Vehicle Logo Recognition and Attributes Prediction by Multi-task Learning with CNN

Yizhang Xia*, Jing Feng and Bailing Zhang
Department of Computer Science and Software Engineering
Xi'an Jiaotong-Liverpool University
SIP, Suzhou, China, 215123
Yizhang.Xia, Bailing.Zhang@xjtlu.edu.cn

***Abstract*—Vehicle Logo Recognition(VLR) has been an important study field in intelligent Transportation system (ITS). This paper proposes to recognize vehicle logo and predict logo attributes by combining Convolutional Neural Network (CNN) with Multi-Task Learning(MTL). In order to accelerate convergence of multi-task model, an adaptive weight training strategy is employed. To verify the algorithm, the Xiamen University Vehicle logo recognition dataset is extended into a larger vehicle logo dataset including 15 brands, 6 visual attributes and 3 no-visual attributes. The experiment results indicate that the proposed multi-task CNN model perform well for both of logo classification and attribution prediction with overall accuracy 98.14%.**

Keywords: Vehicle Logo Recognition (VLR), Convolutional Neural Networks(CNNs), Multi-Task Learning(MTL)

I. Introduction

Intelligent transportation systems includes a important application, Vehicle Manufacturer Recognition (VMR) [1] that has benefit for political or commercial institution from car ownership statistics. VMR system often takes advantage of logo for recognition since vehicle logo is the obviously sign in a vehicle.

A number of papers has been published on vehicle logo recognition. Most of previously proposal approaches combine some hand-crafted features, HOG, with a trainable classifier, SVM, [2]. These methodologies have a number of limitations, e.g. 1) hand-crafted features (e.g., HOG), that is insufficient to concurrently meet various imaging conditions, for example, halflights, rotation, viewpoints, and so on; and 2) indispensable requirement of accurate logo detection bounding box since inexact bounding box will introduce noise and then result in a significantly decreased recognition accuracy.

In the recent years, a new area of machine learning, i.e., deep learning, has attracted world-wide attention, which aims at representational learning. A special kind of deep learning models, i.i., convolutional neural networks(CNNs), has demonstrated outperformance in various computer vision tasks [3]. Resent research also indicated that CNN can extract robust and generic features [4] from raw pixels through several convolutional layers non-linear mapping between inputs and outputs. Due to its hierarchical structure, CNN is strongly robust against illumination variance, car posture, stain, etc. [5], [6] Based on these characteristics, CNN achieves high performance in recognizing vehicle logo.

Describing objects by semantic property, or attribute, is a technique that has been pay much attention by many researchers in computer vision problem [7]. Attributes not only describe object from several aspects, but also act as the bridge between low-level feature and high-level semantic. Various multimedia tasks can benefit from attributes,such as, knowledge transfer and knowledge sharing among different attributes [7]. In our works, we investigated 6 visual attributes and 3 no-visual attributes in CNNs learning framework.

In recent times, MTL has been employed to many visual recognition research, especially when tasks hold some commonality and generally slightly lack various data [8], [9]. This technology attempts to force knowledge sharing among multiple correlated tasks together. MTL aims to improve the performance of each task by sharing the relevant and irrelevant information for enlarging inter-class distance and shortening inner-class distance [9]. The existing MTL with CNN can be simply divided into two kinds. In the first one [10], one task is optimized and other tasks are fixed, then another task is selected to optimize with other tasks remained unchanged, such a procedure is iteratively repeated as training proceeds. However, many researchers employ the second type [11] that it optimizes all the tasks simultaneously. We following the second manner and put forward a new adaptive weighting MTL inspired by [11].

More specifically, the adaptive weighting MTL is proposal to train a CNNs model to recognize vehicle logo and predict its attributes simultaneously. And a large vehicle logo database, including 15 brands, 6 visual attributes and 3 no-visual attributes is extended from Xiamen University Vehicle logo recognition dataset [6] to evaluate the algorithm. According to the experiments result, a satisfied result is obtained.

The rest of this paper is organized as follows. Section II explains the overall system and expounds the adaptive weighting MTL. Experiments are analyzed in Section III, followed by the conclusion and future work in Section IV.

