Vehicle Detection and Classification Based on Convolutional Neural Network

Dongmei He
School of Computer and Information Technology, Beijing Jiaotong University
Beijing, China
amymei@bjtu.edu.cn

Congyan Lang
School of Computer and Information Technology, Beijing Jiaotong University
Beijing, China
cylang@bjtu.edu.cn

Songhe Feng
School of Computer and Information Technology, Beijing Jiaotong University
Beijing, China
shfeng@bjtu.edu.cn

Xuetao Du,Chen Zhang
China Mobile Group Design Institute Co., Ltd.
Beijing, China
{duxuetao,zhangchen}@cmdi.chinamobile.com

ABSTRACT

Deep learning has emerged as a hot topic due to extensive application and high accuracy. In this paper this efficient method is used for vehicle detection and classification. We extract visual features from the activation of a deep convolutional network, large-scale sparse learning and other distinguishing features in order to compare their accuracy. When compared to the leading methods in the challenging ImageNet dataset, our deep learning approach obtains highly competitive results. Through the experiments with in short of training data, the features extracted by deep learning method outperform those generated by traditional approaches.

Categories and Subject Descriptors

I.4.8 [**Image Processing And Computer Vision**]: Scene Analysis—*color, depth cues*

Keywords

Vehicle detection and classification, convolutional neural network, visual feature extraction, SVM

1. INTRODUCTION

Research in object detection and classification is increasingly concerned for contemporary computer vision systems. In contrast to obvious objects classification, in which we need to distinguish obvious categories such as car and desk from each other, the fine-grained [1, 6, 13] categorization problem asks us to distinguish inconspicuous categories such as car and sports car from each other. Here we address the problem of vehicle detection and classification. Moreover, it can be utilized in many aspects including urban traffic, vehicle theft against, large parking lot management, improvement of the road information acquisition and safety management of highway, electronic toll collection (ETC), traffic investigation and so on.

While the vehicle detection and classification is appealing in its usefulness and meaningfulness, several shortcomings have limited its realization: (i) the high similarity between models naturally influence the accuracy, (ii) some models containing only a few images, and (iii) the unified direction like front, side or back in picture causes the inaccurate classification. With limited training data, here we adopt semi-supervised multi-task learning of deep convolutional representations, where representations are learned on a set of related problems but applied to new tasks which have too few training examples to learn a full deep representation. In this paper we address these limitations, providing techniques that are practical for vehicle classification problem.

Considering the above limitations,in feature extraction part, we use the deep convolution neural network. In addition, we adopt three other traditional methods that first one is HSV color histogram based on component model and second one is direction of controllable pyramid characteristic based on component model and the last one is large-scale sparse learning. We divide all the models into eight different labels, and apply the support vector machine (SVM) classifier to perform multi-label classification. The experimental results demonstrate that the depth learning method outperform those traditional methods.

The rest of this article is organized as follows: in section 2, we briefly review the previous estimate of vehicle recognition. In the third part, we analyze the model feature extraction and classification methods. In the fourth part, we make a few experiments to demonstrate our method of performance and compare the results. Finally, section 5 list the conclusion and future work.

2. RELATED WORK

Vehicle detection and classification have received significant attention over the last few years. Then we simply

ICIMCS '15, August 19-21, 2015, Zhangjiajie, Hunan, China

DOI: http://dx.doi.org/10.1145/2808492.2808495

review the previous work of it.

D.Koller et al. [8] put forward a method to get vehicles shape information through detecting and tracking vehicles at the same time. This method applies image frame difference technique to motion segmentation and use Kalman filter to update the background image. This technique gets a better application in the actual system. G.D.Sulhvan [7] adopts three one-dimensional template to detect whether there is a certain type of vehicle. When it discovers some particular vehicle exists, we can track this vehicle according to the corresponding two-dimensional template type [11], namely the validation process of vehicle recognition. The method adopts the thought of multiple templates, which is of certain innovation.

More recently, deep learning has become a hot topic in recognition and object classification area. Convolution neural network (CNN) [10, 12] is a classic network structure of deep learning. It mainly includes two parts: the convolution layer and convergence layer. Convolution layer is a feature map from convolution image which used filter. Convergence layer is the characteristic of figure analysis. It adopts drop sampling operation and filter out the characteristics of the local figure.

LeCun et al. [10] proposes the character recognition system LeNet - 5 based on convolutional neural network. The system has been used for bank handwritten numerals recognition. It has a five layer structure of the convolutional neural network. Krizhevsky [9] proposes a special kind of convolution neural network algorithm, and has received competition-winning numbers on large benchmark datasets consisting of more than one million images, such as ImageNet [4]. The method achieves good effect for large-scale data image classification task. Jia [5] proposes the activation of a deep convolutional network. It has the very good object recognition, fine-grained object recognition and classification.

