2016 12th World Congress on Intelligent Control and Automation (WCICA)
June 12-15, 2016, Guilin, China

A Visual Attention Based Convolutional Neural Network for Image Classification

Yaran Chen, Dongbin Zhao, Le Lv, and Chengdong Li

***Abstract*— This paper presents a visual attention based convolutional neural network (CNN) to solve the image classification problem in the real complex world scene. The presented method can simulate the process of recognizing objects and find the area of interest which is related with the task. Compared with the CNN method in image classification, the model is proficient in fine-grained classification problem and has a better robustness due to its mechanism of multi-glance and visual attention. We evaluate the model on vehicle dataset, where its performance exceeds CNN baseline on image classification.**

I. Introduction

Image classification has already become one of most important branches in computer vision for many applications, such as fingerprint entrance control system, face recognition technology and so on. However, with the requirements of rich applications, scene complexity, and a wide variety of things, the performances of the state-of-art techniques are not satisfied, especially for the fine-grained classification problem for its categories only being discriminated by subtle and local differences.

Vehicle classification is a key means to track the suspect for police and manage vehicles for parking. In modern traffic, the camera installed widely at the crossing is usually an essential equipment. Therefore, images of vehicles in the driveway are easily available. Vehicle classification in these captured images is a very practical and essential technique, which has been extensively studied for decades. However, vehicle classification remains a challenging task. Because some images are not clear due to the strong or dim lights, bad environment or photographic equipments, and some images have interference information, which have pedestrians or other vehicles, shown as Fig. 1. In this paper, we use these captured images to recognize the type of vehicle.

Recognizing the type of vehicle is a typical fine-grained classification task, for the similarity among categories. Fine-grained classification task is a popular research topic in computer vision. Computer vision system processes the whole parts of an image at once and processes each part in a same way, such as the feature extractor (Histogram of Oriented Gradients (Hog) [1], Scale-invariant feature transform (SIFT) [2], Local Binary Patterns (LBP) [3]), common classifiers (template matching [4], Bayesian classification [5], support vector machine (SVM) [6] and random forest (RF) [7], Boosting [8], and convolutional neural network (CNN) [9]). These methods extract local features at all parts of image and they aggregate these features from different spatial regions to get a representation of the image. These methods treat each part of an image in the same way. But we know that every image has the key area, especially for the challenging fine-grained classification task. The fine-grained classification task has a requirement that classifiers are able to catch local differences accurately, which is a challenge for traditional methods.

For fine-grained classification, human has the ability to catch local differences accurately, due to its multi-glance and visual attention mechanism. In detail, human has a fovea area in which vision is acuity. When glancing at an image, a person sees an image with a small part clear and others fuzzy and the image is sent to human brain. After several glances a person knows the image by analyzing these clear parts.

Fig. 1. The vehicle data

The paper proposes a visual attention based CNN model by simulating human vision mechanism. Our model simulates human center fovea to process image with fovea vision, and use the information entropy to evaluate the processed image. The information entropy guides to select the following interest area. We summary the main contributions as follows:

- Inspired by human vision mechanism of multi-glance and visual attention, the paper proposes a visual attention based CNN model for image classification. This

Y. Chen, D. Zhao and L. Lv are with The State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China (e-mail: chenyaran2013@ia.ac.cn; dongbin.zhao@ia.ac.cn; iamlvle@126.com)

C. Li is with the School of Information and Electrical Engineering, Shandong Jianzhu University, Jinan 250101, China (e-mail: lichengdong@sdjzu.edu.cn)

This work is supported by National Natural Science Foundation of China (NSFC) under Grants No. 61273136, No. 61573353, No. 61533017 and No. 61473176, and Natural Science Foundation of Shandong Province for Outstanding Young Talents in Provincial Universities (ZR2015JL021)

model is mainly for fine-grained classification.
- The model prevents exhaustive search or random search when selecting an interest area, by using a feedback (the information entropy) to adjust the recognizing processing.