II. System Overview

The whole system is illustrated in the Fig. 1. The raw image is converted into gray scale and then normalized into uniform size, 64*64. Then the CNNs extracts the features from the image. Finally, several Multi-Layer Perceptrons (MLP) are applied to recognize the vehicle logo and predict some binary

978-1-5090-4093-3/16/$31.00 ©2016 IEEE

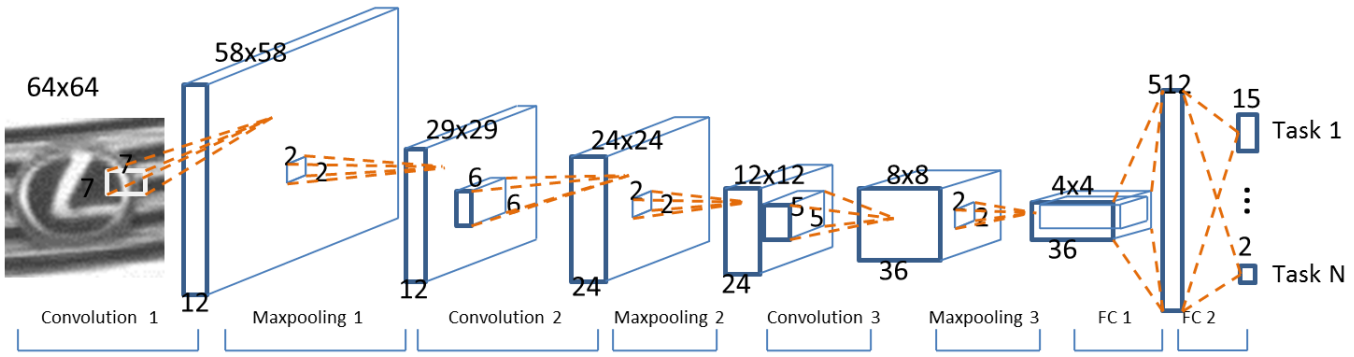


Fig. 1. System Overview

attributes. In the following, the CNNs will be outlined in II-A and the adaptive weighting MLT will be explained in II-B.

A. VLR with CNN

A CNNs can be employed to extract the features hierarchically, from the images, Fig. 1.

C1: In the CNN feed forward procedure, convolution operation employs a set of local receptive field (kernel or filter) to repeatedly slide across the entire visual field from upper-left corner to lower-right corner for extracting the saliency map of object. In order to reduce the parameters of CNN, convolution operation applies the same local receptive field including weight vector and bias do dot product with inputs, which is well-known as parameters sharing. This technology not only make training a convergent CNN become possible in the limited data, but also increase the generalization performance of CNN. In the training processing, kernels of convolution layer will be sensitive to edge, color, or specific patterns detectors of vehicle logo gradually. The convolution operation is formulized as

$$C_1 = f_{act}\left(b + \sum_i w^i * x^i\right) \tag{1}$$

where x^i and C_1 are the i-th input and the i-th output feature map, respectively. For the first convolution operation, the input is raw image and the output is the first layer feature. And for others convolution operation, the input is the output from last layer, pooling layer or activation layer. w^i is the weight vector of kernel. $*$ denotes the dot production operation. b and $f_{act}(\bullet)$ is the bias and nonlinear mapping function. The weights of filter is initialized randomly and then is trained with a well-known back propagation algorithm.

R1: Generally, a nonlinear activation layer follows convolution layer since the convolution operation generates a linear mapping. However, the practical problem often requires a nonlinear complex model to solve the nonlinear problem. In order to introduce nonlinear mapping into CNN, a nonlinear activation layer is indispensable. This layer converts a linear input from convolutional layer into a nonlinear representation. In the past, a sigmoid function play the important role, while it is abandoned in CNN for vanishing gradient problem. In the recent years, Glorot et al. [12] remitted the problems by proposing an new activation function, called rectifier linear unit(ReLU), showed in Eq. 2

$$ReLU(C_1) = max(C_1, 0) \tag{2}$$

Along with deeper CNN model and in consideration of this advantage, recent CNN based approaches [13] employed ReLU as the nonlinear activation layer following both convolution layer and full connection layer, to shorten training time generally as declared in [13].

S1: If CNN extracts feature by convolution layers only, it will lead to the model become too deep or the dimensionality of feature become too high. In order to settles the matter, [13] utilized pooling operation that also can maintain the space-invariance. Widespread pooling functions are max pooling, namely max-pooling layers, as in Equ. 3

$$y^i{}_{m,n} = \max_{0 \le \lambda,\mu < s} \left\{ x^i{}_{m \cdot s+\lambda, n \cdot s+\mu} \right\} \tag{3}$$

Max pooling operation is similar with convolution operation that smooth a rectangle, $s \times s$, across the input feature map, x^i, from upper-left corner to lower-right corner for selecting the local maximum, y^i. And in this paper, a non-overlapping Paradigm is used.

