2016 3rd International Conference on Informative and Cybernetics for Computational Social Systems (ICCSS)

Scene image classification method based on Alex-Net model

Jing Sun, Xibiao Cai
School of Electronics and Information Engineering
Liaoning University of Technology
Jinzhou, Liaoning
694600868@qq.com, xbc1111@126.com

Fuming Sun, Jianguo Zhang
School of Electronics and Information Engineering
Liaoning University of Technology
Jinzhou, Liaoning
sunwenfriend@hotmail.com, 1192102470@qq.com

***Abstract*—Deep convolutional neural network (DCNN) is a powerful method of learning image features with more discriminative and has been studied deeply and applied widely in the field of computer vision and pattern recognition. In order to further explore the superior performance of DCNN and improve the accuracy of the scene image classification, this paper presents a novel algorithm of scene classification, which fully learning the deep characteristics of the images based on the classical Alex-Net model and support vector machine. In the first place, we use the Alex-Net model learning scene image features and extract the last layer with 4096 neurons of the Alex-Net model as the image features in this method; Then, we use the Lib-SVM training model for scene image classification and compare with classification method based on the regression model; Finally, we carried out the experiments on two common datasets in this paper. The experimental results have shown that DCNN can extract the image features effectively. Meanwhile, the trained scene model also has stronger generalization performance and achieves the state-of-the-art classification accuracy.**

Keywords—deep convolutional neural network; scene classification; feature learning; training model

I. INTRODUCTION

With the increasing development of the Internet and digital photography equipment, the image has become an integral part of real life, and how to quickly and effectively retrieve and identify the image information has become a hot topic which needs to be solved urgently at present. Scene image classification is the key to solve the pattern recognition based on image. As one of the important contents of machine learning and pattern recognition, scene image classification has captured increasing attention. The purpose of scene classification is to assign one or more semantic categories to one or a set of images. That is to say, under the condition of giving a set of semantic categories, such as semantic concept: mountain, coast, highway, forest, et al., the images are automatically classified, as shown in Fig. 1.

Traditionally, for the problem of scene image classification, people mainly study the semantic modeling technologies [1-2], and the Bag of Visual Word (BOW) model [3-4] is one of the most widely used methods. Firstly, semantic modeling is carried out on the scene in this method; Secondly, the probability distribution of the semantic subject is learned; Finally, the scene is classified according to the probability distribution. Bosch et al. [5] applied the method of PLSA to achieve the 8 classes scene classification; Fu et al. [6] proposed a progressive approach to generate BOW on the basis of PLSA model, this method has achieved better results in the case of small data sets; Bai et al. [7] used a way of combing the top-down and bottom-up information for visual word weighting, and it have achieved better effect; At the same time, some scholars introduced the formal concept analysis (FCA) model into the scene classification, which has improved the classification performance for some concept.

Fig. 1. Four scene images

However, Hinton et al. [8] proposed deep learning in 2006, and it has become a popular research subject in the field of computer vision and pattern recognition. The basic principle of deep learning is to use neural networks with multiple hidden layers, to form a more abstract high-level representation through the combination of low-level features, and to find the distributed feature representation of the data. Because of its excellent performance, deep learning has aroused widespread concern and in-depth research, and it has gained a great success in computer vision, speech recognition, natural language processing and the other fields.

As one of the representative model of deep learning, one of the most obvious advantages of the deep convolutional neural network model is no longer need to manually design the local features compared with other traditional supervised learning

978-1-5090-3367-6/16/$31.00 ©2016 IEEE

methods, instead of automatically extracting the local features with translation invariance in the stacking convolution and down-sampled operations. Hence, the higher labor caused by designing local features and training time can be greatly reduced. The other advantage of DCNN is that employed a recently-developed regularization method called Dropout which can be prevented the over-fitting.