We organize the paper as follows. First introduce the related work in the next section. In the third part, we design the frame of our model. Then, an experiment of vehicle classification is presented. Finally, we conclude in the fifth part.

II. Related work

CNN has been widely studied more than ten years in computer vision system, for example LeCun uses supervised back-propagation networks in digit recognition [10]. In the early research, CNN has been only applied in small-scale image classification (such as MNIST, CIFAR10/100 and NORB), limited by the computing power. With the development of GPU, the deep CNN with millions of parameters has been used in the scale image classification tasks [9], scene labeling [11] and so on, showing a significant improvement in computer vision. In particular, large-scale CNN trained on ImageNet [12] has an excellent performance [13]. Zeiler studies why large-scale CNN on ImageNet performs so well and shows that convolutional layers are equal to a set of learned filters [14]. Low-level representations are shared among categories, and high-level representations are more global and more distinguishing. When big data is accessible, the deep architecture CNN performs better than shallow architecture CNN. Many researchers have begin to use the deep architecture, but it is difficult to train due to large parameters.

It has a long history for people to pay attention to focus area, and saliency detectors are motivated by human perception in the early time. Bottom-up visual characteristics help guiding eyes movement and the influence of distracter regions might be reduced based on target features [15]. Itti et.al [16] proposes a simple conceptually model for saliency-driven focal visual attention. The above works focus theoretical analysis or based on hardwired. In [17] a method is proposed for improving the run-time of general-purpose object-detection algorithms. This method can help robot cameras to quickly scan scenes of high resolution by simulating digital fovea. In [18], a recurrent model of visual attention is present and has a great performance on the MNIST dataset. Aurelio Ranzato studies where to look for image classification [19]. Our method is also based on visual attention motivated by human vision mechanism.

III. Model Description

In this section we present the Oxford VGG CNN model proposed by [20], and design the visual attention based CNN model by combining human multi-glance and attention mechanism into the Oxford VGG CNN model. Compared to the first model, our proposed model is able to stand out local features like human vision.

A. Oxford VGG model

Oxford VGG model is a deep convolution network for large-scale and complex image classification tasks, which is first used in ImageNet Challenge 2014. The model has five convolutional layers, which can be seen as filters for extracting features. Then three fully-connected (FC) layers follow these convolutional layers. The last second FC layer contains 1000 channels corresponding to the 1000 classes in the ImageNet classification task. The final layer is loss function (softmax) S [20]. Karen Simonyan shows the representations with low feature layers can be generalised to other datasets and perform well [20]. For another task, three Fully-Connected (FC) layers should be changed as the task.

For an input image X, the output of Oxford VGG model is calculated as follow:

$$\begin{aligned} \mathbf{P}(X) &= M(X, \mathbf{w}) \\ \mathbf{P}(X) &= [p_1, ..., p_C]^T,\ p_i \in [0, 1] \end{aligned} \tag{1}$$

where M is the mathematic model of VGG CNN, $\mathbf{w}$ is the parameter vector and p_i is the probability of the image belonging to $i-th$ class. C is the number of possible classes. The maximum probability p_y means that the image is most likely to fall into the $y-th$ class. The recognition result of the image with the model is $y-th$.

$$\begin{aligned} p_y &= \max_i \left(\mathbf{P}(X)\right) \\ &= \max_i \left(p_i\right), i \in [1, C] \end{aligned} \tag{2}$$

If the class of the input image is y^*, our purpose is to maximize the p_{y^*}, so the loss function is cross-entropy error defined by:

$$l(X, y, a, b) = -log(p_{y^*}) \tag{3}$$

Concretely, the training objective is to minimize the loss function $l(X, y, a, b)$ by adaptively tuning the parameter vector $\mathbf{w}$.

B. Visual attention CNN model

For a classification task, we should predict a label of an input image which has some task-unrelated redundant information. In order to find the task-related area, we propose a visual attention based CNN model. Our proposed model is shown as Fig. 2.