For the VLR, the CNN feature extractor can be acquired with raw pixels to automatically learn low-level and mid-level features, easing the need for hand-crafted features and thus achieving a good the recognition performance. The CNN structure unites three constructive ideas to forbear some degree of scale, shift, and distortion invariance, i.e., convolution operation, activation operation and pooling operations. Hence, the VLR boosts a high classification performance and robustness against various complex imaging conditions.

B. Adaptive Weighting MTL with CNN

Inspired by [10], we applied the MTL to classify the vehicle logo brand and predict 9 attributes simultaneously. As showing in the Fig. 1, the sixth layer is a fully connected layer that has 512 neurons and it is split into 10 branches. We proposal a adaptive weighting MLT to increase the inter-class distance.

TABLE I
STATISTICS OF VEHICLE LOGO ATTRIBUTES DATABASE

	Audi	Buick	Chery	Chevrolet	Citroen	Ford	Honda	Hyundai	Lexus	Mazda	Nissan	Peugeot	SGMW	Toyota	VW	Total	Percent
Alphabet	0[1]	0	1[2]	0	0	1	1	1	1	0	1	0	0	0	1	9641	48.7%
X axis symmetry	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	703	3.6%
Y axis symmetry	1	0	0	0	1	0	1	0	0	1	0	0	1	1	1	9374	47.4%
Central symmetry	1	0	0	1	0	0	0	1	0	0	0	0	0	0	0	2974	15.0%
Encircled	1	1	1	0	0	1	0	1	1	1	1	0	0	1	1	13625	68.9%
Animal-like	0	0	0	0	0	0	0	0	0	1	0	1	1	0	0	3918	19.8%
Price[3]	0	1	1	1	1	1	1	1	0	1	1	1	1	1	1	17765	89.8%
Birthplace[3]	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	2295	11.6%
Birth of brand[3]	1	1	0	1	1	1	0	0	0	1	0	1	0	0	1	10360	52.4%
Total	703	1747	1657	635	1616	742	1623	1636	1312	1637	1034	1643	638	1520	1637	19780	100.0%
Percent	3.5%	8.8%	8.3%	3.2%	8.1%	3.7%	8.2%	8.2%	6.6%	8.2%	5.2%	8.3%	3.2%	7.6%	8.2%	100.0%	

[1] 0 means there is not alphabet in the brand of vehicle logo.
[2] 1 means there is alphabet in the brand of vehicle logo.
[3] They are no-visual attributes of vehicle logo.

When MTL is operated, we minimize the linear combination of losses of each task:

$$L = \sum_{i=1}^{10} \alpha_i * L_i \tag{4}$$

$$\sum_{i=1}^{10} \alpha_i = 1 \tag{5}$$

In the equation, L_i is the i-th task's loss and the α_i is the weight of i-th task in the total loss, L. Many researchers [10], [14] just fix the α when they train CNN. On the other hand, Wu [11] initialize the α with 1 and modify it according to the validation performance in the training. We introduce the momentum concept to smooth the training. The adaptive weighting MLT can be formulized below.

$$\alpha_i = \beta * \overline{\alpha_i} + \alpha_{i-1} \tag{6}$$

$$\overline{\alpha_i} = \frac{E_i}{\sum_{j=1}^{10} E_j} \tag{7}$$

where $\beta * \overline{\alpha_i}$ is the variation of task weight in each epoch and β is an experimental parameter to control the quantity that this time error effects the next time's task weight. In our experiment, the β is set as 0.1. In Eq. 7, E_i is the error of i-th task in each epoch. At the initial state, all E_i is initial as 0.1. Experiments proved the adaptive weighting MTL is effective.

III. EXPERIMENT

We evaluated the system performance on the task of vehicle logo classification, III-B, and logo attributes prediction, III-C. And our training of CNNs follows from [10], i.e., stochastic gradient decent with a batch size of 128 examples and learning rate of 0.01. Each epoch of training takes about 20 minutes on an Intel Xeon $E5$ CPU with our inhouse implementation, and the network around converges in 100 epoches.