Due to the good recognition, the activation of a deep convolutional network is applied to vehicle classification in our article. This is different than before.

3. VEHICLE DETECTION AND CLASSIFICATION

3.1 System Framework

For each image presented to the system, preprocessing is applied to transform the image into the template used in database. Then the system extracts visual feature by convolution neural network [10]. Finally the feature is classified by the SVM classifier and compares the classification results.

3.2 Feature Extraction

Convolutional neural network (CNN) [10] is one of the deep learning methods which is widely used in image recognition and speech analysis.

Convolutional neural network (CNN) [10] is a multi-layer neural network system. In each layer there are lots of two-dimensional planes which are composed of many independent neurons. Convolutional neural network (CNN) [10] puts the feature extraction process integration into the multilayer perceptron through restructuring and the same surface weight sharing. At a given level, each neuron input is the local field from previous level, and combined with this level of weight. That is to say, the current layer is a result from a front layer convolution with a pair of convolution kernels. CNN [10] can use more layers, and each layer has many features. The output of a layer constitutes the input of the next layer. The associated neurons in same feature map share the same weight. Figure 1 is a conceptual example of the CNN [10].

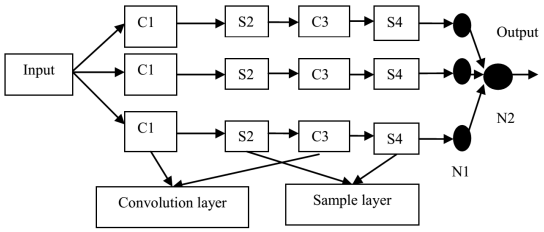

Figure 1: The conceptual example of the CNN

Convolution neural network structure [9] is usually composed of five layers, which are input layer, convolution layer, sample layer, convergence layer and output layer. They are connected by neurons, and form a convolutional neural network.The following is convolution neural network structure [9] used in this article. There are eight layers including five convolution layers and three whole layers.

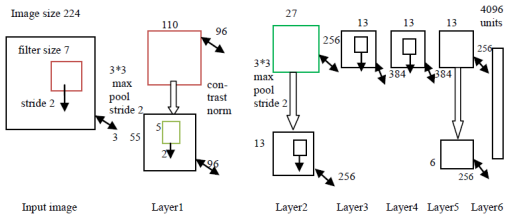

Figure 2: The conceptual example of the CNN

Figure 2 is the structure of the convolutional neural network we use in this paper. This structure requires the input layer image cutting into 224×224 sizes. We divide the image into three planes based on the RGB color system. Then we convolve the input layer by 96 different red convolution kernels with size of 7×7; the sliding window step length is 2, and then we get a convolution layer. It is the first layer of the structure. Next a linear offset function is used for bias operation, and then we get a maximum point in 3×3 areas of convolution layer with sliding mode step length is 2. Finally we get the feature of the first layer after contrast normalization. By the same operation, we can get the second, three, four, five layer features. The sixth and the seventh layer are fully connected layer because the sixth layer is made up of all feature maps that are connected from upper layers. The input of the sixth layer is a $6 \times 6 \times 256$-dimensional vector, and then we can obtain information containing 4096 neurons after pooling operation [14]. The last layer is the output decision layer composed of softmax regression. All the convolution kernels and feature maps in the structure is a square.

What above is introduction of the layer of convolution neural network structure which used in this article. Next we specifically introduce the key methods used in the structure. These methods improve the image classification performance

of the model in some extent.

The general neuron output function of neural network model is $f(x) = \tanh(x)$ or $f(x) = (1 + e^x)^{-1}$. In terms of training speed, the saturation nonlinear method is much slower than unsaturated nonlinear method such as $f(x) = \max(0, x)$. Consequently, the output function of the neuron in this paper is adjusted with the linear unit of nonlinear (ReLU). ReLU has a good performance because the input of ReLU does not need to be normalized to prevent input saturation, so we adopt local normalization method to prevent saturation. We regard $a^i_{x,y}$ as one neuron and make convolution operation with convolution kernels i at (x, y) position. Moreover, this activity is processed by linear ReLU operation. The response normalization can express as $b^i_{x,y}$ and the formula is as follows:

$$b^i_{x,y} = a^i_{x,y} / (k + \alpha \sum_{j=\max(0,i-n/2)}^{\min(N-1,i+n/2)} (a^i_{x,y})^2)^\beta \qquad (1)$$

The sum is process in the same spatial location in the field of convolution kernels n. N is the total number of convolution kernels in the layer of network. The order of convolution kernels is arbitrary, and they have been confirmed before the training. The response normalization is a lateral inhibition.Constant k, n, α, β are super parameters which are determined by the fixed set. The constant values in this article are: $k = 2, n = 5, \alpha = 10^{-4}, \beta = 0.75$. It is helpful to improve the recognition rate via local normalization of the response of the same layer adjacent nodes.