1) Model inference: The working processing is shown as Fig. 3. In the model, every input image is filtered by a digital fovea which imitates the human center fovea. A focussed image X_f processed by the digital fovea is obtained as:

$$\begin{aligned} X_f =& F(X, a, b) = \Phi(a, b) \otimes X \\ =& \begin{pmatrix} \phi_{11}x_{11} & \phi_{12}x_{12} & \cdots & \phi_{1D}x_{1D} \\ \phi_{21}x_{21} & \phi_{22}x_{22} & \cdots & \phi_{2D}x_{2D} \\ \vdots & \vdots & \ddots & \vdots \\ \phi_{D1}x_{D1} & \phi_{D2}x_{D2} & \cdots & \phi_{DD}x_{DD} \end{pmatrix} \end{aligned} \tag{4}$$

where Φ represents digital fovea, (a, b) is the center point which the eye looks at, and $\otimes$ denotes an element-wise multiplication. The digital fovea works like a mapping function

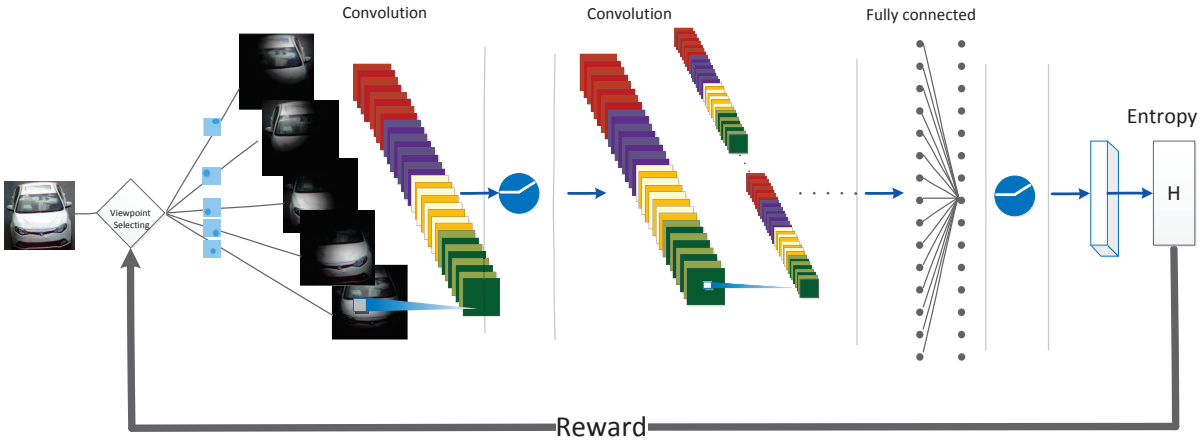

Fig. 2. The model of visual attention based CNN. The system frame consists of three parts: digital fovea, evaluation network and searching a center point.

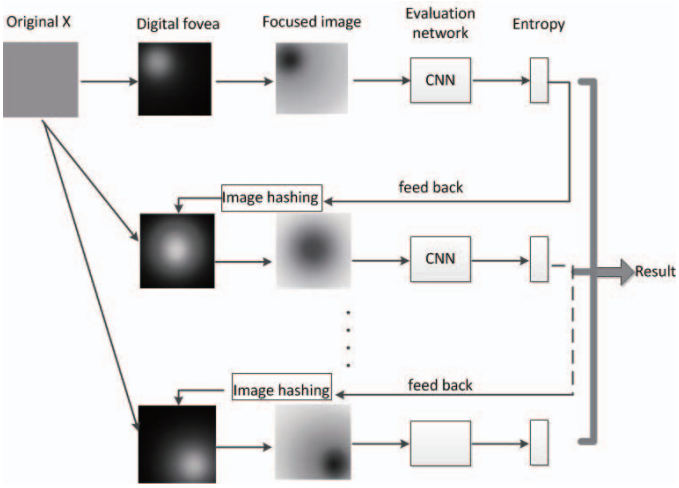

Fig. 3. The mechanism of visual attention based CNN

ϕ :