In this paper, a scene image classification method based on the deep convolutional neural network is proposed. The specific contributions of this paper are as follows: at first, the Alex-Net model is used to learn scene image features, and the last layer with 4096 neurons as the image feature representation. Meanwhile, the 12 scene models are constructed by using the Lib-SVM classifier. Then, the test images are classified by the training scene model and Lib-SVM classifier. Finally, experimental results on two real-world data sets demonstrate that the proposed classification model in this paper can achieve superior results.

II. Convolutional Neural Network Model

CNN is a special neural network model proposed by LeCun, which is used for document image recognition, and has made a great breakthrough in image classification and retrieval [9], target detection [10] and so on. Deep CNN reduces the dimensions of image by increasing the number of hidden layer (convolutional layer and sampling layer), and extracts the sparse image features in low-dimensional space. Because of their weights sharing, CNN has much fewer neurons and parameters, so it is easier to train.

Alex-Net model is the most representative model of CNN, which has three obvious advantages, i.e., superior performance, less training parameters and strong robustness. It is inspired by the principles of human vision, while mimicking the way of dual channel visual transmission, and learning image features by using two channels, as depicted in Fig. 2. Alex-Net is the deep convolutional neural network model with multiple hidden layers, includes an input layer, five convolutional layers, three pooling layers, three fully-connected layers, and an output layer with 1000 output class labels. In the whole model, the learning feature is carried out in the convolutional layer with two channels that are studied separately, and they are crossed only in the third feature extraction layer. The first fully-connected layer is cross-mixing for the features of two groups respectively, and then the next fully-connected layer repeat until the last fully-connected layer combine with the features of two groups together to get a 4096-dimensional feature vector.

The Alex-Net model needs to adopt appropriate methods to make training faster and prevent over-fitting due to the complex structure, many training parameters and a large amount of training data. So Alex-Net building a model by directly using the Rectified Linear Unit (ReLU) nonlinearity in the data structures to make the initialized method is more consistent with the theory, and begin to train the network directly from the starting point to enhance training speed; Local response normalization (LRN) can be performed in a process called "near-suppression" operation, which can effectively improve the generalization performance of the model by normalizing the local input regions. As for data, the data augmentation is completed in the case of label-preserving transformations: the original image is transformed by adding multiple of the data, while enhancing the intensities of the RGB channels in training images, so as to prevent the phenomenon of over-fitting. To improve the robustness in the fully-connected layers we employed a recently-developed regularization method called Dropout, so the learning of hidden layer is not dependent on the features of the upper layer.

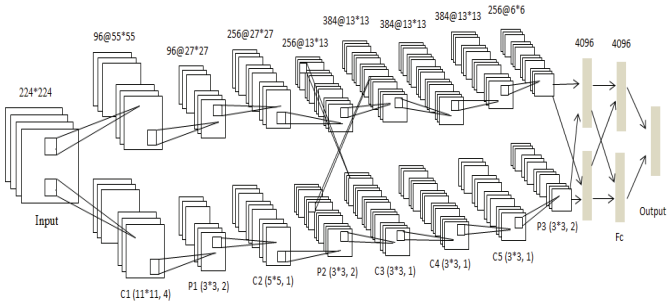

Fig. 2. The framework of Alex-Net model

III. The Specific Process of Deep Learning

We use the Alex-Net model learning scene image features, and the scene image features are obtained through a series of transformation during the training phase, such as convolution, max-pooling, etc. The specific process of feature extraction is shown in Fig. 3.

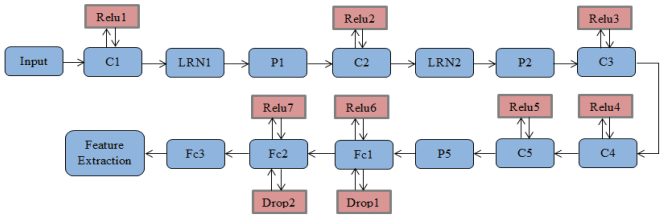

Fig. 3. The flow chart of feature extraction based on CNN

A. Input layer

The size of the raw scene images is 256 × 256. In the first place, the input images are selected: we do this by extracting random 224 × 224 patches from the four corners and the center location of the 256 × 256 images respectively. Then, the horizontal reflections are carried out on these five 224 × 224 patches, hence ten patches in all. Finally, each image of ten patches is selected to be input to the training network, which is used as the input images.