The level pooling of convolution neural network simplifies the adjacent groups output of neurons in the same convolution kernels figure. Traditional pooling operation is processed without overlap, namely we put the $z \times z$ as pooling unit, and if s is equal to z, it represents the traditional process of pooling. Among them, s expresses movement step length of pooling unit. In other case, if $s < z$, it is regarded as overlapping pooling. The value of overlapping pooling and the pool parameters in this article is adopted like this s is equal to 2 and z is equal to 3. This model can reduce error rate about 0.3%. Simultaneously, it can match more valuable feature of hidden layer unit by strengthen learning network structure.

3.3 Deep Convolutional Activation Features

Figure 3 in the below is the convolutional neural network characteristics of the first layer to the seventh layer. In the structure, the original image is adjusted to the size of 224×224 as input layer. The size of the first layer feature map is 55×55, which contains a total of 96 feature image. The second feature map size for 27×27 which contains a total of 256 feature image. The feature of the third and the fourth layer is all 13×13, and the number of the each layer feature map is 384. The size of the fifth layer feature map is 6×6, which contains a total of 256 character image. The sixth and the seventh layer are fully connected layer and each layer is made up of 4096 neurons. In this article, the fully connected layer is a 4096-dimensional feature vector. The eighth layer is fully connected layer as well as policy maker in this structure of the network, and it is no reference to this article, so we only take the front seven layers.

The seventh layer represents the last one hidden layer which is in the front of the fully connected decision-maker layer. The sixth layer is an activation layer before the seventh layer, and the fifth layer is an activation layer before the sixth layer. The fifth layer is the first set of activations that has been fully propagated through the convolutional layers of the network. The convolution layer lying in the front of convolution neural network layer contains the limited semantic representation, and the back layer gets feature from the local feature activation of convolution layer from the lower layer to middle layer. The back layer contains more semantic information, so this article uses the sixth layer feature.

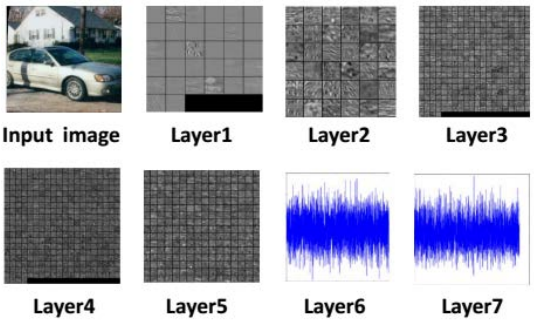

Figure 3: A Deep Convolutional Activation Feature

Vehicle detection and classification can be seen as a multi-class classification problem, so traditional classification algorithms such as the nearest centroid classifier and the support vector machine (SVM) can be directly applied for vehicle detection and classification. After extracting visual features, we need to perform the classification using multi-label classification. We choose the SVM classifier to classify the vehicle image into the right group.

3.4 Other Features

By contrast, three other methods are used in this experiment.

3.4.1 methods based on component model

The first is HSV color histogram based on component model [2]. The HSV color histogram as a statistical histogram is consisting of Hue, Saturation and Value in HSV color space model. It is a description of the proportion in image or a designated area about these three components of HSV color space. The second is direction of controllable pyramid characteristics based on component model(the following article we call it direction) [2]. It divides the image into inequality subbands from different scales and direction different, and therefore, we can accurately detect the characteristics like texture and singular points.

These methods by means of component object model extract the image key component area and middle characteristic, then it extracts image low-level feature in the key component area. Further we fuse the middle and lower characteristics to get the image visual feature, and at last the visual feature is classified by SVM. Here we use the HOG features [3] as a low-level image feature fusion with the image middle characteristic which one is HSV color histogram and another is the direction of controllable subband characteristic of the pyramid. The following is a confusion matrix.

3.4.2 large-scale sparse learning

The last method is large-scale sparse learning. It has re-

cently become a popular research topic because of its ability of conducting simultaneous classification and feature selection. We briefly cover its principle below. It is divided into two steps.

