived from the conventional k nearest neighbor algorithm. First, for each test instance, its k nearest neighbor instances in the training set is obtained. Then, according to statistical information gained from the label sets of these neighboring instances, i.e. the number of neighboring instances belonging to each possible category, maximum a posteriori principle is utilized to determine the label set for the test instance. Finally, we can rank the scores and annotate each image with the top-k labels.



data
conv1 (Convolution) kernel size: 11 stride: 4 pad: 0 output: 96
relu1 (ReLU)
pool1 (MAX Pooling) kernel size: 3 stride: 2 pad: 0
norm1 (LRN)
conv2 (Convolution) kernel size: 5 stride: 1 pad: 2 output: 256
relu2 (ReLU)
pool2 (MAX Pooling) kernel size: 3 stride: 2 pad: 0
norm2 (LRN)
conv3 (Convolution) kernel size: 3 stride: 1 pad: 1 output: 384
relu3 (ReLU)
Feature Extraction
Feature Extraction
prob
prob (Softmax)
fc8 (InnerProduct) output: 1000
fc7 (InnerProduct) output: 4096
fc6 (InnerProduct) output: 4096
pool5 (MAX Pooling) kernel size: 3 stride: 2 pad: 0
conv5 (Convolution) kernel size: 3 stride: 1 pad: 1 output: 256
conv4 (Convolution) kernel size: 3 stride: 1 pad: 1 output: 384
relu5 (ReLU)
relu4 (ReLU)
drop7 (Dropout)
relu7 (ReLU)
drop6 (Dropout)
relu6 (ReLU)

Fig.3: Network structure of the deep CNN model, which is composed of five convolutional layers (conv1 to conv5), two fully connected layers (fc6 and fc7), and a softmax output layer [12]. The layer fc6 is commonly used to extract the deep features.

Experiments

Dataset, Preprocessing and Evaluation Measures. The Tasmania CPC dataset includes 1,258 benthos images captured by an autonomous underwater vehicle (AUV), and it is similar to the natural scene dataset that is used for automatic semantic annotation of images [9]. Each image contains 50 selected randomly points labeled by experts, the image size is 1360×1024. A wide range of category labels is used that indicate biological species, abiotic elements, and types of unknown data. The precise details of the labeling methodology can be found in [18]. A total of 62,900 data points from 1,258 images are included in the data set. However, the labeled categories of some sampling points are relatively vague, such as “*biota*” and “*unknown*”. The similar sampling points labeled with the vague category are discarded.

Finally, we extract 8 categories and a total of 37,636 samples. To obtain the appropriate image size and more image samples, each image of the original dataset is divided into four equal parts. The size of each part is 680×512, a total of 5,032 images are obtained. The image size after segmentation is more suitable as the input data of the deep CNN model. The category of each sampling point is viewed as an image label, each image with multiple sampling points is considered as a multi-label image sample. Each image is labeled with 1 to 5 keywords, on average, each image includes 3.2 labels. Out of the 5,032 images, 4,000 images are used for model training, and the other 1,032 images are used for testing. Figure 3 shows a sample image with multi-label ranking, the numbers in square brackets represent the numbers of sampling point per category, and they are used as a basis for multi-label image ranking.

We adopt four evaluation measures in our experiments. They are used to evaluate the performance of different visual features and algorithms [4], the concrete content is as follows: Hamming Loss, One-error, Ranking Loss and Average Precision. In the above four evaluation criteria, for the first three, a smaller value is better, the optimal value is 0; for the last one, bigger is better, the optimal value is 1.

Setup and Parameters. The multi-label ranking algorithms are implemented based on the Mulan software library. The Caffe framework is deployed to train and extract the deep features on a single

NVIDIA Tesla K40c card. We use the CaffeNet as the base deep CNN model, and it is pre-trained on the ImageNet classification task. The base deep CNN model is fine-tuned toward the training set of the Tasmania CPC dataset. The initial learning rate of the network is set to 0.0001, and it is decreased by a factor of 10 every 2K iterations, for a total of 5K iterations. The deep features extracted from the layer fc7 in the fine-tuned model serve as the final image visual features. For ML-kNN algorithm, the number of the nearest neighbors k is set to be the moderate value of 15, the smoothing parameter is set to be 1. For CRL algorithm, the default setting J48 is used as its base classifier, and the final ranking via soft voting is obtained on the base classifiers' results. The experimental results are compared with the average of five data splits.

Table 1. Comparison of Multi-label Ranking Performance based on Different Visual Features and Algorithms on the Tasmania CPC Dataset.