$$\phi(x, y, \alpha, \beta) = sigmoid(r, \alpha, \beta) = \frac{1}{1 + exp(\alpha(r - \beta))} \quad (5)$$

$$r(x, y) = \sqrt{(x - a)^2 + (y - b)^2} \quad (6)$$

where r is the distance between a pixel position (x, y) and a center point (a, b). The value ϕ of mapping function ranges from 1 to D. The value near (a, b) is close to one, otherwise zero. From these equations, we can see that the focused image has a high resolution near the center point (a, b).

Fig. 4 shows the value of mapping function ϕ changes with the distance. The value keeps larger within a certain distance. So for a focused image, there is a small patch near the center point with high resolution shown as Fig. 5(b).

A CNN following the digital fovea is used to evaluate a focused image X_f. The CNN is called evaluation network, and has a deep structure similar to the Oxford VGG model shown in Sec. III-A. For an input focused image, the output

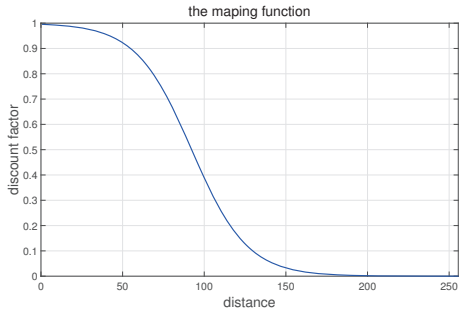

Fig. 4. The mapping function

is a probability distribution $\mathbf{P}(X_f)$ from (1). The maximum probability p_y is the result of the focused image from (2), called mid-result.

The output $\mathbf{P}(X_f)$ is used to calculate the information entropy, which can measures the amount of the uncertainty of a random variable. In detail, if the value of entropy is larger, the confidence of the determination is lower.

The information entropy of a variable s can be calculated by :

$$H(s) = E[-logp_i] = -\sum_{i=1}^{n} p_i logp_i \quad (7)$$

where the variable s has several probable states $s_1, s_2, ..., s_n$, and p_i is the probability of state s_i. When all the probabilities of states is $\frac{1}{n}$ which means that the state of the variable s is equally possible, the information entropy reaches the largest.

The information entropy of a probability distribution $\mathbf{P}(X_f)$ can also measure the discrimination of a focused image X_f and indicates how difficult the focused image can be classified correctly.

Based on the information entropy, the discrimination of a focused image is defined as:

$$E(X_f) = -\frac{1}{logC} \sum_{i=1}^{C} p_i logp_i, \quad \sum_{i=1}^{C} p_i = 1 \quad (8)$$

where $0 \leq E \leq 1$, E equals 1 with $p_i = \frac{1}{C},\ i = 1, 2, ...C$, and E equals 0 with $(p_1, p_2, ..., p_C) = (0, 0, ..., 1, ..., 0)$. We use E to direct the following center point where to look. The following center point is chosen with a feedback B_f:

$$B_f(X_f) = \begin{cases} 1\ , & E(X_f) \leq H_{VPT} \\ -1\ , & E(X_f) > H_{VPT} \end{cases} \tag{9}$$

If the $E(X_f)$ is no more than the given threshold E_{VPT}, the evaluation network gives a positive feedback, otherwise it gives a negative one.

A focused image which has a positive feedback is easier to distinguish and its center point l^k has a key area related to recognition task (k is the k-th center point found by us). The following center point l^{k+1} should be selected in another key area which also has an task-related information.