B. Convolutional layer (C1)

C1 is the first convolutional layer of the feature extraction, and we obtain 96 feature maps of size 55×55. In fact, it is obtained by utilizing the convolutional kernel of size 11×11. Because the size of the receptive field of the neuron is determined by the size of the convolutional kernel, the ideal

size of the convolutional kernel is to extract the effective local features in the range of convolutional kernel with representation ability. Therefore, the proper setting of the convolutional kernel is very important to extract the effective image features and improve the performance of the convolutional neural network. C1 filters the 224×224 input image with 96 convolutional kernels of size 11×11 with a stride of 4 pixels for sampling frequency, namely the convolutional kernel is spread over every unit of size 11×11. Ultimately, we get 96 feature maps with the size of 55×55.

C. Max-pooling layer (P1)

P1 is the first max-pooling layer, which has 96 feature maps of size 27 × 27. The pooling process is to select the maximum in each of the pooling regions as the value of the area after pooling. In this layer, we choose a max-pooling layer over a 3 × 3 region in order to control the speed of dimensionality reduction, because the decline in dimension is decreased exponentially, the speed is falling faster means the image features are more rough, and many image details are lost subsequently. Since the pooling layer has a region of size 3 × 3, while the stride of size 2, so we obtain the overlapping pooling. This scheme reduces the top-1 and top-5 error rates by 0.4% and 0.3%, respectively, as compared with the no-overlapping scheme, which produces output of equivalent dimensions.

D. Convolutional layer (C2)

The second feature extraction layer C2, which has a lot of similarities with the C1, while there is a certain gap. C2 takes as input the output of the first convolutional layer and filters it with 256 convolutional kernels of size 5 × 5 with a stride of 1pixel, and it is feasible to obtain 256 feature maps with the size of 27 × 27. In C1 layer, the receptive field of each neurons is equivalent to the raw image of size 33 × 33. Now C2 layer use a kernel of 5 × 5 to convolution, so that the receptive fields of neurons further enhance, which is equivalent to the original image of size 165 × 165. Each feature maps in C2 is not directly obtained by convoluting in P1 layer, but by combining several or all of the feature maps in P1 as input for convoluting again. The reason for this one is that the sparsely connected mechanism keeps the number of connections in a reasonable range. The other is that the asymmetry of network enables the different combinations can extract various features.

E. The remaining convolutional layers and pooling layers

These layers have the same working principle with the first two layers, but the size and number of the feature maps have changed. However, the size of the C5 layer after the convolution is still 13 × 13. In this process, the max-pooling layers follow the second (C2) and fifth (C5) convolutional layers with the kernels of size 3 × 3. With the continuous increase of the depth of the convolution, the extracted features are more abstract, with more discriminative and expressive power. The experimental results demonstrate that the classification accuracy is only about 40% when only take the first two layers of the convolutional neural network, which proves that depth will have a great affect on the performance of convolutional neural network, and lacks of depth will reduce the abilities of extracting features in the convolutional neural network.

F. Fully-connected layer (Fc)

In the whole process of learning, the remaining three are the fully-connected layers, which make the learning features of two channels cross-mixing to obtain a 4096-dimensional feature vector. We use Dropout technique in the first two fully-connected layers, setting to zero the input of the first two fully-connected layers with the probability 1/2, which do not contribute to the forward pass and do not participate in backward. The hidden layer of learning in this way cannot rely on the presence of particular other features of the previous layer, which makes the learning features can lead to stronger robustness, select more adaptive parameters, and significantly improve the generalization capabilities of the system.

IV. Experiments

To evaluate the approach, we carried out the experiments on two common test databases including ImageNet2012 and NUS-WIDE in this paper. Each of two databases contains a variety of categories of scene images. And the relevant descriptions of scene category and data distribution are described as Fig. 4. There are 53694 scene pictures in the training set, while the test sets ImageNet2012 and NUS-WIDE are composed of 3678 and 3593 pictures respectively.