*Step*1: The function

$$[x, funVal] = LeastR(A, y, \beta, opts)$$

solves the l_1-norm (and the squared l_2-norm regularized least squares problem:

$$\min_x \frac{1}{2} \|Ax - y\|_2^2 + \frac{\alpha}{2} \|x\|_2^2 + \beta \|x\|_1 \quad (2)$$

where $A \in R^{m \times N_{train}}, y \in R^{m \times 1}$,and $x \in R^{N_{train} \times 1}$. The figure m means the number of dimension vector of image feature. The figure N_{train} means the amount of training data. We have illustrate A,x,and y in Figure 4.

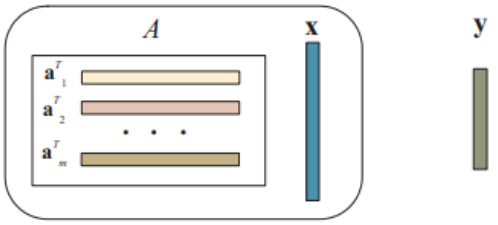

Figure 4: Among them, A is the training data of size $m \times N_{train}$, y is the response data of size $m \times 1$. It is the column vector of a sample in all samples(include training data and testing data) matrix. And x is the solution of size $N_{train} \times 1$.

In equation(2), β is the l_1-norm regularization parameter, and α is the regularization parameter for the squared l_2-norm. Its value is equal to 0 by default. The program uses the input values for β and α by setting $opts.rFlag = 0$. By setting $opts.rFlag = 1$, the program automatically computes β_{max}, above which equation(2) shall obtain the zero solution. In the latter case, the input β whose value lies in the interval [0,1] should be specified as a ratio, and the resulting (ℓ_1) regularization used in the program is $\beta \times \beta_{max}$; similarly, the input α is a ratio larger than 0, and the actual (ℓ_2) regularization used in the program is $\alpha \times \beta_{max}$. Finally,we get the matrix x. When we gathered them, we will get the new feature matrix X_{train}. X_{train} is a matrix of size $N_{train} \times s$, where s means a quantity of all samples including training data and testing data.

*Step*2: After the first learning step, we get the new representation feature matrix X_{train}. We say that $Y \in R^{N_{train} \times k}$ is a label information matrix of training data X_{train}, where k is the number of labels. Based on this, we can formulate a regression function $\hat{W}$:

$$\hat{W} = \underset{W}{argmin} \|Y - ZW\|_F^2 + \|W\|_F^2 \quad (3)$$

Then we can get the label for a testing sample by:

$$h = \underset{h}{argmin} w_i \; s.t \, N_{train} < i \leq N_{total} \quad (4)$$

where w_i is a row vector of W. And the last several rows with the number of total data minus training data are the tag weight values of testing data.

4. EXPERIMENTS

4.1 Vehicle Dataset

In this experiment, we get the images from ImageNet [3]. We download the eight kinds of common models including armored, bicycle, bumper, elevator, motor, railway, sports and tank. Due to the unequal quantity of each model, we take each category with 700 images. In the experiment, we use 600 images for training and 100 images for testing. Figure 5 shows that the eight kinds of images used in experiment.

Figure 5: Eight kinds of models in experiment

4.2 Experimental Results and Analysis

In this part, the confusion matrix show the accuracy result for comparison of these methods. Figure 6 shows confusion matrix on four different methods.

The result of confusion matrix shows that the classification accuracy based on convolutional neural network is the highest. The total classification accuracy followed by chart in order are 84.25%, 83.625%, 52.5% and 46.375%. If we divide some similar models into one class such as car and sports car, the accuracy based on convolutional neural network can reach 90.7143%. The following Table 1 list the accuracy of DeCAF approach and other existing approaches.

Table 1: Comparison of DeCAF and Other Methods

Methods	Accuracy
DeCAF	84.250%
Sparse Learning	83.625%
HSV	52.500%
Direction	46.375%

4.3 Experiment Comparison

In order to make the results more obvious, we take these results into a chart in Figure 7.

Intuitively, our results robustly demonstrate that DeCAF features based on convolution network and large-scale sparse learning outperform other features. Though sparse learning method outperform DeCAF method in some species, but the DeCAF method is Slightly higher than other methods in total accuracy. It is helpful to improve the accuracy of vehicle detection and classification by our method.