Algorithm	CLR			ML-*k*NN		
Features	CIELUV Features	Multi-feature Fusion	Deep Features	CIELUV Features	Multi-feature Fusion	Deep Features
Hamming Loss	0.185±0.021	0.153±0.019	0.118±0.017	0.158±0.087	0.136±0.062	**0.098±0.047**
One-error	0.334±0.051	0.295±0.028	0.238±0.038	0.314±0.022	0.264±0.035	**0.202±0.041**
Ranking Loss	0.127±0.022	0.112±0.012	0.106±0.019	0.120±0.034	0.095±0.015	**0.094±0.031**
Average Precision	0.771±0.057	0.839±0.026	0.853±0.041	0.804±0.036	0.852±0.051	**0.875±0.050**

Experimental Results. As shown in the comparison in Table 1, the performance based on the deep features and ML-kNN algorithms is the best. And the performance based on the deep features is superior to that based on the conventional visual features using different algorithms. The experimental results show that the expression ability of deep features is strong and stable for multi-label image ranking on the Tasmania CPC dataset. Finally, we show some examples of multi-label image ranking in Figure 3. In these images, the predicted labels are more comprehensive than the ground truth labels. Therefore, our work has a certain of practical meaningful.

Conclusion

In this study, we evaluate the performance of the deep features on multi-label ranking problem. These features are extracted from the layer fc7 in the fine-tuned deep CNN model. The performance of the deep features is superior to that of the conventional visual features based on the two representative multi-label ranking algorithms. Our experiments at least show that the global deep features based on the whole raw image are suitable for multi-label image ranking, and the performance of different evaluation measures is the best on the Tasmania CPC dataset. Furthermore, our work appears to evaluate the performance of the deep features on other multi-label image datasets in order to determine the generalization ability of the deep features. We will also compare different deep CNN models to extract better deep features.

Acknowledgements

*Corresponding author. This work is supported by the National Natural Science Foundation of China (Grant No.61379074), the Zhejiang Provincial Natural Science Foundation of China (Grant No.LZ12F02003), the Zhejiang Provincial Natural Science Foundation of China (Grant No.LY15F020035) and the Ningbo Municipal Natural Science Foundation of China (Grant No. 2015A610125).

References

[1] Tsoumakas G, Katakis I, and Vlahavas I, Mining Multi-label Data, Data Mining and Knowledge Discovery Handbook, 2010, pp. 667-685.

[2] Geng X and Luo L, Multi-label Ranking with Inconsistent Rankers, Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference onIEEE, 2014, pp. 3742-3747.

[3] Wei Y C, Xia W, Huang J S, et al, CNN: Single-label to Multi-label, Computer Vision and Pattern Recognition (CVPR),2014 IEEE Conference onIEEE, 2014.

[4] Zhang M L and Zhou Z H, ML-KNN: A lazy learning approach to multi-label learning, Pattern Recognition, 2007, 40(7), pp. 2038-2048.

[5] Fürnkranz J, Hüllermeier E, Meníca E L, et al, Multilabel classification via calibrated label ranking, Machine Learning, 2008, 73(2), pp. 133-153.

[6] Elisseeff A and Weston J, A kernel method for multi-labelled classification, Neural Information processing Systems, Nips, 2001, pp. 681-687.

[7] Makadia A, Pavlovic V, and Kumar S, A New Baseline for Image Annotation, Lecture Notes in Computer Science, 2008, pp. 316-329.

[8] Gong Y C, Jia Y Q, Leung T K, et al, Deep Convolutional Ranking for Multilabel Image Annotation, Computer Vision and Pattern Recognition, 2014 IEEE Conference onIEEE, 2014.

[9] Boutell M R, Luo J, Shen X, et al, Learning multi-label scene classification, Pattern Recognition, 2004, 37(9) , pp. 1757-1771.

[10] Donahue J, Jia Y Q, Vinyals O, et al, DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition, University of California Berkeley Brigham Young University, 2013, pp. 647-655.

[11] Zeiler M D and Fergus R, Visualizing and Understanding Convolutional Networks, Computer Vision - ECCV 2014, Springer International Publishing, 2014, pp. 818-833.

[12] Krizhevsky A, Sutskever I, Hinton G E, ImageNet Classification with Deep Convolutional Neural Networks, Advances in Neural Information Processing Systems, 2012.

[13] Girshick R, Donahue J, Darrell T, et al, Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation, Computer Vision and Pattern Recognition, 2014 IEEE Conference onIEEE, 2014, pp. 580-587.

[14] Iandola F, Deng L, Dollár P, et al, From Captions to Visual Concepts and Back, Eprint Arxiv, 2014.

[15] Vinyals O, Toshev A, Bengio S, et al, Show and Tell: A Neural Image Caption Generator, Eprint Arxiv, 2014.

[16] Gong Y, Ke Q, Isard M, et al, A Multi-View Embedding Space for Modeling Internet Images, Tags, and Their Semantics, International Journal of Computer Vision, 2014, 106(2), pp. 210-233.

[17] Frnkranz J, Hllermeier E, Preference Learning and Ranking by Pairwise Comparison, Preference Learning, 2011, pp. 65-82.

[18] Meyer L, Hill N,Walsh P,et al, Methods for the processing of AUV digital imagery from South Eastern Tasmania, Report, 2011.