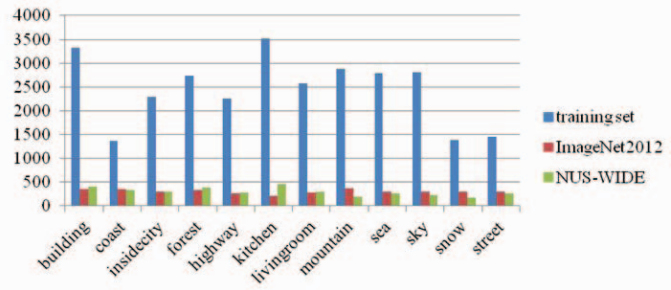

Fig. 4. Each category distribution of databases

In the experiment, we use the Lib-SVM and SoftMax classifier to classify the test images on the two data sets. In the first place, we extract 4096 neurons of the last layer in the Alex-Net model as the scene image features; then, two classifiers are trained by using the image features; finally, the trained classifier is used to classify the test images, and the classification accuracy (AC) as shown in Table I (0:building 1:coast 2:insidecity 3:forest 4:highway 5:kitchen 6:livingroom 7:mountain 8:sea 9:sky 10:snow 11:street).

TABLE I. Accuracy of Different Classifiers On Two Database

AC/(%) on ImageNet2012			AC/(%) on NUS-WIDE		
	Lib-SVM	SoftMax		Lib-SVM	SoftMax
0	91.01	90.73	0	88.25	87.75
1	83.89	83.61	1	80.29	79.41

2	81.17	80.52	2	78.67	78.33
3	80.18	79.87	3	74.15	73.37
4	91.16	91.54	4	86.42	86.78
5	87.14	87.14	5	82.44	81.20
6	92.04	92.04	6	86.00	85.67
7	85.03	85.03	7	81.48	80.42
8	95.02	94.35	8	91.04	89.92
9	93.00	92.33	9	89.32	88.46
10	90.00	90.00	10	87.45	85.47
11	90.07	89.38	11	89.21	86.92
Avg.	88.31	88.03	Avg.	84.56	83.72

From Table Ⅰ, we can see that the average classification accuracy of the two classifiers in the two different data sets are relatively high, which is almost the same accuracy. At the same time, the accuracy is close when the two classifiers are used to classify each class, which is illustrated that the feature extracted from the deep convolutional neural network has stronger discriminative and robustness. In this paper, the scene image classification method based on Alex-Net model is proposed for scene classification, which is extracted scene feature on the basis of the Alex-Net model, by using the Lib-SVM classifier to train model, the classification accuracy of scene images is significantly improved.

Table Ⅱ show the accuracy by three methods that include HOG，KNN and we proposed algorithm on ImageNet2012 and NUS-WIDE, respectively.

TABLE II. ACCURACY OF DIFFERENT METHODS ON TWO DATABASE

AC/(%) on ImageNet2012				AC/(%) on NUS-WIDE			
	HOG	KNN	DCNN		HOG	KNN	DCNN
0	73.26	80.13	91.01	0	71.02	77.32	88.20
1	70.21	71.39	83.89	1	67.76	69.34	80.28
2	73.26	78.34	81.17	2	70.05	76.02	75.58
3	70.03	69.96	80.18	3	67.81	69.06	74.13
4	80.00	84.06	91.16	4	77.83	82.89	86.54
5	81.32	82.65	87.14	5	78.36	79.86	82.43
6	85.34	86.65	92.04	6	81.16	83.10	86.08
7	70.22	70.34	85.03	7	67.22	69.03	81.48
8	75.01	82.45	95.02	8	73.06	79.67	91.03
9	76.65	82.87	93.00	9	73.78	80.00	89.36
10	78.08	83.15	90.00	10	75.54	81.18	87.45
11	76.34	78.23	90.07	11	73.89	76.08	89.01
Avg.	75.81	79.18	88.31	Avg.	73.39	76.96	84.55

From the Table Ⅱ, we can see that the average accuracy of DCNN algorithm is 12.50% higher than HOG and 9.31% higher than KNN in the ImageNet2012 dataset. In the NUS-WIDE database, the average accuracy of DCNN algorithm is 11.16% higher than HOG algorithm and 7.95% higher than KNN algorithm. To sum up, deep convolutional neural network can be more effectively extracted the deep characteristics of the scene image, which makes the extracted features more discriminative, and greatly improves the performance of the scene classification.