A. Vehicle Logo Database

We extended Xiamen University vehicle logo recognition database [6] into a larger dataset consisting of fifteen brands, with six visual attributes and three no-visual attributes. Several thousand outdoor vehicle frontal images were captured from traffic surveillance. Then, the logo is roughly detected by our previous works [15]. Finally, a vehicle logo database with 19780 images from fifteen makers was generated. Some sample images from the fifteen manufacturers in the database are shown in Fig. 2. Furthermore, we labeled nine category-level attributes dividing into visual attributes and no-visual attributes inspired by [16]. All the segmented images are uniformly normalized to $64*64$ pixels. The statistics of vehicle logo recognition database is enumerated in the Table I.

Fig. 2. Sample of Database

B. Vehicle Logo Recognition

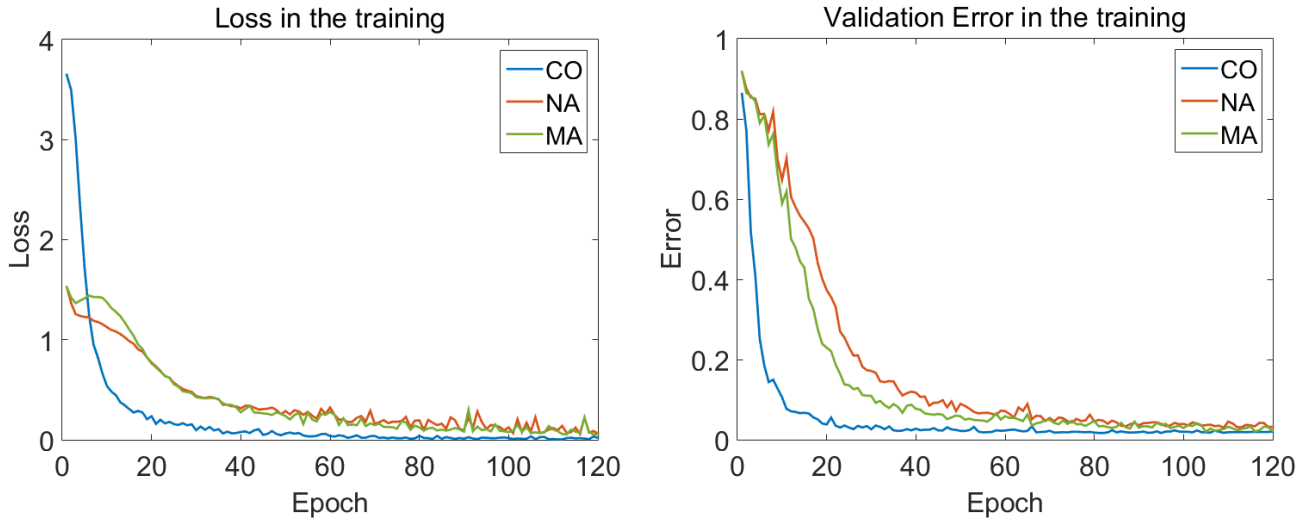
The database described in the Table I is utilized to assess the proposed system. 80%, 10% and the rest of the images are used as train samples, the validation and test samples. In the experiment, we compare four different approaches to demonstrate the benefit of the adaptive weighting MTL CNN in the VLR task:

- Single task CNN
- Dense SIFT and SVM
- PHOG and SVM
- LBP and SVM

The results are shown in Table II. We limited the plot to 120 epoches on the training data in the training process. It can be seen that in general the CNN outperforms the traditional feature extractor, LBP, PHOG and dense SIFT, significantly. Comparing our MLT CNN with the single task CNN, although they all obtain the high performance, our system can predict another nine binary attributes simultaneously.

TABLE II
CLASSIFICATION ACCURACY COMPARED WITH OTHER FOUR APPROACHES

Accuracy %	Audi	Buick	Chery	Chevrolet	Citroen	Ford	Honda	Hyundai	Lexus	Mazda	Nissan	Peugeot	SGMW	Toyota	VW	Average
MTL + CNN + MLP	99.3	98.6	98.3	98.3	96.8	99.4	97.5	98.1	97.2	97.4	100.0	98.4	98.1	94.7	100.0	98.14
CNN + MLP	99.3	99.3	96.6	96.6	95.7	100.0	95.5	98.7	98.9	98.7	97.9	99.7	98.7	96.3	100.0	98.13
Dense SIFT + SVM	96.2	84.4	95.4	98.2	98.6	93.9	95.6	90.7	92.3	97.8	81.6	84.9	82.9	84.8	94.8	91.47
PHOG + SVM	93.0	75.2	94.8	84.4	81.3	71.4	74.9	78.4	91.8	90.6	64.6	78.8	58.0	71.5	96.5	80.35
LBP + SVM	56.5	60.0	62.9	59.4	76.2	70.7	65.9	61.6	63.2	63.9	57.0	57.6	17.5	34.4	86.1	59.53



(a) Loss in the training

(b) Validation Error in the training

Fig. 3. Loss and error in the training

C. Attributes Prediction with adaptive weighting MTL

We also evaluate the adaptive weighting MTL CNN on our data set by predicting nine binary attributes with the VLR. The attributes include six visual attributes and three non-visual attributes. The result is listed in the Table III, which demonstrate that our algorithm obtain high prediction performance for both of the visual attributes and the no-visual attributes.