5. CONCLUSION

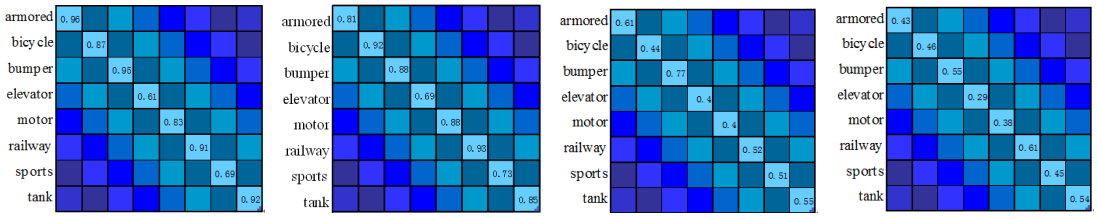

Figure 6: confusion matrix on four different methods.

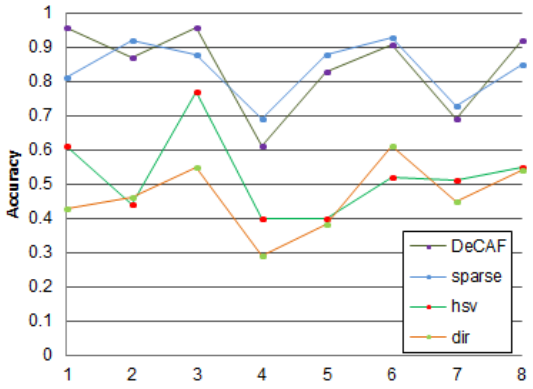

Figure 7: Comparison of methods.Here the data of horizontal ordinate 1,2,3,4,5,6,7,8 respectively stand for armored, bicycle, bumper, elevator, motor, railway, sports, tank. That four lines from top to bottom stand the methods of DeCAF, sparse learning, HSV color histogram based on component model and direction of controllable pyramid characteristics based on component model.

In this paper, a new vehicle detection and classification approach based on convolution activation feature is proposed. We extract the features using the convolutional neural network which are effective for a variety of object recognition tasks. In future, our research is how to continue to improve the accuracy under the poor training samples and better apply to real life.

6. ACKNOWLEDGMENTS

This work was supported in part by the National Natural Science Foundation of China under Grant 61472028, Grant 61372148, Grant 61300071, Grant 61272352, in part by the Fundamental Research Funds for the Central Universities under Grant 2014JBM025, in part by the Beijing Natural Science Foundation under Grant 4142045, and in part by the Beijing Higher Education Young Elite Teacher Project under Grant YETP0547.

7. REFERENCES

[1] A. Angelova and S. Zhu. Efficient object detection and segmentation for fine-grained recognition. *International Journal of Robotics Research*, 30(13):1543–1552, 2011.

[2] T. Berg and P. Belhumeur. Poof: Part-based one-vs.-one features for fine-grained categorization, face verification, and attribute estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 955–962. IEEE, 2013.

[3] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. computer vision and pattern recognition. In *IEEE Computer Society Conference*, pages 886–893. IEEE, January 2005.

[4] J. Deng, W. Dong, and et al. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition*, pages 248–255. CVPR, 2009.

[5] J. Donahue, Y. Jia, O. Vinyals, and et al. Decaf: A deep convolutional activation feature for generic visual recognition. *arXiv preprint*, 1310.1531, 2013.

[6] K. Duan, D. Parikh, and et al. Discovering localized attributes for fine-grained recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. CVPR, 2012.

[7] J. M. Ferryman, A. D. Worrall, G. D. Sullivan, and K. D. Bakerr. A generic deformable model for vehicle recognition. In *Proeeeding of British Machine Vision Conference*, pages 127–136, 1995.

[8] D. Koller. Towards robust automatic traffic scene analysis in real-time. In *Int. Conf. Pattern Reconiton*, pages 126–133, 1994.

[9] A. Krizhevsky, I. Sutskever, and G. E.Hinton. Imagenet classification with deep convolutional neural networks. *NIPS*, 1(2):4, 2012.

[10] Y. LeCun, L. Bottou, and et al. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[11] A. M. Mahdl and J. Mansour. Car type recognition in highways based on wavelet and contour let feature extraction. In *Proceedings of the 2010 International Conference on Signal and Image Processing*, pages 353–356, 2010.

[12] T. N.Sainath, B. Kingsbury, and B. Ramabhadran. Deep convolutional neural networks for lvcsr. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 8614–8618. IEEE, May 2013.

[13] M. Stark, J. Krause, B. Pepik, and et al. Fine-grained categorization for 3d scene understanding. In *Computer Vision and Pattern Recognition*, pages 248–255. CVPR, 2009.

[14] N. Zhang, R. Farrell, and T. Darrell. Pose pooling kernels for sub-category recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*. CVPR, 2012.