There is a small patch near the center point with a high resolution shown as Fig. 5(b). We crop the patch $\bar{X^k} \in R^{d \times d}$. The principle of choosing the following center point is that the most similar patch to the cropped one $\bar{X^k}$ is chosen with a positive feedback $B_f(X_f) > 0$, otherwise a least similar one is chosen. The following center point is located in the the most or least similar patch.

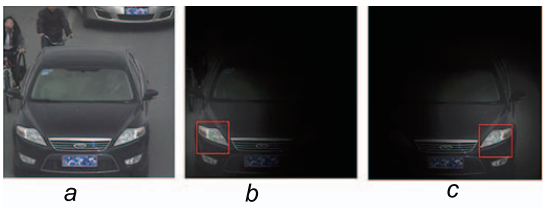

Fig. 5. The performance of Hashing algorithm. a), The original image. b), The center point lies in the left headlight. c), The center point lies is the right headlight which is the result of hashing algorithm searching with a positive feedback

In this paper, we use image hashing algorithm to find the following patch $\bar{X}^{k+1}$ matching with the target patch $\bar{X^k}$, due to its great discriminative ability, and low computational cost [21], [22].

A patch is encoded with the hash function H and projected into low-dimension binary hash codes $h(h_1, h_2..., h_m)$.

$$\begin{pmatrix} h_1 \\ h_2 \\ \vdots \\ h_m \end{pmatrix} = H \begin{pmatrix} \bar{X}^1 \\ \bar{X}^2 \\ \vdots \\ \bar{X}^n \end{pmatrix},\ H \in R^{m \times n}, and\ m < n \tag{10}$$

The similarity between a patch $\bar{X}^i$ and a target patch $\bar{X}^t$ can be expressed by :

$$d(x^i, x^t) = ||h^i - h^t|| = \sqrt{\sum_{j=1}^{m} (h_j^i - h_j^t)^2} \tag{11}$$

The similar sorting $d^1, d^2, ..., d^n$ between the patches $\bar{X}^1, \bar{X}^2, ..., \bar{X}^n$ and the target patch $\bar{X}^t$ is calculated by (11). The most similar patch d^1 is chosen when the feedback is positive, otherwise a least similar one d^n is chosen. The processing of selecting a new center point is shown as Fig. 6. The result of image hashing algorithm is shown in Fig. 5.

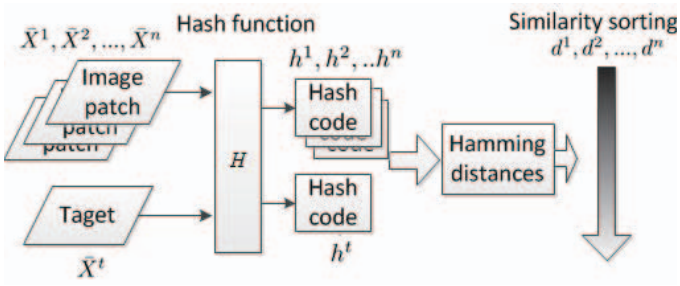

Fig. 6. Image hashing algorithm

We select a center point randomly for simple computation and exploration in the new chosen patch. The new center point l^{k+1} is input to the digital fovea part, and a new focused image is generated. Then the new focused image would repeat the process above. The model stores the focused image whose feedback B_f is positive, and N_{lab} denotes the number of stored images.

$$N_{lab} = \sum_{i=0}^{k} B_f(i),\ B_f(i) > 0 \tag{12}$$

2) Prediction and learning: We get these center points with the positive feedback, then how can we obtain a accurate prediction using these center points.

In this paper we use some combination rules to predict labels. A simple prediction method is majority voting. In the model, the number of center points with the positive feedback is N_{lab}. When N_{lab} reaches the set number N_{lab}^*, the model will stop ($flag = 1$) to find an new center point.

$$flag = \begin{cases} 1\ , & N_{lab} \geq N_{lab}^* \\ 0\ , & N_{lab} < N_{lab}^* \end{cases} \tag{13}$$

For each center point, it generates a focused image X_f^i. Each focused image has a mid-result y^i and an entropy $E(X_f^i)$ ($i \in [1, N_{lab}^*]$), after through the evaluation part. The maximum number of class among these mid-results is the final result by majority voting.