V. CONCLUSIONS

Deep learning has become a hot area of machine learning research, and it has been made a significant breakthrough in image recognition, speech analysis, target detection and other fields. In this paper, we extract the last layer features of the deep convolutional neural network for scene classification by using the Alex-Net model in the deep learning framework, which makes the classification accuracy can be further improved, so a stronger generalization performance and higher efficiency scene image classification model is constructed. In future work, we will introduce the relationship between different scenes in the training set to reduce redundant information, so as to further improve the performance of scene classification.

Acknowledgment

This work is partially supported by National Natural Science Funds of China (No.61572244, No.61472059) and the program for Liaoning Excellent Talents in University (LR2015030).

References

[1] O. Russakovsky, Y. Lin, K. Yu, and F. F. Li, Object-centric spatial pooling for image classification, Computer Vision-ECCV, pp. 1-15, 2012.

[2] A. Angelova and S.H. Zhu, Efficient object detection and segmentation for fine-grained recognition, IEEE Conference on Computer Vision and Pattern Recognition, Portland, pp. 811-818, 2013.

[3] B. Fernando, E. Fronomt, and D. Muselet, Supervised learning of Gaussian mixture models for visual vocabulary generation, Pattern Recognition, vol. 2, pp. 897-907,2012.

[4] T. Li , T. Mei, I.S. Kweon and X.S. Hua, Contextual bag-of-words for visual categorization, IEEE Transactions and Systems for Video Technology, vol. 4, pp. 381-392, 2011.

[5] A. Bosch, A. Zisserman, X. Munoz, Scene classification via pLSA, Computer Vision-ECCV 2006, Berlin Heidelberg: Springer, pp. 517-530, 2006.

[6] P.K. Elango and K. Jayarman, Clustering images using the latent dirichlet allocation model, University of Wisconsin, 2005.

[7] Z. Fu, H. Lu, and W. Li, Incremental visual objects clustering with the growing vocabulary tree, Multimedia Tools and Applications, vol. 3, pp. 535-552, 2012.

[8] G.E. Hinton and O. Simon, A fast learning algorithm for deep belief nets, Neural Computation, vol. 8, pp. 1527-1554, 2006.

[9] K. Alex, S. Ilya, and H. Geoffrey, ImageNet Classification with Deep Convolutional Neural Network. In NIPS, 2012.

[10] R. Girshick, J. Donahue, and T. Darrell T, Rich feature hierachies for accurate object detection and semantic segmentation, IEEE Conference on Computer Vision and Pattern Recegnition. Columbus. USA, 2014.

[11] D. Jia, D. Wei, L.J. Li, L. Kai, and F.F. Li, ImageNet: A Large-scale Hierarchical Image Database, IEEE Conference on Computer Vision and Pattern Recognition, 2009.

[12] F.F. Li and P. Perona, A Bayesian hierarchical model for learning natural scene categories, Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 524-531, 2005.

[13] J.R. Smith and C.S. Li, Image classification and querying using composite region64 templates, Journal of Computer Vision and Pattern Recognition, vol. 2, pp. 165-174, 1999.

[14] S.L. Zhang, J.F. Zhang, and L.H. Hu, Semantic annotation model of image scene based on formal concept analysis, Computer Application, vol. 4, pp. 1093-1096, 2015.

[15] Y. Jiang, Research on scene image content representation and classification, National University of Defense Technology Doctoral Thesis, 2010.