We also compared our algorithm with the fixed weighting training strategy [10] with single task CNN as baseline. As can be seen from the training process, Fig. 3, our algorithm has a lower and more stable loss and validation error. Compared with the single task CNNs, the performance of our algorithm is similar.

TABLE III
CLASSIFICATION ACCURACY OF VRL AND ATTRIBUTES PREDICTION

Task	Accuracy %
VLR	98.14
Alphabet	98.18
X axis symmetry	99.95
Y axis symmetry	98.58
Central symmetry	99.34
Encircled	98.84
Animal-like	98.74
Price	99.60
Birthplace	98.79
Birth of brand	98.79

IV. CONCLUSION

A VMR method with vehicle logo attributes prediction based on a CNN has been proposed in this paper. Our system recognizes vehicle logo and predicts logo attributes at the same time by MTL. We proposed a adaptive weighting MTL to increase the inter-class distance. Through experiment, the algorithm speed up and stabilize the training process of multi-task CNN. The final results demonstrate vehicle logo recognition and logo attributes prediction achieved a satisfactory performance.

There are two directions further work. One is to complete the vehicle logo detection and recognition. Another is to improve the adaptive task weight MTL by combining alternating manner and simultaneous manner.

REFERENCES

[1] L. Figueiredo, I. Jesus, J. A. T. Machado, J. R. Ferreira, and J. L. Martins de Carvalho. Towards the development of intelligent transportation systems. In *Intelligent Transportation Systems, 2001. Proceedings. 2001 IEEE*, pages 1206–1211, 2001.

[2] J. Xiao, W. Xiang, and Y. Liu. Vehicle logo recognition by weighted multi-class support vector machine ensembles based on sharpness histogram features. *IET Image Processing*, 9(7):527–534, 2015.

[3] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[4] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson. Cnn features off-the-shelf: An astounding baseline for recognition. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2014 IEEE Conference on*, pages 512–519, June 2014.

[5] Chun Pan, Zhiguo Yan, Xiaoming Xu, Mingxia Sun, Jie Shao, and Di Wu. Vehicle logo recognition based on deep learning architecture in video surveillance for intelligent traffic system. In *Smart and Sustainable City 2013 (ICSSC 2013), IET International Conference on*, pages 123–126, Aug 2013.

[6] Y. Huang, R. Wu, Y. Sun, W. Wang, and X. Ding. Vehicle logo recognition system based on convolutional neural networks with a pretraining strategy. *IEEE Transactions on Intelligent Transportation Systems*, 16(4):1951–1960, Aug 2015.

[7] C. H. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 951–958, June 2009.

[8] A Evgeniou and Massimiliano Pontil. Multi-task feature learning. *Advances in neural information processing systems*, 19:41, 2007.

[9] Rich Caruana. *Learning to Learn*, chapter Multitask Learning, pages 95–133. Springer US, Boston, MA, 1998.

[10] C. Zhang and Z. Zhang. Improving multiview face detection with multi-task deep convolutional neural networks. In *Applications of Computer Vision (WACV), 2014 IEEE Winter Conference on*, pages 1036–1041, March 2014.

[11] W. Liu, T. Mei, Y. Zhang, C. Che, and J. Luo. Multi-task deep visual-semantic embedding for video thumbnail selection. In *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*, pages 3707–3715, June 2015.

[12] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Deep sparse rectifier neural networks. In *International Conference on Artificial Intelligence and Statistics*, pages 315–323, 2011.

[13] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[14] R. Girshick. Fast r-cnn. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1440–1448, Dec 2015.

[15] B. Zhang and H. Pan. Reliable classification of vehicle logos by an improved local-mean based classifier. In *Image and Signal Processing (CISP), 2013 6th International Congress on*, volume 01, pages 176–180, Dec 2013.

[16] Makoto Ozeki and Takayuki Okatani. Understanding convolutional neural networks in terms of category-level attributes. In *Asian Conference on Computer Vision*, 2014.