While the information entropy reflects the discrimination of a focused image, the mid-result of a focused image with a small E has a high credibility. In the paper, we give each focused image a weight coefficient which is calculated by information entropy.

$$\begin{aligned} F^i &= 1 - E(X_f^i) \\ w_e^i &= \frac{F^i}{\sum_{j=1}^{N_{lab}^*} F_j},\ i = 1, 2, ..., N_{lab}^* \end{aligned} \tag{14}$$

where F denotes the avail value of focused image, and w_e is the weight coefficient of focused images.

Then the second combination rule is weight majority voting. The final result y^* is calculated by:

$$y^* = \max_i \sum_{j=1}^{N^*_{lab}} w^i_e I(y^j = y^i),\ i = 1, 2, ..., N^*_{lab}$$
$$I = \begin{cases} 1\ , & y^j = y^i \\ 0\ , & y^j \neq y^i \end{cases} \tag{15}$$

Through observing dataset and experiment, we find that the useful information usually locates in the lower part of a image. In the upper of the picture there is some useless information such as another vehicle or passerby (shown as Fig. 1), we prefer to believe that the key area is at the lower of each image. Give a priori weight w^i_h to each focused image entropy, and higher weights are assigned to more credibility. The final result y^* is calculated as the following:

$$y^* = \max_i \sum_{j=1}^{N^*_{lab}} w^i_h w^i_e I(y^j = y^i),\ i = 1, 2, ..., N^*_{lab} \tag{16}$$

In the part of selecting a new center point, we choose a similar patch from the candidate set. The number of patches in candidate set are no more than $N-1$, with a image being divided into N patches. It is reduced one in each iteration. When the candidate set is empty, the focused images of positive feedback is not enough, $N_{lab} < N^*_{lab}$. In this case, stop the process and replace N^*_{lab} with N_{lab} in (16).

Model parameters from valuation network CNN need to be tuned at the training time. For each training sample, there are N^*_{lab} focused images with positive feedback. They combine the loss function $L(X, y^*, \mathbf{w})$ shown as:

$$L(X, y^*, \mathbf{w}) = \sum_{i=1}^{N^*_{lab}} l(X, y^*, a^i, b^i) w^i_h w^i_e \tag{17}$$

where $l(X, y^*, a^i, b^i)$ is the cross-entropy error (3). We obtain a set of optimum value or satisfying quasi optimum value of $\mathbf{w}$ by minimize the loss function $L(X, y^*, \mathbf{w})$.

IV. Simulation Results

In this part, we introduce experimental dataset and parameter setting and present quantitative results to validate the effectiveness of the visual attention based CNN against state-of-the-art CNN model.

A. Image Dataset

We test the proposed method's performance on vehicle datasets. The task is vehicle classification discriminated from different brands: Audi, Bavarian Motor Works, Volkswagen and so on. The dataset has 58 classes in all, with 15000 training images and 2000 test images. These images are comprised of front vehicles' perspective, captured by camera from different intersections, in different lighting conditions, with artificial classification and containing a misclassified small portion.

Some classes are too similar in appearance, like Fig. 7. We can see that (a) is belong to MAZDA while (b) is HAIMA AUTO, but they are very similar and we can observe some tiny differences in headlight and logo. So other information of the two images is useless for discriminating them. Thus the task needs more efficient methods and our model can meet its requirements by ignoring useless information and keeping useful information.

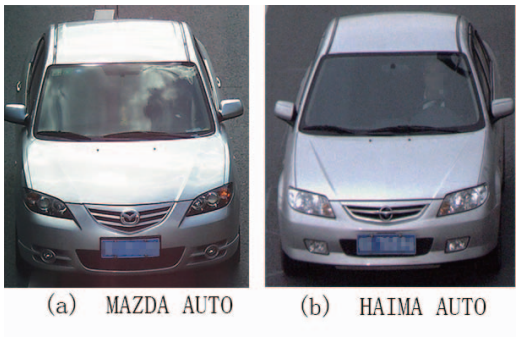

Fig. 7. Similar images among different classes

B. Model parameters setting

In the paper, CNN model, as a standard baseline, is a large-scale network and its parameters initialized by the parameters of Oxford VGG [20]. Oxford VGG has 5 convolution layers and 3 fully-connected layers with 1000 outputs. We choose the front 5 convolution layers, following a new definition fully-connected layer with 100 outputs and the output of the model has 58 channels for 58 classes, called VGG58.

In visual attention based CNN, the size of input image is 224×224 and its evaluation network is the baseline model (VGG58). The setting number N^*_{lab} is 6 and the information entropy threshold E_{VPT} is 0.3.

C. Result

Fig. 8 shows the result of the task with visual attention based CNN and VGG58 separately. (a) and (b) show the error of classification with two model. The sign 'traintop1e' is the error, which is generated in the training dataset with each image being predicted only once. 'traintop5 e' is the error in training dataset with each one being predicted five times. 'testtop1 e' is the error in test dataset with only one time for predicting an image. 'testtop5 e' is the error in test dataset with only five times of prediction for each image. (c) and (d) are objective functions, calculated by (3) and (17) separately.

Compared with the Fig. 8 (a), our method has a cheering result which has an obvious improvement, shown in Fig. 8 (b). In Fig. 8 (c)-(d), the objective of our model is also lower. We can get the same conclusion that it is useful for image classification by visual attention based CNN. And our model gives a smaller loss function than VGG58. That is to say if the class of input image is y^*, p_{y^*} which is the output of our model is larger than that of VGG58 model. Our model gives a predicted label with a large probability and small information entropy, meaning that the prediction of our model has a larger credibility.

Our model can also classify the similar vehicles shown as Fig. 7. The result of the first vehicle is right with a high probability $p_y* = 0.978$, and the second result is also right. The probabilities of two results are larger than VGG58. The loss function of our model is smaller than that of VGG58, indicating our model is sensitive to local information and has a better performance for fine-grained classification task.

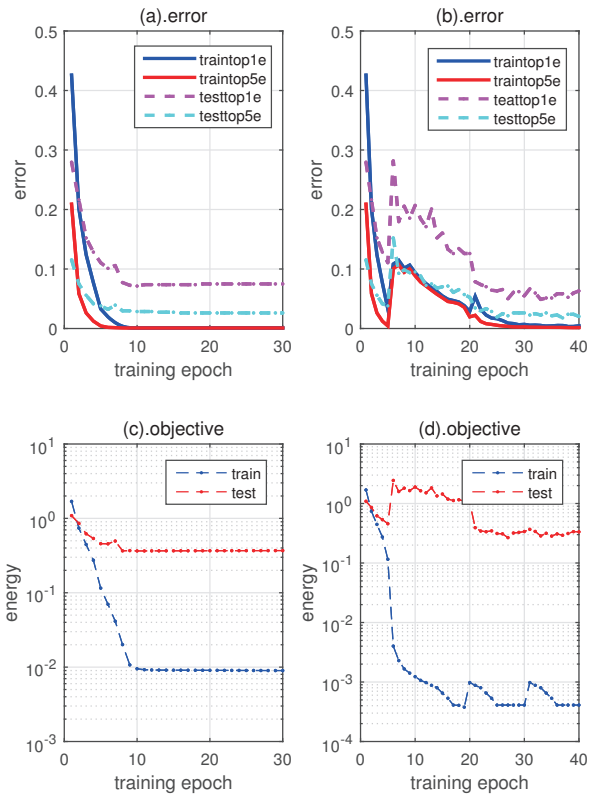

Fig. 8. Train curves tracking the average error and objective (average loss function) on Vehicles-85 task. (a), Each point represents average error on train data and test data per epoch of the total dataset be trained once with VGG58. (b), Average error with visual attention based CNN. (c), Objective on train data and test data with VGG58. (d), Objective with visual attention based CNN.

V. Conclusions

This paper proposes a visual attention based CNN model for fine-grained classification task, and evaluates the model on vehicle dataset. In the training and test process, each image goes through the digital fovea part, creating several images of different center point. These processed images would decide the original image's category together. The digital fovea makes a image clear near the center point and ignoring far area. Then the model saves the valuable area and abandons the redundant information. We have test the system on the realistic data set of 17000 frontal visual images of vehicles. The task has an improvement of accuracy than the VGG58. The result shows that the proposed method is useful for the fine-grained classification task.

References

[1] Navneet Dalal and Bill Triggs, "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*. IEEE, 2005, vol. 1, pp. 886–893.

[2] David G Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[3] Timo Ojala, Matti Pietikäinen, and Topi Mäenpää, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 24, no. 7, pp. 971–987, 2002.

[4] Roberto Brunelli, *Template matching techniques in computer vision: theory and practice*, John Wiley & Sons, 2009.

[5] Daniel Preotiuc-Pietro and Florentina Hristea, "Unsupervised word sense disambiguation with n-gram features," *Artificial Intelligence Review*, vol. 41, no. 2, pp. 241–260, 2014.

[6] Ola Amayri and Nizar Bouguila, "A study of spam filtering using support vector machines," *Artificial Intelligence Review*, vol. 34, no. 1, pp. 73–108, 2010.

[7] Andy Liaw and Matthew Wiener, "Classification and regression by randomforest," *R news*, vol. 2, no. 3, pp. 18–22, 2002.

[8] Jamie Shotton, John Winn, Carsten Rother, and Antonio Criminisi, "Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context," *International Journal of Computer Vision*, vol. 81, no. 1, pp. 2–23, 2009.

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

[10] Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural computation*, vol. 1, no. 4, pp. 541–551, 1989.

[11] Clement Farabet, Camille Couprie, Laurent Najman, and Yann LeCun, "Learning hierarchical features for scene labeling," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 35, no. 8, pp. 1915–1929, 2013.

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 248–255.

[13] Ali S Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson, "Cnn features off-the-shelf: an astounding baseline for recognition," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2014 IEEE Conference on*. IEEE, 2014, pp. 512–519.

[14] Matthew D Zeiler and Rob Fergus, "Visualizing and understanding convolutional networks," in *Computer Vision–ECCV 2014*, pp. 818–833. Springer, 2014.

[15] Geoffrey Underwood, "Cognitive processes in eye guidance: algorithms for attention in image processing," *Cognitive Computation*, vol. 1, no. 1, pp. 64–76, 2009.

[16] Laurent Itti, Christof Koch, and Ernst Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, , no. 11, pp. 1254–1259, 1998.

[17] Nicholas J Butko and Javier R Movellan, "Optimal scanning for faster object detection," in *Computer vision and pattern recognition, 2009. cvpr 2009. ieee conference on*. IEEE, 2009, pp. 2751–2758.

[18] Volodymyr Mnih, Nicolas Heess, Alex Graves, et al., "Recurrent models of visual attention," in *Advances in Neural Information Processing Systems*, 2014, pp. 2204–2212.

[19] Marc Aurelio Ranzato, "On learning where to look," *Eprint Arxiv*, 2014.

[20] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[21] Kave Eshghi, "Hash-based image identification," July 20 2010, US Patent 7,761,466.

[22] Yanyun Qu, Shuyang Song, Jiangjun Yang, and Jianmin Li, "Spatial min-hash for similar image search," in *Proceedings of the Fifth International Conference on Internet Multimedia Computing and Service*. ACM, 2013, pp. 287–290.